

Volume 20
2013

Journal
of
Natural Sciences
Sookmyung
Women's
University

Nano-physics
Chemistry
Medical & Pharmaceutical
Mathematics
Statistics
Computer Science
Multimedia Science



Published by
The Research Institute of Natural Sciences
Sookmyung Women's University
Seoul 140-742, Korea

제20호
(2013년호)

자연과학연구소 2013년도 연구 결과 백서



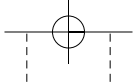
나노물리학
화학
의약과학
수계학
통계학
컴퓨터과학
멀티미디어학

숙명여자대학교 자연과학연구소

자연과학연구소 2013년도 연구 결과 백서

제20호 (2013년호)

숙명여자대학교 자연과학연구소



자연과학연구소 2013년도 연구 결과 백서

편집위원회 위원장	황 선 영 숙명여자대학교 이과대학 통계학과	Editor-in-Chief	Sun-Young Hwang Department of Statistics Sookmyung Women's University
간 사	김 선 홍 숙명여자대학교 이과대학 수학과	Excutive Editor	Seon-Hong Kim Department of Mathematics Sookmyung Women's University
위 원	김 한 칠 숙명여자대학교 이과대학 물리학과		Hanchul Kim Department of Nano-Physics Sookmyung Women's University
	이 한 길 숙명여자대학교 이과대학 화학과		HanGil Lee Department of Chemistry Sookmyung Women's University
	임 광 일 숙명여자대학교 이과대학 의약과학과		Kwang-Il Lim Department of Medical & Pharmaceutical Science Sookmyung Women's University
	서 검 교 숙명여자대학교 이과대학 수학과		Keomkyo Seo Department of Mathematics Sookmyung Women's University
	황 선 영 숙명여자대학교 이과대학 통계학과		Sun-Young Hwang Department of Statistics Sookmyung Women's University
	유 석 중 숙명여자대학교 이과대학 컴퓨터과학과		Seok Jong Yu Department of Computer Science Sookmyung Women's University
	윤 용 익 숙명여자대학교 이과대학 멀티미디어과학과		Yong-Ik Yoon Department of Multimedia Science Sookmyung Women's University

교육이념 우리 대학교의 창학 이념은 정숙, 현명, 정대를 지향하는 여성교육과 민족의 전통성 및 주체성을 확립하는 민족교육의 실현이다. 우리 대학교는 이러한 여성교육과 민족교육을 실현하기 위하여 시대가 요청하는 여성을 양성하고, 더 나아가 국가 민족과 세계 인류 발전에 기여하는 전인적이고 지도적인 여성을 배출하고자 한다.

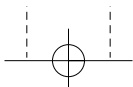
자연과학연구소 2013년도 연구 결과 백서

2014년 2월 8일 인쇄
2014년 2월 10일 발행

발행처 : 숙명여자대학교 자연과학연구소
발행인 : 황 선 혜
편집인 : 황 선 영
서울시 용산구 청파동 2가 53-2

인쇄소 : **한림원(주)**
Tel : 02)2273-4201, FAX : 02)2179-9083
E-mail : hanrim@hanrimwon.com

◆ 숙명여자대학교 자연과학연구소 2013년 연구 결과 백서는 숙명여자대학교 자연과학연구소에서 출판한 것이며, 아울러 본 논문집은 2013년도 숙명여자대학교 재정지원에 의하여 출판되었음.



자연과학연구소 주최 2013 학술행사

- [1] 2013.03.12. 나노물리학과. 핵물리와 원자력발전사업 (연사 : 백춘선 박사)
- [2] 2013.03.15. 수학과. 미분 가능성과 매우 큰 자연수 (연사 : 권도용 박사)
- [3] 2013.03.21. 화학과. EXPLORING THE PROTEIN UNIVERSE WITH TOOLS OF PHYSICAL CHEMISTRY (연사 : Prof.Kurt Wüthrich 박사)
- [4] 2013.03.26. 나노물리학과. 나노 구조 표면 전산모사 (연사 : 이승미 박사)
- [5] 2013.04.02. 나노물리학과. 유령입자가 떨릴 때 내는 빛: 靈光 (연사 : 김지연 박사)
- [6] 2013.04.03. 수학과. 생활 속의 암호, 그리고 수학 (연사 : 지성택 박사)
- [7] 2013.04.29. 통계학과. 콜로퀴엄 (연사 : 임광일 박사)
- [8] 2013.05.03. 수학과. 수학은 정말 유용하긴 한가? CSI로 답하다 (연사 : 박형주 박사)
- [9] 2013.05.06. 컴퓨터과학부. 안드로이드 기반 파일 전송 시스템 개발 프로젝트 및 카네기멜론대학교 소프트웨어공학 석사과정 소개 (연사 : 이지연 박사)
- [10] 2013.05.10. 수학과. 암호의 역할 (연사 : 정상임 박사)
- [11] 2013.05.15. 컴퓨터과학부. 다음 커뮤니케이션 IT 서비스 소개 (연사 : 이해정 동문)
- [12] 2013.05.24. 수학과. 비누막과 비누방울에 담긴 수학 (연사 : 표준철 박사)
- [13] 2013.05.27. 컴퓨터과학과. phone to phone messaging을 위한 통신 프레임워크 (연사 : 윤혜진 박사)
- [14] 2013.05.28. 나노물리학과. 나노소자의 제작과 고해상도 특성 분석 (연사 : 박지용 박사)
- [15] 2013.05.29. 의약과학과. 바이오 의약 연구동향 및 미래 (연사 : 배진검 박사)
- [16] 2013.07.08. 통계학과. 바이오 통계학 연구동향 및 미래 (연사 : 김훈 박사)
- [17] 2013.09.24. 나노물리학과. 스핀트로닉스 : 금속으로부터 유기물까지 (연사 : 김태희 박사)
- [18] 2013.09.25. 화학과. 막 단백질에 존재하는 이온 채널의 전기적인 감응성이 있는 단백질에 대한 연구 (연사 : Prof. Doug J. Tobias 박사)
- [19] 2013.09.27. 수학과. 폐활동 모니터링을 위한 전기단층임피던스에서의 역문제 (연사 : 이은정 박사)
- [20] 2013.10.07. 컴퓨터과학부. 클라우드 컴퓨팅 - 도전과 기회 (연사 : 최재영 박사)
- [21] 2013.10.08. 나노물리학과. Tuning of topological states in Bi_2Se_3 and Bi_2Te_3 with magnetic impurities (연사 : 정명화 박사)
- [22] 2013.10.18~19. 통계학과. 2013 추계 한국데이터정보과학회 학술대회 및 정기총회 (연사 : 노호석 박사 외 46명)
- [23] 2013.10.28. 멀티미디어과학과. 게임 개발 과정 및 개발자 경험 (연사 : 김경헌 선생)
- [24] 2013.10.31. 통계학과. Semi-Markov random evolutions : convergences and applications (연사 : NIKOLAOS LIMNIOS 박사)
- [25] 2013.11.07. 수학과. 선배와의 간담회 (연사 : 임아영, 김민지 동문)
- [26] 2013.11.08. 수학과. Conway의 초현실수(연사 : 박부성 박사)
- [27] 2013.11.12. 나노물리학과. 비주기적 결정구조를 활용한 금속재료의 이론 및 응용

- (연사 : 이진규 박사)
- [28] 2013.11.15. 수학과. 수학 교원 임용고시 (연사 : 김현웅 선생)
- [29] 2013.11.20. 화학과. Total Synthesis of Natural Products via C-C Bond-Forming Hydrogenation (연사 : 우상국 박사)
- [30] 2013.11.27. 의약과학과. A functionalized gold nanoparticles-assisted universal carrier for Oligonucleotide (연사 : 한민수 박사)
- [31] 2013.11.27. 화학과. Optical and Electrochemical Graphene Biosensors (연사 : 서태석 박사)
- [32] 2013.11.29. 멀티미디어과학과. 구글 개발자 경험 2탄-구글 개발자 및 프로그래밍 개발 경험 (연사 : 조성진, 김형욱 선생)
- [33] 2013.11.29. 수학과. 금융과 수학의 만남 (연사 : 김정훈 박사)
- [34] 2013.12.02. 컴퓨터과학과. HTCAAS 서비스 소개 (연사 : 노승원 박사)
- [35] 2013.12.03. 나노물리학과. 신의 입자 힉스 보손에게 돌아간 2013 노벨 물리학상 (연사 : 송정현 박사)
- [36] 2013.12.04. 화학과. Design of highly active nanocatalysts by understanding the growth mechanism of nanoparticles (연사 : 이광렬 박사)

자연과학연구소 연혁

(1) 설립 목적 : 본 연구소는 자연과학의 기초 및 응용분야에 관한 연구를 수행함으로써 자연과학의 발전에 기여함을 목적으로 한다.

(2) 사업

1. 자연과학의 기초 및 응용분야의 연구, 개발 및 조사
2. 학술 강연회와 연구발표회 및 공개 강좌 개최
3. 연구자료의 확보 및 연구 논문집 발간
4. 위탁된 연구조사 및 감정 분석 실험
5. 기타 연구소 설립 목적에 부합되는 사업

(3) 현황

숙명여자대학교는 21세기를 향한 자연과학분야의 기초 및 응용 분야에 관한 연구내용의 질적 향상을 도모하고 특성화 분야 연구의 구심적 역할을 수행하여 사회발전에 공헌하기 위하여 1985년 자연과학 연구소를 설립하였다. 연구소의 활동 및 참여 연구원은 이과대학의 교수진 뿐 아니라 교내 외의 인사들이 대상이었고 현재 물리학 연구부, 화학 연구부, 생명과학 연구부, 수학 연구부, 컴퓨터과학 연구부, 통계학 및 멀티미디어학 연구부 등에서 전임 교원 및 석·박사 학위 소지자를 주축으로 활발한 연구 활동을 수행하고 있다.

향후 예상할 수 있는 기초과학의 획기적인 발전 가능성에 비추어 볼 때 기초과학 연구지원의 증대는 본연구소의 연구수행 및 연구지원 활동에 보다 크게 기여할 것이다. 본 숙명여자대학교는 창학 100주년을 기점으로 새로운 대학으로 거듭나고자 학내 제반 분야에서의 개혁에 강한 의지를 표명하고 지속적으로 학교의 모든 분야에 대한 과감한 투자 및 개혁을 가속화 해 온 바 있다. 특히, 연구 분야에 대한 과감한 체제 개혁 및 투자를 한 결과, 행정 조직으로서 연구지원팀을 신설하여 연구 활동 및 국내외 연구교류를 지원토록 하고, 교내 예산의 일부를 학술연구비로 할당 배분하고, 연구업적 평가제를 도입하여 모든 연구관련 자료를 데이터베이스화하고 있다. 또한, 본 연구소는 기자재 확충 계획을 수립하여 추진하고 있으며 지난 수년 동안 교비 및 교육부 등으로부터 재정지원을 받아 최첨단 연구기자재를 확충하여 활발한 연구 활동을 진행 중에 있다.

역대 소장

1985년 9월 19일 창립

1985년 11월 - 1988년 10월 : 초대 소장 : 박 애 주 교수

1988년 11월 - 1990년 10월 : 2 대 소장 : 박 영 자 교수

1990년 11월 - 1992년 10월 : 3 대 소장 : 장 윤 경 교수

1992년 11월 - 1993년 10월 : 4 대 소장 : 박 영 자 교수

1993년 3월 - 1995년 2월 : 5 대 소장 : 이 장 로 교수

1995년 3월 - 1996년 9월 : 6 대 소장 : 오 성 담 교수

1996년 10월 - 2000년 9월 : 7 대 소장 : 민 경 희 교수

2000년 10월 - 2004년 10월 : 8 대 소장 : 노 광 현 교수

2004년 10월 - 2008년 8월 : 9 대 소장 : 김 재 성 교수

2008년 9월 - 2010년 8월 : 10대 소장 : 정 혁 교수

2010년 9월 - 2012년 8월 : 11대 소장 : 이 원 춘 교수

2012년 9월 - 현재 : 12대 소장 : 황 선 영 교수

Contents

◆ 나노물리학과

Magnetic phases of cobalt atomic clusters on tungsten Jae-Sung Kim	3
Electronic Structure of $p(2 \times 3)$ Ag Films on Si(100) Jae-Sung KIM	5
Role of nonlinearities and initial prepatterned surfaces in nanobead formation by ion-beam bombardment of Au(001): Experiments and theory J.-S. Kim	7
Reduction of charge fluctuation energies in ultrathin NiO films on Ag(001) J.-S. Kim	9
Amorphous Ferromagnetic CoSiB/Pd Multilayer with Perpendicular Magnetic Anisotropy Haein YIM	11
Thermal and Magnetic Properties of Melt-spun Co-Fe-B-Si-(Cr/Ta) Amorphous Alloys Haein Choi YIM	13
Time-dependent magnetization reversal in amorphous CoSiB/Pd multilayers with perpendicular magnetic anisotropy H. I. Yim	15
<i>Ab initio</i> Calculations on the Effect of Mn Substitution in the κ -carbide Fe_3AlC Hanchul KIM	17
Electronic Structures of a Zn Vacancy on the $\text{ZnO}(10\bar{1}0)$ Surface: Density Functional Theory Calculations Hanchul KIM	19
Comment on “Atomistic Picture of Charge Density Wave Formation at Surfaces” Hanchul Kim	21
Electron doping limit in Al-doped ZnO by donor-acceptor interactions Hanchul Kim	23
Cation composition effects on electronic structures of In-Sn-Zn-O amorphous semiconductors Hanchul Kim	25

스핀-궤도 각운동량 상호작용의 구조 최적화에 대한 효과; 비스무스 텔루라이드의 제일원리 계산의 경우 김미영	27
--	----

Electronic Structure and Magneto-optical Properties of Co_2MnX Alloys Where $X=\text{Ge, Sn and Pb}$: a First-principles Investigation in the LDA + U Approach Miyoung KIM	29
--	----

The metal-insulator phase transition in the strained GdBiTe_3 Miyoung Kim	31
---	----

◆ 화학과

Structural heterogeneity in familial Alzheimer's disease mutants of amyloid-beta peptides Sihyun Ham	35
---	----

Conformational Entropy of Intrinsically Disordered Protein Sihyun Ham	37
--	----

Assessing the influence of solvation models on structural characteristics of intrinsically disordered protein Sihyun Ham	39
---	----

Distinct Role of Preferential Hydration in Protein Aggregation Sihyun Ham	41
--	----

Vanadium-induced decrease in the magnetization of a $V_x\text{Co}_{1-x}$ alloy film on the W(110) surface Hangil Lee	43
---	----

Inter-row Adsorption Configuration and Stability of Threonine Adsorbed on the Ge(100) Surfaces Hangil Lee	45
--	----

Annealing Effects after Nitrogen Ion Casting on Monolayer and Multilayer Graphene Hangil Lee	47
---	----

The adsorption selectivity of the functional groups of dithiothreitol and 1,4-butanedithiol 2,3-diamino on the Ge(100) surface Hangil Lee	49
--	----

Adsorption Sequence of Multifunctional Groups: A Study on the Reaction Pathway and the Adsorption Structure of Homocysteine on the Ge(100) Surface Hangil Lee	51
--	----

Variation of adsorption geometries by the influence of nucleophilicity among p-CPA, p-TPA, and p-NPA on Ge(100) Hangil Lee	53
---	----

Confirmation of the coexistence of two tautomers of 2-mercaptothiazoline on the Ge(100) surface Hangil Lee	55
Quantum confinement-induced tunable exciton states in graphene oxide Hangil Lee	57
Coverage Dependent Variation of the Adsorption Structure of 2-Thiophenecarboxaldehyde on the Ge(100)-2×1 Reconstructed Surface Hangil Lee	59
The aniline-to-azobenzene oxidation reaction on monolayer graphene or graphene oxide surfaces fabricated by benzoic acid Hangil Lee	61
Mg Capping Inside <i>p</i> -Tert-butylcalix[4]arene Adsorbed on a Ge(100) Surface Hangil Lee	63
Three Distinct Equilibrium States via Self-Assembly: Simple Access to a Supramolecular Ion-Controlled NAND Logic Gate Jung Su Park	65
Pd(II)-catalyzed direct C-H acylation of <i>N</i> -Boc hydrazones with aldehydes: one-pot synthesis of 1,2-diacylbenzenes Jung Su Park	67
Modulation of Electronics and Thermal Stabilities of Photochromic Phosphino-Aminoazobenzene Derivatives in Weak-Link Approach Coordination Complexes Jung Su Park	69

◆ 수학과

SELF-RECIPROCAL POLYNOMIALS WITH RELATED MAXIMAL ZEROS SEON-HONG KIM	73
TRANSLATION HYPERSURFACES WITH CONSTANT CURVATURE IN SPACE FORMS KEOMKYO SEO	75
An optimal investment, consumption, leisure, and voluntary retirement problem with Cobb-Douglas utility: Dynamic programming approaches Yong Hyun Shin	77
OPTIMAL CONSUMPTION AND INVESTMENT PROBLEM WITH REGIME-SWITCHING AND CARA UTILITY YONG HYUN SHIN	79

◆ 통계학과

Online retailer reputation and consumer response: examining cross cultural differences Bu Yong Kim	83
소지역 실업률의 패널추정을 위한 일반화커널추정방정식 김영원	85
2012년 총선 출구조사 예측오차 분석 김영원	87
Efficiency of Variance Estimators for Two-stage PPS Systematic Sampling Young-won Kim	89
Accuracy of linear approximation for fitted values in nonlinear regression Myung-Wook Kahng	91
Log-density Ratio with Two Predictors in a Logistic Regression Model Myung Wook Kahng	93
CENTRAL LIMIT THEOREMS FOR REDUCED U -STATISTICS UNDER DEPENDENCE AND THEIR USEFULNESS SUN YOUNG HWANG	95
A Graphical Improvement in Volatility Analysis for Financial Series Sun Young Hwang	97
Martingale Estimating Functions for Stochastic Processes: A Review Toward a Unifying Tool S.Y. Hwang	99
두 가지 유형의 보험청구가 있는 확산과정 리스크 모형의 파산확률 이의용	101
Transient and Stationary Analyses of the Surplus in a Risk Model Eui Yong Lee	103
Yukmijihwang-tang for the treatment of xerostomia in the elderly: study protocol for a randomized, double-blind, placebo-controlled, two-center trial Inkwon Yeo	105
Availability of Tongue Diagnosis System for Assessing Tongue Coating Thickness in Patients with Functional Dyspepsia Inkwon Yeo	107
The Effects of <i>Banha-sasim</i> -tang on Dyspeptic Symptoms and Gastric Motility in Cases of Functional Dyspepsia: A Randomized, Double-Blind, Placebo-Controlled, and Two-Center Trial Inkwon Yeo	109

Prediction of the Number of Food Poisoning Occurrences by Microbes In-Kwon Yeo	111
Effect of Korean Herbal Medicine Combined with a Probiotic Mixture on Diarrhea-Dominant Irritable Bowel Syndrome: A Double-Blind, Randomized, Placebo-Controlled Trial Inkwon Yeo	113
An Empirical Characteristic Function Approach to Selecting a Transformation to Symmetry In-Kwon Yeo	115
Nonparametric Two-Sample Tests of Longitudinal Data with Termination Events Yang-Jin Kim	117
Statistical Analysis of Bivariate Recurrent Event Date with Incomplete Observation Gaps Yang-Jin Kim	119
Erythrocyte sedimentation rate and anaemia are independent predictors of survival in patients with clear cell renal cell carcinoma K kim	121
Exploring <i>Trans</i> -acting regulators of gene expression associated with metabolic syndrome: a coupled application of factor analysis and linkage analysis Kyunga Kim	123
Identification of multiple gene-gene interactions for ordinal phenotypes Kyunga Kim	125
Functional Implication of β -Carotene Hydroxylases in Soybean Nodulation Kyunga Kim	127
Prediction of type 2 diabetes in women with a history of gestational diabetes using a genetic risk score Kyunga Kim	129
Quality of Fit Measures in the Framework of Quantile Regression HOHSUK NOH	131
Assessing model adequacy in possibly misspecified quantile regression Hohsuk Noh	133
Copula-Based Regression Estimation and Inference Hohsuk Noh	135

◆ 컴퓨터과학과

다중 인프라 기반의 공력 설계 교육을 위한 과학 워크플로우 통합 환경 김윤희	139
An Allocation and Provisioning Model of Science Cloud for High Throughput Computing Applications Yoonhee Kim	141
하이브리드 클라우드에서 과학 응용을 위한 정책 기반의 적응형 작업 스케줄링 김윤희	143
과학 계산 응용 실행을 위한 하이브리드 클라우드에서의 SLA기반 VM 오토-스케일링 기법 김윤희	145
하둡을 이용한 소셜네트워킹의 TV광고효과 분석 시스템 설계 김윤희	147
워크플로우 응용을 위한 에너지 효율적인 가상 자원 할당 기법 김윤희	149
A SLA driven VM Auto-Scaling Method in Hybrid Cloud Environment Yoonhee Kim	151
협업필터링의 희박성 문제 개선을 위한 통합 유사도 모델 유석중	153
Skyline queries on keyword-matched data Ki Yong Lee	155
Efficient processing of multiple continuous skyline queries over a data stream Ki Yong Lee	157
전자상거래 데이터의 실시간 분석을 위한 데이터 스트림과 다수 릴레이션 간의 효율적인 연속 조인 처리 기법 이기용	159
웹소켓을 이용한 실시간 커뮤니케이션 서비스 시스템 구현 이기용	161
공개 데이터의 프라이버시 안정성: 진료정보 통계자료 HIRA-NPS 2011 사례 분석 이기용	163
맵리듀스에서 집계 질의 스트림의 효율적인 처리 기법 이기용	165
새로운 형태의 사용자 인터페이스를 제공하는 안드로이드용 이미지 합성 애플리케이션 개발 이기용	167

어학 공부와 전자책 읽기 기능을 제공하는 안드로이드용 한영 텍스트-음성 변환 애플리케이션 개발 이기용	169
---	-----

POWER-EFFICIENT CONFIGURATION CACHE STRUCTURE FOR COARSE-GRAINED RECONFIGURABLE ARCHITECTURE YOONJIN KIM	171
---	-----

◆ 멀티미디어과학과

Mobile Advertising Service Based on 4R Mode Yong-Ik Yoon	175
---	-----

스마트 학습 공간 구성을 위한 Timed Button 기반의 다중스크린 동기화 기법 윤용익	177
---	-----

A model of Self-reliant Mobile Cloud for Collaboration Services YongIk Yoon	179
--	-----

Abnormal Behavior Tracking Model Based on Sensors of Mobile User Yong-Ik Yoon	181
--	-----

모바일 환경에서 감성을 기반으로 한 영상 합성 기법 연구 및 개발 윤용익	183
---	-----

Elastic Service Model for Smart Learning based on Cloud Environment YongIk Yoon	185
--	-----

A Study on CCTV-Based Dangerous Behavior Monitoring System Hwa-Jin Park, Yong-Ik Yoon	187
--	-----

다중 객체의 위험 행동 감시 시스템 연구 박화진	189
-------------------------------------	-----

객체 추적을 통한 이상 행동 감시 시스템 연구 박화진	191
--	-----

객체 애니메이션 제작을 위한 스케치 기반의 사용자 인터페이스 기법 임순범	193
---	-----

A sketch-based interface to modify and reproduce motion sequences Soon-Bum Lim	195
---	-----

독서장애인을 위한 음성 도서 어노테이션 검색 기법 임순범, 이종우	197
---	-----

3DTV에서 스테레오 웹 콘텐츠 표현을 위한 CSS3 사양 확장 및 에뮬레이터 연구 임순범	199
An Annotation System for Reading-disabled People on Mobile e-book Soon-Bum Lim, Jongwoo Lee	201
Design a Motion Library with Kinetic Typography for Messengers Soon-Bum Lim	203
POI 데이터 검색을 위한 문자열 유사도 측정 기법 이종우	205
효율적인 집합-기반 POI 데이터 검색 알고리즘 이종우	207
사진 콘텐츠 분류를 위한 태그 클러스터링 기법 및 태그 추천 이종우	209
부호 분포 변동성에 기반한 한국어 음소 ‘ㄸ’, ‘ㅈ’, ‘ㅊ’ 음성인식 이종우	211
안드로이드 NDK(Native Development Kit)를 이용한 3D 큐브 게임 이식 및 구현 이종우	213
중복글자 구분을 지원하는 집합 기반 POI 검색 알고리즘 구현 이종우	215
중복글자 구분을 지원하는 집합 기반 POI 검색 알고리즘 설계 이종우	217
집합 기반 POI 검색 알고리즘을 활용한 In-시스템 검색 엔진 설계 이종우	219
개인 맞춤형 장학금 검색을 위한 알고리즘 연구 이종우	221
HTML5를 이용한 멀티미디어 엽서 하이브리드 어플리케이션 구현 이종우	223
Parameter Optimization for Local Structure-Based Fingerprint Alignment Jongwoo Lee	225
파형의 시퀀스 탐지에 기반한 모음 ‘ㅏ’ 인식 이종우	227

나노물리학

Magnetic phases of cobalt atomic clusters on tungsten

Pavel V Lukashev¹, Ji-Hyun Kim², Seolun Yang², Jae-Sung Kim²,
Xumin Chen¹, Geoffrey Rojas¹, Jan Honolka³, Ralph Skomski¹,
Axel Enders¹ and Renat F Sabirianov⁴

¹ Department of Physics and Astronomy & Nebraska Center for Materials and Nanoscience, University of Nebraska, Lincoln, NE 68588, USA

² Department of Physics, Sookmyung Womens University, Seoul, 140-741, South Korea

³ Max Planck Institute for Solid State Physics, Heisenbergstr. 1, Stuttgart, D-70569, Germany

⁴ Department of Physics, University of Nebraska at Omaha, Omaha, NE 68182, USA

E-mail: pavel.lukashev@unl.edu

Received 21 September 2012, in final form 26 November 2012

Published 10 December 2012

Online at stacks.iop.org/JPhysCM/25/036003

Abstract

First-principle calculations are employed to show that the magnetic structure of small atomic clusters of Co, formed on a crystalline W(110) surface and containing 3–12 atoms, strongly deviates from the usual stable ferromagnetism of Co in other systems. The clusters are ferri-, ferro- or non-magnetic, depending on cluster size and geometry. We determine the atomic Co moments and their relative alignment, and show that antiferromagnetic spin alignment in the Co clusters is caused by hybridization with the tungsten substrate and band filling. This is in contrast with the typical strong ferromagnetism of bulk Co alloys, and ferromagnetic coupling in Fe/W(110) clusters.

(Some figures may appear in colour only in the online journal)

Magnetic nanostructures often exhibit spin configurations that do not naturally occur in bulk materials. For example, iron is a prototypical ferromagnet in its bulk *bcc* phase but exhibits low-spin–high-spin transitions, antiferromagnetism, complex spin structures and even non-collinear magnetic order in thin films and other nanostructures [1–5]. This diversity is especially pronounced for dense-packed *fcc*-like atomic environments [6–8]. By contrast, metallic cobalt tends to exhibit stable ferromagnetism (FM), even in dense-packed atomic arrangements [9] and in nanostructures [10].

The magnetic moment and its average, the magnetization, are the most fundamental magnetic properties. They depend on the atomic environment and can be tuned by alloying and nanostructuring, thereby creating [11], enhancing [12] or reducing [13] the magnetization and yielding phenomena such as high-spin–low-spin transitions [2]. The coupling between magnetic moments is antiferromagnetic (AFM) in a number of materials, especially if the structures contain atoms from the middle of the transition metals series, such as Cr, Mn and Fe. Recent examples of magnetic order that show characteristic properties as a result of nanostructuring are the complex

magnetic ground state in Fe monolayers on Ir(111) [14], antiferromagnetism in Fe layers on W(100) [3], and the non-collinear spin alignment in Mn on W(110) [5]. Electronic hybridization with the substrate often plays an important role, adding complexity to effects such as the non-collinear spin alignment in Cr trimers as a result of geometric frustration of AFM interactions [15].

The frequent occurrence of non-ferromagnetic spin configurations in Cr, Mn, and to a lesser extent in Fe is caused by the general trend towards AFM interactions in metals with nearly half-filled d-bands. By contrast, late iron-series transition-metal elements, especially Co, tend to exhibit robust ferromagnetism in strained systems, nanostructures, and alloys. A good example of the latter is the hexagonal intermetallic Y_2Co_{17} , which maintains a high Curie temperature $T_C = 894^\circ\text{C}$, as compared to $T_C = 44^\circ\text{C}$ in the isostructural alloy Y_2Fe_{17} [9]. Cobalt-poor alloys with non-magnetic metals such as W and Y tend to lose their magnetic moments but become non-magnetic rather than ferrimagnetic or AFM [16]. However, the monolayer

of Co on W(001) surface was surprisingly predicted to have antiferromagnetic ground state [17].

In the present work, we demonstrate from first principles that atomic cobalt clusters of particular size and geometry on the W(110) surface may exhibit AFM coupling, not present in layers of Co on W(110) [18–20] and on 2D tungsten carbide layers [21], and especially in striking contrast to the FM coupling in bulk Co, ultrathin films and nanoclusters of Co on various *fcc* and *bcc* substrates [10]. For comparison, we also show that such a behavior is absent in the Fe clusters of the similar geometry.

Figure 1(a) shows the structural model used for the calculations. We consider Co_N atomic clusters containing $N = 3$ –12 Co atoms on *bcc*(110) tungsten supercell. We perform our calculations using the projector augmented wave (PAW) method [22], implementation of PAW in Vienna *ab initio* simulation package (VASP) [23] within Perdew–Burke–Ernzerhof (PBE) generalized gradient approximation (GGA) [24] of the density functional theory (DFT). The simulations were performed using periodic boundary conditions in the 3 layer slab geometry. In particular, the unit cell for Co_4 clusters has lattice parameters $a = 8.94 \text{ \AA}$, $b = 6.32 \text{ \AA}$, $c = 20 \text{ \AA}$ (this corresponds to 2 by 2 translations of W(110) unit cell), with vacuum layer of $\sim 12 \text{ \AA}$ in *c*-direction. The unit cell for all other clusters has lattice parameters $a = 8.94 \text{ \AA}$, $b = 12.64 \text{ \AA}$, $c = 20 \text{ \AA}$. We use $6 \times 6 \times 1$ *k*-point sampling for smaller and $4 \times 4 \times 1$ *k*-point sampling for larger cells, and the Blöchl’s tetrahedron integration method [25]. We relax the atomic positions of all atoms in the unit cells until Hellmann–Feynman forces were less than $0.005 \text{ eV \AA}^{-1}$. Based on the relaxed positions, we use second-order perturbation theory and the Green’s function approach [7, 26] to calculate the exchange interaction parameters. The effective exchange interaction is calculated from

$$J_0 = -\frac{1}{4\pi} \sum_{LL'} \text{Im} \int_{-\infty}^{E_F} \{ \Delta_l^0 (T_{LL'}^{00\uparrow} - T_{LL'}^{00\downarrow}) + \Delta_l^0 T_{LL'}^{00\uparrow} \Delta_{l'}^0 T_{LL'}^{00\downarrow} \} dE$$

where $\Delta = t_{0\uparrow}^{-1} - t_{0\downarrow}^{-1}$, and t^{-1} is the inverse single-site scattering matrix and T is the matrix of the scattering-path operator.

Figure 1(b) summarizes the calculated results for the net magnetic moments $\langle m \rangle$ of Co and Fe clusters as a function of the number N of atoms per cluster. The magnetization of the rhombic Co_4 and Co_{12} clusters is strongly reduced by AFM exchange inside the clusters. In all clusters, the spin density is localized at the Co atoms but slightly extended towards the nearest W neighbors. This indicates the hybridization of Co and W states, which, however, does not yield a significant spin polarization in the W substrate. The present calculations indicate the AFM exchange in the system, although the sign and strength of the exchange coupling vary with cluster configuration, and we also obtain FM (Co_3 , Co_5) and non-magnetic (rhomboidal Co_4 , Co_6 , Co_8) spin structures. Most striking is the transition from ferrimagnetic Co_4 to non-magnetic Co_6 , where the atomic arrangement with

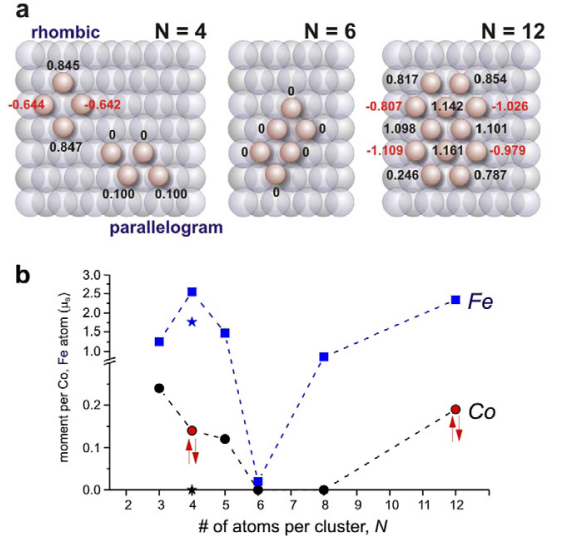


Figure 1. Structure and magnetism of Co and Fe clusters on W(110): (a) Assumed arrangement of the atoms in clusters Co_4 , Co_6 and Co_{12} , with calculated moments for each atom. Moments labeled in black and red correspond to spins pointing into and out of the surface plane, respectively. (b) Calculated magnetic moments per Co (black circles) and Fe atoms (blue squares). Shown numbers also include contributions from the small induced moment on W. Red symbols and arrows indicate the clusters that exhibit AFM intra-cluster exchange. Stars are used to denote the moments in tetramers by parallelogram configuration.

respect to the substrate is similar, and the different behavior of the rhombic and parallelogram Co_4 clusters (see figure 1(a)).

Next we tested our results for the different configurations of the initial magnetic moments. In particular, for the rhombic Co_4 system we performed calculations for the different sets of initial moments (ranging from 0 to $5 \mu_B$, both ferro- and antiferro-magnetic), including non-spin-polarized case. The lowest energy case corresponds to the ferrimagnetic structure reported in the paper. In particular, the non-magnetic state has $\sim 37 \text{ meV}$ higher energy than the ferrimagnetic structure. All our calculations are performed for collinear magnetic states.

The calculated energy difference between parallel and antiparallel spin arrangements for the rhombic Co_4 configuration, chosen in the following as a case example, is $\Delta E = 240 \text{ meV}$. The Heisenberg exchange coupling parameter J_{ij} for nearest Co neighbors, estimated from this value, is 27 meV . The magnitude of the interatomic exchange constants J_{ij} is comparable to that in bulk Co, which is characterized by an interatomic exchange parameter $J_0 = \sum_j J_{0j}$ of 15 meV and a Curie temperature $T_C = 1117^\circ\text{C}$, but the sign of J_{ij} is negative (AFM). To determine the thermal stability of the spin structures of the tetramers, we have performed Monte Carlo simulations using the Metropolis scheme. The simulations show that the intra-cluster spin correlations are very stable in fields up to 10 T and at temperatures below 100 K .

Figure 2 shows densities of states (DOS) per Co atom for the ferrimagnetic rhombic Co_4 clusters on W(110) and

Electronic Structure of $p(2 \times 3)$ Ag Films on Si(100)

Soon-Ki KIM and Jae-Sung KIM*

Department of Physics, Sook-Myung Women's University, Seoul 140-742, Korea

Geunseop LEE

Department of Physics, Inha University, Incheon 402-751, Korea

Jun NOGAMI

*Department of Materials Science and Engineering,
University of Toronto, Toronto, Ontario M5S 1A4, Canada*

Ki-jeong KONG

Advanced Materials Division, Korea Research Institute of Chemical Technology, Taejeon 305-600, Korea

Byung Deok YU

Department of Physics, University of Seoul, Seoul 130-743, Korea

Doyeol AHN

IQUIPS, University of Seoul, Seoul 130-743, Korea

(Received 24 July 2012, in final form 12 October 2012)

The electronic structure of $p(2 \times 3)$ Ag films on Si(100) is studied by using electron energy loss spectroscopy (EELS) and scanning tunneling spectroscopy (STS). We observe three energy loss peaks with their loss energies around 1.28, 1.63, and 4.0 eV, respectively. They are assigned to the interband transition in the $p(2 \times 3)$ islands, the interface plasmon between Si(100) and Ag crystallites formed concomitantly with the $p(2 \times 3)$ islands, and a combination of the surface plasmon and interband transition of the Ag crystallites, respectively. STS over the $p(2 \times 3)$ surface also reveals a band gap around 1 eV in good agreement with the EELS observation of the interband transition. The present observation of the band gap suggests that the $p(2 \times 3)$ surface is semiconducting. Furthermore, we examine two models previously proposed for the $p(2 \times 3)$ surface against the present experimental observation by *ab initio* band structure calculations.

PACS numbers: 73.20.Mf, 71.45.Gm

Keywords: Ag, Si(100), Electronics structure, EELS, First principles calculations

DOI: 10.3938/jkps.62.86

I. INTRODUCTION

Ultrathin metal films on semiconductor surfaces have been widely studied especially with their possible applications to microelectronics in mind. From the viewpoint of fundamental research, they also serve as a model for two dimensional electronic systems [1], if a Schottky barrier is formed between the metal film and the substrate, and if the interface is sharply defined. Ag film on Si surface meets both requirements [2] and has drawn vast attention of researchers from many different disciplines.

Ag grows on silicon in the Stranski-Krastanow mode,

with 3D Ag crystallites coexisting with a Ag terminated 2D surface. At elevated growth temperatures, the 2D surface shows an Ag induced surface reconstruction. Most of the prior work on Ag growth on silicon has focused on the Ag/Si(111) system, and in particular, there is a large body of work concerned with the $(\sqrt{3} \times \sqrt{3})$ Ag/Si(111) surface reconstruction. The determination of the structure of this Ag terminated Si surface has taken many years and the integration of theory and data from many different surface analytical techniques [3–6].

In contrast, comparatively little work has been done on the Ag/Si(100) system. In this growth system, two surface reconstructions seem to be related; $p(2 \times 3)$ and $c(2 \times 6)$, and of these two, the $p(2 \times 3)$ is the more stud-

*E-mail: jskim@sookmyung.ac.kr

ied. Although a range of structure models have been proposed for the $p(2 \times 3)$ surface, two models are strongly supported by prior experimental work. However, there is a major discrepancy between these two models with respect to their prediction of the surface electronic structure.

Michely *et al.* [7,8] proposed a model (Michely model) with three Ag atoms per surface unit cell and predicted a metallic surface according to simple electron counting. On the other hand, Shivaprasad *et al.* [9] contend from X-ray photoelectron diffraction (XPD) work that there are an even number of Ag atoms in a surface unit cell, and the surface is semi-conducting. Recently, they studied the system by angle resolved ultraviolet photoelectron spectroscopy (ARUPS) and, in support of their model observed no states at the Fermi level [10] (Kono model).

In the present work, we study the electronic structure of $p(2 \times 3)$ Ag grown on Si(100) by using electron energy loss spectroscopy (EELS) and scanning tunneling spectroscopy (STS). We also investigate the electronic structures of the $p(2 \times 3)$ surface by *ab initio* band structure calculations for both Michely and Kono models. This article is comprised of four sections. In the following section is given an overview of our experimental procedures. Then, the results and discussion will be detailed. A summary and conclusions will be given in the last section.

II. EXPERIMENT

EELS experiments were performed in an ultrahigh vacuum chamber with a base pressure of 4×10^{-11} Torr. The Si(100) sample (p -doped with resistivity $\rho \sim 20 \Omega \text{cm}^{-1}$) was cleaned by repeated resistive heating up to 1200 °C. During the cleaning procedure, we kept the chamber pressure below 2×10^{-9} Torr to avoid possible silicon carbide formation. The cleanliness of the sample was checked by using vibration spectroscopy with a commercial EELS spectrometer (LK2000) [11]. The structural order of the clean Si was monitored by low energy electron diffraction (LEED). We found sharp LEED spots for double domain, $p(2 \times 1)$ Si(100) with low background intensity.

For Ag deposition, a Ag wire (99.999 % purity) was wound around a tungsten (99.99 % purity) filament, and was thermally evaporated to a Si substrate held at 500 °C. The deposition rate of Ag was 1 monolayer (ML) per 4 minutes. The thickness of the Ag film was estimated by using a quartz microbalance placed next to the sample. During the Ag deposition, the chamber pressure was kept below 4×10^{-10} Torr.

The electronic excitation spectra were obtained by using the same EELS spectrometer as that employed for vibration spectroscopy. The spectrometer was operated intentionally at low resolution, 20 meV, to increase its signal-to-noise ratio. The angle of incidence, Θ_{in} , was in the range from 30° to 60° from the surface normal,

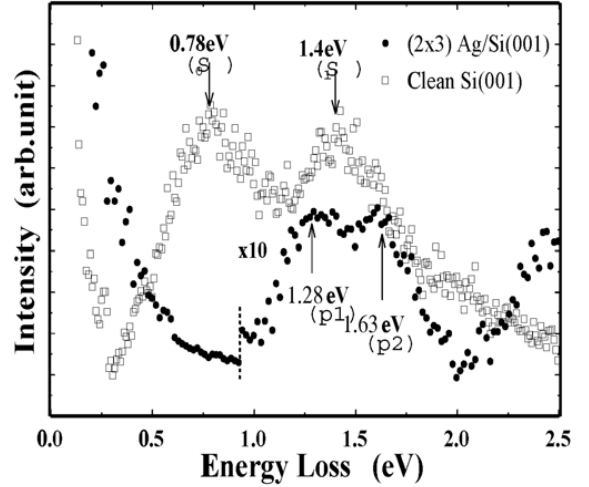


Fig. 1. Electron energy loss spectra for both clean Si(001) (open square) and 1ML Ag film grown on Si(001) at 500 °C (filled circle). The p1 peak is clearly visible for the 2.5 ML film showing a well-defined $p(2 \times 3)$ LEED pattern. On the contrary, p1 peak is not discernable from the p2 for 9ML film that shows a very poorly-defined $p(2 \times 3)$ LEED pattern. The primary energy is 11.3 eV. Both spectra were obtained in specular geometry at an incidence angle of 40 °C from the surface normal.

and all the presented spectra were taken in the specular geometry.

Scanning tunneling spectroscopy was done in a separate chamber equipped with LEED and scanning tunneling microscope (STM). The $p(2 \times 3)$ Ag surfaces were prepared by thermal deposition of Ag on Si(100) at room temperature and subsequent annealing at 450 K. All STM and LEED measurements were done at room temperature.

III. RESULT

For completeness of the presentation of the EELS results, electronic excitations from a clean Si(100) surface are presented first. In Fig. 1 is an energy loss spectrum from clean Si(100). Two energy loss peaks are found: S_0 at 0.7 eV and S_1 at 1.4 eV. The S_0 peak is commonly assigned to the electronic excitation from the bulk valence band to an empty surface band, while the S_1 is from an occupied surface band to the conduction band [12]. Except for these two peaks, no pronounced loss feature is observed up to a loss energy of 5 K eV.

In Fig. 2 is shown energy loss spectra of Ag films grown on Si(100) at 500 °C. The nominal thicknesses of the Ag films range from 0.5 ML to 3 ML in 0.5 ML increments. In this thickness regime, LEED shows $p(2 \times 3)$ patterns with varying degrees of order. The spectra show quite

Role of nonlinearities and initial prepatterned surfaces in nanobead formation by ion-beam bombardment of Au(001): Experiments and theory

J.-H. Kim,¹ J.-S. Kim,^{1,*} J. Muñoz-García,² and R. Cuerno^{2,†}¹*Department of Physics, Sook-Myung Women's University, Seoul 140-742, Korea*²*Departamento de Matemáticas and Grupo Interdisciplinar de Sistemas Complejos, Universidad Carlos III de Madrid, 28911 Leganés, Spain*

(Received 10 July 2012; revised manuscript received 15 October 2012; published 25 February 2013)

Au(001) surfaces that have been prepatterned into a rippled morphology develop one-dimensional nanodot arrays (nanobeads) selectively along the ripples when bombarded with energetic ions at an angle that is normal to the average surface orientation. By quantifying the shape and morphology of these arrays, we show experimentally and by numerical simulations of an extended Kuramoto-Sivashinsky equation that the degree of one-dimensional order of the nanobeads can be optimized by considering initial rippled surfaces with various wavelength and roughness values. Our simulations employ physical units and use the experimental topographies as initial conditions. Such nonideal shapes are key to elucidating the influence of nonlinear effects (like conformal interface motion and local redeposition) since the early stages of the dynamics for these prepatterned systems. In spite of the fact that the evolution of the surface morphology becomes far from trivial under these circumstances, our continuum model is able to reproduce the experimental results quantitatively, in contrast to relevant alternative models in the context of surface nanopatterning by ion-beam bombardment.

DOI: [10.1103/PhysRevB.87.085438](https://doi.org/10.1103/PhysRevB.87.085438)

PACS number(s): 79.20.Rf, 81.16.Rf, 68.35.Ct, 05.45.-a

I. INTRODUCTION

For the last few decades, ion-beam sputtering (IBS) has proven to be a promising tool to fabricate highly ordered, nanoscale patterns on the surfaces of diverse materials via self-organization.^{1–3} Under oblique incidence of the ion beam, for example, IBS fabricates periodic ripple patterns,⁴ while under normal incidence it can generate hexagonal^{5,6} or square-symmetric^{7,8} patterns of dots or holes.

Development of such ordered patterns has been explained theoretically by a number of models. Bradley and Harper (BH) proposed a linear evolution equation for the height of the eroded target in which pattern formation by IBS results from counteracting processes such as roughening due to erosion by the ion beam and smoothening by diffusion of thus-created adatoms and vacancies.⁹ This model reproduces some features of the ripple pattern,² although a number of nonlinear continuum equations were later introduced to improve upon its limitations.^{7,10–12} Moreover, the interplay between sputtering and surface diffusion has been recently challenged experimentally as the main physical mechanism for pattern formation in targets such as semiconductors that become amorphous under irradiation.¹³ Still, the BH mechanism is expected to dominate for metallic targets in the so-called erosive regime in which it is the ion beam, rather than crystalline anisotropies in surface diffusion, that controls the pattern formation process.^{1,2}

Recently, multiple-ion-beam sputtering, in which two or more beams are projected onto the target surface, has been theoretically proposed to fabricate more sophisticated patterns via superposition of those produced by each ion beam.^{14,15} Dual-ion-beam sputtering, the *simultaneous* projection of two ion beams at a grazing angle, but perpendicular to each other in azimuth, has been experimentally shown to generate square-symmetric patterns of nanodots/holes.¹⁶ Also *sequential* ion-beam sputtering (SIBS) has been reported, in which the direction of an ion beam with respect to the

surface is changed sequentially. Two types of SIBS have been examined; after forming an initial ripple pattern in one direction by sputtering at an oblique angle, the rippled surface is subsequently sputtered after changing the azimuthal angle by 90°,^{17–19} or normal to the surface.²⁰ The patterns generated experimentally by SIBS disprove theoretical predictions¹⁵ on the mere superposition of the patterns formed by each beam, although the latter has been observed in Monte Carlo simulations of discrete models of SIBS for an admittedly short simulation time window.²¹

Most of the experimental SIBS patterns^{17–19} are similar to those that can be produced by a single beam. However, SIBS as in Ref. 20 does produce a novel nanostructure, such as a nanobead pattern, that is a highly ordered one-dimensional array of nanodots, formed along each initial ripple when these are subject to irradiation at normal incidence.²⁰ The nanobead morphology thus demonstrates the capability of SIBS to fabricate elaborate nanostructures via the sequential formation of basic patterns, in a form of hierarchical self-assembly.²² This complex pattern formation process is driven by the competition of two length scales, namely, the wavelength of the initial ripple and the nanobead size. In particular, it provides a touchstone against which models of IBS nanopatterning can be examined. The extended Kuramoto-Sivashinsky (eKS) model^{7,12} has been found to reproduce most of the observed experimental results, such as formation of the nanobead structure and its preservation during SIBS, in contrast to other models, such as the KS and BH equations.²⁰ Hence, the additional nonlinear effects present in the eKS model are found to be essential in order to explain the evolution of nanobead morphologies during SIBS.

In this work, we study experimentally and theoretically how the shape of the initial ripple pattern affects the formation and dynamics of the nanobead structure by IBS normal to the initial rippled surface. While in Ref. 20 a single rippled condition was explored, here we address the interplay between the two relevant length scales by performing experiments on

several different initial conditions. The degree of ordering, and thus the quality, of the nanobead structure is seen to depend on the properties of the initial surface. Nevertheless, for large enough ion doses, the morphologies become similar, largely independent of those features. An additional difference of the present work with respect to Ref. 20 is to perform simulations of the eKS model employing physical units and for a range of different nonflat initial conditions. This enables successful quantitative comparison with experiments and allows us to develop a unified picture of the process that self-consistently elucidates the observed temporal evolution of the nanobead patterns.

II. EXPERIMENTS

All our ion-beam-sputtering experiments have been carried out in a custom-built ultrahigh vacuum chamber with base pressure of about 5×10^{-10} Torr. For the fabrication of the initial ripple patterns, IBS was performed along the densely packed [110] direction of the Au(001) targets at a polar incidence angle of 72° with respect to the surface normal. The ion energy ε and partial pressure of Ar^+ P_{Ar} were 2 keV and 1.2×10^{-4} Torr, respectively, generating an ion flux $f \simeq 0.3 \text{ ions nm}^{-2} \text{ s}^{-1}$. The largest ion fluence achieved, defined as the ion flux multiplied by the total sputter time, was 4500 ions nm^{-2} . In order to vary the mean wavelength and roughness of the initial ripple pattern, we sputtered substrates at different temperatures (Table I), while keeping other conditions fixed. For higher substrate temperatures, the mean wavelength of the initial ripple pattern becomes larger. The temperature was measured by a K -type thermocouple slightly away from the sample and can be a little lower than the real one. *In situ* x-ray photoelectron spectroscopy (XPS) of sputtered Au samples reveals no traceable amount of contaminants, implying little unintended deposition of impurities by sputtering (figures not shown). The initial ripple patterns were then analyzed *ex situ* by an atomic force microscope (AFM) in the contact mode.

Table I summarizes the features of the rippled surfaces thus generated, such as the mean ripple wavelength λ_R^i and the global surface roughness or surface width W^i . Here, the superscript i refers to initial and the subscript R refers to the ripple pattern. The roughness $W(t)$ is defined as $W(t) \equiv \sqrt{[h(\mathbf{r}, t) - \bar{h}(t)]^2}$, where $\bar{h}(t)$ is the mean height at time t

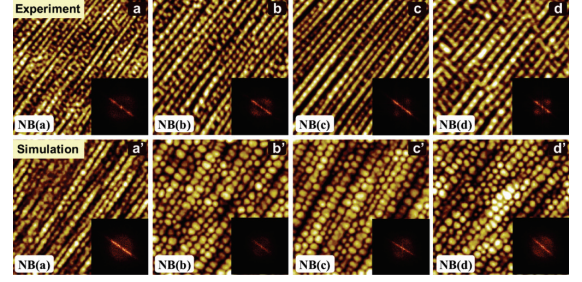


FIG. 1. (Color online) (top) Optimal nanobead patterns formed by IBS normal to the initial rippled surfaces in Table I: (a) NB(a), (b) NB(b), (c) NB(c), and (d) NB(d). The size of each image is $1 \times 1 \mu\text{m}^2$. (bottom) Nanobead patterns obtained by numerical simulations of Eq. (1) for corresponding experimental initial rippled surfaces. For each image, the inset provides the corresponding height power spectrum (squared modulus of the Fourier transform).

and $\langle \cdot \rangle$ denotes the space average. The mean ripple wavelength λ_R was estimated as the distance from the central ripple to the neighboring one in the two-dimensional height-height correlation function $G(\mathbf{r}, t)$, which is defined as $G(\mathbf{r}, t) \equiv \langle h(\mathbf{r}_1, t)h(\mathbf{r}_1 + \mathbf{r}, t) \rangle$. For each experimental condition NB(a), NB(b), etc., values quoted in Table I for the various quantities are average results over more than ten samples.

We subsequently sputtered each rippled surface at normal incidence. Arrays of nanobeads self-organize along each initial ripple. During this sputtering process, P_{Ar} , ε , and f were 1.2×10^{-4} Torr, 2 keV, and $1.1875 \text{ ions nm}^{-2} \text{ s}^{-1}$, respectively, corresponding to erosive conditions for IBS of Au.^{1,2,20} The mean wavelength of the beads λ_B is extracted by measuring the mean distance between adjacent beads on numerous line profiles along the ripples. Since nanobead patterns are closely packed along each ripple, λ_B can be considered the mean nanobead diameter D . In order to estimate the mean length of uninterrupted nanobeads arrays, denoted by ℓ_B , this was extracted from line profiles along ripples until convergence of the mean bead length is reached. In these symbols, the subscript B refers to the bead pattern.

Figures 1(a)–1(d) show the *optimal* nanobead patterns formed by IBS normal to the surfaces of the initial ripple patterns described in Table I. The insets provide the corresponding

TABLE I. Features of initial rippled surfaces and optimal nanobead patterns after IBS normal to the former. Here, λ_R^i and W^i are, respectively, the mean ripple wavelength and surface roughness of the initial ripple pattern; T is the temperature during sputtering to fabricate initial ripple pattern, and $\bar{D}$, σ , $\bar{A}_B$, and ℓ_B are the mean bead diameter, the standard deviation of the distribution of nanobead diameters, the mean bead amplitude, and the mean length of uninterrupted nanobead arrays along each ripple of the optimal nanobead pattern. Superscript i and subscript B refer to initial ripple pattern and nanobead pattern, respectively.

	Initial ripple pattern			Optimal bead pattern				
	λ_R^i (nm)	W^i (nm)	T (K)	$\lambda_B \simeq \bar{D}$ (nm)	$\sigma/\bar{D}$	$\bar{A}_B$ (nm)	ℓ_B (μm)	Time (min)
NB(a)	27	1.1	298	36.5	0.24	1.8	0.46	10
NB(b)	30	1.8	303	47.8	0.21	3.1	>0.95	25
NB(c)	47	2.7	313	47.9	0.18	2.8	>2.5	25
NB(d)	69	1.0	333	51.3	0.20	3.0	>0.8	25



Reduction of charge fluctuation energies in ultrathin NiO films on Ag(001)

Seolun Yang^a, H.-K. Park^a, J.-S. Kim^{a,*}, S.-H. Phark^b, Young Jun Chang^d, T.W. Noh^b, H.-N. Hwang^c, C.-C. Hwang^c, H.-D. Kim^{c,**}

^a Department of Physics, Sookmyung Women's University, Seoul 140-742, Republic of Korea

^b Department of Physics, Seoul National University, Seoul 151-747, Republic of Korea

^c Pohang Accelerator Laboratory, Pohang 790-784, Republic of Korea

^d Department of Physics, University of Seoul, Seoul 130-743, Republic of Korea

ARTICLE INFO

Article history:

Received 8 October 2012

Accepted 19 May 2013

Available online 24 May 2013

Keywords:

Ultrathin NiO film

Ag

Coulomb correlation energy

Charge transfer energy

Image charge potential

Polarization energy

ABSTRACT

As the film becomes atomically thin, the on-site Coulomb interaction energy between two 3p holes of the NiO films on Ag(001) U (Ni 3p) significantly decreases as revealed by both X-ray photoelectron and Auger electron spectroscopies. The reduction of U (Ni 3p) for the ultrathin films is well accounted for by varied image potentials and polarization energies in the films from their bulk values. The present results confirm a previous model predicting the reduction of charge fluctuation energies in ultrathin oxide films on highly polarizable substrates due to the extra-atomic relaxations.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

The electronic properties of ultrathin films are significantly different from those of their bulk states due to their reduced dimensionality and the influence of substrates. In the case of oxides on metallic substrates, it has been predicted that characteristic features such as the on-site Coulomb interaction energy U and the charge transfer energy Δ from a ligand to a neighboring cation would be substantially altered for atomically thin films [1,2]. Altieri et al. [2] observed that for ultrathin MgO films on Ag(001), both U and Δ decreased monotonically as the film became thin. They attributed the decrements of U and Δ in the film from the bulk values, δU and $\delta \Delta$, to the varied extra-atomic relaxation energies E_{rb} . As the major sources of E_{rb} , they considered both the image charge potential energy E_{image} between an extra charge and its image induced in the metal substrate and the polarization energy E_{pol} of the oxide caused by the extra charge also known as the Madelung potential energy. The magnitude of the image potential energy should be larger for thinner films due to the smaller mean distance between an extra charge in the film and its image in the substrate. The polarization energy of the film should be different from that of the bulk oxide because the volume of the oxide in the film is reduced, whereas the polarizability is enhanced at the surface of the oxide. The resulting variation between E_{rb} in the film and in the bulk state, δE_{rb} , even quantitatively reproduced the experimental δU for a 1-monolayer (ML) MgO film on Ag(001) [2].

Nevertheless, the suspicion was raised that the successful reproduction of the experimental δU by δE_{rb} could have been fortuitous, as there were many other effects such as dipole–dipole interactions that were not taken into account as well as unjustified assumptions e.g. $1/r$ dependence of the image potential on the atomic distance [2]. Moreover, Chambers and Droubay [3] reported that both Fe_2O_3 and Cr_2O_3 films on Pt(111) exhibited negligible δU and $\delta \Delta$. This contrasting observation was attributed to effective intrinsic screening of charge transfer, which reduced the extra-atomic relaxation to an undetectable level. Thus, no comprehensive picture of the electronic properties of ultrathin oxide films on highly polarizable substrates seems to exist, and the number of experimental studies is very limited to properly assess the existing hypotheses.

The objective of the present work is to assess the existing hypotheses by comparing their predictions with experimental results for a different system: ultrathin NiO films on Ag(001). Bulk NiO is prototypical as a charge-transfer insulator [4], and its charge fluctuation energies have already been studied [5,6]. Furthermore, the lattice mismatch between NiO(001) and Ag(001) is only ~2%, and the pseudomorphic growth of an NiO film is well established [7–9]. NiO films grown on Ag(001) are, thus, well suited for studying the thickness dependency of charge fluctuation energies such as U and Δ .

It is, however, difficult to obtain the Coulomb interaction energy U (Ni 3d) between Ni 3d electrons via the method of Altieri et al. [2], because the Ni 3d spectrum is difficult to isolate due to its overlap with the Ag 4d band of the substrate. Instead, we study the interaction energy between Ni 3p holes U (Ni 3p); As the film becomes atomically thin, U (Ni 3p) exhibits a substantial reduction from its bulk

* Corresponding author. Tel.: +82 27109406.

** Corresponding author. Tel.: +82 119933 0352.

E-mail addresses: jskim@sm.ac.kr (J.-S. Kim), hdkim@pal.postech.ac.kr (H.-D. Kim).

value. Moreover, the extra-atomic relaxation energies represented by both E_{image} and E_{pol} well reproduce the change in $U(\text{Ni } 3p)$ from bulk to thin film, $\delta U(\text{Ni } 3p)$, for a 1 ML NiO film on Ag(001). Using the observed values of δU and $\delta \Delta$, we estimate the Néel temperature T_N in the mean field approximation that is found to be compatible with the experimental value of T_N for a 3 ML NiO film [10]. These results reinforce the idea that E_{rel} represented mainly by E_{image} and E_{pol} determines δU and $\delta \Delta$ for ultrathin oxide films of NiO on highly polarizable substrates.

2. Experiment

We performed in situ scanning tunneling microscopy (STM), photoelectron spectroscopy (PES), and Auger electron spectroscopy (AES) of ultrathin NiO films on Ag(001). The STM work was carried out with a variable-temperature STM (Omicron). The NiO films were grown in an attached preparation chamber, where the preliminary characterization of both Ag substrate and NiO film was accomplished by X-ray photoelectron spectroscopy (XPS) and low-energy electron diffraction (LEED).

The PES and AES works were carried out at a soft X-ray beamline (7B1) of Pohang Light Source in Korea. The end station of the beamline is composed of both an analysis chamber and a preparation chamber. The analysis chamber is equipped with a hemispherical electron energy analyzer with a multichannel detector (SPECS). For the PES, the photoelectrons are collected at a take-off angle of 45° with respect to the surface normal of the sample. The PES resolving power is ~ 4000 [11]. The zero point of the binding energy is determined in reference to the binding energy of the Ag 3d (368.3 eV) of the clean Ag substrate. All spectra presented in this work were recorded with the sample maintained at room temperature. For both STM and PES, no charging effects were observed.

The NiO films were grown in preparation chambers for both STM and PES. The base pressures were $< 5 \times 10^{-10}$ Torr for both chambers. Wedge-shaped NiO films were grown by e-beam evaporation of high purity (5 N) Ni rod onto clean Ag(001) at room temperature at an ambient O_2 pressure (P_{O_2}) of $1\text{--}3 \times 10^{-6}$ Torr. The films were then thermally annealed at 430–450 K at $P_{\text{O}_2} \sim 5 \times 10^{-7}$ Torr. In the present work, we are especially interested in films within the monolayer limit, as this enables a definite comparison of experimental δU with theoretical values obtained by considering extra-atomic relaxation energies. However, for films less than 2 ML, the growth mode is somewhat complicated due to the (2×1) reconstruction and the bilayer growth of the NiO film [7–9]. Under the aforementioned growth conditions, we were able to grow nickel oxide films made of mainly (1×1) 1 ML NiO, as assessed by a combination of techniques, including STM, LEED, and XPS. (Further details are given in the following section.) According to our previous extensive PES of NiO films, such growth conditions also minimize the chemical defects [12].

The thickness of each film was mainly determined by the ratio of the peak intensity of the Ag 3d with the NiO film covered to that of the clean Ag substrate, assuming layer-by-layer (LBL) growth of the film. Because the growth of a NiO film does not follow LBL in an ideal fashion, the thicknesses described in the present work are nominal. For coverage ~ 0.5 ML, the film is mainly composed of monolayer-high islands and can be taken as a model for a 1 ML film (Fig. 1(a)). (Further discussion is presented below.) Moreover, up to 0.5 ML, the coverage recorded by a quartz microbalance is in reasonable agreement with the nominal coverage estimated by the Ag 3d intensity, assuming layer-by-layer growth of the NiO film. Based on this estimation of the film thickness, the growth rate is adjusted to be ~ 0.25 ML/min throughout the experiments.

3. Result

Fig. 1(a)–1 shows a typical image of a nickel oxide film with ~ 0.5 ML coverage, consisting of nickel oxide patches. The line profile

(Fig. 1(a)–3) across a typical patch in Fig. 1(a)–2 displays a plateau of apparent height ~ 0.15 nm, which corresponded to 1 ML in our previous STM study of NiO films on Ag(001) under similar tunneling conditions [13]. The embedded layer growth of the NiO films on Ag(001) has been reported [14]. Such a growth mode, however, looks not prevailing in our growth condition according to our STM study as shown in Fig. 1(a). Fig. 2 shows Ag 3d spectra before and after depositing 0.5 ML of NiO film. For direct comparison of the two spectra, we rescaled the spectrum of Ag 3d after depositing the NiO film and shifted it vertically. Both are almost identical, indicating the inertness of the Ag substrate to the NiO film. As the film is mostly composed of 1 ML-thick islands, its Ni and O spectra are taken as representative of the electronic properties of a 1 ML NiO film.

During further deposition to the nominal coverage around 1 ML, the second layer is preferably occupied (Fig. 1(b)), and the film is almost bilayered (Fig. 1(c)). All the films exhibit (1×1) LEED pattern (Fig. 1(b) and (c)), whereas the well-known (2×1) reconstruction is observed only sporadically in the STM images (figure not shown). The (2×1) area, however, becomes abundant if P_{O_2} during growth is lowered below 10^{-6} Torr.

$U(\text{Ni } 3p)$ can be obtained by comparing the energy of a two-hole state, $\text{Ni } 3p^4$, to that of two one-hole state, $\text{Ni } 3p^5$ as follows,

$$U(\text{Ni } 3p) = E(3p^4) + E(3p^6) - 2E(3p^5). \quad (1)$$

The variation of $U(\text{Ni } 3p)$ from that of bulk $\delta U(\text{Ni } 3p)$ can then be determined by the following relation according to Altieri et al. [2],

$$\delta U(\text{Ni } 3p) = \delta E^{\text{bind}}(\text{Ni } 2p) - 2\delta E^{\text{bind}}(\text{Ni } 3p) - \delta E^{\text{kin}}(\text{Ni}_{\text{LMM}}). \quad (2)$$

In order to estimate $\delta U(\text{Ni } 3p)$, we measured the XPS spectra of Ni 2p, Ni 3p, and the Ni_{LMM} Auger transition as a function of the thickness of the NiO film. We also utilized the O 1s spectra to estimate $\delta \Delta(\text{O } 2p \rightarrow \text{Ni } 3d)$, as described below. Fig. 3(a), (b), (c), and (d), respectively, shows the Ni 2p, Ni 3p, Ni_{LMM} Auger transition, and O 1s spectra of the NiO films. The nominal thicknesses of the films range from 0.5 ML to 15.00 ML.

The core-level spectra of Ni comprise many peaks of various origins, such as final state effects and non-local screening [12], making identification of the main peak uncertain. This complication is also transferred to the Ni_{LMM} Auger transition. Thus, we estimate shifts of the peak positions of Ni 2p, Ni 3p, and the Ni_{LMM} Auger transition with the decrease of the film thickness in terms of shifts of the centroids of their spectra, anticipating that if shifts of the peak positions are mainly caused by extra-atomic relaxation, then all of the constituent peaks should shift by the same amount.

The constant relative spectral weight irrespective of the film thickness is a prerequisite for the centroid argument to estimate the peak shift. However, for the films in the monolayer limit, the oxygen defects, dimensional effects, and the bonding with the substrate modify the relative spectral weight from that of the thicker films. Most notably for the 0.5 ML film in Fig. 3(a), the spectral weight of the satellite peak of Ni 2p peak is smaller than that for the higher NiO coverage. The abrupt increase of the centroid energy by 0.5 eV for the 0.75 ML film with respect to that of 0.5 ML in Fig. 4 (uppermost graph) could be partly attributed to the increase of the relative intensity of the satellite peak with respect to the main peak. The jump of the centroid energy could also be accounted for in part by the fact that the second layer becomes populated from the coverage of 0.5 ML. The centroid shift due to the spectral weight change is, then, less than 0.5 eV. The change of the spectral weight is, however, a relatively minor source of the shift of the centroid energy, because it increases further by 0.75 eV for the thicker films, even though relative spectral weight for those films is almost constant. Similar argument should be applied to the other peaks for which even such abrupt shift of the centroid energies as for Ni 2p is not observed.

Amorphous Ferromagnetic CoSiB/Pd Multilayer with Perpendicular Magnetic Anisotropy

Sol JUNG and Haein YIM*

Department of Physics, Sookmyung Women's University, Seoul 140-742, Korea

Jungbum YOON

Department of Physics, Inha University, Incheon 402-751, Korea

(Received 19 October 2012, in final form 2 November 2012)

Perpendicular magnetic anisotropy has been studied for possible use in high-density spin transfer torque magnetic random access memories. The key issues of high-density spin transfer torque magnetic random access memories are decreasing the switching current and have the high thermal stability. In order to solve these problems, the approach of developing a new amorphous ferromagnetic material for use as a pinned layer for a multilayer with a low saturated magnetization value has been suggested. In a previous study, amorphous ferromagnetic material CoSiB was shown to have a low saturated magnetization ($M_s = 470 \text{ emu/cm}^3$) and a modest anisotropy constant ($K_u = 1,500 \text{ erg/cm}^3$). Then, we investigated the perpendicular magnetic anisotropy of a $[\text{CoSiB } t_{\text{CoSiB}} \text{ nm/Pt } 1.4 \text{ nm}]_5$ multilayer, where t_{CoSiB} are 0.1, 0.2, 0.3, 0.4, and 0.5 nm. In this study, we present a new multilayer type based on the previous study, and this multilayer, CoSiB/Pd multilayer has the perpendicular magnetic anisotropy. In this letter, we have investigated the perpendicular magnetic anisotropies of a $[\text{CoSiB } t_{\text{CoSiB}} \text{ nm/Pd } 1.3 \text{ nm}]_5$ multilayer, where t_{CoSiB} are 0.1, 0.2, 0.3, 0.4, 0.5, and 0.6 nm and $[\text{CoSiB } 6.3 \text{ nm/Pd } t_{\text{Pd}} \text{ nm}]_5$ multilayer, where t_{Pd} are 1.1, 1.2, 1.3, 1.4, 1.5, and 1.6 nm.

PACS numbers: 75.70.Cn, 75.70.-i, 75.50.Kj

Keywords: Perpendicular magnetic anisotropy, Amorphous, CoSiB, Magnetic multilayer, Pd

DOI: 10.3938/jkps.62.L10

I. INTRODUCTION

Perpendicular magnetic anisotropy is the phenomenon of magnetic thin film which is preferentially magnetized in a direction perpendicular to the film's plane. This is currently of interest in relation to high-density magnetic recording, and the Co/Cr thin film where anisotropy has a magneto-crystalline origin is a popular example. In studies of other various candidates for recording media, multilayers (thin films) offer the possibility that the anisotropy is dominated by the existence of interface anisotropy. Magnetic multilayers with perpendicular magnetic anisotropy have been intensively studied as possible candidates for spin transfer torque magnetic random access memories (STT-MRAMs) [1–4]

Because exchange-coupled super-lattice films of alternating ferromagnetic and non-magnetic layers can be engineered to show strong perpendicular magnetic anisotropy, some materials with magnetic multilayers, including metal alloys based on Co/Pt and Co/Pd, have been investigated [5–7]

The amorphous metal alloy CoSiB/Pt multilayer is expected to be a potential candidate material for the free layer in a high-density magnetic random access memory, and amorphous magnetic materials are required for high-density magnetic devices because such materials show better interface roughness than crystalline materials [8–13]. Mixing of this kind has been demonstrated for the amorphous ferromagnetic CoSiB/Pd multilayer by the Magnetic Materials Science Laboratory (MMS Lab) in Sookmyung Women's University. In other letters, values of the saturated magnetization (M_s) for CoFe, NiFe, and CoFeSiB were $1,400 \text{ emu/cm}^3$, 800 emu/cm^3 , and 560 emu/cm^3 , respectively [8]. In a previous study, the amorphous ferromagnetic material CoSiB has a low saturated magnetization ($M_s = 470 \text{ emu/cm}^3$) and a modest anisotropy constant ($K_u = 1,500 \text{ erg/cm}^3$).

In this search, we investigated materials, including amorphous metal alloys and magnetic multilayers with amorphous ferromagnetic CoSiB/Pd. In this letter, we present new data for the $\text{Co}_{75}\text{Si}_{15}\text{B}_{10}$, which was studied in a previous study. Moreover, we propose a new idea: CoSiB/Pd multilayer.

*E-mail: haein@sm.ac.kr; Fax: +82-30-3079-90362

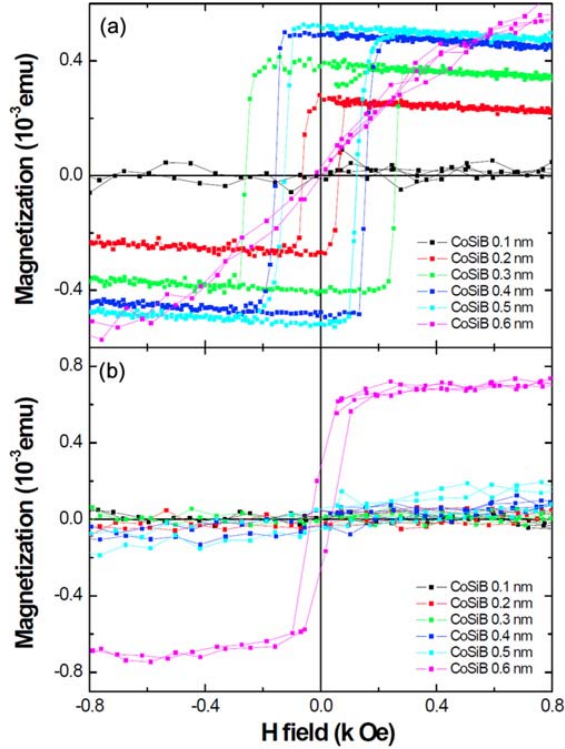


Fig. 1. (Color online) Hysteresis loops for the $[\text{CoSiB } t_{\text{CoSiB}} \text{ nm}/\text{Pd } 1.3 \text{ nm}]_5$ multilayer, where t_{CoSiB} take on values of 0.1, 0.2, 0.3, 0.4, 0.5, and 0.6 nm: (a) out-of-plane and (b) in-plane.

II. EXPERIMENTS AND DISCUSSION

In this experiment, a thin film was fabricated through the system of 3-inch, 6-gun-type DC magnetron sputtering. A SiO_2 (110) wafer was prepared, placed into an ultrasonic washer with ethanol for 10 minutes, acetone for 10 minutes and DI-water for 10 minutes in order to remove impurities on the SiO_2 (100) wafer's surface. We used three sputtering targets, Ta, CoSiB and Pd: these were manufactured by mixing high-purity metal powders and used for the fabrication of the thin films.

The chamber's base pressure was up to 2.0×10^{-7} Torr, and the working pressure was 2×10^{-3} Torr after flowing Ar gas (46 sccm). All thin films were uniform in size $1.4 \text{ cm} \times 1.4 \text{ cm}$, and were deposited at room temperature. The film stack included a substrate SiO_2 100 nm, a buffer layer [Ta 5 nm/Pd 3 nm] and capping layer Ta 5 nm. The magnetic properties of thin films were measured by a VSM (Vibrating Sample Magnetometer) at magnetic fields from -1 k Oe to $+1 \text{ k Oe}$.

Co/Pt and CoFe/Pt multilayer films have been studied as candidates for perpendicular magnetic anisotropy, and CoSiB has been studied as a quite soft magnetic mate-

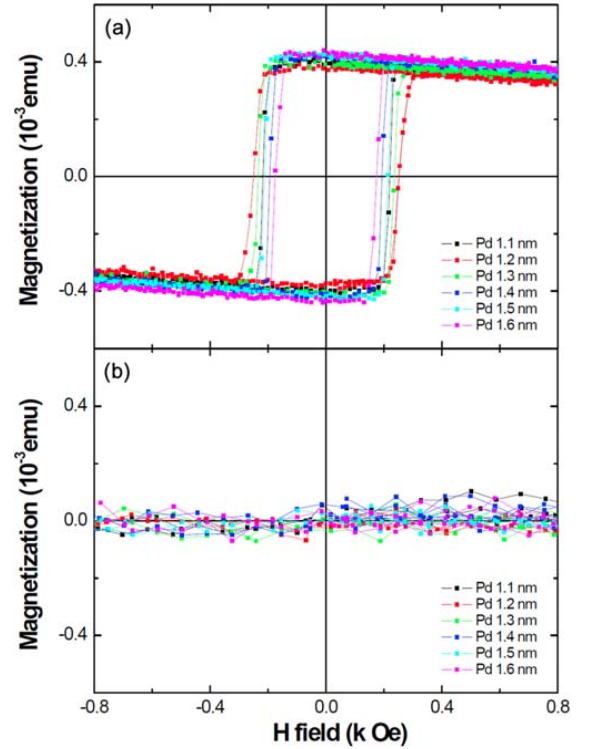


Fig. 2. (Color online) Hysteresis loops for the $[\text{CoSiB } 0.3 \text{ nm}/\text{Pd } t_{\text{Pd}} \text{ nm}]_5$ multilayer, where t_{Pd} take on values of 1.1, 1.2, 1.3, 1.4, 1.5, and 1.6 nm: (a) out-of-plane and (b) in-plane.

rial [8-10,14,15]. In a previous study, we investigated the magnetic properties of CoSiB, CoSiB/Pt and CoSiB/Au multilayers with various thicknesses of the CoSiB, Pt and Au layers, respectively. We obtained the coercivity (H_c) and the saturated magnetization (M_s) of for a CoSiB single layer, 1.6 Oe and $407 \text{ emu}/\text{cm}^3$, respectively. This amorphous ferromagnetic CoSiB is confirmed to be beneficial for applications of thin films with perpendicular magnetic anisotropy. Therefore, we propose a new idea in this study: CoSiB/Pd multilayer.

Figure 1 shown the hysteresis loops for a CoSiB/Pd multilayer with various thicknesses of CoSiB. The CoSiB thickness range were 0.1 – 0.6 nm but the Pd thickness was fixed at 1.3 nm. Figure 1(a) shown the out-of-plane hysteresis loop and Fig. 1(b) shown the in-plane hysteresis loop. In the case of CoSiB thicknesses of 0.1 nm and 0.6 nm, these samples do not show perpendicular magnetic properties. Other hysteresis loops show good squareness, and some data tend to shown the perpendicular magnetic anisotropy. Especially, a $[\text{CoSiB } 0.3 \text{ nm}/\text{Pd } 1.3 \text{ nm}]_5$ multilayer film shows the best tendency of the perpendicular magnetic anisotropy in this series. The H_c and the M_s depend on the CoSiB thickness in

Thermal and Magnetic Properties of Melt-spun Co-Fe-B-Si-(Cr/Ta) Amorphous Alloys

Hye-in JO, Bokyeong HAN and Haein Choi YIM*

Department of Physics, Sookmyung Women's University, Seoul 140-742, Korea

Woojun KWON

Department of Physics, Kookmin University, Seoul 136-702, Korea

(Received 23 November 2012, in final form 4 December 2012)

The ferromagnetic amorphous alloys Co-Fe-B-Si-Cr and Co-Fe-B-Si-Ta were systemically investigated. $[(\text{Co}_{1-x}\text{Fe}_x)_{0.75}\text{B}_{0.2}\text{Si}_{0.05}]_{96}\text{Cr}_4$ ($x = 0, 0.1, 0.4, 0.7, 1.0$) and $[(\text{Co}_{1-x}\text{Fe}_x)_{0.75}\text{B}_{0.2}\text{Si}_{0.05}]_{96}\text{Ta}_4$ ($x = 0, 0.1, 0.3, 0.5, 1.0$) ribbons were prepared by using the melt-spinning technique. Amorphous structures were confirmed by the presence of a broad and halo peak in X-ray diffraction (XRD). The thermal characteristics, such as the glass transition temperature (T_g), the crystallization temperature (T_x) and the supercooled liquid region ($\Delta T_x = T_x - T_g$), were measured by means of a differential scanning calorimeter (DSC). For the Co-Fe-B-Si-Cr alloys, the values of ΔT_x were in the range of 51.6 – 61.4 K, and for Co-Fe-B-Si-Ta alloys, they were in the range of 39.3 – 55.5 K. The magnetic properties were examined by means of a vibrating sample magnetometer (VSM) at a maximum field strength of 3000 Oe. These glassy ribbons exhibited soft magnetic properties. The saturation magnetization (M_s) and the coercive force (H_c) were in the ranges of 307.85 – 854.52 emu·cm³ and 0.64 – 0.89 Oe, respectively.

PACS numbers: 81.05.Kf, 81.40.Rs, 81.70.Pg

Keywords: Co-based, Fe-based, Amorphous, Ferromagnetic, Soft magnetic, Magnetic properties

DOI: 10.3938/jkps.62.523

I. INTRODUCTION

Since metallic glasses were discovered by Duwez *et al.* [1] in 1960, they have become a class of new materials with unusual physical properties: high yield strength (>2 Gpa), high elastic strain limit (–2%), high hardness, and high corrosion resistance. In particular, metallic glasses composed of transition metals, such as Fe, Co, and Ni, and of metalloids, such as B and Si, have good soft magnetic properties, and these alloys have attracted great attention from many researchers over the past few decades [2–5]. Amorphous Fe- and Co-based materials are used in diverse applications, such as power devices, information handling technology and magnetic sensors [6]. Fe-based soft magnetic glassy alloys have been commercialized as consolidated magnetic cores for power supplies, choke coils, common modes and noise filter [7,8].

We studied some Co- and Fe-based materials, and we reported preliminary results for Co-Fe-B-Si-Mo and Co-Fe-B-Si-Nb ferromagnetic alloys [9,10]. A small amount of Nb addition to $[(\text{Co}_{0.5}\text{Fe}_{0.5})_{0.75}\text{B}_{0.2}\text{Si}_{0.05}]_{100-x}\text{Nb}_x$ ($x = 5, 6, 7, 8$ at.%) glassy ribbons had beneficial ef-

fects on improving the structural properties and the thermal stability while maintaining excellent soft magnetic properties. Furthermore, the mechanical properties were not harmed. Therefore, as the proportion of Nb in Co-Fe-B-Si-Nb increases, the supercooled liquid region (ΔT_x) in the amorphous ribbons has a tendency to be enlarged, the coercivity (H_c) to be higher, and the saturation magnetization (M_s) to be decreased a little [9]. Increasing the proportion of Fe in $[(\text{Co}_{1-x}\text{Fe}_x)_{0.75}\text{B}_{0.2}\text{Si}_{0.05}]_{96}\text{Mo}_4$ ($x = 0, 0.2, 0.5, 0.8, 1.0$) and $[(\text{Co}_{1-x}\text{Fe}_x)_{0.75}\text{B}_{0.2}\text{Si}_{0.05}]_{96}\text{Nb}_4$ ($x = 0, 0.2, 0.5, 0.8, 1.0$) amorphous ribbons tends to enlarge the supercooled liquid region, and a small amount of Mo or Nb addition will have beneficial effects on thermal stability. Moreover, Co-Fe-B-Si-Mo and Co-Fe-B-Si-Nb have semi-hard magnetic properties: high coercivity and low saturation magnetization [10].

In addition to previous studies, we make an attempt to modify the thermal or the magnetic properties by adding a small amount of Cr or Ta addition with a different view of the composition. In this research, the thermal and the magnetic properties of a series of $[(\text{Co}_{1-x}\text{Fe}_x)_{0.75}\text{B}_{0.2}\text{Si}_{0.05}]_{96}\text{Cr}_4$ ($x = 0, 0.1, 0.4, 0.7, 1.0$) and $[(\text{Co}_{1-x}\text{Fe}_x)_{0.75}\text{B}_{0.2}\text{Si}_{0.05}]_{96}\text{Ta}_4$ ($x = 0, 0.1, 0.3, 0.5, 1.0$) melt-spun amorphous alloys was investigated in detail.

*E-mail: haein@sookmyung.ac.kr; Fax: +82-2-2077-7320

Table 1. Thermal and magnetic properties of $[(\text{Co}_{1-x}\text{Fe}_x)_{0.75}\text{B}_{0.2}\text{Si}_{0.05}]_{96}\text{Cr}_4$ ($x = 0, 0.1, 0.4, 0.7, 1.0$) and $[(\text{Co}_{1-x}\text{Fe}_x)_{0.75}\text{B}_{0.2}\text{Si}_{0.05}]_{96}\text{Ta}_4$ ($x = 0, 0.1, 0.3, 0.5, 1.0$) glassy alloys.

Alloys	Thermal Properties			Magnetic Properties	
	T_g (K)	T_x (K)	ΔT_x (K)	H_c (Oe)	M_s (emu·cm ³)
$\text{Co}_{72}\text{B}_{19.2}\text{Si}_{4.8}\text{Cr}_4$	728.3	789.7	61.4	0.72	352.71
$\text{Co}_{64.8}\text{Fe}_{7.2}\text{B}_{19.2}\text{Si}_{4.8}\text{Cr}_4$	751.5	803.1	51.6	0.68	416.95
$\text{Co}_{43.2}\text{Fe}_{28.8}\text{B}_{19.2}\text{Si}_{4.8}\text{Cr}_4$	763.6	818.5	54.9	0.68	563.48
$\text{Co}_{21.6}\text{Fe}_{50.4}\text{B}_{19.2}\text{Si}_{4.8}\text{Cr}_4$	764.7	821.9	57.2	0.85	854.52
$\text{Fe}_{72}\text{B}_{19.2}\text{Si}_{4.8}\text{Cr}_4$	764.7	823.1	58.5	0.88	687.01
$\text{Co}_{72}\text{B}_{19.2}\text{Si}_{4.8}\text{Ta}_4$	834.5	890.0	55.5	0.69	307.85
$\text{Co}_{64.8}\text{Fe}_{7.2}\text{B}_{19.2}\text{Si}_{4.8}\text{Ta}_4$	792.1	835.4	43.3	0.68	332.45
$\text{Co}_{50.4}\text{Fe}_{21.6}\text{B}_{19.2}\text{Si}_{4.8}\text{Ta}_4$	834.2	884.0	49.8	0.64	537.70
$\text{Co}_{36}\text{Fe}_{36}\text{B}_{19.2}\text{Si}_{4.8}\text{Ta}_4$	820.3	859.6	39.3	0.89	600.83
$\text{Fe}_{72}\text{B}_{19.2}\text{Si}_{4.8}\text{Ta}_4$	794.1	836.8	42.7	0.71	835.50

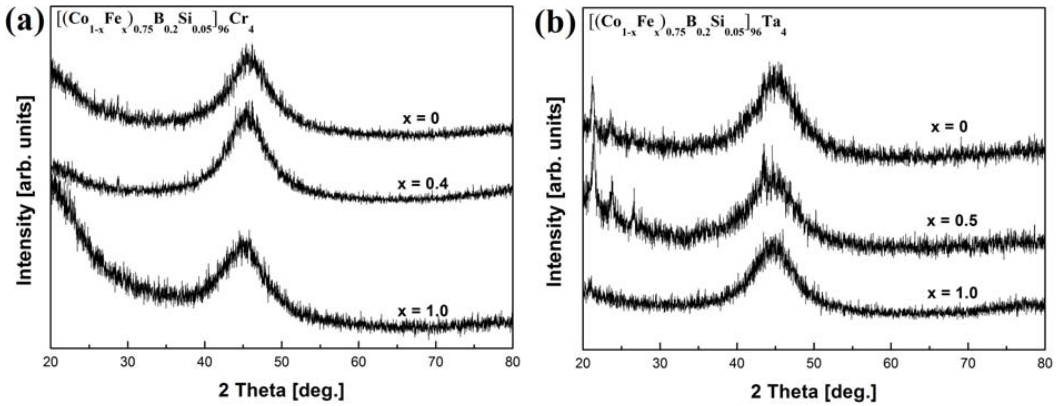


Fig. 1. XRD patterns of (a) $[(\text{Co}_{1-x}\text{Fe}_x)_{0.75}\text{B}_{0.2}\text{Si}_{0.05}]_{96}\text{Cr}_4$ ($x = 0, 0.1, 0.4, 0.7, 1.0$) and (b) $[(\text{Co}_{1-x}\text{Fe}_x)_{0.75}\text{B}_{0.2}\text{Si}_{0.05}]_{96}\text{Ta}_4$ ($x = 0, 0.1, 0.3, 0.5, 1.0$) samples

II. EXPERIMENTS AND DISCUSSION

Alloy ingots of $[(\text{Co}_{1-x}\text{Fe}_x)_{0.75}\text{B}_{0.2}\text{Si}_{0.05}]_{96}\text{Cr}_4$ ($x = 0, 0.1, 0.4, 0.7, 1.0$) and $[(\text{Co}_{1-x}\text{Fe}_x)_{0.75}\text{B}_{0.2}\text{Si}_{0.05}]_{96}\text{Ta}_4$ ($x = 0, 0.1, 0.3, 0.5, 1.0$) were prepared by using an arc process under a Ti-gettered argon atmosphere. To achieve compositional homogeneity, we remelted the high-purity ingots at least three times. Rapidly solidified ribbons with thicknesses of 30 – 40 μm and a width of 3 mm were produced by using the melt-spinning technique. A microstructure characterization of the ribbons was performed by using X-ray diffraction (XRD) with Cu-K α radiation. The thermal stability associated with the glass transition, the supercooled liquid region and the crystallization were measured by using a differential scanning calorimeter (DSC) at the heating rate of 0.67 K/s. Room-temperature hysteresis loops were acquired by using a vibrating sample magnetometer (VSM) at a maximum field of 3000 Oe.

Figure 1 shows the XRD patterns for the $[(\text{Co}_{1-x}$

$\text{Fe}_x)_{0.75}\text{B}_{0.2}\text{Si}_{0.05}]_{96}\text{Cr}_4$ ($x = 0, 0.1, 0.4, 0.7, 1.0$) and the $[(\text{Co}_{1-x}\text{Fe}_x)_{0.75}\text{B}_{0.2}\text{Si}_{0.05}]_{96}\text{Ta}_4$ ($x = 0, 0.1, 0.3, 0.5, 1.0$) ribbons. As shown in Fig. 1, broad and halo X-ray diffraction patterns corresponding to an amorphous structure were observed. However, the $[(\text{Co}_{1-x}\text{Fe}_x)_{0.75}\text{B}_{0.2}\text{Si}_{0.05}]_{96}\text{Ta}_4$ ($x = 0, 0.5$) samples have some sharp diffraction peaks at 20 – 30° corresponding to a crystalline phase that has not yet been identified. As shown in Fig. 2, the melt-spun ribbons $[(\text{Co}_{1-x}\text{Fe}_x)_{0.75}\text{B}_{0.2}\text{Si}_{0.05}]_{96}\text{Cr}_4$ ($x = 0, 0.1, 0.4, 0.7, 1.0$) and $[(\text{Co}_{1-x}\text{Fe}_x)_{0.75}\text{B}_{0.2}\text{Si}_{0.05}]_{96}\text{Ta}_4$ ($x = 0, 0.1, 0.3, 0.5, 1.0$) exhibit clear glass transition and crystallization events during heating. These indicate that all the ribbons have an amorphous structure. In the results, the differences between the glass temperature and the crystallization temperature ($\Delta T_x = T_x - T_g$) were found to have values ranging up to 39.3 – 61.4 K, and T_g and T_x were in the ranges of 728.3 – 834.5 K and 789.7 – 890 K, respectively. Table 1 summarizes the results. The $\text{Co}_{72}\text{B}_{19.2}\text{Si}_{4.8}\text{Cr}_4$ ($x = 0$) and the $\text{Co}_{72}\text{B}_{19.2}\text{Si}_{4.8}\text{Ta}_4$ ($x = 0$)

Time-dependent magnetization reversal in amorphous CoSiB/Pd multilayers with perpendicular magnetic anisotropy

Jungbum Yoon,¹ Sol Jung,² Youngha Choi,³ Jaehun Cho,¹ Chun-Yeol You,^{1,a)} Myung-Hwa Jung,³ and H. I. Yim²

¹Department of Physics, Inha University, Incheon 402-751, South Korea

²Department of Physics, Sookmyung Women's University, Seoul 140-742, South Korea

³Department of Physics, Sogang University, Seoul 121-742, South Korea

(Presented 15 January 2013; received 5 November 2012; accepted 7 February 2013; published online 12 April 2013)

We investigated the perpendicular magnetic anisotropy of amorphous CoSiB/Pd multilayer systems. We replaced the crystalline Co layer with an amorphous CoSiB layer in a Co/Pd multilayer system in order to reduce the domain wall pinning sites, which are mainly located at the grain boundaries and interfaces. We also found the correlation between the coercivity and the thermal activation volume of the magnetization reversal in the amorphous CoSiB/Pd multilayers. The time-dependent magnetization reversal was investigated using the magneto-optical Kerr effect image and the magnetic field sweep-rate-dependent coercivity. © 2013 AIP Publishing LLC. [<http://dx.doi.org/10.1063/1.4801425>]

I. INTRODUCTION

Magnetic materials with perpendicular magnetic anisotropy (PMA) are applicable to new generation devices such as high-density data storage and spin-transfer torque magnetic random access memory (STT-RAM).^{1,2} The nano-size structure of devices, demanded for high-density applications, has been developed through the evolution of nanofabrication techniques. Due to the superparamagnetism of the nanostructure, the thermal stability of the free layer becomes an important issue. To overcome the thermal stability problem, PMA materials with strong magnetic anisotropy are required. Crystalline PMA materials, for example, Co,^{3,4} and Fe,⁵ based on superlattice systems such as Co/Pd,^{6,7} Co/Pt,⁸ or Fe/Au,⁹ are being extensively investigated. Compared to crystalline PMA systems, amorphous PMA systems^{10,11} have two strong advantages. First, there are no grain boundaries, which act as domain wall (DW) pinning sites during the magnetization reversal process. Since the grain boundary leads to irregular switching, the amorphous system guarantees a high squareness. Second, the roughness of the amorphous system is generally less than that of crystalline systems. Interface roughness is important in superlattice systems, since the origin of the PMA is the broken symmetry at the interface in multilayer films.

In this study, we investigated the PMA of CoSiB/Pd multilayer systems by replacing the crystalline Co layer with an amorphous CoSiB layer. We note that the PMA multilayers with the amorphous CoSiB thin films are rarely investigated. In addition, even if the Co/Pd system shows similar magnetic properties to the Co/Pt system, the Pd is different from Pt in the spin pumping and Rashba effect. The Pd is weaker than these effects of the Pt, since the spin-orbit coupling of the Pt is small.^{12,13} Furthermore, since the PMA

multilayers with the Pt have the strong second order uniaxial anisotropy, the magnetic hysteresis of the system including the Pt is more complicated than the case of the Pd.^{10,11} Therefore, exploring the magnetic properties of the CoSiB/Pd system is interesting. In order to investigate the role of the amorphous CoSiB layer, the thermally activated reversal mechanism was studied by magneto-optical Kerr effect (MOKE) microscopy and magnetic field sweep-rate-dependent coercivity measurements.

II. EXPERIMENTS

The amorphous multilayers are deposited by a DC magnetron sputtering system at room temperature in Ar gas at a pressure of 2 mTorr onto a Co₇₅Si₁₅B₁₀ target. The multilayers consist of SiO₂ (100)/Ta (5)/Pd (3)/[CoSiB (*t*_{CoSiB})/Pd (1.3)]₅/Ta (5) with various thicknesses, *t*_{CoSiB} = 0.1, 0.2, 0.3, 0.4, 0.5, and 0.6 nm. To verify the PMA of the amorphous multilayers, the hysteresis loop is systematically measured using a vibrating sample magnetometer (VSM). The DW motion due to the thermal activation reversal is confirmed with MOKE microscope measurements at room temperature. During the MOKE microscope measurements, a saturation magnetic field is applied in the upward direction perpendicular to the film plane, and then, a magnetic field smaller than the coercivity is applied in the opposite (downward) direction. For the last process, MOKE images are obtained in a fixed magnetic field. The magnetic field sweep-rate-dependent coercivities are determined using the VSM with various sweep rates at room temperature. The sweep rate of the applied magnetic field is defined by the waiting time per step. In the sweep process, one magnetic field step is 5 Oe with a waiting time of 1–15 s.

III. RESULTS AND DISCUSSION

The PMA of amorphous CoSiB/Pd multilayers with various CoSiB thicknesses is confirmed by hysteresis loop

^{a)}Author to whom correspondence should be addressed. Electronic mail: cyyou@inha.ac.kr.

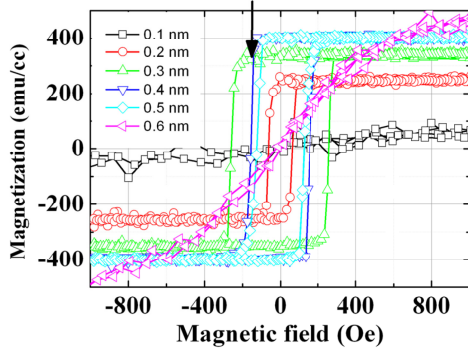


FIG. 1. Hysteresis loops of amorphous $[\text{CoSiB } (t_{\text{CoSiB}} \text{ nm})/\text{Pd } (1.3 \text{ nm})]_5$ multilayers with various CoSiB thicknesses of the out-of-plane field ($t_{\text{CoSiB}} = 0.1, 0.2, 0.3, 0.4, 0.5$, and 0.6 nm). The magnetic field sweep rate is 5 Oe/s in this measurement. The black vertical arrow indicates the field for the domain image measurement.

measurements using the VSM with a magnetic field sweep rate of 5 Oe/s , as shown in Fig. 1. For samples with $t_{\text{CoSiB}} = 0.2, 0.3, 0.4$, and 0.5 nm , the PMA manifests itself by the squareness of the hysteresis loops for the perpendicular field. The 0.1-nm sample shows paramagnetic behavior, because the formation of the 0.1-nm -thick layer is imperfect to preserve ferromagnetic order. When the sample thickness is 0.6 nm , the sample has in-plane anisotropy. The maximum coercivity ($H_c = 261 \text{ Oe}$) is observed for a CoSiB thickness of 0.3 nm , and the coercivity decreases with increasing t_{CoSiB} . The measured magnetization is calculated with the total volume of the samples, including the non-magnetic Pd layer volume.

Since the amorphous CoSiB/Pd multilayers have a low density of pinning sites, they are expected to show smoother DW motion than crystalline multilayers. In order to confirm the DW motion, we obtained domain images with a MOKE microscope. Figure 2 shows the MOKE microscope images of a $[\text{CoSiB } (0.3 \text{ nm})/\text{Pd } (1.3 \text{ nm})]_5$ multilayer with waiting times of (a) 30 s and (b) 180 s with an applied field of -150 Oe , corresponding to 57% of the coercivity. The measurement point is indicated by the black vertical arrow in Fig. 1. At 57% of the coercivity, the fast expansion of the dark region implies that the thermally activated DW motion is dominant.⁴ The DW propagation velocity is approximately $1 \mu\text{m/s}$, which is larger than a typical value for thermally activated DW velocity.⁶ We emphasize here that the domain expansion occurs in a very large area, as shown in Fig. 2.

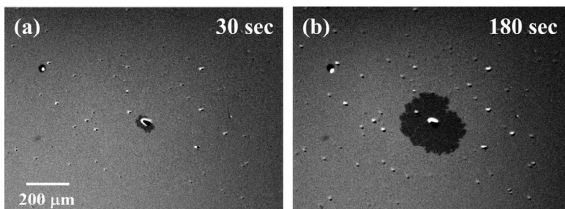


FIG. 2. Magneto-optical Kerr effect image for a $[\text{CoSiB } (0.3 \text{ nm})/\text{Pd } (1.3 \text{ nm})]_5$ multilayer after (a) 30 s and (b) 180 s at -150 Oe , as indicated by the black vertical arrow in Fig. 1.

The length scale of the expanded domain is on the order of $150 \mu\text{m}$ for 150 s . Such a long propagation of DW with only thermally activated motion and rather isotropic propagation indicate that only weak pinning centers exist. This can be easily overcome by the thermally activated DW motion. This is one of the main advantages of the amorphous system, as we mentioned in the earlier text.

To confirm the thermal activation process, we performed magnetic field sweep-rate-dependent measurements. We varied the field sweep rate from 0.33 to 5.0 Oe/s , and we found that the coercivities of the amorphous CoSiB/Pd multilayers were changed, as shown in Fig. 3. The variation of coercivity as a function of the logarithm of dH/dt , which is defined as the sweep rate in Oe/s units. Bruno *et al.*¹⁴ deduced the relation between H_c and the logarithm of dH/dt for the thermally activated reversal process

$$H_c(dH/dt) = k_B T / (v_a M_s) \ln(dH/dt) + \text{const}, \quad (1)$$

where k_B is Boltzmann's constant, T is the absolute temperature, v_a is the activation volume, and M_s is the saturation magnetization. Here, the fluctuation field, H_f , is defined by

$$H_f = k_B T / v_a M_s \quad (2)$$

and is determined from the linear fitting by Eq. (1) for the magnetic field sweep-rate dependence of the coercivity data, as shown Fig. 3. The physical meaning of H_f is clearly related to v_a in Eq. (2). The Zeeman energy, with H_f for the volume of v_a , is comparable to the thermal energy, $k_B T$. The thermal fluctuation of the magnetization is converted into the term H_f in Eq. (2). Therefore, it is a useful parameter for experimental measurements to understand the thermal reversal process.^{15,16} Here, another important physical parameter is v_a , which is determined from Eq. (2). The activation volume is related to the Barkhausen volume, which determines the time-dependent coercivity in the magnetization reversal process, and it is closely related to the density of the pinning sites. In Fig. 4, H_c and $1/v_a$ are plotted together as functions of t_{CoSiB} . Here, H_c is obtained from the hysteresis loop with a magnetic field sweep rate of 5 Oe/s . The correlation between H_c and $1/v_a$ can be clearly seen in Fig. 4, which shows that they have a close relationship. Generally, the magnitude of H_c is determined by not only the perpendicular

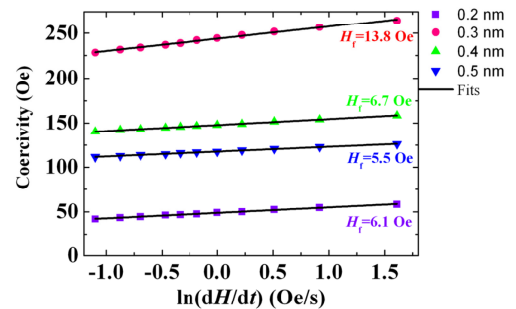


FIG. 3. Field sweep rate, $\ln(dH/dt)$, dependence of the coercivity in $[\text{CoSiB } (t_{\text{CoSiB}} \text{ nm})/\text{Pd } (1.3 \text{ nm})]_5$ multilayers with various thicknesses, $t_{\text{CoSiB}} = 0.2, 0.3, 0.4$, and 0.5 nm . The solid lines represent linear fits with Eq. (1).

Ab initio Calculations on the Effect of Mn Substitution in the κ -carbide Fe_3AlC

Ji-Young NOH and Hanchul KIM*

Department of Physics, Sookmyung Women's University, Seoul 140-742, Korea

(Received 4 January 2013)

We have performed spin-polarized density functional theory calculations on κ -carbides, $(\text{Fe}, \text{Mn})_3\text{AlC}$, to investigate both the effect of different exchange-correlation functionals and the effect of Mn substitution. Differently from our previous calculations using the Perdew-Wang functional, Fe_2MnAlC is found to be the most stable crystal among crystalline κ -carbides by using the Perdew, Burke, and Ernzerhof functional. Supercell calculations to simulate low Mn concentration show that substitutional Mn atoms hardly interact with each other and suggest that a random alloy model can be applied. The stabilization of Fe_2MnAlC and its enhanced magnetization are attributed to the formation of a $-\text{Mn}-\text{C}-$ linear chain structure.

PACS numbers: 71.20.-b, 71.20.Be, 75.50.Bb, 75.30.Hx

Keywords: Density functional theory, κ carbide, Substitution, Alloy, Density of states, Magnetic moment

DOI: 10.3938/jkps.62.481

I. INTRODUCTION

Fe-Mn-Al-C alloys are based on austenite (fcc γ -iron) and are known to show better strength than the ferrite-based alloys. This is attributed to the closely packed arrangement of atoms in austenitic alloys [1]. In addition to the enhanced strength, Fe-Mn-Al-C alloys have further excellent mechanical properties, such as light weight, superior ductility, and high-temperature oxidation resistance [2,3]. Such desirable properties lead to potential applicability in advanced transportation systems, automotive vehicles, structural materials, and alternatives to Cr-Ni stainless steels [2–5]. These Fe-Mn-Al-C alloys are made up of an austenitic matrix with regularly-dispersed nano-sized κ -carbide, $(\text{Fe}, \text{Mn})_3\text{AlC}$ [2,6]. The uniformly-distributed κ -carbide within the austenitic matrix plays a role in enhancing the mechanical properties of Fe-Mn-Al-C steels [2,5,7,8].

The stoichiometric κ -carbide Fe_3AlC has a perovskite-type structure classified as $E2_1$ as shown in Fig. 1. It is formed by inserting one C atom at the octahedral (bcc) site surrounded by six Fe atoms of $\text{Fe}_3\text{Al}-L1_2$. In reality, stoichiometric Fe_3AlC has not been observed experimentally [7,9]; instead, κ -carbides are known to have a continuous spectrum with varying contents of Mn from Fe_3AlC to Mn_3AlC , *i.e.*, $(\text{Fe}, \text{Mn})_3\text{AlC}$ [4,10].

Only a few previous density functional theory (DFT) calculations exist for κ -carbide. In contrast with the extensive experimental research on Fe-Mn-Al-C steels,

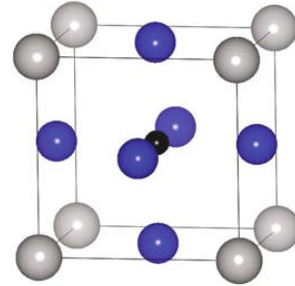


Fig. 1. (Color online) Cubic unit cell of Fe_3AlC showing the crystal structure of $E2_1$. Blue (face centers), gray (corners), and black (body center) balls indicate Fe, Al, and C atoms, respectively. Fe atoms will be replaced by Mn atoms as substitutional Mn atoms are introduced.

these theoretical works mainly focus on the influence of C atoms. Connétable and Magis [7] investigated the influence of the C atom by comparing $\text{Fe}_3\text{Al}-L1_2$ and $\kappa\text{-Fe}_3\text{AlC}$. They found that the insertion of a C atom decreased the magnetic moment of the neighboring Fe atoms and modified the heat capacity and the elastic constants significantly. Kellou *et al.* found that the presence of κ -carbide precipitate enhanced the mechanical properties and that the insertion of C atoms produced a softening of the magnetic properties [8]. In a recent paper [11], we reported DFT calculations for κ -carbides by employing the generalized-gradient approximation (GGA) of Perdew and Wang (PW91) [12]. We found that the

*hanchul@sookmyung.ac.kr

Table 1. Lattice constants ($\text{\AA}$) and formation energy (eV/unit cell) of crystalline κ -carbides, $(\text{Fe,Mn})_3\text{AlC}$. Numbers in parentheses are the lattice constant along the z -axis of the tetragonal structure.

	Lattice constants				E_{form}	
	PBE ^a	PW91 ^b	Other calc. ^c	Exp.	PBE ^a	PW91 ^b
Fe_3AlC	3.76	3.75	3.75	3.78 ^{d,e}	0.977	1.259
Fe_2MnAlC	3.78 (3.79)	3.76 (3.77)	—	—	1.013	1.176
FeMn_2AlC	3.80 (3.77)	3.79 (3.76)	—	—	0.984	1.066
Mn_3AlC	3.81	3.80	—	3.88 ^f	0.963	0.940

^aPresent work, ^bRef. 11, ^cRef. 7, ^dRef. 4, ^eRef. 9, ^fRef. 24.

insertion of carbon at the bcc site reduced the magnetic moment of Fe and enhanced the rigidity compared with the host material $\text{Fe}_3\text{Al-L1}_2$. Increasing the Mn content in $\text{Fe}_{3-x}\text{Mn}_x\text{AlC}$ ($x = 0, 1, 2, 3$), we found that the formation energy decreased monotonically, the magnetization increased with an exceptional minimum value for Fe_2MnAlC , and the Mn-C bond was harder than the Fe-C bond.

In this work, we performed spin-polarized DFT calculations to further investigate the effects of Mn substitution for Fe in κ -carbides. First, we employed a different GGA functional by Perdew, Burke, and Ernzerhof (PBE) [13] to compare with our previous results that were obtained using the PW91 functional. Then, we used different supercells to investigate the behavior of substitutional Mn at low Mn concentration and examined the energetics, as well as the electronic and the magnetic properties.

II. COMPUTATIONAL METHODS

We performed DFT calculations [14, 15] on $(\text{Fe,Mn})_3\text{AlC}$ in the spin-polarized manner by using the Vienna *Ab initio* Simulation (VASP) package [16–19] with the plane-wave basis set. The electron-ion interaction was described by using the projector-augmented-wave (PAW) potentials [20, 21], and the electron-electron exchange-correlation interaction was described within the GGA-PBE [13]. Brillouin zone integrations were done by employing the linear tetrahedron method [22], with the Γ -centered Monkhorst-Pack $\mathbf{k}$ -point mesh equivalent to $24 \times 24 \times 24$ for the BZ of a $1 \times 1 \times 1$ cubic unit cell [23]. Plane waves up to the kinetic energy cutoff of 500 eV were used to expand single-electron wave functions. The total energy was converged to guarantee an accuracy smaller than 1 meV/(unit cell). In order to simulate low Mn concentrations, we used supercells of $2 \times 2 \times 2$ and $2 \times 2 \times 4$, which yield Mn concentrations of $x = 0.125$ and 0.0625, respectively, if a single Fe atom is replaced by Mn.

III. RESULTS AND DISCUSSION

1. Structural Properties and Energetics of Crystalline κ -carbides

We performed both spin-polarized and spin-compensated calculations to determine the equilibrium properties of bulk crystalline κ -carbides, $\text{Fe}_{3-x}\text{Mn}_x\text{AlC}$ ($x = 0, 1, 2, 3$), by using the PBE exchange-correlation functional. In accordance with our previous calculations using the PW91 functional [11], the ferromagnetic (FM) phases are energetically preferred to the non-magnetic phases. The calculated equilibrium lattice constants are shown in Table 1 and are compared with previous calculations and available experimental data. The lattice constants increase as the Mn content increases from Fe_3AlC to Mn_3AlC , and the PBE results are slightly larger and, thus, closer to experimental data than those of PW91. Notice that Fe_2MnAlC and FeMn_2AlC have a tetragonal symmetry and are characterized by two different lattice constants. The increase in the lattice constants with increasing Mn substitution is thought to originate from the atomic radius of Mn being larger than that of Fe [11, 25]. The difference between the lattice constant between Fe_3AlC and Mn_3AlC is almost the same as the difference between the atomic radii of Fe (1.56 $\text{\AA}$) and Mn (1.61 $\text{\AA}$). The substitution of the lighter Mn atom for the heavier Fe atom and the increase in lattice constants work together to produce the light steel Fe-Mn-Al-C.

Using the equilibrium lattice constants, we performed total energy and electronic structure calculations with the PBE functional to reexamine the previous PW91 results [11] for the thermodynamic stability and the magnetic properties of κ -carbides. The formation energy is obtained by using DFT total energies as in the following equation:

$$E_{\text{form}} = \sum_X n_X \times \varepsilon_X - E_{\text{DFT}}, \quad (1)$$

where n_X and ε_X stand for the number of atoms of element X in the simulation cell and the atomic chemical potential (*i.e.*, the total energy per atom of the elemental crystal) of element X , respectively. Notice that a larger

Electronic Structures of a Zn Vacancy on the ZnO(10 $\bar{1}$ 0) Surface: Density Functional Theory Calculations

Kisung CHAE

*Department of Materials Science and Engineering,
Rensselaer Polytechnic Institute, Troy, New York 12180, USA*

Hanchul KIM*

Department of Physics, Sookmyung Women's University, Seoul 140-742, Korea

(Received 7 January 2013)

We perform spin-polarized density functional theory calculations for a Zn vacancy on the ZnO(10 $\bar{1}$ 0) surface. Two stable configurations of the surface Zn vacancy are found, and the activation energy barrier is estimated to be ~ 0.01 eV. The lower energy configuration has a newly formed surface Zn-O bond to restore the bulk-like structure on the surface. Due to the newly formed bond, the vacancy state in the band gap is characterized by a complicated hybridization of neighboring surface and subsurface atoms and by a more extended electron density. Despite such a hybridization, the surface Zn vacancy is found to have a robust magnetic moment of $1 \mu_B$, implying that surface Zn vacancies may be responsible for the ferromagnetism observed in ZnO thin films and nanoparticles. Simulated scanning tunneling microscope images show that the two structures of the surface Zn vacancy can be distinguished in the filled-state images.

PACS numbers: 75.70.-i, 73.22.-f, 73.20.Hb, 73.61.Ga

Keywords: Density functional theory, Zinc oxide, (10 $\bar{1}$ 0) surface, Zn vacancy, Energy barrier, Defect state, Scanning tunneling microscopy images

DOI: 10.3938/jkps.62.508

I. INTRODUCTION

In recent experiment, nonmagnetic oxide materials (II-VI semiconductors) such as HfO₂, TiO₂, and ZnO were found to show ferromagnetism (FM) without doping at room temperature (RT) [1,2]. Other than the observation of room-temperature FM, which opens up a potential application in spintronics, ZnO is very attractive among oxides for its applicability in optoelectronic devices because it is a wide-band-gap (~ 3.4 eV) semiconductor and has a large exciton binding energy (~ 60 meV) [3].

Regarding the origin of FM in ZnO, different scenarios have been suggested in varying perspectives. Sundaresan and Rao attributed the FM to an exchange interaction of O vacancies (V_O) located at the surfaces of the nanoparticles [2]. On the other hand, more recent studies revealed that the Zn vacancy (V_{Zn}) had a magnetic moment both experimentally and theoretically by considering various structures such as nanowires and defects on surfaces and/or near-surface regions [4–7]. In a very recent publication, we found that V_{Zn} on a non-polar (10 $\bar{1}$ 0) surface had a significantly smaller formation en-

ergy than it did in the bulk [8]. The surface V_{Zn} was found to have two different structures, and the two structures showed a finite magnetic moment of $1 \mu_B$. There seems to be an increasing consensus that cation vacancies play a role in stabilizing the observed RT-FM in undoped, inorganic, compound semiconductors.

In this research, we extended our previous work on V_{Zn} on the ZnO (10 $\bar{1}$ 0) surface by employing first-principles calculations. A non-polar (10 $\bar{1}$ 0) surface is known to have the smallest surface formation energy among low-index surfaces [9]; thus, {10 $\bar{1}$ 0} facets would be the major constituting facet of a crystalline ZnO nanoparticle. We investigated the energetics involved in the structural change between the two local minimum structures of V_{Zn} and obtained a detailed electronic structure of V_{Zn} by examining the angular-momentum-projected density of states in relation with the hybridized defect states. The scanning tunneling microscopy (STM) images of V_{Zn} were simulated to provide characteristic fingerprints of V_{Zn} that could be used in future experimental STM studies.

*E-mail: hanchul@sookmyung.ac.kr

II. COMPUTATIONAL METHODS

First-principles calculations based on density functional theory (DFT) were performed by using the Vienna *ab initio* simulation package [10]. Projector-augmented-wave potentials [11] were employed to describe the ions, and the generalized gradient approximation [12] was used for the exchange-correlation functionals. A plane-wave basis set with a kinetic energy cut-off of 400 eV was used to expand the wave functions. The stopping criterion for ionic relaxation was that the Hellmann-Feynman forces exerted on the ions were less than 0.01 eV/Å. The DFT+ U method [13] with the on-site Coulomb correlation parameter $U = 6.0$ eV was applied to the localized Zn(3d) orbitals in order to correct the underestimated band-gap from conventional DFT calculations. The (10 $\bar{1}$ 0) surface of wurtzite ZnO was modeled by using a four-bilayer slab. Valence charge-modified hydrogen atoms were used to compensate the dangling bonds at the bottom surface of the slab. The V_{Zn} was generated by removing one of the Zn atoms on the top surface of the slab. In order to describe an isolated V_{Zn} , we used a 4×3 supercell. Using the relaxed atomic structure and converged Kohn-Sham eigenfunctions, we simulated constant-current STM images following the Tersoff-Hamann scheme [14]. The approximate tunneling current (I) in the Tersoff-Hamann scheme can be written as

$$I(\mathbf{r}, V) = \frac{16\pi^3 C^2 \hbar^2 e}{\kappa^2 m^2} \rho^T \int_{E_F^S}^{E_F^S + eV} \rho^S(\mathbf{r}, \epsilon) d\epsilon, \quad (1)$$

where the superscripts S and T stand for the sample and the tip, respectively, V is the applied sample bias voltage, and ρ is the local charge density.

III. RESULTS AND DISCUSSION

1. Stable Structures and Energetics

After a structural relaxation, we found two stable structures of V_{Zn} on the (10 $\bar{1}$ 0) surface as shown in Fig. 1. One structure (C_1) remains more or less the same as the clean (10 $\bar{1}$ 0) surface except for one missing Zn atom on the surface while the other structure (C_2) shows a significant relaxation of a nearest-neighbor (NN) O atom (O_{0S}) on the surface. Upon relaxation, O_{0S} moves along the [0001] direction and forms a new bond with another surface Zn atom (Zn_{1S}) in a neighboring Zn-O chain. The O_{0S} atom is displaced towards the [0001] direction by 0.98 Å and moves upwards by 0.07 Å compared with its position on the clean surface. This displacement pattern agrees with a previous report [15], but the amounts of displacement are somewhat smaller than the previous values of 1.04 Å in the [0001] direction and 0.15 Å in the

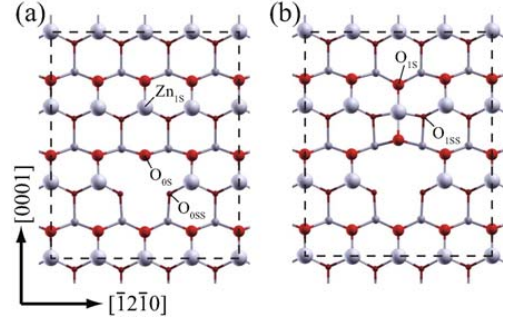


Fig. 1. (Color online) Relaxed atomic structures of V_{Zn} on the (10 $\bar{1}$ 0) surface: (a) C_1 and (b) C_2 . Grey and red balls are for Zn and O atoms, respectively, with the larger balls for surface atoms and the smaller ones for subsurface atoms. A (4×3) supercell is designated by the dashed rectangle.

[10 $\bar{1}$ 0] direction. Similarly, Zn_{1S} also moves towards the O_{0S} by 0.42 Å and is pushed up by 0.05 Å. As a consequence, Zn_{1S} attains four NN O atoms, and the chemical environment of Zn_{1S} becomes similar to the bulk environment except for the distorted tetrahedron formed by the surrounding O NNs. While the Zn-O bonds on the clean (10 $\bar{1}$ 0) surface are shortened to 1.82 (1.89) Å between a surface Zn and O atoms at the surface (subsurface) compared with the bulk bond length of 1.94 Å, the bond lengths in the newly-formed distorted O tetrahedron of C_2 are altered to 1.94, 1.87, and 1.95 Å for Zn_{1S} - O_{0S} , Zn_{1S} - O_{1S} , and Zn_{1S} - O_{1SS} , respectively. A driving force for the transition from C_1 to C_2 seems to be the electronic energy gain due to the formation of an additional Zn-O bond, which results in the local bulk-like environment of Zn_{1S} . In fact, C_2 has a significantly lower vacancy formation energy (E_{form}) than C_1 by 0.36 eV [8]. The electronic energy gain by forming the new bonds seems to exceed the strain energy cost of O_{0S} . The details of the electronic structure change will be discussed later.

Because V_{Zn} on ZnO(10 $\bar{1}$ 0) has two locally stable structures, figuring out the energy barrier for the C_1 -to- C_2 transition is interesting. To get the potential energy landscape between the two structures, we linearly interpolated C_1 and C_2 with nine images including the C_1 and C_2 configurations and calculated the total energies without structural relaxation. The resulting potential energy surface is plotted in Fig. 2. The result shows that C_1 is a very shallow local minimum and that the activation energy barrier of the transition is as small as ~ 0.01 eV. Thus, the transition from C_1 to C_2 is expected to occur readily as soon as V_{Zn} is created on the surface. The observed ferromagnetism in pristine ZnO thin films and nanoparticles has been reported to be retained even at room temperature, so C_2 would play a significant role in the unexpected room-temperature ferromagnetic properties relevant to V_{Zn} .

Comment on “Atomistic Picture of Charge Density Wave Formation at Surfaces”

In a recent Letter [1], Wall *et al.* investigated the melting and refreezing behavior of the charge density wave (CDW) in In wires on Si(111) by using ultrafast electron diffraction and density functional theory calculations. They observed that the CDW (a low-temperature 8×2 structure) melted (transformed into a high-temperature 4×1 structure) by optical pumping at 20 K and refreezed slowly due to an energy barrier, suggesting a first-order transition. From the observation of the shortened refreezing time for the aged samples, the authors claimed that the adsorbates not only trigger the recovery of the CDW but also raise the transition temperature (T_c). In this Comment, we point out that (1) the conclusion of the adsorbates facilitating the 4×1 to 8×2 transformation, in analogy to the impurity-induced freezing of the supercooled liquid, is not substantiated and (2) the role of the adsorbates claimed in the Letter is not consistent with the previous findings [2–4].

We highlight two things that were observed but not properly addressed by the authors: (i) only a fraction of the 8×2 surface turned into the 4×1 phase by the laser-induced excitation for the aged surface and (ii) the amount of the melted CDW decreases with increasing adsorption time, as shown in Fig. 1(a). To establish the facilitating role of the adsorbates as nucleation seeds, one should show that the *same* amount of melted CDW becomes refrozen in shorter time with the adsorbates. However, the observation of partial melting of the CDW indicated that significant numbers of $8 \times 2/4 \times 1$ phase boundaries should be present, and the adsorbates simply increased them [see Fig. 1(b)]. It is readily conceivable that refreezing of the melted CDW takes place at these $8 \times 2/4 \times 1$ interfaces, not from immediate vicinities of the adsorbates, and propagates into the melted (4×1) region along the wire.

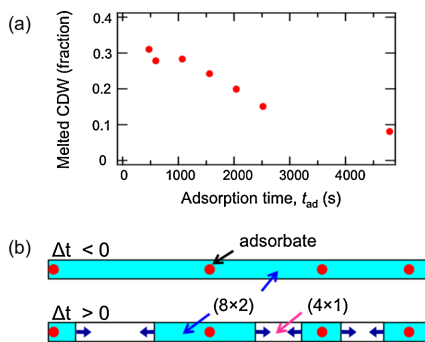


FIG. 1 (color online). (a) Amount of the melted CDW as a function of adsorption time [reconstructed from Fig. 3(a) of Ref. [1]]. (b) Persistence of the 8×2 patches (shaded) within the 4×1 (white) phase on the surface with adsorbates after optical pumping ($\Delta t = 0$). The propagation of the 8×2 phase fronts is indicated by arrows.

This interpretation is natural to explain the complete recovery of the two-dimensional 8×2 correlation by refreezing. If the refreezing propagates from the randomly located adsorbates as the authors suggested, the transverse correlation between the neighboring 4×2 chains cannot be achieved. In this sense, the refreezing is not necessarily related to the presence of the adsorbates. Therefore, the data in the Letter [1] do not support the authors' claim that the adsorbates expedite the freezing into the 8×2 structure, acting as the nucleation seeds. Instead, the decreased amounts of the melted CDW with the increasing adsorbates shows that the adsorbates prevent the neighboring 8×2 region from melting.

The authors also reported that the T_c value was 100 K for the 480 s aged surface and raised with increasing aging time. We note that the value $T_c \sim 100$ K is rather too low compared with that of the pristine surface [5]. This suggests possible contamination by T_c -decreasing adsorbates [2–4]. It is, however, self-contradictory to the increase of T_c with further aging. The raised T_c with adsorbates implies that the relevant adsorbate is the exceptional T_c -raising adsorbates, i.e., oxygen [3]. In fact, the authors implicitly assumed this by citing the previous work [3] to estimate the amount of the adsorbates. The most likely residual gas contaminating a cold surface is water. However, the water molecules are not likely to be dissociated into oxygen atoms at a temperature as low as 20 K. In addition, we found that the adsorption of the residual gases at 80 K for 12 h [compared with 80 min at 20 K in the Letter [1]] hardly changed the T_c [6], contrasting with the authors' observation. Therefore, the result of T_c increase by aging, claimed in the Letter, is questionable.

This work was supported by a National Research Foundation grants funded by the Korean Government (Grants No. 2010-0012231, No. 2008-0059595, and No. R11-2008-053-02001-0 through the Quantum Metamaterials Research Center).

Hyungjoon Shim, Jonghoon Yeo, and Geunseop Lee*

Department of Physics
Inha University
Inchon 402-751, Korea

Hanchul Kim

Department of Physics
Sookmyung Women's University
Seoul 140-742, Korea

Received 22 April 2013; published 1 October 2013

DOI: 10.1103/PhysRevLett.111.149601

PACS numbers: 68.35.Rh, 61.05.jh, 68.43.Bc, 71.45.Lr

*glee@inha.ac.kr

- [1] S. Wall, B. Krenzer, S. Wippermann, S. Sanna, F. Klasing, A. Hanisch-Blicharski, M. Kammler, W. G. Schmidt, and M. Horn-von Hoegen, *Phys. Rev. Lett.* **109**, 186101 (2012).

-
- [2] H. Shim, S.-Y. Yu, W. Lee, J.-Y. Koo, and G. Lee, *Appl. Phys. Lett.* **94**, 231901 (2009).
- [3] G. Lee, S.-Y. Yu, H. Shim, W. Lee, and J.-Y. Koo, *Phys. Rev. B* **80**, 075411 (2009); T. Shibusaki *et al.*, *ibid.* **81**, 035314 (2010).
- [4] W. Lee, H. Shim, and G. Lee, *J. Korean Phys. Soc.* **56**, 943 (2010).
- [5] The reported values of T_c for the pristine surface are 120–130 K in the literature. See the references in Ref. [1].
- [6] H. Shim *et al.* (unpublished).

Electron doping limit in Al-doped ZnO by donor-acceptor interactions

Ji-Young Noh,^{1,2} Hanchul Kim,¹ Yong-Sung Kim,^{2,3,a)} and C. H. Park⁴

¹Department of Physics, Sookmyung Women's University, Seoul 140-742, South Korea

²Korea Research Institute of Standards and Science, Daejeon 305-340, South Korea

³Department of Nano Science, University of Science and Technology, Daejeon 305-350, South Korea

⁴Research Center for Dielectric and Advanced Matter Physics, Pusan National University, Pusan 609-735, South Korea

(Received 17 January 2013; accepted 27 March 2013; published online 15 April 2013)

We investigate the maximum available free electron carrier density in Al-doped *n*-type ZnO, based on density-functional theory calculations. The Coulomb interactions between the Al dopants and the Zn-vacancy native acceptors are found to limit the carrier density. In typical growth conditions, the *n*-type doping limit is found to be in the range of $10^{19} \sim 10^{21} \text{ cm}^{-3}$. © 2013 AIP Publishing LLC
[<http://dx.doi.org/10.1063/1.4801533>]

I. INTRODUCTION

Heavy doping has been a frontier of semiconductor science and engineering, and recently, it has been another pioneering work in solar cells and flat panel displays with the emerging transparent conducting oxide (TCO) technology. Requirements for a good TCO are (i) high mobility and density of carriers, (ii) high transparency, and (iii) high process compatibility for its manufacturing. Although In-Sn-oxide (ITO) has been widely used as a representative TCO, a problem is that In is contained in the ITO as an important constituent, whose natural resources are limited on the earth crust. Nowadays, finding an In-free TCO becomes an important topic in material science communities.

The Al-doped ZnO (AZO) is one of the candidates.¹ The mobility of carriers has been reported to be comparable to or even higher than that of the ITO.² The natural resources are abundant, and it is easily *n*-type doped.^{3,4} However, the density of free carriers has been known to be limited to about $10^{19} \sim 10^{21} \text{ cm}^{-3}$ even with a higher concentration of the dopants,^{5–12} which has been known as the *n*-type doping limit in ZnO. Although a lot of efforts have been carried out to extend the doping limit, no clear mechanism for the origin of the *n*-type doping limit has been investigated yet.

There is a similar doping limit problem in the Si technology. For ultra-integrated circuits such as the state of the art of micro-processing unit and dynamic random access memory, achieving a high concentration of free carriers is an important but a difficult job. In heavily *n*-type doped Si, it is highly possible for one dopant to meet another one at a nearby site, and it has been suggested that the atomic structure around the two dopants are seriously distorted (forming a donor-pair defect) with capturing two electron carriers.^{13,14} Then, the Fermi level and the carrier concentrations can be no more increased.

There are several notable factors that can determine the doping limit in *n*-type ZnO. (i) The dominant acceptor-like

native defect is Zn-vacancy (V_{Zn}), which is a double acceptor,^{15–18} and can act as a compensating center in *n*-type ZnO. (ii) Since the density of states near the conduction band edge is low, the Fermi level shift is sensitive to the increase of the electron carrier concentration,¹⁹ by which the acceptor-like defect formation can be enhanced. (iii) Coulomb interactions between the charged centers can be strong, due to the low dielectric constant of ~ 8.65 (resulting from the wide band gap and the high ionicity of ZnO) and the small lattice constants (by which more atomic sites are available close to a charged center).

In this study, we investigate the *n*-type doping limit in ZnO by using density-functional theory (DFT) calculations. We find that the doping limit comes from the attractive interactions between the positively charged Al^{+} donors and the negatively charged V_{Zn}^{2-} native acceptors. Without considering the donor-acceptor interactions, we could not find an intrinsic *n*-type doping limit, but the carrier concentrations were increased with the Al^{+} dopant concentrations. The *n*-type doping limit with considering the donor-acceptor interactions is found to be in the range of $10^{19} \sim 10^{21} \text{ cm}^{-3}$ in typical growth conditions of ZnO.

II. METHODS

We performed DFT calculations as implemented in the Vienna *ab initio* simulation package (VASP) code.^{20,21} The projector augmented wave (PAW) pseudopotentials,²² a kinetic energy cutoff of 495 eV, a $2 \times 2 \times 2$ *k*-point mesh in the Brillouin zone, and a 96-atomic supercell in the wurtzite structure were used. The local density approximation²³ plus *U* (LDA + *U*) with the *U* = 6 eV on-site Coulomb interaction energy²⁴ for the Zn 3*d* states was used for the exchange-correlation energy. With *U* = 6 eV, the position of the Zn-3*d* bands in energy level is found near -7 eV below the valence band maximum (VBM), which is close to the experiments.²⁵ The calculated LDA + *U* band gap was 1.74 eV. The theoretical equilibrium lattice constants of $a = 3.148 \text{ \AA}$ with $c/a = 1.612$, calculated within the LDA + *U*, were used. The atomic structures were relaxed within the LDA + *U* until the Hellmann-Feynman forces were less than $0.01 \text{ eV/\AA}$.

^{a)}yongsung.kim@kriss.re.kr

III. RESULTS

A. Non-interacting case

We plot the calculated formation energies of the *isolated* native defects of V_{Zn} , O-interstitial (O_i), O-vacancy (V_O), and Zn-interstitial (Zn_i) in ZnO, as a function of the O chemical potential (μ_O) in Fig. 1. Since we are interested in *n*-type ZnO, the Fermi level (ε_f) is supposed to be at the conduction band minimum (CBM), and then the stable charge states of the native defects are V_{Zn}^{2-} , O_i^{2-} , V_O^0 , and Zn_i^{2+} . The O_i^{2-} is in a split-interstitial configuration, and the Zn_i^{2+} is at the octahedral site.^{15–18} The formation energies (E_f^α) of the isolated native defects are calculated from

$$E_f^\alpha = E_{t,\alpha} - (N_{\text{Zn}}\mu_{\text{Zn}} + N_{\text{O}}\mu_{\text{O}}) + q(\varepsilon_f + \varepsilon_{\text{VBM}}), \quad (1)$$

where α is either the V_{Zn}^{2-} , O_i^{2-} , V_O^0 , or Zn_i^{2+} , and $E_{t,\alpha}$ is the total energy of the supercell containing the α defect, N_{Zn} and N_{O} are the numbers of Zn and O atoms, respectively, in the supercell, μ_{Zn} and μ_{O} are the Zn and O chemical potentials, respectively, q is the charge state of the α defect, and ε_{VBM} is the VBM energy level of the perfect ZnO. The μ_{Zn} and μ_{O} chemical potentials are correlated by $\mu_{\text{Zn}} + \mu_{\text{O}} = \mu_{\text{ZnO}}^{\text{bulk}}$. For the O-rich limit condition, $\mu_{\text{O,max}} = E_t(\text{O}_2)/2$, where $E_t(\text{O}_2)$ is the DFT total energy of an isolated O_2 molecule. For the Zn-rich limit condition, $\mu_{\text{Zn,max}} = E_t(\text{Zn})$, where $E_t(\text{Zn})$ is the DFT total energy per atom of the bulk Zn in the hexagonal closed packed structure. Then, $\mu_{\text{O,min}} = \mu_{\text{ZnO}}^{\text{bulk}} - \mu_{\text{Zn,max}}$. The heat of formation of ZnO (H_{ZnO}) can be calculated from $H_{\text{ZnO}} = \mu_{\text{O,max}} - \mu_{\text{O,min}}$, and it is 3.3 eV. We set $\mu_{\text{O,max}} = 0$ eV for convention, and then the μ_{O} range for stable ZnO formation is -3.3 eV (Zn-rich limit) $< \mu_{\text{O}} < 0$ eV (O-rich limit). The ε_f is set to zero, when it is at the ε_{VBM} .

As shown in Fig. 1, the dominant acceptor-like native defect in ZnO is the V_{Zn}^{2-} , and thus the V_{Zn}^{2-} is expected to act as a main compensating center for donors in *n*-type ZnO. As the stoichiometry goes to the O-rich conditions (as μ_{O} increases), the V_{Zn}^{2-} is more stabilized. In Zn-rich conditions, since the formation of V_{Zn}^{2-} is less effective, the compensation of donors by V_{Zn}^{2-} would not be effective. On the other hand, in O-rich conditions, the formation of V_{Zn}^{2-} is highly

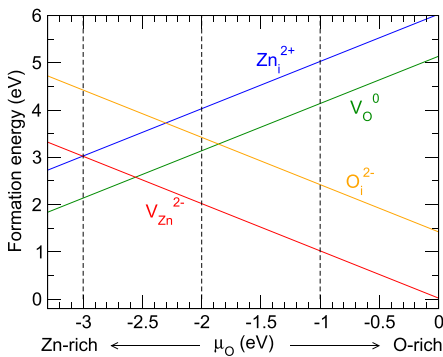


FIG. 1. Calculated formation energies of the isolated V_{Zn}^{2-} , O_i^{2-} , V_O^0 , and Zn_i^{2+} in *n*-type ZnO (when ε_f is at the LDA + U CBM), as a function of the O chemical potential, μ_{O} .

probable and the V_{Zn}^{2-} is considered to effectively compensate the donors in *n*-type doped ZnO.

Quantitatively, we calculate the equilibrium concentrations of the V_{Zn}^{2-} compensating centers, as a function of the Al doping concentrations for four different stoichiometry conditions of $\mu_{\text{O}} = 0, -1, -2$, and -3 eV. The Al dopant is supposed to be located at a Zn substitutional site in the (1+) charge state ($\text{Al}_{\text{Zn}}^{1+}$). The equilibrium concentration of the V_{Zn}^{2-} defect is calculated through $[V_{\text{Zn}}^{2-}] = N_0 \times \exp(-E_f/k_B T)$, where N_0 is the number of available sites for the defect formation ($4.59 \times 10^{22} \text{ cm}^{-3}$ for a substitutional defect in ZnO), E_f is the formation energy of the isolated V_{Zn}^{2-} defect, and T is the growth temperature (supposed to be 900 K). The formation energy of V_{Zn}^{2-} depends on the Fermi level, and the equilibrium Fermi level ($\varepsilon_{f,\text{eq}}$) is determined by solving the charge neutrality condition of the system: $[\text{Al}_{\text{Zn}}^{1+}] + n_h = n_e + 2 \times [V_{\text{Zn}}^{2-}]$, where the n_e and n_h are the electron and hole concentrations in the bulk bands, respectively.²⁶ The bulk conduction and valence bands are approximated by the quadratic forms with the electron and hole effective masses of $m_e^* = 0.23$ and $m_h^* = 0.5$, respectively.

Figure 2 shows the calculated equilibrium concentrations of V_{Zn}^{2-} and n_e as a function of the $\text{Al}_{\text{Zn}}^{1+}$ doping density. In Zn-rich conditions ($\mu_{\text{O}} = -3$ and -2 eV), compensation of the donors by V_{Zn}^{2-} is not effective as expected above. As shown in Figs. 2(a) and 2(b), the V_{Zn}^{2-} concentrations are far less than the $\text{Al}_{\text{Zn}}^{1+}$ density, and the electron carrier concentrations are found to increase simply with the $\text{Al}_{\text{Zn}}^{1+}$ density. On the other hand, in O-rich conditions ($\mu_{\text{O}} = -1$ and 0 eV), the V_{Zn}^{2-} formation is highly probable, and the V_{Zn}^{2-} concentrations are found to be comparable to the $\text{Al}_{\text{Zn}}^{1+}$ doping density. As a result, the electron carrier concentrations are largely reduced far less than the $\text{Al}_{\text{Zn}}^{1+}$ doping density, especially in the highly O-rich condition of $\mu_{\text{O}} = 0$ eV [Fig. 2(d)]. Nevertheless, we could not find a limit of the carrier concentrations ($n_{e,\text{max}}$) even in the O-rich conditions, and the n_e increases monotonically with the $\text{Al}_{\text{Zn}}^{1+}$ density. If this compensating behavior was the case, the *n*-type doping limit of ZnO would come from the solubility limit of the dopants in ZnO, which is known to be about 10^{21} cm^{-3} for Al.^{27–29} However, the *n*-type doping limit of Al-doped ZnO has been reported experimentally to be in the range of $10^{19} \sim 10^{21} \text{ cm}^{-3}$.^{5–12}

B. Interacting case

We consider an additional effect of the defect interactions between the $\text{Al}_{\text{Zn}}^{1+}$ dopants and the V_{Zn}^{2-} native defects. Based on DFT calculations, we put two defects, $\text{Al}_{\text{Zn}}^{1+}$ and V_{Zn}^{2-} , into the supercell at various sites [for an example, as shown in Fig. 3(a)], and calculated the total energies. With respect to the total energy of the isolated $\text{Al}_{\text{Zn}}^{1+}$ and V_{Zn}^{2-} , the energies of the $\text{Al}_{\text{Zn}}^{1+}$ and V_{Zn}^{2-} pairs are plotted as a function of the defect separation (d) in Fig. 3(b), which can be defined as an interaction energy $[U^{\text{DFT}}(d)]$ between the two defects. As shown in Fig. 3(b), the $|U^{\text{DFT}}(d)|$ is found to increase as the defect separation d is closer. We also plot the Coulomb interaction energy, $U^{\text{Coul}}(d) = \alpha q q' / 4\pi \varepsilon d$, where α is a constant, q is (1+) for $\text{Al}_{\text{Zn}}^{1+}$, q' is (2−) for V_{Zn}^{2-} , and ε is the

Cation composition effects on electronic structures of In-Sn-Zn-O amorphous semiconductors

Ji-Young Noh,^{1,2} Hanchul Kim,¹ Ho-Hyun Nahm,² Yong-Sung Kim,^{2,3,a)} Dae Hwan Kim,⁴ Byung-Du Ahn,⁵ Jun-Hyung Lim,⁵ Gun Hee Kim,⁵ Je-Hun Lee,⁵ and Junho Song⁵

¹Department of Physics, Sookmyung Women's University, Seoul 140-742, South Korea

²Korea Research Institute of Standards and Science, Daejeon 305-340, South Korea

³Department of Nano Science, University of Science and Technology, Daejeon 305-350, South Korea

⁴School of Electrical Engineering, Kookmin University, Seoul 136-702, South Korea

⁵Samsung Display Co., Ltd., Yongin 446-711, South Korea

(Received 11 January 2013; accepted 16 April 2013; published online 9 May 2013)

Based on density-functional theory calculations, the effects of cation compositions on electronic structures of In-Sn-Zn-O amorphous semiconductors were investigated. We considered various composition ratios of In, Sn, and Zn in O stoichiometric condition, and found that the conduction band minimum (CBM) energy level decreases and the valence band tail (VBT) energy level extent increases as the sum of In and Sn ratios ($R_{\text{In}}+R_{\text{Sn}}$) increases. The CBM lowering is attributed to the increased overlap of the In-5s and Sn-5s orbitals as the $R_{\text{In}}+R_{\text{Sn}}$ increases, and correspondingly the electron effective masses (m_e^*) are found to be reduced. The VBT increase is found to be due to the increased density of the In and Sn atoms, near which the O-2p inter-site $pp\sigma^*$ coupling is larger than that near the Zn atoms. The acute O-(In,Sn)-O angles are suggested to be structurally important, giving the stronger O-O $pp\sigma^*$ coupling. © 2013 AIP Publishing LLC. [<http://dx.doi.org/10.1063/1.4803706>]

I. INTRODUCTION

Amorphous oxide semiconductors (AOS) have high electron mobility ($>10 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$) even in the amorphous phase.¹ The amorphous InGaZnO_4 is a prototype AOS, and various kinds of AOS with alternative cations (Sn, Hf, Zr, Al, Si, and so on) have been suggested.^{2–6} Basically, the high electron mobility originates from the cation s -like orbitals of the conduction band (CB) states. Since the s orbitals are spherical, their inter-site overlap is robust to the orientational disorder of the amorphous structure.⁷ It is in stark contrast to the hydrogenated amorphous Si (a-Si:H), which has the sp^3 -hybridized conduction bands and exhibits severe localization of the sp^3 orbitals in amorphous phase (giving the conduction band tail states). Thus, the electron mobility of a-Si:H is degraded to about $\sim 1 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ from that of the crystalline Si of about $\sim 200 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$.

With the high mobility, the AOS has process compatibility. The low (room) temperature thin-film deposition process (usually sputtering) makes it possible to manufacture large-area thin-films on a variety choice of substrates, such as soda-lime glasses and flexible organic plastics, with low-cost and a high production (deposition) rate.^{8,9} Electronics on a large-area transparent substrate for commercial flat-panel displays have used a-Si:H thin-film transistors (TFTs), and large-area low-cost solar cells have been made by using a-Si:H pn junctions. The requirement of high speed TFTs for the next-generation high-resolution larger-area flat-panel displays makes the use of a-Si:H TFTs doubted due to its fundamental limit of electron mobility, and a new way by utilizing the AOS-based TFTs is nowadays intensively pursued.¹⁰

However, the facing bottleneck of the AOS-based TFTs for display applications is their instabilities under bias and/or illumination stress (IS). By positive gate-bias stress (PBS) in AOS-based TFTs, the threshold voltage (V_{th}) of the TFTs is known to be positive-shifted,^{11–17} and by negative bias and illumination stress (NBIS)^{4,5,17–24} or only by IS,^{17,25–29} the V_{th} is largely negative-shifted. The NBIS is known to be the most serious among the stresses,¹⁸ and it is important in active matrix liquid crystal displays (AM-LCD), since it is applied to the TFTs in AM-LCD steadily in the normal off-state stand-by condition. On the other hand, the PBS is steadily applied to the TFTs in active matrix organic light emitting diode (AM-OLED) displays during the normal state.

Understanding the microscopic mechanisms of the instabilities in AOS-based TFTs has been recently advanced. For the PBS instability, (i) it has been suggested that the electrons attracted to the gate electrode are trapped at the interface between the AOS channel and the gate insulator and/or injected into the gate insulator.^{11–13,17} (ii) The electron traps at H_2O -related trap sites have been also suggested.¹⁴ The electron traps can make a positive-shift of the V_{th} electrostatically, but their meta-stability is not clear yet. (iii) Surface doping induced by O_2 and H_2O molecules has been suggested as another mechanism.¹⁵ (iv) More importantly, it has been proposed that acceptor-like deep states is generated in AOS by the PBS.¹⁶ For the NBIS and IS instabilities, a few plausible models have been suggested. (i) Holes are generated by the light illumination and trapped at the interface and/or injected into the gate insulator.^{4,22,23} However, the IS instability has been observed not only in AOS TFTs but also in AOS thin films without a gate insulator through persistent photo-conductivity (PPC),^{26,29} and thus there has been considered to be an intrinsic mechanism in AOS materials.^{22,26}

^{a)}yongsung.kim@kriss.re.kr

(ii) NBIS or IS desorbs surface O atoms (reduction) and accordingly dope electrons,⁵ and an overlying passivation layer on AOS TFTs has been adopted to avoid it.⁵ Although somewhat improved, the NBIS and IS instabilities have been still serious even with avoiding the surface effect.^{4,19,20,22} (iii) Photo-ionized O-vacancies (V_O) persist with slow electron recombination.^{19–22,24,30,31} The V_O can give photo-carriers and be a hole-trap, but the meta-stability of the ionized V_O remains unclear. The recovery time of the negative-shifted V_{th} in AOS TFTs^{19,26–28} and the persistency of the PPC in AOS thin films^{26,29} are an order of days at room temperature, and the energy barrier from the meta-stable to the stable state should be very high. The activation energy has been reported to be in the range of 0.9–1.0 eV experimentally.²⁹ (iv) Meta-stable O_2^{2-} peroxide defect formation in AOS by NBIS or IS has been recently proposed.³² With excited holes in the valence band tail (VBT) states in AOS, the empty $pp\sigma^*$ -like VBT states induce a driving force to form an O-O bond through the $pp\sigma$ bonding interaction. Along with the O-O bond formation, the $pp\sigma^*$ state is heightened up into the CBs, and the electrons excited are doped with occupying the conduction band minimum (CBM) state. The energy barrier from the meta-stable O_2^{2-} peroxide to the normal amorphous oxide configuration is 0.97 eV in the hybrid density-functional theory (DFT) calculations.³²

In order to improve the stability of AOS-based TFTs against the bias and illumination stresses, a variety of attempts, such as a novel TFT architecture (including adopting a passivation layer),^{5,23,33,34} changing the gate insulator material,^{4,24} process design (for examples, O_2 or water annealing),^{21,24,35} impurity doping in AOS,² and optimizing the cation species and compositions of the AOS materials,^{26,36} have been made. The improvement of stability should not accompany (or at least minimize) any degradation of the electron mobility in the AOS. In this study, we investigate the cation composition effects on the electronic structures of In-Sn-Zn-O amorphous semiconductors to optimize both the stability and the mobility. We find that the CBM and VBT show nearly monotonic changes with respect to the In and Sn cation composition ratios. The In-5s and Sn-5s orbital overlaps and the stronger O-O $pp\sigma^*$ orbital coupling near the In and Sn atoms are found to be responsible for the CBM and VBT changes, respectively. The results indicate that the electron mobility and the stability are in a trade-off relation, when they are controlled by the In and Sn contents in amorphous In-Sn-Zn-O.

II. METHODS

We performed DFT calculations as implemented in the Vienna *ab initio* simulation package (VASP) code.^{37,38} The projector-augmented wave pseudopotentials³⁹ and the plane-wave basis set with a kinetic energy cutoff of 400 eV were used. The local-density approximation⁴⁰ plus U (LDA+ U)⁴¹ was used for the exchange correlation energy with the on-site Coulomb energy U of 6 eV for the In, Ga, and Zn d states. The Sn $4d$ were treated as core states. For the Brillouin zone integration in DFT molecular dynamics simulations, a single k -point at (0.25, 0.25, 0.25) was used. The

total energy difference between the single k -point and the $2 \times 2 \times 2$ Monkhorst-Pack mesh is only a few meV. In electronic density-of-states calculations, we used the $3 \times 3 \times 3$ mesh including the Γ k -point. The LDA+ U method with $U = 6$ eV was tested for crystalline ZnO. The major role of U for the Zn-3d states in ZnO is correction of the energy level position of the Zn-3d bands.⁴² The Zn-3d bands are located near -7 eV below the valence band maximum (VBM) in our LDA+ U calculations, and is close to the X-ray emission spectroscopy measurements.⁴³ With this correction, the band gap underestimation problem of the LDA is partially improved by suppressing the spurious Zn-3d and O-2p hybridization in the LDA. The calculated LDA+ U band gap of ZnO was 1.74 eV, while the LDA band gap was 0.80 eV. The calculated LDA and LDA+ U gaps are consistent with the previous LDA+ U calculations,⁴⁴ but still lower than the experimental,⁴⁵ hybrid functional,^{46,47} and GW^{48–53} results. Thus, here we focus only on the *changes* of the band gaps with respect to the cation compositions.

The considered metal compositions of the In-Sn-Zn-O amorphous semiconductors are listed in Table I. We used cubic supercells. The number of atoms for each atom species (N_{In} , N_{Sn} , N_{Zn} , and N_O) and the total number of atoms (N_t) in the supercell are shown in Table I. We only considered O stoichiometric condition as in the binary oxides of In_2O_3 , SnO_2 , and ZnO . The supercell volumes (V_t) are derived from the experimental mass densities of In_2O_3 (7180 kg/m³), SnO_2 (6900 kg/m³), and ZnO (5600 kg/m³). We also considered amorphous $InGaZnO_4$ as a reference AOS. The number of atoms in the $InGaZnO_4$ cubic supercell is $N_t = 168$ with $N_{In} = 24$, $N_{Ga} = 24$, $N_{Zn} = 24$, and $N_O = 96$. The $InGaZnO_4$ supercell volume is also derived from the experimental mass densities [Ga_2O_3 (6000 kg/m³)] ($V_t = 1972.9 \text{ \AA}^3$). The average volume per unit formula obtained by this *a priori* approach for the amorphous phase $InGaZnO_4$ is 9.6% larger than the theoretically optimized volume per unit formula for the crystalline phase $InGaZnO_4$. The amorphous structures were generated by melt-and-quenching molecular dynamics simulations within the Nosé canonical ensemble. A random amorphous oxide structure was melted at 3000 K for 6 ps and then quenched down to 300 K at a quenching rate of 900 K/ps. The amorphous structure was then optimized with the static DFT calculations until the Hellmann-Feynman forces were less than 0.02 eV/Å.

TABLE I. Seven considered amorphous In-Sn-Zn-O semiconductors in this work are listed. N_{In} , N_{Sn} , N_{Zn} , and N_O are the numbers of In, Sn, Zn, and O atoms, respectively, in the supercell. N_t is the total number of atoms in the supercell, and V_t is the total supercell volume. $R_{In}+R_{Sn}$ is the sum of the In and Sn composition ratios.

Sample	N_{In}	N_{Sn}	N_{Zn}	N_O	N_t	$V_t (\text{\AA}^3)$	$R_{In} + R_{Sn}$
ISZO#1	24	24	24	108	180	2221.0	0.667
ISZO#2	12	24	24	90	150	1835.6	0.600
ISZO#3	6	27	27	90	150	1824.2	0.550
ISZO#4	10	16	24	71	121	1481.1	0.520
ISZO#5	6	8	16	41	71	869.2	0.467
ISZO#6	10	12	28	67	117	1432.5	0.440
ISZO#7	6	18	36	81	151	1714.9	0.400

스핀-궤도 각운동량 상호작용의 구조 최적화에 대한 효과; 비스무스 텔루라이드의 제일원리 계산의 경우

Tran Van Quang

아주대학교 물리학과, 경기 수원시 원천동 산 5-1, 443-749

김미영*

숙명여자대학교 나노물리학과, 서울시 용산구 청파로 47길, 140-742

(2012년 11월 21일 받음, 2013년 2월 5일 최종수정본 받음, 2013년 2월 5일 게재확정)

스핀 궤도 각운동량의 상호작용은 저차원계 자성물질에서 나타나는 여러 가지 특이한 현상들의 물리적 원인을 제공하는 것으로 알려져 있다. 최근 들어 자성 도핑을 이용한 열전 물질의 합성에 대한 관심이 높아지면서, 열전 및 위상 절연체(Topological Insulator) 등의 물리적 성질 결정에 중요한 역할을 하는 페르미 에너지 준위 부근에서의 전자구조에 대한 스핀 궤도 각운동량의 효과 연구가 관심을 끌고 있다. 본 연구에서는 가장 일반적인 열전 호스트 물질인 비스무스 텔루라이드의 격자 상수 및 부피 팽창률에 대한 스핀 궤도 각운동량 상호작용의 효과를 연구하기 위하여 모든 전자(all-electron) FLAPW(full-potential linearized augmented plane wave) 방법을 이용하여 전자구조 계산을 수행하였다. 국소밀도 근사법 및 일반 기울기 보정법의 서로 다른 교환상호작용 퍼텐셜을 채용하고, 수평격자 및 수직격자를 분리하여 변화시키는 구조최적화 계산을 통하여, 스핀-궤도 각운동량 상호작용의 효과가 격자상수 평형 값을 약하게 증가시키는 반면, 부피탄성률을 크게 감소시키는 영향을 주며, 그 효과는 구조적 이방성이 뚜렷한 비스무스 텔루라이드의 특성에 의하여 격자방향에 대한 의존성을 보인다는 것을 확인했다.

주제어 : 스핀 궤도 각운동량 상호작용, 열전성질, 전자구조 계산

I. 서 론

열전효과(thermoelectric effect)의 물리적 메커니즘과 그 응용을 통해 일상에서 소모되는 열에너지를 전기에너지로 전환하여 사용할 수 있는 가능성이 알려진 것은 수십 년 전의 일이지만, 최근 들어서 탄소 저감 및 그린 에너지에 대한 관심과 맞물려서 고효율 열전 물질의 개발 및 그 실제 응용을 위한 이론 및 실험 연구가 더욱 많은 관심을 끌고 있다. 물질의 열전 성능을 나타내는 열전 감도 지수(figure of merit)인 ZT 값은 $ZT = S^2 \sigma T / (\kappa_L + \kappa_e)$ 로 계산할 수 있으며, 따라서 물질의 전기전도도(σ), 열전도도의 격자적 및 전기적 기여($\kappa_L + \kappa_e$) 및 Seebeck 계수(S) 등에 의존하여 결정된다[1]. 물질의 전기 전도도와 열전도도는 밀접하게 연결되어 있어서 동시에 증가하거나 감소하는 경우가 일반적이므로, 열전 성능을 증가시키기 위해 전기전도도를 증가시키면서 낮은 열전도도 값을 유지시키는 물질을 디자인 하기 위해 이론과 실험 양쪽 분야에서 많은 노력이 있어왔다. 그 일환으로서 전도의 방향에 수직으로 서로 다른 물성의 물질을 교대로 쌓아서 형

성된 초격자 구조를 이용하여 포논막음/전자투과(phonon-blocking/electron transmission) 기작을 이용하는 여러 형태의 연구가 보고되어, 300 K 온도에서 p -형 반도체인 $\text{Bi}_2\text{Te}_3/\text{Bi}_2\text{Se}_3$ 초격자 물질로부터, 고전적인 열전물질인 비스무스 텔루라이드의 열전 감도 지수($ZT \sim 1.0$)의 2~3배의 ZT 값이 보고되기도 하였으나 열적 안정성 등의 기술적인 문제로 인하여 실용화되지 못하고 있는 실정이다[2]. 특히 $\text{Bi}_2\text{Te}_3/\text{Sb}_2\text{Te}_3$ 초격자 구조에 대한 제일원리 계산은 열전 성질의 기작을 설명하는데 있어서 초격자를 이루는 성분 물질의 격자 상수 부 조합에 따른 변형이 중대한 영향을 준다는 것을 보임으로써 텔루라이드 계열의 열전 성질 계산에 있어서 구조 최적화의 역할이 중요함을 보고하고 있다[3].

한편 고효율 열전물질 개발에 대한 노력은 자성 도핑을 이용한 연구로도 이어지고 있다. Fe, Co, Mn 등의 전이금속 도핑된 비스무스 텔루라이드 및 셀레나이드 합금에서의 강자성 상태의 안정화에 대한 보고와 함께 3차원 체적 위상 절연체(Topological Insulator)의 전형적인 디랙 형태(Dirac-like)의 전도 전자 구조를 갖는다는 최근 일련의 연구결과들은 자성 도핑된 열전 물질에 대한 관심을 더욱 증가 시킨 바 있다[4-10]. 전이금속 도핑에 비하여 희토류 물질의 도핑에 의한 열전 물질의 전자구조 및 자성 상변화에 대한 연구는 상대적으로

로 미미한 형편인데, 특히 강하게 국소화되어있는 f 전자띠로 인해 기존의 자성물질에서 보여온 특성을 고려할 때 열전 성질 및 이에 의한 페르미 에너지 준위에서의 전자구조에 대한 영향은 흥미로운 주제가 아닐 수 없다. 강한 자성이방성 및 큰 자기모멘트 등의 고유한 자기적 특성과 함께, 희토류를 통한 자성도핑의 경우 도핑 원자의 큰 사이즈로 인하여 호스트 물질의 대체 원자 종류에 따라 기존 물질의 구조 및 격자 상수의 변화가 크게 나타나고, 이에 따른 새로운 물리적 현상을 보이는 것이 보고되어, 희토류 도핑 열전물질의 연구에서도 역시 합금의 정확한 최적 구조를 얻는 것이 중요한 요인임을 확인하고 있다[11, 12].

이에 우리의 연구에서는, 상온 작동 고효율 열전 물질 개발의 기본 재료가 되고 있는 비스무스 텔루라이드(Bi_2Te_3)를 호스트로 한 열전 물질 연구에 있어서 구조 최적화의 중요성에 주목하고, 스핀-궤도 각운동량 상호작용이 최적 격자 상수 및 부피 탄성률 등의 구조관련 물성에 미치는 영향에 대한 연구를 목적으로 제일원리 전자구조 계산을 수행하였다.

II. 연구방법 및 모형

Bi_2Te_3 는 실험적으로 162 meV로 알려진 작은 에너지띠 틈 (energy band gap)값과 다중의 전도띠극대점/바깥띠극소점 (conduction band minimum/valence band maximum) 구조를 갖는 절연체이다[12]. 밀도범함수법(DFT; Density functional Theory) 기반의 제일원리 계산을 통해 실험값에 유사한 에너지띠 틈값과 전자구조를 얻기 위해서는 금속성 및 가벼운 원자들로 이루어진 광범위한 물질들을 효과적으로 기술해 온 국소밀도 근사법(LDA; local density approximation)을 뛰어넘는 보완적인 교환상호작용 퍼텐셜 및 스핀-궤도 각운동량 상호작용을 도입하는 것이 필요하다는 것이 알려져 있다[14, 15]. 이러한 보완적인 교환상호작용 퍼텐셜은 에너지띠 틈값 뿐만 아니라 열전성질 결정에 중요한 요인이 되는 페르미 에너지 준위 근처에서의 에너지띠 위상 특성에도 영향을 주는 것으로 나타났으나, 이러한 변화는 실제 열전성질 계산에서는 결정적인 변화를 주지 않음으로 인하여, 복잡한 열전물질의 경우, 보완 퍼텐셜 대신 강에너지띠근사(rigid-band approximation) 내에서 전도 에너지띠 이동으로 에너지띠 틈값을 실험값에 근사시킨 후 계산하기도 한다. 본 연구에서도 구조 최적화와 그에 따른 최적 격자 상수 및 부피 탄성률의 스핀-궤도 각운동량 상호작용의 의존성에 연구의 주된 목적이 있으므로, 정확한 에너지띠 틈값을 얻기 위한 보완 퍼텐셜은 채택하지 않았으며 대신하여 가장 일반적인 두가지 교환상호작용 퍼텐셜인 국소밀도 근사법과 PBE형 일반기울기 근사법(GGA, generalized gradient approximation)[16]을 채택하여 비교하

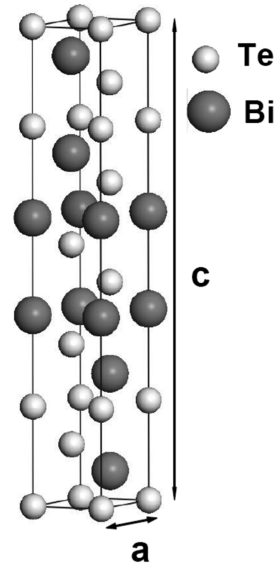


Fig. 1. The atomic structure of Bi_2Te_3 with five atoms in its hexagonal unit cell with the in-plane and the perpendicular lattice constant, a and c , respectively.

였다.

Bi_2Te_3 는 한 단위셀 당 다섯 개의 원자를 가진 삼방정계 (trigonal) 격자구조를 가지고 있으며 Fig. 1에서 보인 바와 같이 육방격자(hexagonal lattice)로 나타내었을 때 수평격자상수(in-plane lattice constant) a 와 수직격자상수(perpendicular lattice constant) c 의 실험값은 각각 원자단위(atomic unit)로 $a = 8.28$ a.u.와 $c = 45.56$ a.u.로 알려져 있다[17]. 본 연구에서는 전자구조 계산과 구조 최적화를 위하여 FLAPW(full-potential augmented plane-wave) 방법[18]을 이용하였다. Bi 원자와 Te 원자의 머핀 띠(MT, muffin tin) 구 영역으로서는 핵을 중심으로 각각 3.0 및 2.4 a.u. 반지름을 갖는 원으로 잡고, 이 MT 영역 안에서의 구면조화함수는 $L_{\max} = 8$ 의 각운동량 절단값을 채택하여 계산하였으며, 그 밖의 격자틈새(interstitial) 영역에서의 보강평면파는 원자당 약 200개의 기저 함수를 사용하도록 하는 에너지 차단값을 사용하였다. 핵심 전자들은 디랙 방정식을 직접 풀어냄으로써 완전 상대론적으로 취급하는 한편 원자가전자들은 준상대론적으로 취급한 후, 제 2 변분법적 접근법(second variational approach)을 사용하여 스핀-궤도 각운동량의 상호작용을 자체 충족적인 되풀이법으로 포함하는 방식으로 계산하였다. 격자상수의 평형 값을 결정하기 위하여 각 격자상수 값에 대하여 원자적 힘 계산을 통하여 원자의 평형위치를 얻었으며 이렇게 얻어진 총 에너지 계산 값을 최소화 하여 격자 상수 평형 값을 얻었다. 결과 및 논의 부분에서 자세히 설명하였듯이 부피탄성률을 계

Electronic Structure and Magneto-optical Properties of $\text{Co}_2\text{Mn}X$ Alloys Where $X = \text{Ge, Sn and Pb}$: a First-principles Investigation in the LDA+U Approach

Tran Van QUANG

Department of Physics, Ajou University, Suwon 443-749, Korea

Jae Il LEE

Department of Physics, Inha University, Incheon 402-751, Korea

Miyoung KIM*

Department of Nano Physics, Sookmyung Women's University, Seoul 140-742, Korea

(Received 5 June 2012, in final form 26 July 2012)

We report our *ab-initio* calculational results for the electronic structures and the magneto-optical (MO) properties of the ferromagnetic $\text{Co}_2\text{Mn}X$ full Heusler alloys where $X = \text{Ge, Sn and Pb}$. Employing the +U corrections for the $3d$ bands of transition-metal atoms in addition to the conventional local density approximation (LDA), we investigate the correlation effect on the MO spectra in a polar geometry and the detailed electronic structures. Results show that the correlation effect results in a blue-shift of the peak positions and a large enhancement of the MO spectra up to double in magnitude compared to the LDA results. We find that our results can be attributed to the increased $t_{2g} - e_g$ splitting of the spin-minority d -bands of both Co and Mn atoms, which indicates a suppression of the diagonal elements of the optical conductivity in the energy region of 1-2 eV where the interband transitions are forbidden.

PACS numbers: 73.20.At, 71.20.Be

Keywords: First-principles, Electronic structure, Magneto-optics, Surface states

DOI: 10.3938/jkps.62.2184

I. INTRODUCTION

Heusler alloys are known to be the essential ingredient for spintronics devices because some of them have been predicted to exhibit half-metallicity, which maximizes the spin-injection rate from ferromagnetic to semiconducting materials [1]. The Co-based Heusler alloy in the form of Co_2YX , where $Y = \text{Fe or Mn}$ and $X = \text{group III or IV element}$, is one of the most promising half-metallic ferromagnets for this usage and has advantageous properties such as a structural compatibility with semiconducting substrates, a high magnetization, and a high Curie temperature (T_c) up to 1100 K. In spite of great efforts in both experiment and theory to realize the predicted high spin-injection rate [2], recent experiments have reported spin-injection rates of only a few percent at room temperature for the Co-based Heusler alloy/semiconducting heterostructures [3], indicating that there is a lot to be desired in these materials. Along with the electronic and the magnetic properties,

the magneto-optical (MO) properties of various Heusler alloys have been intensively studied for decades, especially after the early report of a huge Kerr rotation (up to 5° at 80 K) for the half-Heusler PtMnSb alloy [4], with an aim to benefit the field of magneto-optical recording technology. While the theoretical descriptions of the MO properties within *ab-initio* band structure calculations provide qualitatively reasonable agreement with experiment, some discrepancies, which have been partly ascribed to an insufficient treatment of electronic correlations by the local density approximation (LDA), exist. An improved description of correlation effect has recently been demonstrated in the framework of the LDA combined with dynamical mean-field theory (DMFT) to lead to a better agreement of the theoretical and the experimental MO spectra for NiMnSb [5]. The on-site Coulomb correlation effect through LDA+U approach has been pointed out to play an important role in electronic structure (and thus the half-metallicity) of the Co_2FeSi and Co_2MnSi [6] while its effect on the MO properties of these alloys has not reported yet.

*E-mail: mykim.nu@gmail.com

II. METHODOLOGY

In this work, we performed first-principles calculations for the magneto-optical Kerr spectra of ferromagnetic Co₂MnX alloys, where $X = \text{Ge, Sn, and Pb}$, within the LDA+U approach, aiming to reveal the correlation effect, via the on-site Coulomb interaction, on the MO properties. The precise all-electron full-potential linearized augmented plane-wave (FLAPW) [7] method is employed where the exchange-correlation potential is treated within LDA+U where the on-site Coulomb interaction is characterized by a spherically averaged Hubbard U and is implemented by using the second-variation procedure to self-consistently solve the Kohn-Sham equation [8]. An energy cut off of 220 eV for the interstitial plane wave basis and $10 \times 10 \times 10$ special k-point meshes are used for the 3D Brillouin zone integral. The spin-orbit coupling, which is a crucial source for MO Kerr effect, is also treated in the second-variation technique. The equilibrium lattice constant of Co₂MnPb is obtained from total energy calculations while the experimental values are used for Co₂MnGe and Co₂MnSn [9].

The magneto-optical polar Kerr spectra can be obtained from the optical conductivity tensor. For the present crystal with a four-fold rotational symmetry

about the z -axis along which the wave vector of incident light is assumed to be oriented, the complex conductivity tensor elements can be simplified by using $\sigma_{xx} = \sigma_{yy}$ and $\sigma_{xy} = -\sigma_{yx}$. Then, applying the macroscopic electromagnetic boundary problem, we can calculate the polar Kerr angles, $\theta_K + i\varepsilon_K$, as a function of the photon energy ω by using

$$\theta_K(\omega) + i\varepsilon_K(\omega) = \frac{-\sigma_{xy}}{\sigma_{xx} \sqrt{1 + i \left(\frac{4\pi\sigma_{xx}}{\omega} \right)}}, \quad (1)$$

where θ_K and ε_K are the Kerr rotation and ellipticity, respectively. While Eq. (1) accounts for the *interband* transitions, the contribution from *intradband* transitions also can be considered by adding the so-called Drude term, $\sigma_0 = \omega_p^2/[4\pi(1 - i\omega\tau_D)]$, into the diagonal element of the conductivity tensor by adopting empirical values of the plasma frequency (ω_p) and the relaxation time (τ_D) [10]. The optical conductivity tensor can be determined from the microscopic formalism in terms of Kubo's linear response theory [11]. In the long-wavelength limit of the random phase approximation, the diagonal and the off-diagonal optical conductivities are evaluated using the relativistic Kohn-Sham eigenvalues and wave functions as [12]

$$\begin{aligned} \sigma_{xx}(\omega) &= \frac{ie^2}{m^2\hbar} \sum_k \sum_{ij} \frac{|\Pi_{ij}^x|^2}{\omega_{ij}} \left(\frac{1}{\omega - \omega_{ij} + i\delta} + \frac{1}{\omega + \omega_{ij} + i\delta} \right), \\ \sigma_{xy}(\omega) &= \frac{ie^2}{m^2\hbar} \sum_k \sum_{ij} \frac{1}{\omega_{ij}} \left(\frac{\Pi_{ji}^x \Pi_{ij}^y}{\omega - \omega_{ij} + i\delta} - \frac{\Pi_{ij}^x \Pi_{ji}^y}{\omega + \omega_{ij} + i\delta} \right), \end{aligned} \quad (2)$$

where i and j go over the occupied and the unoccupied states, respectively, ω_{ij} is the energy difference between the i and the j states and Π_{ij}^x is the x -component of the relativistic momentum operator calculated with the wave functions of the i and the j states.

III. RESULTS AND DISCUSSION

Before presenting the calculated results of the MO spectra, we briefly summarize the results for structural, electronic, and magnetic properties calculated within LDA+U, as well as the LDA, for comparison in Table 1: the equilibrium lattice constants, the total and the atomic spin magnetic moments, the energy gap between the valence band maximum (VBM) and the conduction band minimum (CBM), $E_C^\downarrow - E_V^\downarrow$, and the density of states (DOS) at the Fermi energy (E_F), $D^\downarrow(E_F)$, where the downward arrow stands for a spin minority channel

are included. A finite energy gap between the VBM and the CBM with a completely vanished DOS at E_F in the minority spin channel indicates half-metallicity because the majority spin channels show metallic behavior as well be seen later from the DOS plot. Among the six sets of calculations presented in the table, only Co₂MnGe in the LDA exhibits half-metallicity; Co₂MnSn in the LDA is “nearly half-metallic”, as reported elsewhere, in the LDA and the GGA [2]. The LDA+U calculations result in a metallic character for all three alloys because, despite the minority energy gaps, $E_C^\downarrow - E_V^\downarrow$, being greatly enhanced compared to the LDA values, E_F is located not in the energy gap but in the middle of the minority top valence band - as can be seen in Fig. 1 where the spin-resolved projected DOS in the LDA+U approach for the Co and the Mn d -bands are plotted. Comparing with the LDA DOS for Co₂MnGe shown in Fig. 1(d), we find how the half-metallicity is lost upon adding the U correlation. With the U inclusion, the occupation of the spin majority channel increases because the majority

The metal-insulator phase transition in the strained GdBiTe₃

Tran Van Quang¹ and Miyoung Kim^{2,a)}

¹Department of Physics, Ajou University, Suwon 443-749, South Korea

²Department of Nano Physics, Sookmyung Women's University, Seoul 140-742, South Korea

(Presented 18 January 2013; received 10 November 2012; accepted 12 December 2012; published online 9 April 2013)

In this work, we investigated the electronic structures and magnetic properties of the GdBiTe₃ alloy employing a first-principles all-electron density-functional approach, aiming to understand the magnetic phase stability and electronic structure dependences on the exchange correlation potential and the strain. The results show that the ferromagnetic phase is energetically more stable over the paramagnetic phase and the metal-insulator phase transition occurs upon the lattice distortion via the strain along the perpendicular *c* direction, which is not influenced by the strength of correlation energy introduced to describe the localized *f* orbitals. Thermoelectric transport properties are also investigated to reveal that the compressive strain markedly enhances the Seebeck coefficient, which is reduced in comparison with the Bismuth telluride due to the Gd doping. © 2013 American Institute of Physics. [<http://dx.doi.org/10.1063/1.4795743>]

The bismuth-based chalcogenides such as Bi₂Te₃ and Bi₂Se₃ and their compounds have been considered as the best thermoelectric (TE) materials operating at room temperature for decades.^{1,2} In spite of the extensive studies in both experiment and theory, it still remains as a huge challenge to design a novel material with a high TE efficiency exceeding that of the conventional bismuth telluride alloys. Recently, magnetically doped Bi₂Te₃ alloys also have been studied intensively with the transition metal dopings such as the Mn, Fe, and Co atoms. They were reported to exhibit the unique electronic structures of the Dirac-like conducting bands as the bulk topological insulators (TIs), where the magnetic impurities and ferromagnetism play the important roles on the TI surface states.^{3–5} The discoveries in both theory and experiment on Bi₂Te₃ and Bi₂Se₃ as well as their compounds so far offer that the good TI materials are also the good TE materials.^{6–10}

However, the studies on the rare earth element dopings on the bismuth telluride are relatively lacking probably due to the difficulties in experimental synthesization and theoretical description of the strongly localized *f*-states. Recently, the ternary rare earth chalcogenides in the form of LBiTe₃, where *L* is La or Y have been reported theoretically to exhibit the electronic structure of TI.¹¹ Among many rare earth elements, Gd has been referred as an intriguing magnetic dopant due to the completely filled spin majority 4*f* orbitals, which cause many exotic magnetic properties. Various compounds with Gd have been reported to show interesting phenomena such as GdX, where X = N, P, As, Sb, and Bi (Ref. 12) and the gadolinium monochalcogenides.^{13,14} Among them, we focus on the strain induced transition from the half-metallic to the semiconducting transport property as predicted in GdN,¹⁵ where the strain was found to significantly influence on the magnetic phase transition temperature and the energy gap. The pressure induced metal-insulator phase transition has been observed in the rare earth nickelates

as well.¹⁶ Also interestingly, the significant influences of the volume¹⁷ and the strain¹⁸ on the electronic structure as well as the TE property for Bi₂Te₃ and Bi₂Se₃ were reported by first-principles studies. These all noteworthy phenomena suggest that GdBiTe₃ may have peculiar electronic and magnetic properties highly sensitive to the strain or the lattice distortion. Therefore, in this work, we performed the first-principles electronic structure calculations of GdBiTe₃ with the -5% – 5% of strain along the perpendicular *c*-axis, aiming to investigate the strain dependence of the electronic structures, magnetic properties, and TE efficiency.

The electronic structures were calculated using the all-electron full-potential linearized augmented plane-wave (FLAPW) method¹⁹ employing the exchange-correlation potential provided by the local density approximation (LDA). To determine the relatively strong correlation interaction of the Gd 4*f* states properly, we also implemented the +*U* method²⁰ in addition to the LDA with the correlation energies of *U* = 7.7 eV and *J* = 0.7 eV, which are known to be suitable to describe the electronic structures of the pristine single Gd crystal.^{20–24} The spin-orbit coupling (SOC) is treated within the second variational way in a self-consistent manner. The crystal structure of GdBiTe₃ in its rhombohedral unit cell of the space group $R\bar{3}m$ is formed by substitutionally replacing one of the two Bi atoms of the conventional Bi₂Te₃ unit cell with the Gd atom. The experimental lattice constants are $a_{exp} = 4.16$ Å and $c_{exp} = 31.20$ Å.²⁵

Within the LDA, we performed the total energy calculation where the atomic positions were fully relaxed through the atomic force calculations and found both in-plane and perpendicular lattice constants underestimated by 8% compared to the experimental values. Total energy comparison revealed that the ferromagnetic (FM) phase is energetically more stable over the paramagnetic (PM) phase. The FM phases both with- and without-SOC give the magnetic moment of 7.0 μ_B per unit cell, which comes mostly from the Gd atom. The calculated band structure of GdBiTe₃ is shown in Fig. 1 with that of Bi₂Te₃ for comparison. As is well known, the Bi₂Te₃ is a semiconductor with a small band gap of 0.162 eV (in experiment²⁶)–0.154 eV

^{a)} Author to whom correspondence should be addressed. Electronic mail: mykim.nu@gmail.com.

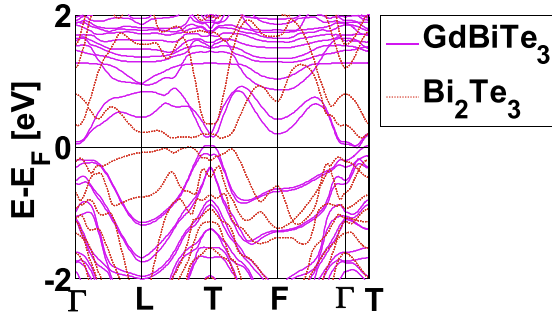


FIG. 1. The calculated band structure for GdBiTe₃ is shown together with that of Bi₂Te₃ for comparison.

(in the first-principles calculation²⁷). Upon the Gd doping, we found that GdBiTe₃ becomes metallic as can be seen from Fig. 1. The valence band maximum (VBM) which is located around the T point slightly crosses the Fermi level (E_F) making a tiny hole pocket. Also, the conduction band minimum (CBM) which can be found along the symmetry line between T and Γ crosses the E_F . The effect of SOC on the GdBiTe₃ electronic structure near the E_F was found to be crucial. Without SOC, we found that it is an insulator with a small band gap of 0.08 eV and the locations of VBM and CBM are also different with those shown in Fig. 1. On the other hand, the additional correlation interaction via +U correction for the Gd 4f bands was found to affect markedly the 4f bands but not the bands near E_F . Within the LDA, the majority 4f band is located at 4.5 eV below E_F and the minority 4f band is at 0.5 eV above the Fermi level. Now introducing the additional correlation interaction via +U, we found the majority 4f band was shifted down to the energy of 10 eV below E_F and the minority band up to the energy of 1.5 eV above E_F . While not shown here, we examined the electronic structure change in the wide ranges of U (4.0–8.0 eV) and J (0.5–1.3 eV) values and found that the band structures around E_F and its vicinity do not show crucial change, while the Gd 4f bands show significant dependence. Therefore, we show here the calculated results employing the fixed values of U = 7.7 eV and J = 0.7 eV to discuss the properties mostly due to the electronic structures near the Fermi level.

Now, in order to study the strain dependence of the electronic structure, we introduced the strain along the c -direction δc , which is given by $\delta c = (c - c_{exp})/c_{exp}$. Five different strains, which are $\delta c = -5.0\%$, -2.5% , 0% , 2.5% , and 5.0% , are adopted with the in-plane lattice fixed as the experimental value. With the fixed lattice constants according to a given strain, the internal atomic positions were fully relaxed via the atomic force calculations. The calculated band structures and the total density of states (TDOS) are shown in Fig. 2. The band structures for $\delta c = -2.5\%$ (a) and for $\delta c = 2.5\%$ (b) are shown (in solid lines) together with the result of no strain, i.e., $\delta c = 0$ (in dotted lines) for comparison. Especially, the band structures along the symmetry line from the T to the Γ point, where the VBM and CBM were found for $\delta c = 0$, are blown up and shown on the right side of each figure. The TDOS for all five strains are shown in Fig. 2(c). As pointed out above in Fig. 1, GdBiTe₃ is found to be metallic for $\delta c = 0$; the VBM as well as the CBM are slightly crossing E_F yielding a hole-pocket and an electron-pocket,

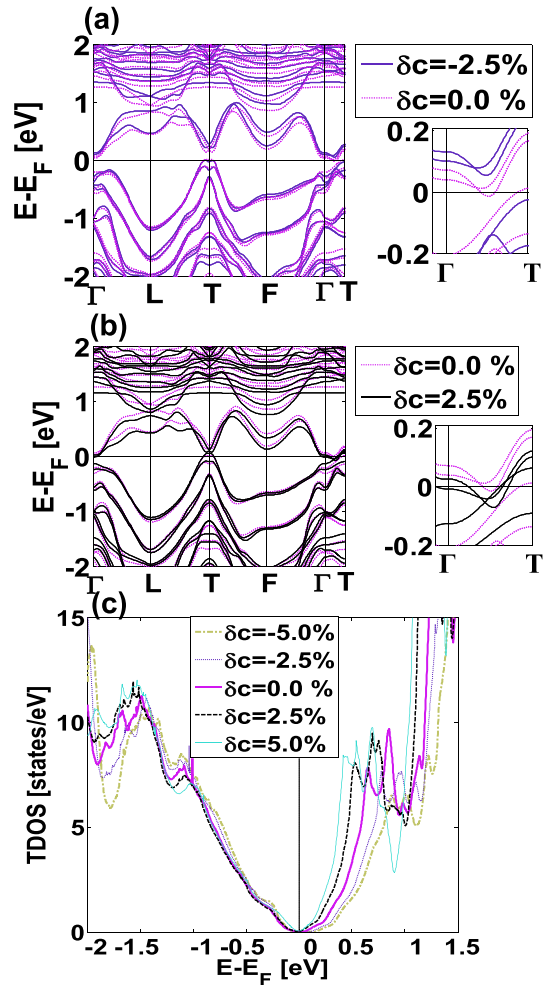


FIG. 2. The calculated band structure and TDOS of GdBiTe₃. With the strain, δc , defined by $\delta c = (c - c_{exp})/c_{exp}$, five different strains of -5% , -2.5% , 0% , 2.5% , and 5% are calculated. The band structures for $\delta c = -2.5\%$ (a) and $\delta c = 2.5\%$ (b) are shown (solid lines) together with the result of no strain (i.e., $\delta c = 0$) (dotted lines) for comparison. The bands along the Γ to T points are blown up to show clearly in the small boxes on the right side of each plot. Also, the calculated TDOS of $\delta c = -5\%$, -2.5% , 0% , 2.5% , and 5% are shown in (c).

respectively. When the perpendicular lattice c is decreased by 2.5% with the in-plane lattice fixed, both the VBM and the CBM are found to move further from E_F as shown in Fig. 2(a) for $\delta c = -2.5\%$, resulting in a finite energy gap. The band shift induced by the compressive strain occurs mostly at the k -points along Γ to T for the valence band, while the whole conduction bands suffer the upward band shift. From the TDOS in the Fig. 2(c), we can see that the conduction band shift to the high energy region due to the compressive strain becomes more obvious upon increasing the strain. As a result, the induced band gap increases with the increasing strain; it is found to be 0.05 eV for $\delta c = -2.5\%$ and 0.1 eV for $\delta c = -5.0\%$. On the other hand, the strain in opposite direction ($\delta c = 2.5\%$ and 5%) implies the band structure change in opposite way, as can be seen from Figs. 2(b) and 2(c). The tensile strain of $\delta c = 2.5\%$ causes an

화학과

Structural heterogeneity in familial Alzheimer's disease mutants of amyloid-beta peptides†

Cite this: *Mol. BioSyst.*, 2013, **9**, 997

Song-Ho Chong, Janghyun Yim and Sihyun Ham*

Alzheimer's disease is a neurodegenerative disorder characterized by progressive deposition of amyloid-beta (A β) peptides in brain parenchyma and cerebral blood vessels. Several pathogenic familial mutations of A β peptides have been identified that exhibit enhanced neurotoxicity and aggregative ability. However, knowledge of the structural characteristics of those A β mutants is still limited. Here, we report multiple all-atom molecular dynamics simulations of the wild-type 42-residue A β peptide (A β 42) and its Flemish (A21G), Arctic (E22G), Dutch (E22Q), Italian (E22K), and Iowa (D23N) familial mutants in explicit water. After validating our simulations by comparison with available experimental data, we examined common/different features in the secondary and tertiary structures of the wild-type and five familial mutants of A β 42. We found that A β 42 peptides display quite heterogeneous secondary and tertiary structure ensembles. Such structural heterogeneity in the monomeric state would facilitate interconversions between various secondary structures during the formation of a β -sheet-rich amyloid fibril, and may also serve as a structural basis of the amyloid polymorphism.

Received 19th October 2012,
Accepted 4th January 2013

DOI: 10.1039/c2mb25457c

www.rsc.org/molecularbiosystems

Introduction

A number of human diseases are associated with aggregation of misfolded or intrinsically disordered proteins into insoluble amyloid fibrils or plaques.^{1,2} Aggregation of amyloid-beta (A β) peptides is linked to Alzheimer's disease (AD).^{3,4} The 40- and 42-residue forms (A β 40 and A β 42) are the major products derived through endoproteolysis of transmembrane amyloid-beta precursor protein in the brain parenchyma and cortical blood vessels. While A β 40 and A β 42 differ in only two amino acid residues at the C-terminal end, A β 42 aggregates faster and its neurological toxicity is stronger than A β 40.⁵ Several pathogenic familial mutations of A β peptide have also been described.⁶ In particular, a set of mutations that occur at amino acid positions 21 through 23 have drawn special attention. Some of the most well studied familial AD mutants include the Flemish (A21G),⁷ Arctic (E22G),⁸ Dutch (E22Q),⁹ Italian (E22K),¹⁰ and Iowa (D23N)¹¹ mutants that exhibit distinct neurotoxicity and aggregative ability.^{8,12}

Knowledge of the high-resolution structures of monomeric and aggregate species would serve as a basis for elucidating aggregation mechanisms of A β peptides. However, while the

amyloid-fibril structures of A β peptides have been well resolved,^{13,14} experimental investigations on A β monomers have been hindered by their intrinsically disordered nature as well as by their strong tendency to aggregate in physiological aqueous environments.¹⁵ Hence, molecular dynamics (MD) simulations have been extensively used to characterize atomic-level structural features of A β monomers. For an efficient sampling of A β conformations, a number of groups have studied A β peptides in implicit solvent.^{16–20} Explicit-water simulations have also been performed to fully take into account the detailed peptide–water interactions, including those for some of the familial AD mutants.^{21–32} Yet, no systematic exploration of the mutation effects has yet been carried out.

In this paper, we investigate structural characteristics of the full-length A β 42 peptide and its Flemish, Arctic, Dutch, Italian, and Iowa familial mutants based on fully atomistic, explicit-water MD simulations. For intrinsically disordered peptides such as A β 42, how to efficiently sample their potentially wide conformational space is a challenging issue. The replica-exchange method has often been adopted,^{23,28–31} but explicit handling of solvent water molecules requires a large number of replicas since the energetic fluctuation becomes relatively smaller for a larger system. This will limit the length of the simulation at a desired temperature. In the present work, we conduct multiple explicit-water MD simulations at 300 K and 1 bar starting from carefully prepared independent peptide structures by heating and conformational annealing simulations.

Department of Chemistry, Sookmyung Women's University, Cheongpa-ro 47-gil 100, Yongsan-Ku, Seoul 140-742, Korea. E-mail: sihyun@sookmyung.ac.kr;

Fax: +82-2-2077-7321; Tel: +82-2-710-9410

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c2mb25457c

Our simulated structures are validated by comparison with the available experimental data. We then analyze the secondary and tertiary structures of the familial AD mutants. Conformational fluctuations and structural heterogeneity are explored based on the multiple simulation trajectories.

Methods

MD simulation procedure

We carried out all-atom, explicit-water MD simulations at 300 K and 1 bar for the wild-type (wt) A β 42 and its five pathogenic familial mutants listed in Fig. 1 using the *sander* module of the AMBER11 simulation package.³³ We employed the ff99SB force field³⁴ for peptides and the TIP4P-Ew model³⁵ for water. The simulations were performed under neutral pH, where Lys and Arg residues are positively charged and Glu and Asp are negatively charged. The wt-A β 42 is in the -3 charged state, whereas the total peptide charge varies for the Arctic (E22G), Dutch (E22Q), Italian (E22K), and Iowa (D23N) mutants (see Fig. 1). The particle mesh Ewald method³⁶ was applied for treating long-range electrostatic interactions, while a 10 Å cut-off was used for the short-range non-bonded interactions. The hydrogen atoms were constrained to the equilibrium bond length using the SHAKE algorithm,³⁷ which enables simulations with a 2 fs time step. Temperature and pressure were controlled using Berendsen's thermostat and barostat with coupling constants of 1.0 and 2.0 ps, respectively.³⁸

Since A β peptides in the monomeric state are intrinsically disordered and no atomic-level structures in aqueous environments are available experimentally, the initial peptide systems for our simulations were carefully prepared as follows. The starting conformation for the wt-A β 42 was taken from the NMR structure determined in an apolar solvent (PDB code: 1IYT³⁹). The starting structures for the mutants were generated using Swiss PDB Viewer⁴⁰ based on the 1IYT structure of the wt-A β 42. Each system was then solvated by explicit water and neutralized by counter Na⁺ ions. The rest of the simulation procedures is common to all the peptide systems. The starting system was initially subjected to 500 steps of steepest descent minimization followed by 500 steps of conjugate gradient minimization while the A β peptide was constrained by 500 kcal mol⁻¹ Å⁻² harmonic potential. Then, the system was minimized using 1000 steps of steepest descent minimization

followed by 1500 steps of conjugate gradient minimization without harmonic restraints. The system was subsequently subjected to a 20 ps equilibration process in which the temperature was gradually raised from 0 to 300 K with a constant volume. This is followed by 200 ps constant-pressure (NPT) ensemble simulation of 1 bar to achieve the correct water density. We then heated the system to 600 K and carried out 20 ns canonical (NVT) ensemble simulation. From this high-temperature simulation, we took 10 peptide structures with a 2 ns time interval. The C α RMSD (root-mean-squared deviation) between any pair of these 10 structures was ~ 10 Å, and they were considered independent configurations. Each of the 10 structures was then gradually annealed to 300 K: the temperature was decreased with a 50 K interval, and 1 ns NVT simulation was performed at each of the intervening temperatures between 600 K and 300 K. Subsequently, additional 5 ns equilibration NPT simulation was performed at 300 K and 1 bar. The 10 final structures obtained in these annealing simulations were used as the initial structures for the 10 independent production runs, each 100 ns in length. The length of the production run was chosen in view of the estimation reported in ref. 29, where it was suggested that a simulation longer than ~ 60 ns per replica is necessary to achieve convergence in the ensemble for wt-A β 42. Since the replica exchange method is not adopted here, we instead carried out 10 independent production runs of 100 ns in length each starting from heated-then-annealed conformations. The minimization, heating, annealing, and production runs were all carried out under fully hydrated conditions with $\sim 14\,000$ water molecules in the cubic periodic box with a side length of ~ 75 Å. In total, the accumulative simulation time is 1.13 μ s for each A β peptide, and ~ 7 μ s simulations were performed as a whole.

Structural analysis

For each system, a total of 200 000 peptide conformations were selected and analyzed from the production runs of 1 μ s equilibrium simulations with a 5 ps time interval. To validate our simulations by comparison with experimental NMR data, *J*-coupling constants were calculated from the backbone ϕ angles using the Karplus equation:⁴¹

$$J(\phi) = A\cos^2(\phi - 60^\circ) + B\cos(\phi - 60^\circ) + C \quad (1)$$

We used the parameter set of Vuister and Bax⁴² in which $A = 6.51$, $B = -1.76$, and $C = 1.60$. The secondary-structure contents were calculated by using the DSSP program⁴³ implemented in

Wild-type (wt) 42-residue form of A β peptide

wt-A β 42 (-3) DAEFRHDSGY EVHHQKLVFF AEDVGSNKG A IIGLMVGGV IA

Pathogenic familial mutants of A β 42

Flemish	A21G	(-3)	DAEFRHDSGY EVHHQKLVFF	GEDVGSNKG A	IIGLMVGGV IA
Arctic	E22G	(-2)	DAEFRHDSGY EVHHQKLVFF	AGDVGSNKG A	IIGLMVGGV IA
Dutch	E22Q	(-2)	DAEFRHDSGY EVHHQKLVFF	AQDVGSNKG A	IIGLMVGGV IA
Italian	E22K	(-1)	DAEFRHDSGY EVHHQKLVFF	AKDVGSNKG A	IIGLMVGGV IA
Iowa	D23N	(-2)	DAEFRHDSGY EVHHQKLVFF	AENVGSNKG A	IIGLMVGGV IA

Fig. 1 Sequences of the wild-type (wt) A β 42 and its familial AD mutants studied in the present work. The mutated sites are indicated in red. The number in parentheses refers to the total peptide charge.

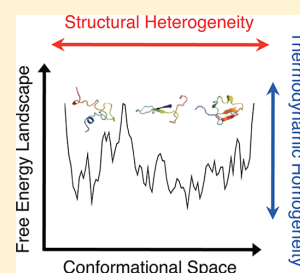
Conformational Entropy of Intrinsically Disordered Protein

Song-Ho Chong and Sihyun Ham*

Department of Chemistry, Sookmyung Women's University, Cheongpa-ro 47-gil 100, Yongsan-Ku, Seoul 140-742, Korea

Supporting Information

ABSTRACT: Intrinsically disordered proteins (IDPs), though lacking stable tertiary structures, are known to possess a certain amount of residual structure. Conformational disorder plays a crucial role through the conformational entropy in regulating protein–protein and protein–ligand interactions involved in signaling and regulation, and also modulates protein aggregation and amyloidogenesis associated with a number of human diseases. However, a direct and quantitative connection between the residual structure and the conformational entropy remains to be established. Here we show using a novel computational approach that the conformational entropy of amyloid-beta protein, an IDP whose aggregation is associated with Alzheimer's disease, is significantly correlated with the contents of the residual helical structure, β -sheet structure, and salt-bridge network. Identification of the thermodynamically significant residual structure is of fundamental importance for a comprehensive understanding of the relationship between the functional conformational disorder and the protein activity regulation, and will also serve the thermodynamic basis of the amyloid polymorphism.



INTRODUCTION

It is now widely acknowledged that a wide range of proteins involved in gene regulation and signal transduction are intrinsically disordered or comprise inherently unstructured regions.^{1–3} Intrinsically disordered proteins (IDPs) are also associated with a number of human diseases such as cancer, neurodegenerative disorders, and diabetes.^{4,5} It has been quite a challenge to detect and characterize conformational disorders in IDPs, since the secondary structure and long-range contacts in IDPs are inherently transient and comprise only a small fraction of the total protein sequence. The past several years, however, have witnessed significant improvements in sensitivity and resolution of spectroscopic methods, and the application of those advanced experimental techniques has revealed that IDPs are far from homogeneous statistical random coil polymers and instead exhibit a rich variety of local and even long-range residual structure.^{6,7}

Intrinsic disorder is a unique feature that enables IDPs to bind with multiple partners.^{8–10} While the binding affinity with different partners obeys general thermodynamic principles, a potentially important role of the conformational entropy is expected to play in determining the binding free energy of IDPs. Historically, the protein conformational entropy has remained one of the most elusive thermodynamic parameters to estimate. However, a recent NMR analysis method that employs conformational dynamics as a proxy for conformational entropy is breaking new ground.¹¹ In fact, this method has not only uncovered the significance of the conformational entropy in the protein–ligand binding free energy but also revealed how the binding affinity with different partners is tuned through the changes in the conformational entropy.^{11,12} However, a quantitative connection between the protein structural order/disorder and the conformational entropy remains elusive.

Clarifying those structural factors that influence the protein conformational entropy will open the possibility of a direct interpretation of the conformational entropy in structural terms.

Here, we investigate a quantitative connection between the residual structure and the conformational entropy of amyloid-beta ($A\beta$) protein, an IDP whose aggregation is linked to Alzheimer's disease,¹³ by using extensive explicit-water molecular dynamics (MD) simulations and the molecular theory of solvation. The full-length, 42-residue form of the wild-type $A\beta$ protein (wt- $A\beta$ 42) as well as its pathogenic familial mutants^{14–16} such as Flemish (A21G), Arctic (E22G), Dutch (E22Q), Italian (E22K), and Iowa (D23N) were studied. We also analyzed synthetic mutants termed RR (I41R, A42R) and DQ (I41D, A42Q) whose aggregation propensity was investigated *in vivo*.¹⁷ MD simulations were used for the characterization of heterogeneous conformational ensembles of these IDPs. On the other hand, the estimation of the protein conformational entropy from simulations still remains a considerable challenge.¹⁸ The applicability of the well-adopted quasi-harmonic approximation^{19–21} is questionable, since it assumes a Gaussian distribution of protein conformations about an average structure, which does not hold for IDPs lacking stable three-dimensional structures. In this respect, we recently developed an energy-based approach,²² implemented by combining MD simulations with the liquid integral-equation theory, that does not require the Gaussian distribution of protein conformations and thus is applicable to IDPs. Through the characterization of the conformational entropy in terms of residual structural properties,

Received: January 30, 2013

Revised: March 25, 2013

Published: March 26, 2013



Table 1. Structural and Thermodynamics Characteristics of the Wild-Type (wt) A β 42 and Its Mutants Studied in the Present Work

system	helix ^a	β -sheet ^b	SB ^c	SB network ^d	$\Delta\Delta\mu^e$	TS_{conf}^f	agg. ^g
wt-A β 42	4.9 $\pm$ 6.1	9.1 $\pm$ 7.1	5.7 $\pm$ 3.8	0.6 $\pm$ 2.1	0.0	382.3	
Flemish (A21G)	6.8 $\pm$ 7.2	4.1 $\pm$ 5.7	4.7 $\pm$ 3.4	0.5 $\pm$ 1.8	-5.5	404.2	↓
Arctic (E22G)	4.8 $\pm$ 5.7	7.7 $\pm$ 6.3	5.5 $\pm$ 3.7	1.4 $\pm$ 3.2	142.0	372.0	↑
Dutch (E22Q)	6.9 $\pm$ 6.3	3.2 $\pm$ 3.3	4.4 $\pm$ 3.2	0.7 $\pm$ 2.1	116.1	392.7	↑
Italian (E22K)	5.0 $\pm$ 7.3	7.8 $\pm$ 6.9	6.0 $\pm$ 4.2	0.5 $\pm$ 1.8	164.7	376.2	↑
Iowa (D23N)	6.5 $\pm$ 6.1	9.3 $\pm$ 8.3	5.2 $\pm$ 3.4	1.3 $\pm$ 2.8	112.1	361.3	↑
RR (I41R, A42R)	4.7 $\pm$ 7.4	5.3 $\pm$ 6.8	7.4 $\pm$ 4.3	0.9 $\pm$ 2.5	-14.1	405.6	↓
DQ (I41D, A42Q)	8.0 $\pm$ 8.4	2.9 $\pm$ 3.5	5.6 $\pm$ 3.9	1.1 $\pm$ 2.7	-166.8	376.6	↓

^aAverage population (%) and root-mean-squared fluctuation of the helical structure. ^bAverage population (%) and root-mean-squared fluctuation of the β -sheet structure. ^cAverage population (%) and root-mean-squared fluctuation of the salt-bridge (SB). ^dAverage population (%) and root-mean-squared fluctuation of the salt-bridge (SB) network. ^eChange in the solvation free energy upon mutation, $\Delta\Delta\mu = \Delta\mu(\text{mutant}) - \Delta\mu(\text{wt})$, in kcal/mol. ^fProtein conformational entropy S_{conf} multiplied by the temperature T in kcal/mol. ^gExperimental aggregation propensity change (decrease, ↓; increase, ↑) upon mutation.

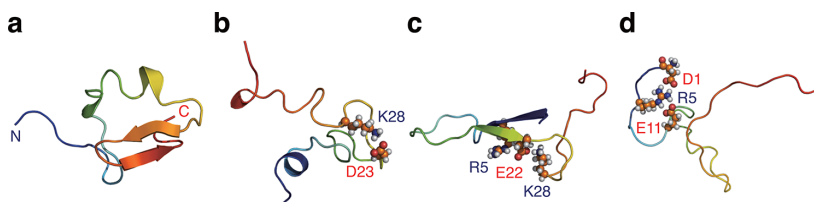


Figure 1. Representative structures of the wild-type A β 42 protein. Each structure is color-coded according to the sequence, ranging from blue to red at the N- and C-termini, respectively. Side chains of charged residues forming salt-bridges are drawn with sphere representation (C, N, O, H: orange, blue, red, white). Structures in panels c and d exhibit the salt-bridge network (R5–E22–K28 and D1–R5–E11, respectively) consisting of two salt-bridges that include a common charged residue (E22 and R5, respectively).

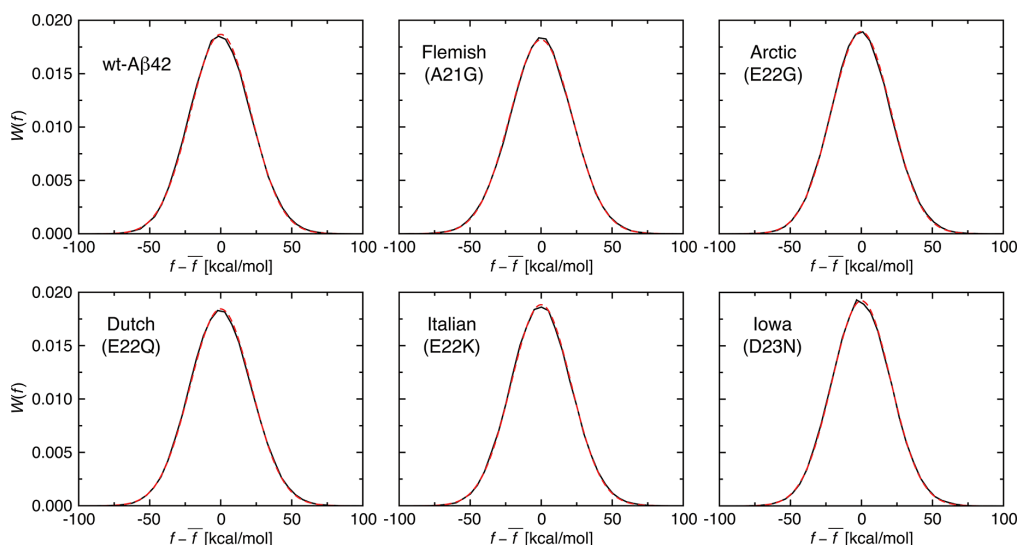


Figure 2. Probability distribution function $W(f)$ (solid curves) of the solvent-averaged protein potential energy f of the wild-type (wt) A β 42 and its pathogenic familial mutants indicated in each panel as a function of the deviation $f - \bar{f}$ from the mean value $\bar{f}$. Dashed red curves denote the fit by the Gaussian function. Corresponding results for RR (I41R, A42R) and DQ (I41D, A42Q) mutants are presented in Figure S1 of the Supporting Information.

we explore a quantitative link between the conformational disorder and entropy in IDPs.

RESULTS

Structurally Heterogeneous Ensemble of A β Protein.

Extensive explicit-water MD simulations were carried out at 300

K and 1 bar for the wt-A β 42 protein and each of its mutants listed in Table 1 to investigate their structural characteristics (see the Methods section for details on the simulation procedures). While no stable tertiary structure was found from our simulations reflecting the intrinsically disordered nature in aqueous environments, A β protein does not exhibit completely random



Assessing the influence of solvation models on structural characteristics of intrinsically disordered protein



Song-Ho Chong, Sihyun Ham*

Department of Chemistry, Sookmyung Women's University, Cheongpa-ro 47-gil 100, Yongsan-Ku, Seoul 140-742, Republic of Korea

ARTICLE INFO

Article history:

Received 9 April 2013

Received in revised form 23 May 2013

Accepted 23 May 2013

Available online 5 June 2013

Keywords:

Molecular dynamics simulation

Intrinsically disordered protein

Water model

Solvation

Amyloid-beta protein

ABSTRACT

Intrinsically disordered proteins are associated with functions of signaling and regulation as well as with a number of diseases such as cancer and neurodegenerative conditions. Molecular dynamics simulation is potentially a powerful tool to study those protein conformational disorders in full atomic details that would be difficult to gain from experimental measurements. However, the accuracy of the simulation critically relies on the quality of underlying potential energy functions or force fields. In particular, structural properties of intrinsically disordered proteins are expected to depend more sensitively on protein–water interactions because of their increased solvent exposure than folded globular proteins. Here, we investigate the impact of solvation models on structural characteristics of full-length amyloid-beta ($A\beta$) protein, an intrinsically disordered protein whose aggregation is linked to Alzheimer's disease. For this purpose, we performed extensive molecular dynamics simulations by combining AMBER ff99SB force field for protein with differing solvent water models as well as solvation procedures with different solvent box sizes. We find that the structural characteristics of $A\beta$ protein differ remarkably depending on the water model adopted as well as on the solvent box size. Our results demonstrate the significant sensitivity on the solvation procedure including the solvent force field in characterizing structural properties of intrinsically disordered proteins.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

Recent efforts to refine force fields for biomolecules have contributed to develop molecular dynamics (MD) simulations into a more reliable tool for studying structure and dynamics of proteins associated with their functions [1–3]. However, relatively less attention has been paid to the validity of models for solvent water. Several deficiencies have been reported for standard three-site water models such as TIP3P [4] and SPC [5], in particular, concerning the temperature dependence of various water properties [6,7]. These models may therefore be inadequate in simulating protein behavior at thermodynamic conditions other than standard temperature and pressure. Furthermore, it is expected that more accurate models for water are necessary to elucidate structural and dynamic properties of intrinsically disordered proteins (IDPs) [8]. This is because IDPs are more solvent-exposed than folded globular proteins reflecting the unfolded nature, and their conformational fluctuations may depend more sensitively on protein–water interactions. In fact, recent simulation studies have demonstrated that the use of more refined water models such as TIP4P-Ew [9] and TIP4P/2005 [10] yields more accurate conformational ensembles,

particularly of disordered peptides [11–14]. However, these studies focused on model peptides or shortened fragments of disease-related proteins. In this paper, we investigate the effects of solvation models on the structural characteristics of full-length, 42-residue form of amyloid-beta protein ($A\beta$ 42), an IDP whose aggregation is associated with Alzheimer's disease [15].

Several groups have carried out explicit-water MD simulations for $A\beta$ 42 protein [16–27], but dissimilar results among those simulations are reported concerning, e.g., representative conformations, distribution of the radius of gyration, and average secondary-structure populations. These inconsistencies in the reported results may partly arise from the heterogeneous conformational ensemble of disordered $A\beta$ 42 protein and an insufficient sampling due to limited computational times. Another contribution to the discrepancies would be the sensitivity of $A\beta$ 42 conformational fluctuations on protein–water interactions since those simulations have been performed with a variety of different combinations for protein force fields and water models. But would this be all? Among characteristic features of IDPs is the large net charge that is primarily responsible for their inability to fold into well-defined structures [8]. In view of the long-range nature of the electrostatic interaction as well as of possible sensitivity of IDPs' structure on their interaction with water, the solvation procedure such as a choice of the solvent box size in the preparation stage of

* Corresponding author. Tel.: +82 2 710 9410; fax: +82 2 2077 7321.

E-mail address: sihyun@sookmyung.ac.kr (S. Ham).

simulations can be another source of discrepancies. It would therefore be important to examine how the solvation procedure affects structural characteristics of IDPs.

Here we report extensive explicit-water MD simulations to investigate the impact of solvation models on the structural characteristics of A β 42 protein. Three distinct sets of simulations that differ in the treatment of solvent water were carried out with the force field for protein fixed to AMBER ff99SB [28]. The first set of simulations was performed employing the TIP4P-Ew water model with a larger solvent box size than the one in the previous simulation studies for A β 42. The second set of simulations differs from the first set only in that the standard TIP3P water model is adopted. The third set of simulations was conducted also with the TIP3P water model, but with a smaller solvent box size that is comparable to the one in the previous simulations. For IDPs such as A β 42, how to efficiently sample their potentially wide conformational space is a challenging issue. Replica-exchange method [29] is a powerful tool in this regard, and has often been adopted in the explicit-water MD simulations for A β 42 [18,23–26]. However, explicit handling of solvent water molecules requires a large number of replicas since the energetic fluctuation becomes relatively smaller for larger systems. This is problematic especially for the first and the second sets of our simulations that include a large number of explicit water molecules, which will significantly limit the length of the simulation at a desired temperature. In the present work, we conducted multiple independent simulations that start from carefully prepared protein structures obtained by heating and conformational annealing simulations. We then examined the performance of the water model and the dependence on the solvation procedure in characterizing structural features of A β 42 through the comparison with available experimental data.

2. Material and methods

2.1. MD simulations

Three distinct sets of MD simulations for A β 42 protein were carried that differ in the water model and the solvation procedure detailed below and summarized in Table 1. The sequence of A β 42 protein is ⁽¹⁾DAEFR HDSEY EVHHQ KLVFF AEDVG SNKGA IIGLM VGGVV IA⁽⁴²⁾, and the protein force field employed here is fixed to ff99SB. All the simulations were performed at 300 K and 1 bar using AMBER11 simulation package [30]. Each set of simulations consists of ten independent production runs of 100 ns length (i.e., an accumulative simulation time of 1 μ s). The simulations were done under neutral pH, where Lys and Arg residues are positively charged and Glu and Asp are negatively charged, and A β 42 is in the -3 charged state. The particle mesh Ewald method [31] was applied for treating long-range electrostatic interactions, while a 10 Å cutoff was used for the short-range non-bonded inter-

actions. The hydrogen atoms were constrained to the equilibrium bond length using the SHAKE algorithm [32], which enables simulations with a 2-fs time step. Temperature and pressure were controlled by Berendsen's thermostat and barostat with coupling constants of 1.0 and 2.0 ps, respectively [33].

2.2. A β 42/TIP4P-Ew system

We employed the TIP4P-Ew model for solvent water in the first set of our simulations, and this system is referred to as A β 42/TIP4P-Ew in the following. Since A β 42 protein in the monomeric state is intrinsically disordered and no atomic-level structure in aqueous environment is available experimentally, the initial conformations for independent production runs were carefully prepared as follows. The starting conformation was taken from the NMR structure determined in an apolar solvent (PDB ID: 1IYT [34]), which was solvated by explicit water and neutralized by counter Na⁺ ions. The starting system was initially subjected to 500 steps of steepest descent minimization followed by 500 steps of conjugate gradient minimization while A β 42 protein was constrained by 500 kcal/(mol Å²) harmonic potential. Then, the system was minimized using 1000 steps of steepest descent minimization followed by 1500 steps of conjugate gradient minimization without harmonic restraints. The system was subsequently subjected to 20 ps equilibration process in which the temperature was gradually raised from 0 to 300 K with a constant volume. This is followed by 200 ps constant-pressure (*NPT*) ensemble simulation of 1 bar to achieve the correct water density. We then heated the system to 600 K and carried out 20 ns canonical (*NVT*) ensemble simulation. From this 600-K simulation of 20 ns length, we took ten protein structures with 2 ns time interval. The C α RMSD (root-mean-squared deviation) between any pair of these ten structures was $\sim$ 10 Å, and they were considered as independent configurations. Each of the ten structures was then gradually annealed to 300 K: the temperature was decreased from 600 K with 50 K interval, and 1-ns *NVT* simulation was performed at each of the intervening temperatures between 600 K and 300 K. Subsequently, additional 5-ns equilibration *NPT* simulation was performed at 300 K and 1 bar. The final structures obtained from the annealing simulations starting from the ten different 600-K structures were used as the initial structures for the ten independent *NPT* production runs of 100 ns each at 300 K and 1 bar. The length of the production run was chosen in view of the estimation in Ref. [24], where it was suggested that the simulation longer than $\sim$ 60 ns/replica is necessary to achieve the convergence in the ensemble for A β 42. The minimization, heating, annealing, and production runs were all carried out under the fully hydrated condition with 13,417 water molecules dissolved in the cubic periodic box of the average side length of 74.3 Å. The size of the solvent box is larger than a typical one ($\sim$ 50 Å) in the recent simulation studies for A β 42 [24–27].

Table 1
Summary of simulation procedures for the systems studied.

Systems	Preparation protocols	Production runs	# of water	Box size ^a (Å)
A β 42/TIP4P-Ew	1. Heating in water at 600 K 2. Annealing to 300 K 3. Equilibration at 300 K (5 ns)	100 ns $\times$ 10	13,417	74.3
A β 42/TIP3P-I	1. Heating in water at 600 K 2. Annealing to 300 K 3. Equilibration at 300 K (5 ns)	100 ns $\times$ 10	13,666	75.0
A β 42/TIP3P-II	1. Heating <i>in vacuo</i> at 600 K 2. Solvation by explicit water 3. Equilibration at 300 K (5 ns)	100 ns $\times$ 10	4290 ^b	51.3

^a Average side length of the cubic simulation box in the production runs.

^b Average number of water molecules upon solvating A β 42 protein with 10 Å buffer.

Distinct Role of Preferential Hydration in Protein Aggregation**

Song-Ho Chong and Sihyun Ham*

Understanding the molecular determinants of the relative propensity for proteins to aggregate in a cellular environment has been a central issue in attacking protein-aggregation diseases^[1–5] and in the development of human therapeutics.^[6,7] Despite the expectation that the protein aggregation can largely be attributed to the direct protein-protein interactions within an aggregate or in solution, we here unveil a crucial role of hydration water in ruling the aggregation propensity of proteins both in vitro and in vivo. The protein overall hydrophobicity, defined solely by the hydration free energy of a protein in its monomeric state sampling its equilibrium structures, was shown to predominantly dictate the protein aggregation propensity in aqueous solutions. We also find striking discrimination by the hydration water of positively and negatively charged residues depending on the protein net charge in regulating the solubility of a protein, which establishes novel design strategies for the biotechnological generation of aggregation-resistant proteins as biotherapeutics.

Hydration water plays a central role in many biological processes.^[8,9] Yet, the role of water in protein aggregation has been largely unexplored due to the perception that protein adhesion (protein-protein interaction) is the main actor and the surrounding water is just a spectator and can be disregarded in understanding the protein aggregation phenomena. Indeed, almost all studies on the prediction of protein aggregation have mainly focused on the sequence characteristics such as the intermolecular β -structure forming propensity.^[10–12] Such a protein-centric view, however, ignores water as an active constituent of biological systems.^[13–15] In fact, what determines a protein to remain soluble or to be aggregation-prone in aqueous environments should intrinsically rely on its affinity toward water in its monomeric state. The main focus of the present work is to uncover water's controlling role in discriminating aggregation-prone proteins from soluble ones and thereby to contribute to the understanding and prediction of the protein aggregation propensity.

An interaction induced by water such as the hydrophobic interaction is often invoked to explain biomolecular self-assembly in aqueous media.^[8,16] According to the statistical thermodynamics, the water-mediated force is quantified by the hydration free energy G_{hyd} – transfer free energy from gas to aqueous phases – that characterizes the affinity toward the solvent water.^[17] Molecules of larger G_{hyd} are called more hydrophobic, between which the hydrophobic interaction is more effective. A key question here is whether this concept on hydrophobicity applies beyond the conventional levels of small molecules or individual amino acids, and whether a protein of higher hydrophobicity is more prone to self-aggregate in aqueous environments. While the protein overall hydrophobicity (protein hydration free energy G_{hyd}) is not directly accessible from experiments,^[17] recent developments in computational tools make it possible to approach this quantity. First, owing to advances in computing power and refinements of force fields, molecular dynamics (MD) simulations have developed into a powerful and accurate tool for protein conformational studies.^[18,19] Second, molecular theory of solvation has been much advanced in recent years, allowing us to compute^[20] and analyze^[21–23] protein hydration free energy based on simulated protein conformations.

In this study, we quantified the protein overall hydrophobicity based on these advanced computational tools by applying them to intrinsically disordered proteins, twenty-two mutants of amyloid- β (A β 42) protein and twenty-one mutants of the N-terminal domain of the *Escherichia coli* protein HypF (HypF-N), as well as to natively folded proteins, fifteen mutants of human muscle acylphosphatase (AcP), whose aggregation propensities have been experimentally characterized in vitro^[24,25] and in vivo.^[26–28] (See Figures 1b and 1c for snapshot structures and sequences of wild-type proteins. All the mutants studied are specified in Tables S1–S3 in the Supporting Information.) For each of these proteins, we performed extensive all-atom, explicit-water MD simulations to sample their equilibrium solution structures. (Details of the computational methods are provided in the Supporting Information.) The hydration free energy G_{hyd} for each protein was then computed by applying the integral-equation theory of liquids^[20–23] to simulated protein conformations, and the resulting average G_{hyd} is identified as the protein overall hydrophobicity: a larger (i.e., more positive) G_{hyd} value is associated with more increased protein hydrophobicity.

The correlation between the experimental aggregation propensity and the computed protein hydrophobicity is striking (Figure 1a) considering a large variation in the aggregation propensity, distinct nature of underlying protein structures, and several types of mutations involved. This highlights that the hydration water surrounding a protein, rather than a protein itself, is a predominant factor that controls to what extent a protein would exhibit aggregation-prone nature in aqueous environments. This is corroborated by the observation that representative sequence-based

[*] Dr. S.-H. Chong, Prof. S. Ham
Department of Chemistry
Sookmyung Women's University
Cheongpa-ro 47-gil 100, Yongsan-Ku, Seoul 140-742, Korea
Fax: (+) 82-2-2077-7231
E-mail: sihyun@sookmyung.ac.kr
Homepage: <http://nbcc.sm.ac.kr>

[**] This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (Nos. 2012-0003068 and 2012R1A2A01004687).



Supporting information for this article is available on the WWW under <http://dx.doi.org/10.1002/anie.201309317>

models focusing on the intermolecular β -structure forming propensity do not consistently yield such high correlation with the experimental data (Figure S1 in the Supporting Information). Furthermore, we found that it is the electrostatic term in the protein-water interaction that dictates the protein hydrophobicity (Figure S2 in the Supporting Information). This implies that the water-induced interaction that causes proteins to aggregate is of long-range nature. In fact, we have recently demonstrated that the long-range water-mediated force is primarily responsible for two aggregating proteins to approach each other from a large separation to a contact distance.^[23] After proteins make atomic contacts, direct protein-protein interactions will then come into play. The intermolecular β -sheet forming propensity, taken into account in the sequence-based algorithms to identify aggregation-prone regions,^[10–12] will play a major role in the subsequent structural reorganization step toward the formation of an ordered aggregate called amyloid fibril. Methods to identify interaction regions through the estimation of desolvation free energy^[29] will also be useful since the protein contact formation involves removal of interfacial water.

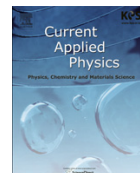
What are, then, molecular determinants of the protein overall hydrophobicity? The dominance of the electrostatic interaction in regulating the protein hydrophobicity implies more relevance of the charged residues over the neutral ones. In fact, we find from the site-directed thermodynamic analysis that the charged residues more significantly influence the hydration free energy than the neutral ones (Figure 2). Quite unexpectedly, we observe a contrasting role of positively (Lys and Arg) and negatively (Glu and Asp) charged residues embedded in the protein context (Figure 2). The negatively charged residues exhibit much more negative hydration free energy (i.e., much more affinity toward water) than the positively charged ones in both A β 42 (Figure 2b) and HypF-N (Figure 2c) whose total charge is negative. Just the opposite trend is observed for AcP whose total charge is positive: the positively charged residues display significantly more negative hydration free energy than the negatively charged ones (Figure 2d). It is well recognized that chaotropes (Lys and Arg side chains) bind water more weakly than kosmotropes (Glu and Asp side chains).^[30] In addition, the 4 and 3 methylenes in Lys and Arg side chains may play more significant hydrophobic roles compared to the 2 and 1 methylenes in Glu and Asp side chains. These are in fact reflected in the hydration free energy for free charged amino acids (Figure 2a). However, this argument on the free charged amino acids does not rationalize the strikingly distinct behavior of positively and negatively charged residues embedded in the protein.

The key to understand this contrasting behavior lies in the long-distance hydration structure surrounding the protein surface (Figure 3). Due to the long-range nature of the electrostatic term dominating the charged residue-water interaction, the long-distance hydration structure of a charged residue is affected also by neighboring charged residues: the equilibrium water distribution is determined in leading order by a net charge produced by those residues. In the case of the protein whose net charge is negative as in A β 42 and HypF-N, this yields such a water orientational distribution at long-distance layers in which the net dipole moment of water is directed toward the protein (blue water molecules shown in Figures 3b and 3c). The resulting interaction between charges on the protein surface and the water dipole at long

distances becomes unfavorable (favorable) between the positively (negatively) charged residues and water. Thus, the long-distance hydration structure surrounding a protein of negative net charge provides thermodynamic preference for negatively charged residues over positively charged ones on the protein surface. Applying the same argument, just the opposite discrimination of positively and negatively charged residues results on the protein surface whose net charge is positive as in AcP since the net dipole moment of solvent water at long distances in this case will be reversed, i.e., directed outward from the protein (red water molecules displayed in Figure 3d). Thus, the contrasting behavior of positively and negatively charged residues is caused by the “cross talk” between charged residues on the protein surface through the solvent water in the long-distance hydration layers.

The striking discrimination of positively and negatively charged residues depending on the protein net charge has significant implications in various fields including biotechnological and medical sciences. One can control the aggregation propensity of a protein by site-directed mutagenesis using the following “design principles”. When the net charge of a protein is negative, mutating neutral amino acid to negatively (positively) charged one will in principle increase (decrease) the solubility in water. Just the opposite will result when the protein net charge is positive, i.e., mutating neutral amino acid to negatively (positively) charged one will diminish (enhance) the protein solubility in water. These design principles are consistent with the recent solubility measurements for ribonuclease Sa having a negative net charge that reported^[31] (i) Glu and Asp contribute more favorably to protein solubility than do any of the other amino acids, (ii) Lys and Arg contribute unfavorably to protein solubility, and (iii) Lys turns to provide a favorable contribution to solubility when there is an excess of positive charge on the protein compared to when there is an excess of negative charge. The contrasting behavior of charged amino acids also provides a detailed rationalization of the previous experimental observation^[32] that the proteins with higher contents of negatively charged residues tend to be more soluble than those of positively charged ones in the entire *Escherichia coli* proteins since the net charge of soluble proteins is mostly negative in physiological conditions.

The identification of the factors that control whether a protein exhibits aggregation-prone nature or remains soluble in aqueous environments is of central importance in preventing protein aggregation diseases. Major findings of the present study are that the protein overall hydrophobicity quantified by hydration free energy predominantly determines protein aggregation propensity both in vitro and in vivo and that the protein hydrophobicity is significantly influenced by a contrasting behavior of positively and negatively charged residues depending on the net charge of a protein. Our computational tools that deal with the protein three-dimensional structure as well as its hydration thermodynamics enable an *ab initio* prediction of mutation effects on the protein aggregation propensity, and the design principles therefrom for developing aggregation-resistant proteins will find a wide range of applications in biotechnology and biotherapeutics.



Vanadium-induced decrease in the magnetization of a V_xCo_{1-x} alloy film on the W(110) surface



Hangil Lee

Department of Chemistry, Sookmyung Women's University, Seoul 140-742, Republic of Korea

ARTICLE INFO

Article history:

Received 27 July 2012

Received in revised form

27 November 2012

Accepted 29 November 2012

Available online 2 January 2013

Keywords:

SRPES

Magnetic phase transition

Structural phase transition

LEED

Alloy film

ABSTRACT

The magnetic and structural properties of a series of V_xCo_{1-x} alloy films prepared in a variety of compositions on a W(110) surface were investigated using spin-resolved photoemission spectroscopy (SRPES) and low-energy electron diffraction (LEED) measurements. The epitaxial structures of the films were preserved over all V concentrations, despite the observation of a structural phase transition (SPT: hcp/bcc at $x = 0.38$), which was apparent from the LEED images. A magnetic phase transition (MPT) from ferromagnetic to paramagnetic occurred in the Co-rich region at $x = 0.38$. The SPT (from hcp and bcc) and MPT (from ferromagnetic to paramagnetic) were clearly correlated at $x = 0.38$, possibly due to changes in the electronic and magnetic structures of the valence band, as monitored by SRPES and LEED. The magnetization in the V_xCo_{1-x} alloy films on a W(110) surface varied systematically with the V concentration, and correlations between the structural changes and magnetic properties were revealed using SRPES and LEED.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

The composition-dependent magnetic and structural properties of magnetic thin alloy films have been actively investigated because this degree of freedom provides a means by which film properties may be controlled beyond what can be achieved using magnetic alloy films on a single metal substrate. The physical properties of magnetic thin alloy films can be tailored widely toward optimization for specific applications [1–4].

The characteristics of alloy films that are fabricated from ferromagnetic metals and non-magnetic species, such as V, display unusual paramagnetic behavior in the bulk state due to the induction of magnetic moments at the non-magnetic sites. A detailed understanding of these properties warrants theoretical and experimental investigations that could guide the tailoring of magnetic materials toward specific applications. Binary alloy films containing V in combination with 3d ferromagnetic materials, such as V_xFe_{1-x} and V_xNi_{1-x} alloy films, display magnetic properties that can vary depending on the alloy composition [5]. The electronic and magnetic properties of these systems depend on the presence or absence of a Stoner structure in the conventional strictly exchange-split bands [6]. Loss of the Stoner structure results in antiferromagnetic coupling between the early transition metals and the late

transition metals, yielding a ferromagnetic state indicative of a decrease in the magnetic moment [7,8].

We previously reported that the saturation magnetization of a V_xFe_{1-x} alloy film decreased continuously as the V content increased, and the saturation magnetization completely vanished at V concentrations in excess of 70%. The magnetic moment of this system behaved in a more complicated manner [3,9]. In light of these results, a V_xCo_{1-x} alloy film deposited on a W(110) surface was expected to display similar properties, such that the total magnetization would be determined by the Co and V moments and concentrations. An additional advantage of studying alloy films is that one can mix not only ferromagnetic and paramagnetic metals, but also elements that crystallize with different Bravais lattices, as in the case discussed here, bcc-V and hcp-Co.

These cases have presented an opportunity for observing correlations between the structural phase transition (SPT) and the magnetic phase transition (MPT). In a bulk V_xCo_{1-x} alloy, the SPT, from hcp to bcc, and the MPT, from ferromagnetic to paramagnetic, occur at the same V concentration ($x > 0.38$), raising the question of a possible correlation between the two phenomena, particularly in epitaxial films. A recent study of the magnetization in V_xCo_{1-x} alloy films showed that the magnetization decreased with increasing V content, and the films were paramagnetic for $x > 0.35$. This study did not, however, suggest a correlation between the structural phase transition and the magnetic phase transition [10,11]. The easy magnetization direction in bulk Co, hcp-[0001], differed from that in a Co film on a W(110) surface, which was

E-mail address: easyscan@sookmyung.ac.kr.

oriented along the $[1-10]$ -axis of the substrate [12]. The magnetic properties, therefore, changed upon application of a magnetic field along the $[1-10]$ direction.

Here, the magnetization of a V_xCo_{1-x} alloy film on a W(110) surface was measured as a function of x , and the correlations between the structural changes and the magnetic properties were examined using spin-resolved photoemission spectroscopy (SRPES) and low-energy electron diffraction (LEED) studies. To our knowledge, the composition-dependent variations in the magnetic properties of epitaxial V_xCo_{1-x} alloy films on a W(110) surface have yet to be systematically reported.

2. Experimental

All experiments were performed under ultrahigh vacuum conditions (base pressure 5×10^{-11} Torr) [13]. The W(110) substrates were prepared using repeated cycles of high-temperature oxygen exposure (1800 K, 1×10^{-8} Torr), followed by flash heating at higher temperatures (2500 K) under UHV conditions [14]. The W(110) surface cleanliness was monitored prior to deposition by inspecting the surface states of the W 4f core-level spectra, wherein the quality of the atomic order was determined based on visual inspection of the sharpness of the (1×1) LEED pattern. The V_xCo_{1-x} films were prepared *in situ* by co-deposition of the Co and V metals on a clean W(110) surface.

During electron beam evaporation of the high-purity metals (0.9999), the pressure increased by at most 1.5×10^{-10} Torr. During the deposition and any subsequent measurements, the samples were kept at room temperature to prevent phase segregation at the W(110) surface. The film thickness was calibrated using a thickness monitor, and the deposition rate was maintained at an approximately constant value of 4 Å/min. Relatively thick films (40 Å) were prepared. The alloy films were characterized by the complete absence of the W(110) signal (particularly the sharp W 4f core-level spectrum) from the photoemission spectra.

The W(110) crystal was oriented along the *in-plane* $\langle 1-10 \rangle$ direction, along the vertical. The incident light, which was linearly polarized parallel to the horizontal plane, was incident at a 45° angle with respect to the W(110) surface normal. All photoemission spectra were recorded in normal emission mode using a commercial hemispherical electron analyzer (Omicron EA125). The magnetic measurements were performed under conditions in which the samples were in remanence, after saturation using magnetizing coils connected to a current power supply. The spin polarization of the photoemitted electrons were measured using a 30 kV mini-Mott detector mounted on the back of the dispersive optic of the electron spectrometer.

3. Data and results

Fig. 1 displays the core-level spectra (left panel), the corresponding LEED patterns (middle panel), and the valence band spectra (right panel) of V_xCo_{1-x} alloy films with various alloy compositions. The core-level spectra displayed in the left panel of Fig. 1 were sensitive to the evolution of the Co 3p core level, located at 59 eV, and the V 3p and V 3s core levels, located at 37 eV and 66 eV, respectively (bottom: pure Co and top: V rich). These figures revealed that the relative intensities of the Co 3p and V 3p core-level peaks were remarkably uniform, suggesting the formation of random alloyed films, at least on the macroscopic scale probed by the photoemission experiments (the light spot dimensions were $500 \times 200 \mu\text{m}$). The intensity ratio between the Co 3p and V 3p core-level peaks was used to independently verify the concentrations, which were determined by the relative evaporation fluxes and fixed using the thickness monitor.

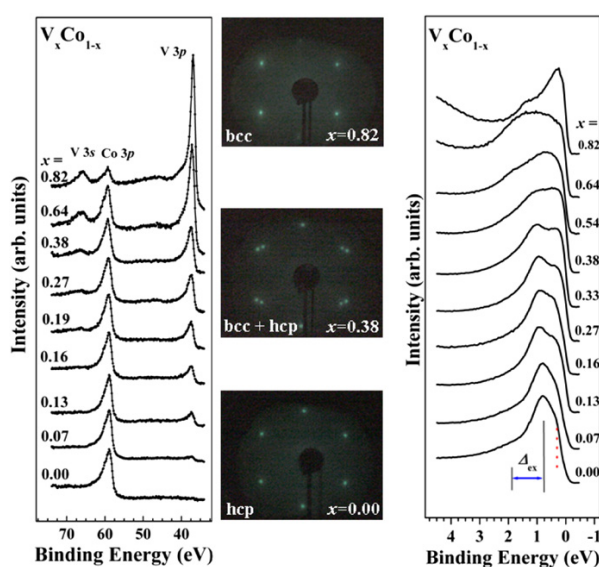


Fig. 1. Left panel: core-level photoemission spectra collected from V_xCo_{1-x} alloy films in the region encompassing the Co 3p, V 3p, and V 3s core levels as a function of the V concentration (x), under 160 eV of photon energy. Middle panel: LEED images acquired at representative V concentrations. Right Panel: Valence band photoemission spectra obtained from the same conditions, under 40 eV of photon energy.

The middle panel of Fig. 1 reveals sharp LEED patterns obtained from a variety of V concentrations (x), indicating the formation of epitaxial films over the entire concentration range. The V-rich alloy films displayed LEED patterns with rectangular spots indicative of bcc films with an in-registry relationship to the (110) lattice of the W(110) surface. In the Co-rich alloys, the LEED pattern was hexagonal, indicating the formation of an hcp lattice that exposed the hexagonal plane and adopted a Nishiyama–Wasserman relationship, $\langle 11-20 \rangle$ Co $\parallel$ $\langle 001 \rangle$ W. Similar properties were reported previously for the epitaxial growth of pure Co and pure Ni on a W(110) surface [15,16].

Interestingly, the region at a V concentration of 38% displayed a mixed phase of two LEED patterns (bcc and hcp) collected over a relatively narrow range ($\delta x < 4\%$). These LEED spots appeared quite sharp, possibly indicating phase separation between the hcp and bcc lattices. It may be interesting to investigate this system using spectro-microscopy techniques to determine whether these two structures corresponded to local inhomogeneity in the film stoichiometry.

The valence band spectra were collected from the normal emission and are presented in the right panel of Fig. 1. The lower spectrum, obtained from clean Co, was dominated by emission from the 3d bands, which extended to the first 0.8 eV below E_F at photon energy of 40 eV. The addition of V shifted a feature at 0.8 eV in the clean Co spectrum (bottom) toward lower binding energies. This shift transformed the broad Co spectrum into a spectrum with two distinct features: one at 0.8 eV that filled the region between the two peaks, and the other at higher binding energies (about 2.0 eV). This transformation appeared to reach completion at a concentration corresponding to the bcc/fcc mixture concentration ($V_{0.38}Co_{0.62}$). Remarkably, the valence states underwent a major redistribution in the structural phase transition region. The feature between the two peaks became dominant at higher V concentrations. In the pure V film, the spectrum was again dominated by the 3d state, although this dominance was less than that observed in the pure Co film. At a photon energy of 40 eV, the valence band of

Inter-row Adsorption Configuration and Stability of Threonine Adsorbed on the Ge(100) Surfaces

Myungjin Lee, Youngchan Park, Hyuk Jeong, and Hangil Lee*

Department of Chemistry, Sookmyung Women's University, Seoul 140-742, Korea. *E-mail: easyscan@sookmyung.ac.kr

Received December 7, 2012, Accepted January 8, 2013

The adsorption structures of threonine on the Ge(100) surface were investigated using core-level photoemission spectroscopy (CLPES) in conjunction with density functional theory (DFT) calculations. CLPES measurements were performed to identify the experimentally preferred adsorption structure. The preferred structure indicated the relative reactivities of the carboxyl and hydroxymethyl groups as electron donors to the Ge(100) surface during adsorption. The core-level C 1s, N 1s, and O 1s CLPES spectra indicated that the carboxyl oxygen competed more strongly with the hydroxymethyl oxygen during the adsorption reaction. Three among six possible adsorption structures were identified as energetically favorable using DFT calculation methods that considered the inter- and intra-bonding configurations upon adsorption onto the Ge(100) surface. These structures were O-H dissociated N dative inter bonding, O-H dissociated N dative intra bonding, O-H dissociation bonding. One of the adsorption structures: O-H dissociated N dative inter bonding was predicted to be stable in light of the transition state energies. We thus confirmed that the most favorable adsorption structure is the O-H dissociated N dative-inter bonding structure using CLPES and DFT calculation.

Key Words : Threonine, Ge(100), Inter-row adsorption configuration, DFT calculation, CLPES

Introduction

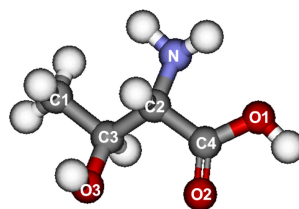
Studies of the reactions between organic molecules and semiconductor surfaces have become increasingly important over the past few years. Technologies that rely on interactions between organic materials and semiconductor technologies are being developed for molecular electronics, biosensor, and nanotechnology applications.¹⁻⁸ An understanding of the reactions of amino acids, which act as zwitterions, on semiconductor surfaces may guide the design of these new technologies.⁹⁻¹¹ Studies in this field have focused primarily on the covalent attachment of organics to group IV semiconductors.¹² Methods for providing appropriate chemical functionalization of group IV semiconductor materials have been actively pursued.¹³⁻¹⁶

Our group has studied the adsorption structures and properties of a variety of amino acids on the Ge(100) surfaces in an effort to lay the groundwork for fabricating new bio-electronic hybrid devices.¹⁷⁻²¹ In this study, we have investigated the adsorption structures of threonine on the Ge(100) surface using core-level photoemission spectroscopy (CLPES) experiments and density functional theory (DFT) calculations. Threonine is a model molecule for studying the effects of multiple functional groups on the adsorption chemistry at a semiconductor surface. Threonine includes 4 carbon atoms bonded to a carboxyl group (-COOH), a hydroxyl group (-OH), an amine group (-NH₂), a methyl group (-CH₃), and a hydrogen atom.²²⁻²⁵ Several adsorption structures are possible for threonine on the Ge(100) surface, making threonine a very interesting adsorption test case.

We previously reported the adsorption structures of serine on the Ge(100) surface, which is structurally similar to

threonine, with the exception of the side chain length. Serine was predicted to adsorb to the Ge(100) surface via a stable O-H dissociated N dative bonding.²⁶ The similarities of threonine and serine suggest that the compounds should yield similar experimental and theoretical results. However, previous studies of serine did not consider the selectivity and stability associated with inter bonding and intra bonding. It is, therefore, necessary to examine the selectivity and stability of the threonine inter bonding and intra bonding adsorption structures to evaluate the possibility that threonine can interact with adjacent the Ge(100) surface dimers due to its chain length.

The primary purpose of this study was to investigate the competition between hydroxyl groups from hydroxymethyl and carboxyl groups during the adsorption of threonine on Ge(100). In addition, we also wanted to show the adsorption selectivity of inter bonding and intra bonding in the adsorption structure of threonine on the Ge(100) surface. The most stable adsorption configuration was identified, and the electronic structure was characterized experimentally using CLPES



Scheme 1. Ball and stick model for chemical structure of L-threonine. The blue, red, gray, and white colored balls indicate nitrogen (N), oxygen (O), carbon (C), hydrogen (H), respectively.

experiments and DFT calculations.

Experimental Section

The Ge(100) surface (*p*-type, $R = 0.10\text{--}0.39\ \Omega$) was cleaned using several cycles of sputtering with 1 keV Ar^+ ions for 20 minute at 700 K, followed by annealing at 900 K for 10 minutes. The cleanness of the Ge(100) surface was checked using low-energy electron diffraction (LEED) and the existence of surface state in Ge 3*d* core level spectrum using CLPES. Threonine ($\text{HO}_2\text{CCH}(\text{NH}_2)\text{CHOHCH}_3$, 99% purity) was purchased from Aldrich and further purified through several sublimation and pumping cycles to remove all dissolved gases prior to exposure to the Ge(100) surface. We defined 1.0 monolayer (ML) threonine coverage as the maximum density of threonine that could be deposited on the Ge(100) surface without yielding the N 1*s* CLPES peak characteristic of multilayer formation.

CLPES experiments were performed at the 10D beamline of the Pohang Accelerator Laboratory. The C 1*s*, N 1*s*, and O 1*s* core-level spectra were obtained using a PHOIBOS 150 electron energy analyzer equipped with a two-dimensional charge-coupled device (2D CCD) detector (Specs GmbH) using photon energies of 345, 460, and 600 eV to enhance the surface sensitivity. Four binding energies in the core-level spectra were calibrated with respect to the binding energy of a clean Au 4*f* core-level spectrum (84.0 eV) at the same photon energy. The base pressure of the chamber was maintained below 9.5×10^{-11} Torr. All spectra were recorded in the normal emission mode. The core-level spectra were

carefully analyzed using a standard nonlinear least squares fitting procedure with Voigt functions.²⁷

DFT calculations were used to predict the energetics of the reaction pathways for threonine adsorption onto the Ge(100). All DFT calculations of the adsorption energies were performed using the JAGUAR 9.1 software package, which applied a hybrid density functional method and included Becke's three-parameter nonlocal-exchange functional with the Lee–Yang–Parr (B3LYP) correlation functional.^{28–30} The calculations considered a four-dimer ($\text{Ge}_{35}\text{H}_{32}$) cluster models. The geometries corresponding to the important local minima on the potential energy surface were determined at the B3LYP/LACVP** level of theory. The LACVP** basis set is a mixed basis set that uses the LACVP basis set to describe Ge atoms and the 6-31G basis set for the remaining atoms. Moreover, the LACVP basis set describes atoms beyond Ar in the periodic table using the Los Alamos effective core potentials developed by Hay and Wadt.^{31,32} For each cluster, optimization was performed by fixing the bottom two layers of the Ge atoms in ideal Ge crystal positions while allowing the top layer of the Ge atoms (including the dimer atoms) and the atoms of the chemisorbed adsorbate to relax. The geometries of important local minima and the transition states on each energy diagram were also calculated. The local minima and transition states were verified as being optimized structures using the same basis sets.^{33–36}

Results and Discussion

CLPES data were acquired for the core-levels in a threo-

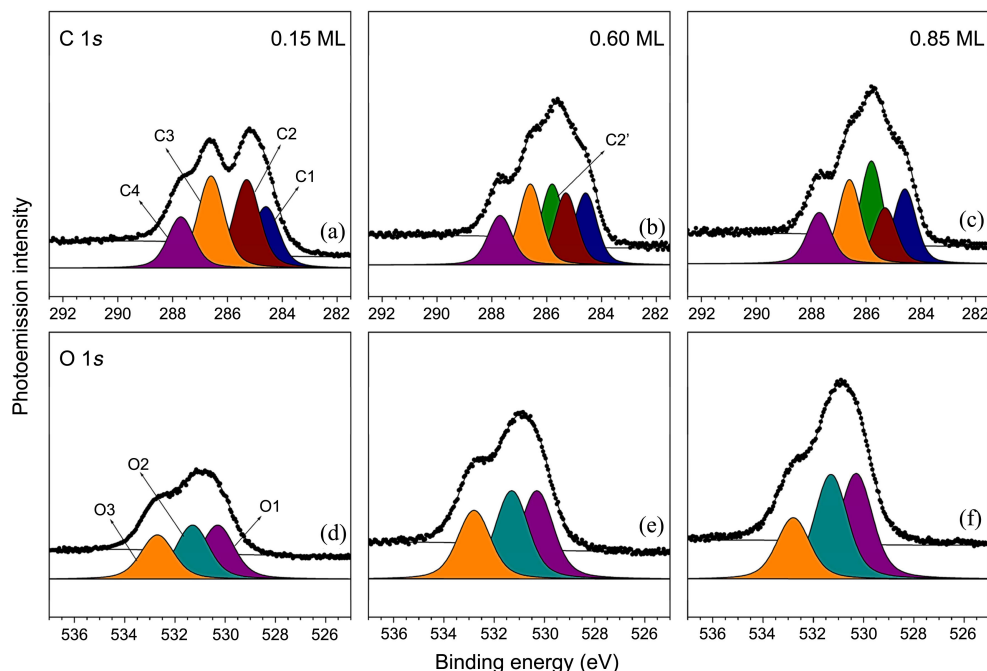
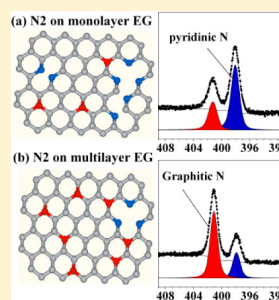


Figure 1. C 1*s* and O 1*s* core-level spectra for (a, d) 0.15 ML, (b, e) 0.60 ML, (c, f) 0.85 ML threonine on the Ge(100) at 300 K. The dots indicate the experimental values and the solid lines represent the results of peak fitting.

Annealing Effects after Nitrogen Ion Casting on Monolayer and Multilayer Graphene

Ki-jeong Kim,[†] Sena Yang,[‡] Youngchan Park,[‡] Myungjin Lee,[‡] BongSoo Kim,^{*,†} and Hangil Lee^{*,‡}[†]Beamline Research Division, Pohang Accelerator Laboratory (PAL), Pohang 790-784, Republic of Korea[‡]Department of Chemistry, Sookmyung Women's University, Seoul 140-742, Republic of Korea

ABSTRACT: The modification of the electronic properties of a coverage-dependent graphene layer by nitrogen ions irradiation was investigated using core-level photoemission spectroscopy (CLPES). Here we describe preparation of monolayer and multilayer epitaxial graphene (EG) functionalized by nitrogen ions irradiation with 100 eV to minimize the damage of graphene layer as we annealed up to 1300 K to track the surface property changes using CLPES. As a result, on the monolayer EG, we found that pyridinic nitrogen mainly existed on the surface. On the multilayer EG, we observed the formation of graphitic nitrogen remaining as a major species confirmed using N 1s core-level spectra. Through a work function change ($\Delta\Phi$) measurement, both systems indicated p-type doping properties with 4.71 (monolayer EG) and 4.87 eV (multilayer EG) of work function values after N₂ ion irradiation. Interestingly, we observed that monolayer EG maintained its p-type doping character, whereas multilayer EG changed the doping character from p-type to n-type as we annealed both systems up to 1300 K due to the discrepancy of the electron charge transfer. We will systematically demonstrate these variations of electronic properties for both monolayer and multilayer EG.



INTRODUCTION

In 2010, A. K. Geim and K. S. Novoselov were awarded the Nobel Prize in Physics for their landmark work on graphene, which demonstrated the outstanding electronic and mechanical properties of this material.¹ Its potential applications are expected to yield new electrical devices, and industries have supported significant research in graphene-related studies seriously.^{2–8} By controlling the band gap effectively, which can be achieved through a variety of processes, its versatility can be improved.^{9,10} The zero band gap in graphene is a consequence of the identical environments of the two carbon atoms in the graphene unit cell. Unfortunately, it is not easy to manipulate the atoms in a unit cell or manipulate them using the lithographic modification techniques because the distance between the carbon atoms is only 1.4 Å apart.³ Instead, chemical-modified reactions that form superstructures using chemical functionalized molecule and atom may be used to introduce a potential difference into the band gap.^{11–13} Recently, many research groups have focused on chemical doping using nitrogen for various applications such as biosensors, fuel cells, and so on.^{14–16}

Recently, L. Zhao et al. grew N-doped graphene films using chemical vapor deposition on copper foil substrates, indicating that the individual nitrogen atoms were incorporated as graphitic dopant, and a fraction of the extra electron on each nitrogen atom was delocalized into the graphene lattice. The electronic structure of nitrogen-doped graphene was strongly modified only within a few lattice spacings of the site of the nitrogen dopant.^{17–22}

In this work, we introduced low-energy nitrogen ions (100 eV) as a method for changing the monolayer and multilayer

graphene surfaces properties to minimize the surface damage. In addition, we identified three nitrogen species that induced such changes on few-layer epitaxial graphene (EG) on SiC(0001): N₂, pyridinic nitrogen, and graphitic nitrogen, described previously.²³ Through the consecutive sample annealing up to 1300 K, we could pursue that the surface property changes and these properties were monitored by core-level photoemission spectroscopy (CLPES).

EXPERIMENTAL SECTION

Monolayer and multilayer EG were grown on 6H-SiC(0001) substrates by thermal annealing under ultrahigh vacuum (UHV) conditions. The nitrogen-doped ($N_D = 9 \times 10^{17} \text{ cm}^{-3}$) Si-terminated 6H-SiC(0001) sample was purchased from Cree Research (USA). Samples were prepared 2 mm $\times$ 12 mm in size with a thickness of 0.25 mm. The 6H-SiC(0001) annealing temperature was monitored by infrared pyrometer, assuming an emissivity of 0.90. After overnight sample outgassing at 1200 K, samples were annealed at 1200 K under a Si flux (1 Å/min) to make a Si-rich surface on 6H-SiC(0001) for 5 min. After annealing at 1450 K, a carbon-rich ($6\sqrt{3} \times 6\sqrt{3}$)R30° electron diffraction pattern was observed. A well-reconstructed 1×1 surface with a 12 Å thick (about four layers) graphene layer was found after further annealing at 1500 K. The average thickness of the EG was consistent with the monolayer EG, as confirmed by the attenuation of the Si 2p

Received: October 9, 2012

Revised: January 13, 2013

Published: January 17, 2013

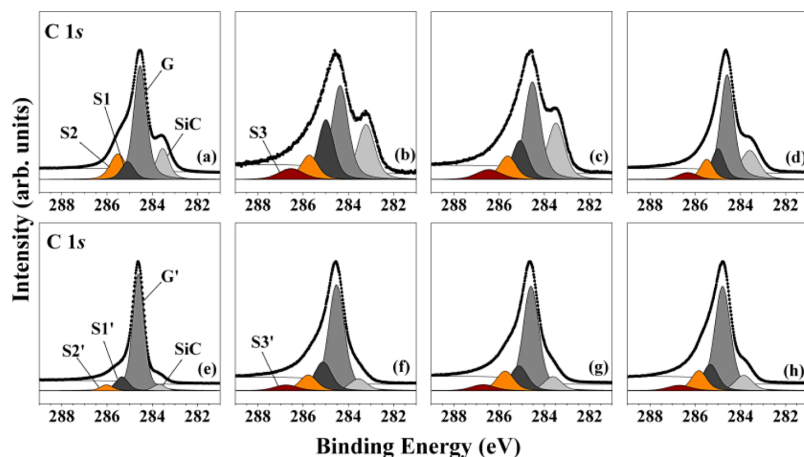


Figure 1. C 1s CLPES spectra measured at photon energy of 500 eV for clean monolayer EG (a) and multilayer EG (e). (b,f) After the nitrogen ion irradiation with the energy of 100 eV on panels a and e. Panels c (and g) and d (and h) are after annealing of panel b (and f) at the temperature of 750 K and at the temperature of 1300 K for 2 min, respectively.

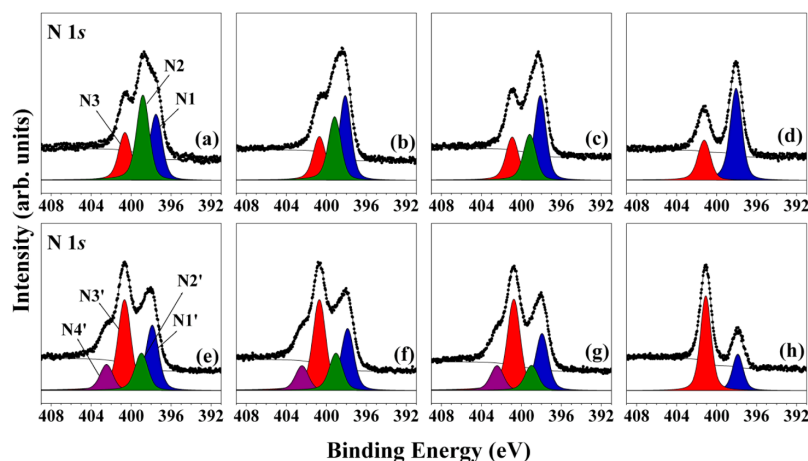


Figure 2. N 1s CLPES spectra measured at photon energy of 500 eV. (a) Monolayer EG and (e) multilayer EG after irradiation with 100 eV nitrogen ions. N 1s CLPES spectra acquired after annealing at 750 K (b) (and (f)), 1000 K (c) (and (g)), and 1300 K (d) (and (h)) for 2 min, respectively.

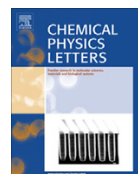
signal and low-energy electron diffraction (LEED) measurements.^{11,24,25}

The graphene/SiC surface was irradiated with 100 eV energetic nitrogen ions for 10 min at 300 K in a UHV chamber. Nitrogen ions (N_2^+) were generated by an ion gun (PHP 3000), and the sample current was measured to be $0.4 \mu\text{A}/\text{cm}^2$ during ion beam irradiation. The estimated number of nitrogen ions was 2.5×10^{12} ions/ $\text{cm}^2\cdot\text{sec}$. The effects of the nitrogen ion irradiation on the graphene/SiC surface properties were investigated using core-level photoelectron spectroscopy (CLPES). The C 1s and N 1s core-level spectra were measured taken before and after the nitrogen ion irradiation at photon energy of 500 eV with a total resolution of 200 meV at the 8A2 HRPES Beamline at the Pohang Accelerator Laboratory (PAL). The binding energy (BE) and the spectral resolutions of the N 1s core-level were calibrated by measuring the Au $4f_{7/2}$ core-level. The surface properties immediately after nitrogen ion irradiation have been previously described. Secondary electron

emission spectra (-20 V sample bias) and valence band spectra were measured at photon energies of 130 eV. All spectra were recorded in the normal emission mode. The photoemission spectra were carefully analyzed using a standard nonlinear least-squares fitting procedure with Voigt functions.^{26,27}

RESULTS AND DISCUSSION

We focused on describing the changes of C 1s and N 1s core-level spectra according to the annealing temperature because Si 2p spectra were not shown the remarkable spectral changes depending on nitrogen ion irradiation and annealing effect. Figure 1 shows the C 1s core-level spectra obtained before and after nitrogen ion irradiation of the monolayer or multilayer EG as a function of annealing temperature. As shown in Figure 1a (monolayer EG) and Figure 1e (multilayer EG), four distinct peaks were observed at 285.5, 285.1, 284.7, and 283.6 eV, which were well-assigned, respectively, to sp^2 carbon atoms at



The adsorption selectivity of the functional groups of dithiothreitol and 1,4-butanedithiol 2,3-diamino on the Ge(100) surface

Myungjin Lee, Youngchan Park, Hangil Lee*

Department of Chemistry, Sookmyung Women's University, Seoul 140-742, Republic of Korea

ARTICLE INFO

Article history:

Received 8 November 2012

In final form 4 March 2013

Available online 13 March 2013

ABSTRACT

The adsorption selectivities between dithiothreitol (DTT) and 1,4-butanedithiol 2,3-diamino (BDD) were revealed through density functional theory (DFT) calculations. DFT calculations predict that DTT does not stably adsorb onto the Ge(100) surface because both hydroxyl and thiol group in DTT serve as similar nucleophilic groups, therefore they are not selective. However, BDD is predicted to adsorb stably onto the Ge(100) surface because the amine and thiol groups in BDD operate relatively different nucleophilic group, and it is selective. The adsorption energies of BDD onto the Ge(100) surface suggested that an S–H dissociated N2–H dissociation bonded structure is the most stable structure.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

Recent interest in biochip and biosensor applications has motivated studies of the adsorption structures and reaction pathways associated with bioorganic molecular adsorption onto semiconductor surfaces [1–9]. A molecular-level understanding of the selectivity and stability of interactions between biologically relevant organic molecules and semiconductor surfaces is important for bio-related applications such as biochip and biosensor [10].

For several years, our group has studied the Ge(100) surface adsorption properties of biologically relevant organic molecules, such as amino acids, in an effort to characterize the selectivity and stability properties of the adsorption structures [11,12]. From among the various commercially relevant semiconductors, we selected the Ge(100) surface for examination because its superior electronic properties; zwitter ionic character [13–16]. The Ge(100)-2 × 1 reconstructed surface provide an excellent substrate for a variety of adsorption reactions involving multi-functional molecules because the zwitterionic dimers can act as Lewis bases and acid [17–19]. The zwitter ionic characteristics of the Ge(100)-2 × 1 surface have been also demonstrated in measurements of the selectivity and stability properties associated with the adsorption of a variety of functional groups that are commonly to amino acids [20].

The adsorption properties of molecules with sulfur functional groups have been extensively studied for their relevance to a variety of chemical processes, including polymer cross-linking, wettability, corrosion, and microelectronic fabrication [21]. Thiol groups are useful in photo-induced detectors because their reactions may

be induced upon photo-excitation with light of wavelength at 510 nm [22,23]. With this application in mind, we examined the adsorption properties of sulfur-containing compounds on the Ge(100) surface using DTT (which includes hydroxyl and thiol functional groups: see Scheme 1(a)) and BDD (which includes amine and thiol functional groups: see Scheme 1(b)) as test molecules. Hydroxyl and thiol functional groups generally serve as similar nucleophilic groups, whereas amine groups generally operate relatively different nucleophilic groups. The multi-functional DTT and BDD molecules can adsorb to the Ge(100) surface through different modes.

Here, the relative stabilities of these modes were examined, along with the corresponding adsorption transition states and reaction pathways, for all possible adsorption structures using DFT calculations that considered the relatively nucleophilicity or electrophilicity of the molecules and the Ge(100) surface [24,25].

This study using DFT calculation methods will serve the important information for the relative adsorption selectivity due to the properties of functional groups. Here, we will demonstrate the adsorption selectivities as we perform the model study of DTT and BDD upon adsorption onto the zwitterionic Ge(100)-2 × 1 surface.

2. Experimental section

Experimental DFT calculations were conducted to predict the energetics of the reaction pathways and the geometrically optimized structures of BDD and DTT derivatives adsorbed onto the Ge(100) surface. All DFT adsorption energy calculations were carried out using the JAGUAR 9.1 software package employing a hybrid density functional method that included the Becke's three-parameter nonlocal exchange functional with the correlation functional of Lee–Yang–Parr (B3LYP). These calculations considered six

* Corresponding author. Fax: +82 2 2077 7321.

E-mail addresses: mopoyo@sookmyung.ac.kr (M. Lee), easyscan@sookmyung.ac.kr (H. Lee).

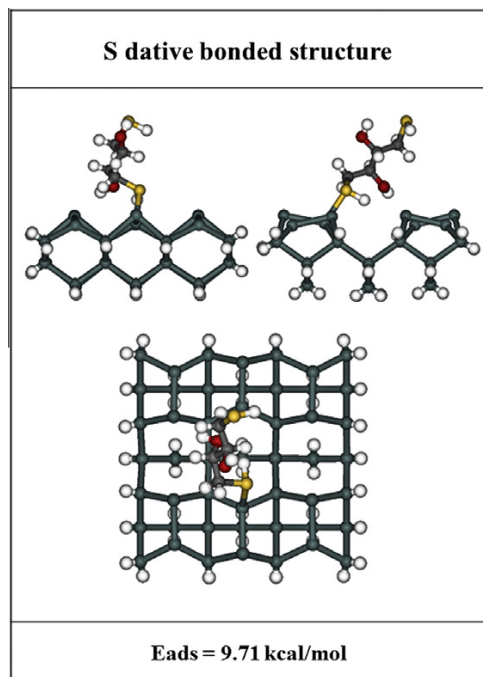


Figure 1. S dative bonded structure of DTT which shows positive adsorption energy value.

dimer ($\text{Ge}_{49}\text{H}_{40}$) cluster models. The geometries corresponding to the important local minima on the potential energy surface were determined at the B3LYP/LACVP** level of theory [26–28]. The LACVP** basis set is a mixed basis set that uses the LACVP basis set to de-

scribe Ge atoms and the 6-31G basis set to describe the remaining atoms. The LACVP basis set describes atoms beyond Ar in the periodic table using the Los Alamos effective core potentials developed by **Hay and Wadt [29,30]. For each cluster, optimization was performed by fixing the bottom two layers of the Ge atoms in ideal Ge crystal positions while allowing the top layer of the Ge atoms (including the dimer atoms) and the atoms of the chemisorbed adsorbate to relax. The geometries of important local minima and the transition states on each energy surface were calculated. Optimization of the local minima and transition states was confirmed using the same basis sets [31–34].

3. Results and discussion

We first calculated DTT adsorbed on the Ge(100) surface system using DFT calculations to identify a stable adsorption structure. DTT has two kinds of pathway to absorb on the Ge(100) surface: S dative bonded structure first pathway and starting from O dative bonded structure pathway. At first, we calculated an initial structure for starting from S dative bonded structure pathway (see Figure 1). However, S dative bonded structure has positive energy value (+9.71 kcal/mol) indicating that O dative bonded structure should be a starter for DTT absorption pathway. O dative bonded structure first pathway has a variety different final adsorption structures. It can proceed from O dative bonded structure to O–H dissociated O–H dissociation bonded structure, O–H dissociated S1–H dissociation bonded structure (on the thiol group next to the dissociated the O–H group) and O–H dissociated S2–H dissociation bonded structure (on the thiol group next to the undissociated the O–H group).

First of all, we calculated O–H dissociated O–H dissociation bonded structure for DTT adsorbed on the Ge(100) surface. The adsorption energy of this structure was substantially reasonable considering the adsorption energy value of –56.69 kcal/mol. However, this would not reasonable due to high energy of transition

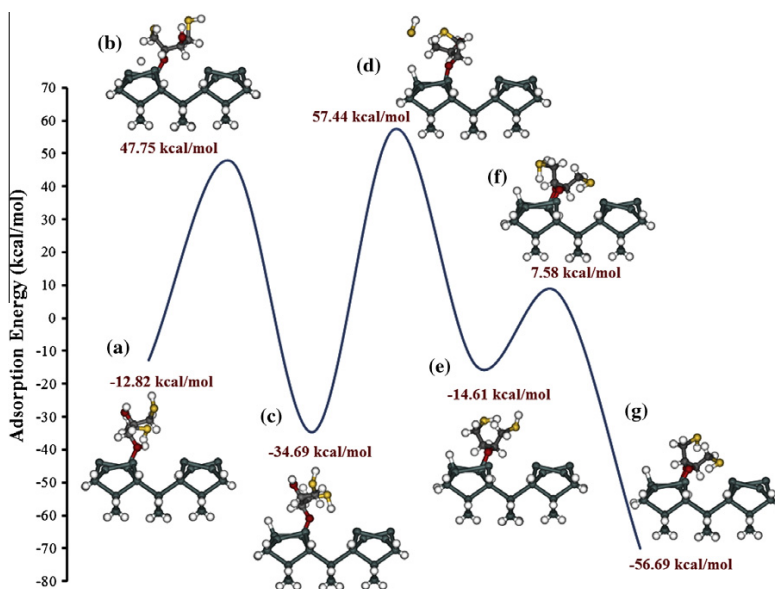


Figure 2. Calculated reaction pathways from the O dative bonded structure to the O–H dissociated O–H dissociation bonded structure, including three transition states, for DTT adsorbed onto the Ge(100) surface. (a) O dative bonded structure, (b) transition state from the O dative bonded structure to the O–H dissociated bonded structure, (c) O–H dissociated bonded structure, (d) transition state from the O–H dissociated bonded structure to the O–H dissociated O dative bonded structure, (e) O–H dissociated O dative bonded structure (f) transition state from the O–H dissociated O dative bonded structure to the O–H dissociated O–H dissociation bonded structure, and (g) O–H dissociated O–H dissociation bonded structure for DTT adsorbed onto the Ge(100) surface.

DOI: 10.1002/cphc.201300124

Adsorption Sequence of Multifunctional Groups: A Study on the Reaction Pathway and the Adsorption Structure of Homocysteine on the Ge(100) Surface

Youngchan Park, Myungjin Lee, and Hangil Lee^{*[a]}

We investigated the adsorption mechanism of homocysteine ($\text{HS-CH}_2\text{-CH}_2\text{-CH(NH}_2\text{)-COOH}$) on the Ge(100) surface along with its electronic structures and adsorption geometries to determine the sequence of adsorption of this amino acid's functional groups using core-level photoemission spectroscopy (CLPES) in conjunction with density functional theory (DFT) calculations. We found that the "SH-dissociated OH-dissociated N-dative-bonded structure" and the "SH-dissociated OH-dissocia-

tion-bonded structure" were preferred at a monolayer (ML) coverage of 0.30 (lower coverage) and 0.60 (higher coverage), respectively. The "SH-dissociated OH-dissociated N-dative-bonded structure" was the most stable structure. Moreover, we systematically confirmed the sequence of adsorption of the functional groups of the homocysteine molecule on the Ge(100) surface, which is thiol group ($-\text{SH}$), carboxyl group ($-\text{COOH}$), and amine group ($-\text{NH}_2$).

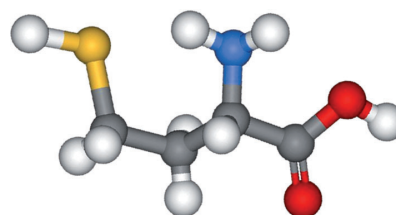
1. Introduction

Functional groups of biomolecules such as amino acids which interact with various surfaces have been intensively studied in the field of surface science to understand the stable adsorption configurations of biomolecules on, for example, semiconductor or metal surfaces.^[1–6] Potential applications in nano-biotechnology and biomedical sciences require an understanding of the adsorption process at the molecular level.^[7,8] As amino acids contain two or more functional groups, such as carboxyl groups, amine groups, thiol groups, and other groups that are present in side chains, a variety of adsorption structures are possible for each molecular unit. Adsorbed on a semiconductor surface, such as the Ge(100) or Si(100) surfaces, the atomic and electronic nature of surface dimer atoms differentiates plausible adsorption configurations. Compared to the bulk the Ge(100) surface is built up by pairs of atoms with weaker π bonds between them. Thus, these Ge dimers are buckled away from a symmetric configuration at room temperature. The tilted dimers produce an unequal distribution of charge, resulting in zwitterion-like properties.^[9,10] The "up"-Ge atom is an electron-rich nucleophile, and the "down"-Ge atom is an electron-poor electrophile. For instance, if the functional group of the landing molecules has lone-pair electrons, it can easily form a dative bond by donating these lone-pair electrons to the electron-deficient down atom of a surface dimer.^[11]

In other words, elucidating the behavior of biomolecules at a molecular level at an interface with a solid surface is desirable because such interfacial phenomena lie at the heart of bio-

logical interactions and, hence, have implications for industrial applications such as biochips or biosensors. Especially, bio-thiol groups can play crucial roles in maintaining biological systems as they are involved in numerous important functions in metabolism and homeostasis.^[12–14]

Herein, we focus on homocysteine ($\text{HS-CH}_2\text{-CH}_2\text{-CH(NH}_2\text{)-COOH}$, Scheme 1) adsorbed onto the Ge(100) surface. The electronic structures of homocysteine on a Ge(100) surface were



Scheme 1. Structure of the homocysteine molecule. The gray, blue, red, yellow, and white colored balls indicate carbon (C), nitrogen (N), oxygen (O), sulfur (S), and hydrogen (H), respectively.

characterized experimentally by using core-level photoemission spectroscopy (CLPES), and the stable adsorption geometries were modeled using density functional theory (DFT) calculations. Moreover, we performed DFT calculations to study the mechanism of the homocysteine adsorption pathway in regard to the adsorption sequence of the functional groups. This work provides a better understanding for the adsorption behavior of biomolecules and furthermore, it may contribute to the laying of a foundation for the fabrication of new biosensors.

To our knowledge, to date no systematic experimental and theoretical studies have been performed to examine the ad-

[a] Y. Park, M. Lee, Prof. H. Lee
Department of Chemistry
Sookmyung Women's University
Seoul 140-742 (Republic of Korea)
E-mail: easyscan@sookmyung.ac.kr

Supporting information for this article is available on the WWW under <http://dx.doi.org/10.1002/cphc.201300124>.

sorption structure and reaction pathway of homocysteine on the Ge(100) surface.

2. Results and Discussion

First we obtained core-level spectra for homocysteine adsorbed on a Ge(100) surface for two different coverage levels, 0.30 and 0.60 monolayers (ML), which show distinctly different adsorption structures. Our goal was the understanding of the electronic structures of homocysteine adsorbed on the Ge(100) surface by using CLPES. Figure 1 shows a series of C 1s, S 2p, N 1s, and O 1s core-level spectra as a function of coverage. We first acquired the C 1s core-level spectra of a Ge(100) surface covered by 0.30 ML homocysteine. As shown in this figure, four chemically distinct types of carbon are present. Considering the relationship between electronegativity and binding energy, we assigned the bonding features C1, C2, C3, and C4 to $-\text{CH}_2-\text{CH}_2(\text{CH})$ (284.4 eV), $-\text{CH}_2-\text{SH}$ (285.2 eV), $-\text{CH}-\text{NH}_2^+$ (285.9 eV), and $-\text{COO}^-$ (288.7 eV), respectively. These results

show that the homocysteine molecules are adsorbed on the surface at this coverage level, as the four distinct peaks indicate distinguishable carbon environments (see Figure 1a). Increasing the coverage of homocysteine to 0.60 ML (Figure 1e) produces different features from those acquired at 0.30 ML. A new peak (marked C3') emerges at 286.2 eV, indicating a change in bonding due to the formation of a new adsorption structure. This result is similar to those obtained for the cysteine system.^[15] This observation suggests that the electronic state (or electronic charge) of the amine group changes from positive (C3 and N1) to neutral (C3' and N2). If the electronic charge of the amine group becomes neutral upon formation of the adsorption structure, the corresponding charge of the adjacent α -carbon becomes more positive because of the reduced electron-withdrawing ability of the neutral nitrogen atom. Therefore, the peak is expected to shift to higher binding energies.^[16–18]

Figure 1b,f shows the analysis of the S 2p core-level spectra which were obtained at the same two distinct coverage levels. In these figures, the spectrum is characterized by a doublet structure due to the presence of the $\text{S}2\text{p}_{3/2}$ and $\text{S}2\text{p}_{1/2}$ peaks. All spectra could be fitted by using a 2:1 peak-area ratio, and a 1.2 eV spin-orbit splitting and a full-width-at-half maximum (FWHM) of 0.60 eV were observed.^[19] At a coverage of 0.30 ML, a single S peak is observed at the $\text{S}2\text{p}_{3/2}$ binding energy of 161.7 eV. This binding energy for the S1 bonding feature is consistent with the formation of a thiol ($-\text{SH}$) by S–H bond scission.^[20,21] Hence, we speculated that the homocysteine molecules had adsorbed onto the Ge(100) surface by forming a S–Ge bond. At 0.60 ML, the observed features differ from those described at 0.30 ML. As shown in Figure 1f, a new peak emerges (marked as S2) at 162.7 eV. Considering the binding energy of S2, we concluded that this peak also represents a type of thiol. Even though this peak is shifted to a higher binding energy, it is still in the range which is typical for thiols. This result is similar to that of methionine adsorbed on a Ge(100) surface system.^[22]

Next, the changes in the N 1s core-level spectra at the two coverage levels were determined using the same analysis procedure described above. For 0.30 ML homocysteine (Figure 1c) a single N 1s peak is recorded at 400.1 eV, which corresponds to the charged N atom ($\text{C}-\text{NH}_2^+$: dative-bonded amine) in the homocysteine molecule. Peak assignments of previous studies involving N-dative-bonded adsorption structures on Ge(100) surfaces clearly confirm our conclusions.^[15–17] Figure 1f shows the N 1s core-level spectrum obtained at a coverage of 0.60 ML. A new peak emerges at 398.7 eV (marked N2), providing critical evidence that a new adsorption structure is present at this coverage level. As mentioned in the interpretation of the C 1s spectra, the new N 1s peak (N2) arises from the electronic changes of the amine group (from positive to neutral) with the increase in the homocysteine coverage level, and it appears that the nitrogen atom is breaking its initial dative bonding to the Ge(100) surface.

Finally, we examined the changes in the O 1s peaks as a function of exposure to homocysteine molecules. Figure 1d shows the O 1s core-level spectrum obtained after deposition of

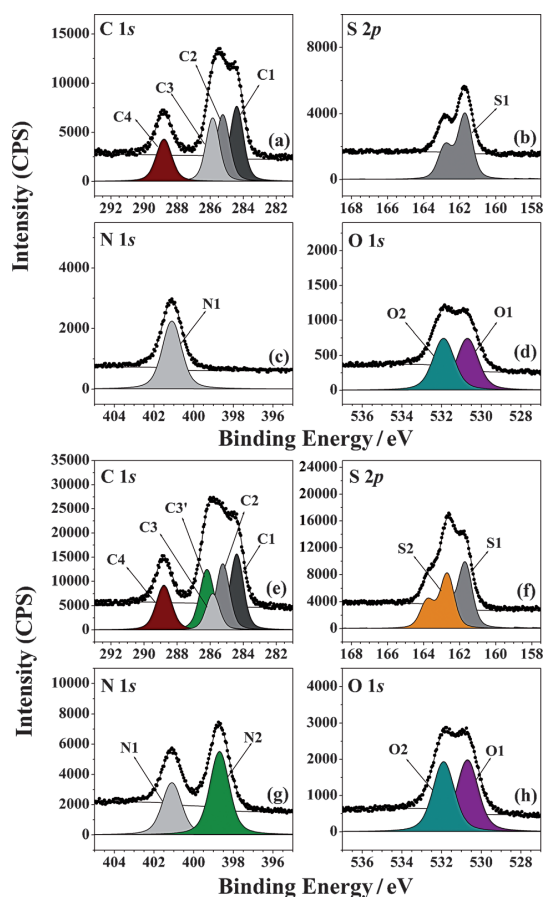


Figure 1. a,e) C 1s, b,f) S 2p, c,g) N 1s, and d,h) O 1s core-level spectra for homocysteine adsorbed at coverages of 0.30 ML (a–d) and 0.60 ML (e–h) on the Ge(100) surface. The dots indicate experimental values, whereas the solid lines represent the results of peak fittings.



Variation of adsorption geometries by the influence of nucleophilicity among p-CPA, p-TPA, and p-NPA on Ge(100)



Heeseon Lim^a, Sena Yang^a, Myungjin Lee^b, Sehun Kim^{a,*}, Hangil Lee^{b,*}

^a Molecular-Level Interface Research Center, Department of Chemistry, KAIST, Daejeon 305-701, Republic of Korea

^b Department of Chemistry, Sookmyung Women's University, Seoul 140-742, Republic of Korea

ARTICLE INFO

Article history:

Received 10 May 2013

In final form 4 June 2013

Available online 13 June 2013

ABSTRACT

The geometric configurations among phenylalanine derivatives (p-chlorophenylalanine (p-CPA) thiophenylalanine (p-TPA), and p-nitrophenylalanine (p-NPA)) adsorbed on Ge(100) surfaces were investigated using density functional theory (DFT) calculations. We focused on describing the role of nucleophilic group (–Cl, –SH, and –NO₂) being included in p-CPA, p-TPA, and p-NPA molecules, which can affect the stable adsorption structure energies and geometric configurations when they adsorb on the Ge(100) surface. We confirmed that geometric differences in the adsorption configurations indicated that their phenyl rings were tilted with respect to the Ge(100) surface by the effect of nucleophilicity and different sizes among –Cl, –SH and –NO₂ groups.

© 2013 Published by Elsevier B.V.

1. Introduction

With the growing demand for semiconductor surface based industrial applications, a study of bio-molecular adsorption structures on semiconductor surface has become immensely important because of their potential in the development of bio-chips, bio-sensors, and microelectronic devices [1–9]. Especially, the adsorption of multifunctional molecules such as amino acids on semiconductor surface is an useful and interesting topic to study chemical modifications and molecular adsorption structures [10,11]. In addition, various functional groups being included in bio-molecules can play a key role in adsorption structures and semiconductor functionalization due to their properties such as a nucleophilicity. Nucleophilicity is a significant property to describe the electronic structure of atoms during the reaction between molecule and semiconductor surface and to determine the most stable adsorption structure on semiconductor surface depending on how well it interacts with surface [12–14].

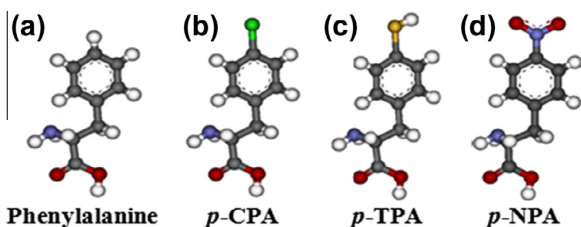
However, there has been little research about other molecules except for amino acid yet. Based on the results from the previous research into adsorption configuration of various amino acids, we examined the adsorption of other molecules to understand the extended trend of other molecules being included in diverse functional groups. Thus, we can expect that our theoretical studies of various molecules as well as amino acid are anticipated as future work for application such as biosensor and biochip fabrication.

Although many kinds of semiconductor surfaces have been studied, we focused on the Ge(100)-2 (1 surface among various semiconductor surface as a semiconductor substrate due to its unique surface features of zwitterionic character leading to the observation of a Lewis acid/base reactions at room temperature [2,15–17]. Therefore, the Ge(100)-2 × 1 surface is an important substrate for this bio-molecular adsorption research.

Our group recently examined the electronic configurations and adsorption structures of various amino acids adsorbed onto Ge(100)-2 × 1 surfaces including phenylalanine and tyrosine [18]. Based on the result from our previous research into adsorption configuration of various amino acids and amino acid derivatives on the Ge(100)-2 × 1 surface, we can easily expect that the 'O–H dissociated-N dative bonded structure' is the most stable structure at lower coverage [11,19]. However, the adsorption geometry of molecules on the Ge(100)-2 × 1 surface can be different depending on the functional groups within the molecules. For example, a recent study of adsorption geometry for phenylalanine and tyrosine on the Ge(100)-2 × 1 surface reveals that there is a significant difference due to influence of the hydroxyl group (nucleophilic group) only existing in tyrosine, although the molecular structure between phenylalanine (see Scheme 1a) and tyrosine is very similar [18]. Therefore, in order to verify the role of the nucleophilic groups in bio-molecules, the relative stabilities of the adsorption products were examined by performing DFT calculations of phenylalanine derivatives such as p-chlorophenylalanine (p-CPA) including –Cl group (see Scheme 1b), p-thiophenylalanine (p-TPA) including –SH group (see Scheme 1c), and p-nitrophenylalanine (p-NPA) including –NO₂ group (see Scheme 1d), on the Ge(100) surface.

* Corresponding authors. Fax: +82 2 2077 7321 (H. Lee).

E-mail addresses: lim2628@kaist.ac.kr (H. Lim), sehun-kim@kaist.ac.kr (S. Kim), easyscan@sookmyung.ac.kr (H. Lee).



Scheme 1. Schematic diagrams showing the models of the chemical structures of (a) phenylalanine, (b) p-chlorophenylalanine (p-CPA), (c) p-thiophenylalanine (p-TPA), and (d) p-nitrophenylalanine (p-NPA). The gray, blue, red, white, green, and yellow colored balls indicate carbon (C), nitrogen (N), oxygen (O), hydrogen (H), chlorine (Cl), and sulfur (S), respectively.

In this paper, we will systematically demonstrate the effect of nucleophilic groups being included in p-chlorophenylalanine (p-CPA) (see Scheme 1b), p-thiophenylalanine (p-TPA) (see Scheme 1c), and p-nitrophenylalanine (p-NPA) (see Scheme 1d) as we compare the adsorption structures and geometric configurations among p-CPA, p-TPA and p-NPA with those of phenylalanine (see Scheme 1a), and thus we will confirm the correlation between the effect of nucleophilic group and the variation of geometric configuration.

2. Computational methods

DFT calculations were conducted to predict the energetic of the reaction pathways and geometrically optimized structures of the phenylalanine derivatives adsorbed onto Ge(100) surfaces. All DFT adsorption energy calculations were carried out using the JAGUAR 9.1 software package employing a hybrid density functional method that included the Becke's three-parameter nonlocal exchange functional with the correlation functional of Lee–Yang–Parr (B3LYP) [20]. These calculations considered four dimer ($\text{Ge}_{35}\text{H}_{32}$) cluster models. The geometries corresponding to the important local minima on the potential energy surface were determined at the B3LYP/LACVP** level of theory. The LACVP** basis set is a mixed basis set that uses the LACVP basis set to describe Ge atoms and the 6-31G basis set to describe the remaining atoms. The LACVP basis set describes atoms beyond Ar in the periodic table using the Los Alamos effective core potentials developed by Hay and Wadt [21]. For each cluster, optimization was performed by fixing the bottom two layers of the Ge atoms in the ideal Ge crystal positions while allowing the top layer of the Ge atoms (including the dimer atoms) and the atoms of the chemisorbed adsorbate to relax. The geometries of important local minima and transition states on each energy surface were calculated. Local minima and transition states were confirmed to be optimized using the same basis sets [22].

3. Results and discussion

DFT calculation results for the adsorption structures of three molecules with different nucleophilic groups, p-chlorophenylalanine (p-CPA (fencloine): see Scheme 1b), p-thiophenylalanine (p-TPA: see Scheme 1c), and p-nitrophenylalanine (p-NPA: see Scheme 1d), on the Ge(100) surface were performed to examine the influence of the nucleophilic functional groups on the adsorption geometry. According to previous study relating to adsorption of amino acids on the Ge(100) surface, they can have many possible adsorption structures. Among their possible structures, the most stable adsorption structure of these possible structures has been considered as the 'O–H dissociated–N dative bonded structure' [18].

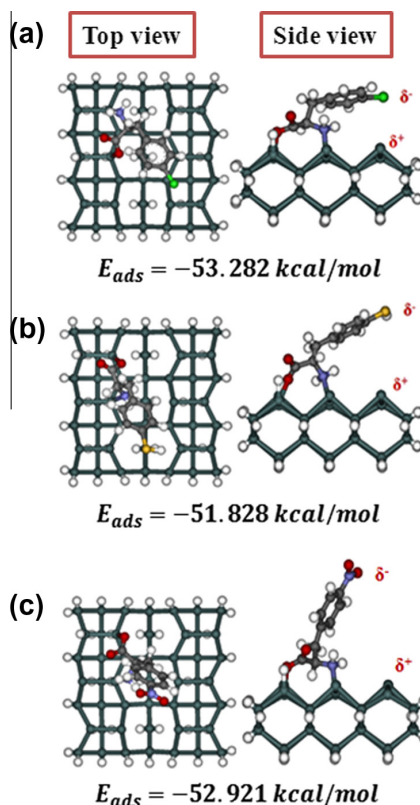


Figure 1. (a)–(c) show the 'O–H dissociated–N dative bonded structure' of p-chlorophenylalanine (p-CPA), p-thiophenylalanine (p-TPA), and p-nitrophenylalanine (p-NPA) on Ge(100) surfaces, predicted using DFT calculations, respectively.

For this reason, we expected that p-CPA, p-TPA, and p-NPA molecules have the same adsorption configuration with other molecules. To obtain more clear information, we performed DFT calculations for these molecules. Figure 1 shows DFT results about the geometrical adsorption structures of three phenylalanine derivative molecules on the Ge(100) surface, respectively.

The adsorption energy values of p-CPA, p-TPA, and p-NPA on the Ge(100) surface are shown in Figure 1. As shown in Figure 1, the adsorption energies of p-CPA, p-TPA, and p-NPA are -53.282 kcal/mol, -51.828 kcal/mol, and -52.921 kcal/mol, respectively. Thus, the 'O–H dissociated–N dative bonded structure' of p-CPA, p-TPA, and p-NPA on the Ge(100) surface is also quite reasonable configuration. Interestingly, we can find that the adsorption configurations of p-CPA, p-TPA, and p-NPA on the Ge(100) surface are tilted with respect to the Ge(100) surface, as determined by the corresponding DFT calculations. We suggested that this discrepancy was attributed to the nucleophilic groups such as $-\text{Cl}$, $-\text{SH}$, and $-\text{NO}_2$ group, respectively. In general, $-\text{Cl}$, $-\text{SH}$, and $-\text{NO}_2$ are highly nucleophilic. The nucleophilicity indices of $-\text{Cl}$, $-\text{SH}$ are 0.042 eV, 0.145 eV, respectively [23]. Also, the nucleophilicity indices of $\text{C}_6\text{H}_5-\text{Cl}$ and $\text{C}_6\text{H}_5-\text{NO}_2$ are 2.41 eV, 1.53 eV, respectively [24]. Therefore, we can clearly expect that the nucleophilicity trend of the $-\text{Cl}$, $-\text{SH}$, and $-\text{NO}_2$ groups is followings; $-\text{SH} > -\text{Cl} > -\text{NO}_2$. In addition, we also compared the adsorption configurations among p-CPA, p-TPA, and p-NPA with phenylalanine to investigate the role of nucleophilic groups.

Since we confirmed that the most stable adsorption structures in three cases are 'O–H dissociated–N dative bonded structures', we next track to model the reaction pathways for its structure

Confirmation of the coexistence of two tautomers of 2-mercaptothiazoline on the Ge(100) surface†

Cite this: *Phys. Chem. Chem. Phys.*, 2013, **15**, 16594

Youngchan Park,^a Sena Yang,^b Myungjin Lee,^a Heeseon Lim,^b Yaewon Kim,^a Sehun Kim^{*b} and Hangil Lee^{*a}

We confirmed the coverage dependent variation of tautomers of 2-mercaptothiazoline (the thiolate and thione forms) adsorbed on the Ge(100) surface under UHV conditions by using HRXPS measurements in conjunction with the DFT calculation method, which was studied before only in aqueous systems. The C 1s, S 2p, and N 1s core-level spectra obtained using HRXPS revealed the simultaneous presence of two distinct adsorption structures in different proportions at both low (0.15 ML) and high (0.65 ML) coverages. Moreover, we modelled the adsorption structures and geometric configurations of the bond states of 2-mercaptothiazoline on the Ge(100) surface by using the DFT calculation method, and found that the S dative bonded structure is the most stable adsorption structure for the thione form of 2-mercaptothiazoline and that the S–H dissociated–N dative bonded structure is the most stable adsorption structure for the thiolate form.

Received 2nd July 2013,
Accepted 31st July 2013

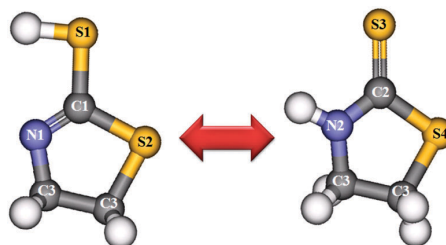
DOI: 10.1039/c3cp52760c

www.rsc.org/pccp

1 Introduction

In recent years, studies of the interactions between the functional groups of organic molecules and semiconductor-based materials have been carried out by many researchers because of their potential for various applications such as biosensors and biochips.^{1–5} The functionalization of organic molecules containing heteroatoms such as sulfur, oxygen, and nitrogen can provide an in-depth understanding of molecular level behavior at an interface with a solid surface. In particular, functionalized sulfur-containing compounds have been widely studied for their various applications such as cross-linkers of polymers in biological events, effective corrosion inhibitors, and in micro-electronic fabrication.^{6–10} As a diffusion inhibitor and organic corrosive molecule, 2-mercaptothiazoline (see Scheme 1) is commonly used because it contains nitrogen and sulfur atoms that easily donate electrons, which results in its adsorption on the surfaces.^{11–13}

In this study, we observed the interactions between the hetero-organic molecule, 2-mercaptothiazoline, and the Ge(100)-2 × 1 semiconductor surface. There are weak π -bonds between pairs of surface dimer atoms on the Ge(100) surfaces, so Ge dimers



Scheme 1 The structures of the two tautomers of 2-mercaptothiazoline: the thiol form (left) and the thione form (right). The yellow, blue, gray, and white balls indicate sulfur (S), nitrogen (N), carbon (C), and hydrogen (H), respectively.

are buckled away from the symmetric configuration at room temperature. The tilted dimers produce an unequal distribution of charge, resulting in zwitterion-like properties.^{14–16} These properties provide a framework for understanding the adsorption reactions of several functional groups on a semiconductor surface.

We begin by discussing the adsorption structures and geometric configurations of 2-mercaptothiazoline on the Ge(100) surface. There are several possible binding sites for 2-mercaptothiazoline on the Ge(100) surface because it contains two sulfur atoms and one nitrogen atom in its heterocyclic ring, which can act as functional groups. Moreover, 2-mercaptothiazoline exists in two tautomeric forms, the thiol form and the thione form. In the thione form, the nitrogen atom is protonated, whereas one sulfur atom in the thiol form is protonated and the

^a Department of Chemistry, Sookmyung Women's University, Seoul 140-742, Republic of Korea. E-mail: easyscan@sookmyung.ac.kr; Fax: +82-2-2077-7321; Tel: +82-2-710-9409

^b Department of Chemistry, KAIST, Guseong-dong, Yuseong-gu, Daejeon-si, Gyeongbuk, 790-784, Korea. E-mail: sehun-kim@kaist.ac.kr

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c3cp52760c

nitrogen atom is unprotonated. In other words, this tautomerization of the 2-mercaptothiazoline molecule highly affects its chemical reactivity, which can result in various adsorption structures on the Ge(100) surfaces.^{17–21} Many researchers have studied the tautomers in aqueous solution, particularly pH value induced structural changes.^{22–25}

Here, we focused on determining the coverage dependence of the tautomerization of 2-mercaptothiazoline on the Ge(100) surface and the stabilities of the adsorbed tautomers. Hence we confirmed the stable adsorption structures and geometric configurations experimentally by using high-resolution X-ray photoemission spectroscopy (HRXPS) and corroborated these results theoretically with density functional theory (DFT) calculations.

To our knowledge, no previous study has investigated the adsorption of 2-mercaptothiazoline on the Ge(100) surface under UHV conditions.

2 Experimental and computational methods

The Ge(100) surfaces (p-type, $R = 0.10\text{--}0.39\ \Omega$) were cleaned by several cycles of Ar⁺ ion sputtering at 1 keV for 20 minutes at 700 K, followed by annealing at 900 K for 10 minutes. The cleanness of the Ge(100) surfaces was checked by performing low-energy electron diffraction (LEED) measurements. 2-Mercaptothiazoline (C₃H₃NS₂, 98% purity) was purchased from Aldrich and further purified with several sublimation and pumping cycles to remove dissolved gases prior to exposure to the Ge(100) surface.

The HRXPS measurements were conducted at the 10D beamline of the Pohang Accelerator Laboratory. The S 2p, C 1s, and N 1s core-level spectra were obtained with a PHOIBOS 150 electron energy analyzer equipped with a two-dimensional charge-coupled device (2D CCD) detector (Specs GmbH). Photon energies of 230, 340, and 460 eV were used to enhance the surface sensitivity. The binding energies of the three core-level spectra were calibrated with respect to that of the clean Au 4f core-level spectrum (84.0 eV) for the same photon energy. The base pressure of the chamber was maintained below 1.2×10^{-10} Torr. All spectra were recorded in the normal emission mode. The photoemission spectra were carefully analyzed by using a standard nonlinear least squares fitting procedure with Voigt functions.²⁶ We have estimated the coverage based on the N 1s multi-layer peak. The peak of multi-layer structures of nitrogen-containing molecules such as amino acids is usually obtained at near 399 eV. 2-Mercaptothiazoline is deposited every 30 seconds until the multi-layer peak is observed. We consider the amount that we obtained around the time when we observed the multi-layer peak as 1.1 ML, considered as the standard amount.

DFT calculations were conducted to predict the energetics of the reaction pathways and the geometrically optimized structures of 2-mercaptothiazoline adsorbed on the Ge(100) surfaces. All the DFT adsorption energy calculations were carried out with the JAGUAR 9.1 software package by employing a hybrid density functional method that included Becke's three-parameter non-local exchange functional and the correlation functional of

Lee–Yang–Parr (B3LYP).²⁷ The geometries corresponding to the important local minima and transition states on the potential energy surface were determined at the B3LYP/LACVP** level of theory in order to maximize the accuracy on the chemically active electrons of the reactions while minimizing computational time. The LACVP** basis set is a mixed basis set that uses the LACVP basis set to describe Ge atoms which are heavier than Ar and the 6-31G basis set used for chemically active remaining atoms. For each cluster, optimization was performed by fixing the bottom two layers of the Ge atoms in ideal Ge crystal positions while allowing the top layer of Ge atoms (including the dimer atoms) and the atoms of the chemisorbed adsorbate to relax.

3 Results and discussion

First, we carried out HRXPS measurements of the carbon (C 1s), sulfur (S 2p), and nitrogen (N 1s) core-levels of 2-mercaptothiazoline to elucidate the coverage dependence of its adsorption structures when it adsorbed on the Ge(100) surface, as shown in Fig. 1. The top row shows the three core-level spectra obtained after deposition of 0.15 ML 2-mercaptothiazoline on the Ge(100) surface at room temperature, and the bottom row shows the spectra obtained after deposition of 0.65 ML 2-mercaptothiazoline on the Ge(100) surface under the same conditions. As shown in Fig. 1(a), the C 1s peak can be deconvoluted into three distinct peaks marked C1, C2, and C3. These peaks can be assigned to C=N (286.9 eV; in thiolate form), C–NH (285.9 eV; in thione form), and C in the heterocyclic ring (284.8 eV) by considering the differences in electronegativity between the carbon and nitrogen atoms and their peak intensities. The C3 peak due to the four carbon atoms in the heterocyclic ring could not be distinguished because they have similar environments and are not in direct contact with the Ge(100) surface.²¹ The nitrogen atom has greater electronegativity than the carbon atom, so the electron cloud of the C–N covalent bonds tends to move toward the N atom. Thus, as the number of

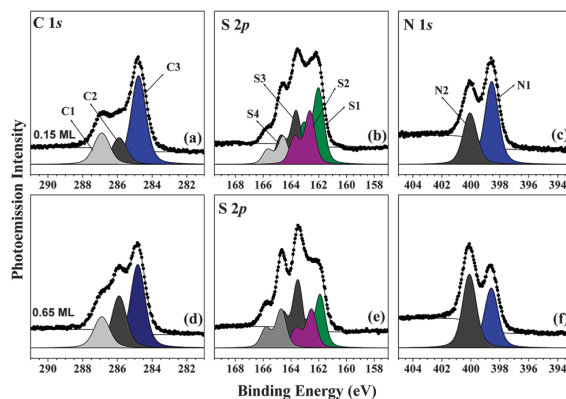


Fig. 1 Core-level spectra of the C 1s, S 2p, and N 1s as a function of 2-mercaptothiazoline coverages on the Ge(100) surfaces at room temperature. 0.15 ML (top row: lower coverage), 0.65 ML (bottom row: higher coverage).



OPEN

Quantum confinement-induced tunable exciton states in graphene oxide

Dongwook Lee¹, Jiwon Seo^{1,2}, Xi Zhu³, Jiyoul Lee⁴, Hyeon-Jin Shin⁴, Jacqueline M. Cole^{1,5,6}, Taeho Shin⁴, Jaichan Lee², Hangil Lee⁷ & Haibin Su^{3,8}

SUBJECT AREAS:

ELECTRONIC STRUCTURE

ELECTRONIC PROPERTIES
AND MATERIALS

NANOPARTICLES

CHARACTERIZATION AND
ANALYTICAL
TECHNIQUES

Received
25 April 2013

Accepted
4 July 2013

Published
22 July 2013

Correspondence and
requests for materials
should be addressed to
D.L.

(dongwooklee@324@
gmail.com)

¹Cavendish Laboratory, University of Cambridge, J. J. Thomson Avenue, Cambridge CB3 0HE, United Kingdom, ²School of Advanced Materials Science and Engineering, Sungkyunkwan University, Suwon 440-746, Republic of Korea, ³Division of Materials Science, School of Material Sciences and Engineering, Nanyang Technological University, 639798 Singapore, ⁴Samsung Advanced Institute of Technology Giheung-gu Yongin 446-712, South Korea, ⁵Department of Chemistry, University of New Brunswick, P. O. Box 4400, Fredericton, NB, E3B 5A3, Canada, ⁶Department of Physics, University of New Brunswick, P. O. Box 4400, Fredericton, NB, E3B 5A3, Canada, ⁷Department of Chemistry, Sookmyung Women's University, Seoul 140-742, South Korea, ⁸CINTRA CNRS/NTU/THALES, UMI 3288, Research Techno Plaza, 50 Nanyang Drive, Border X Block, Level 6, Singapore 637553.

Graphene oxide has recently been considered to be a potential replacement for cadmium-based quantum dots due to its expected high fluorescence. Although previously reported, the origin of the luminescence in graphene oxide is still controversial. Here, we report the presence of core/valence excitons in graphene-based materials, a basic ingredient for optical devices, induced by quantum confinement. Electron confinement in the unreacted graphitic regions of graphene oxide was probed by high resolution X-ray absorption near edge structure spectroscopy and first-principles calculations. Using experiments and simulations, we were able to tune the core/valence exciton energy by manipulating the size of graphitic regions through the degree of oxidation. The binding energy of an exciton in highly oxidized graphene oxide is similar to that in organic electroluminescent materials. These results open the possibility of graphene oxide-based optoelectronic device technology.

The discovery of graphene in 2004 ignited prolific research owing to its potential for diverse applications, such as high speed semiconductor chips, flexible devices, optoelectronics applications, and energy storage^{1–3}. Unlike graphene, which is a zero band-gap semiconductor, nano-graphene materials have a band gap. Accordingly, some of these materials exhibit optical emission due to quantum confinement effects⁴. In this way, nano-sized graphene and related materials have attracted much attention since they are expected to replace conventional cadmium-based quantum dots.

Among graphene-based materials, graphene oxide (GO) has stimulated a large body of research due to its luminescent properties. However, the origin of these properties is still controversial due to the complex structure of GO^{5–8}. Here, we report for the first time that core/valence excitonic states are induced by quantum confinement in the graphitic regions of GO. Furthermore, the exciton binding energies are tuned by the degree of oxidation, i.e. the size of sp² domains.

Results

Figure 1 shows transmission electron microscope (TEM) images of two GO samples (GO1 and GO2) with different degrees of oxidation. The O:C ratios of GO samples obtained by energy dispersive X-ray (EDX) analysis are 0.35 and 0.38 for GO1 and GO2, respectively. The GO images appear as dim fringes, which are nearly straight lines, owing to the presence of overlapping GO sheets. These fringes are readily observed in sample GO1 (Figure 1a). The stripes appearing in GO2 (Figure 1b) alternate black and white, similar to that found for graphene ripples whose edges are functionalized with chemical groups^{9,10}. The stripes represent flat regions that are wrinkled at their boundaries. The average stripe width in the image is 20 nm (5 stripes per 100 nm). The insets of Figure 1a and Figure 1b are the corresponding electron diffraction patterns of GO1 and GO2, respectively. These diffraction patterns can clearly be distinguished, implying that the crystalline phase is dominant. A small circular diffuse feature is also visible at the center of these diffraction patterns, revealing that GO also harbors amorphous regions. These TEM images therefore demonstrate that two spatially separated phases coexist in GO: a reacted amorphous region and an unreacted crystalline region. Figure 1c illustrates the schematic structure of GO based on these TEM images.

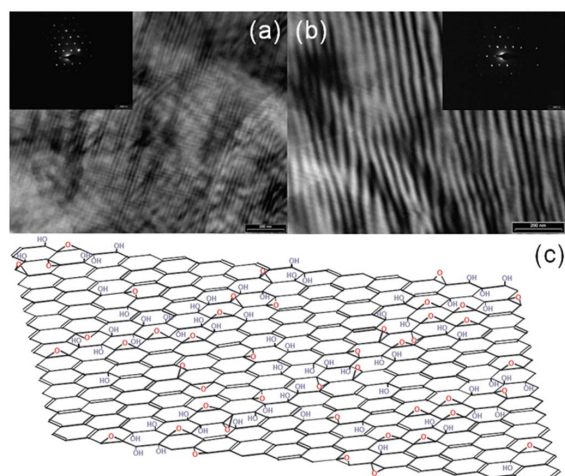


Figure 1 | TEM images of GO samples. (a) GO1 and (b) GO2. The scale bars of (a) and (b) are 200 nm. (c) Schematic structure of GO based on the TEM.

The unreacted regions are not perfectly flat, i.e. there are some functionalized chemical groups within such regions. The reacted regions act as electronic potential energy barriers between unreacted regions. Owing to the structural distinction of GO, as determined by its specific domain and oxidation characteristics, we surmise that each unreacted region acts like an individual graphene nanoribbon (GNR) or graphene nano-diamond separated by these boundary walls. If unreacted regions are properly separated by reacted regions that function as potential energy barriers, electrons in graphitic (unreacted) regions can be confined to generate electron-hole coupling, while new electronic structures can be induced by controlling the degree of oxidation in GO.

In order to confirm the presence of core excitons in GO, we performed C K-edge X-ray absorption near edge structure (XANES) measurements. Given the dielectric characteristics of GO, there was a need to circumvent potential charging problems in this XANES experiment. In order to avoid charging, the entrance to the electron detector was biased to a positive potential, effectively raising the potential of the sample surface above the ground potential. In this way, charge-compensating electrons could fill the vacant

states in GO more efficiently. The total electron yield (TEY) signal was collected as the bias voltage increased from 0 to +500 V. At 0 V bias, the peaks at 285 eV, 287 eV, and 291 eV were unclear due to charging effects. At +250 V, GO1 in Figure 2a shows four clear peaks at 285 eV, 287 eV, 288 eV, and 291 eV. The peak at 285 eV corresponds to the π^* antibonding state of C=C¹¹. The peak at 287 eV was assigned to the π^* antibonding of C-OH¹². The third peak at 288 eV may originate from the π^* antibonding of the epoxy group (C-O-C). The fourth peak at 291 eV denotes the σ^* state of C=C^{11,13}. At +500 V, the peak at 287 eV (π^*) becomes much sharper and increases in intensity. The shape and sharpness of this peak is typical of a core exciton, and is assigned as E_{ex} . The binding energy of the core exciton (E_B) was estimated by evaluating E_{CB} , the conduction band edge of GO1. E_{CB} was extracted by fitting the XANES spectrum with Lorentzian and arctangent functions, which describe the exciton state and the continuous step¹⁴, respectively (Figure 2b) such that $E_B = E_{CB} - E_{ex}$. This yielded an E_B of 500 meV for GO1. Generally, the binding energy of the core exciton is larger than that of the valence exciton. The binding energy found here is much larger than that found for excitons in Si and GaAs, corresponding to 14.7 meV and 4.2 meV, respectively. However, the binding energy is similar to that found in organic electroluminescent (EL) materials (e.g. polydiacetylene (PDA) 500 meV, polythiophene (PT) 500 meV, and Tris(8-hydroxyquinolato) aluminum (Alq3) 1400 meV)¹⁵.

The effective mass (μ) of the core exciton in GO1 with a binding energy was obtained by,

$$\mu = \frac{E_B}{R_H} * \epsilon_r^2 * m_0, \quad (1)$$

where R_H is the Rydberg energy (13.6 eV), ϵ_r is the dielectric constant of GO (ca. 2.31 at room temperature¹⁶), and m_0 is the electron rest mass. Thus, the effective mass of a core exciton in GO1 is $\mu = 0.196 m_0$. The Bohr radius of this core exciton (a_x) can also be calculated with the following equation,

$$a_x = \frac{\epsilon_r}{\mu} * a_H * m_0, \quad (2)$$

where a_H is the Bohr radius (0.529 Å). Equation (2) leads to an exciton Bohr radius (a_x) in GO1 of about 6.23 Å, implying that core excitons are localized in GO1.

In the same way as for GO1, the Bohr radius of the core exciton in GO2 was obtained. The Bohr radius, the effective mass, and the binding energy of these core excitons are presented in Table 1, together with the O:C ratios of GO samples. As the O:C ratio increases, the binding energy of the exciton increases, while the Bohr radius decreases. An increasing O:C ratio produces more chemical groups

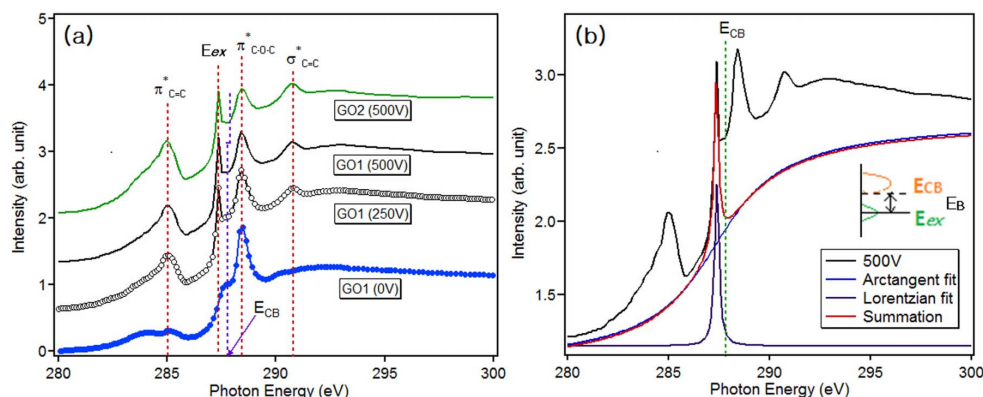


Figure 2 | XANES spectra of GOs. (a) High-resolution C K-edge XANES spectra of GO. (b) Fit of GO1 spectrum (500 V) with Lorentzian and Arctangent functions.

Article

Coverage Dependent Variation of the Adsorption Structure of 2-Thiophenecarboxaldehyde on the Ge(100)-2 × 1 Reconstructed Surface

Myungjin Lee, Minjeong Shin and Hangil Lee *

Department of Chemistry, Sookmyung Women's University, Seoul 140-742, Korea;

E-Mails: mopoyo@sookmyung.ac.kr (M.L.); smj5757@sookmyung.ac.kr (M.S.)

* Author to whom correspondence should be addressed; E-Mail: easyscan@sookmyung.ac.kr;
Tel.: +82-2-710-9409; Fax: +82-2-2077-7321.

Received: 22 July 2013; in revised form: 16 August 2013 / Accepted: 19 August 2013 /

Published: 26 August 2013

Abstract: High-resolution photoemission spectroscopy (HRPES) measurements were collected and density functional theory (DFT) calculations were performed to track the exposure-dependent variation of the adsorption structure of 2-thiophenecarboxaldehyde (C₄H₃SCHO: TPCA) on the Ge(100) 2 × 1 reconstructed surface at room temperature. In an effort to identify the most probable adsorption structures on the Ge(100)-2 × 1 reconstructed surface, we deposited TPCA molecules at low exposure and at high exposure and compared the differences between the electronic features measured using HRPES. The HRPES data suggested three possible adsorption structures of TPCA on the Ge(100)-2 × 1 reconstructed surface, and DFT calculations were used to determine the plausibility of these structures. HRPES analysis corroborated by DFT calculations, indicated that an S-dative bonded structure is the most probable adsorption structure at relatively low exposure levels, the [4 + 2] cycloadduct structure is the second most probable structure, and the [2 + 2]-C=O cycloadduct structure is the least probable structure on the Ge(100)-2 × 1 reconstructed surface at relatively high exposure levels.

Keywords: HRPES; DFT calculation; 2-thiophenecarboxaldehyde; Ge(100)-2 × 1 reconstructed surface; adsorption structure

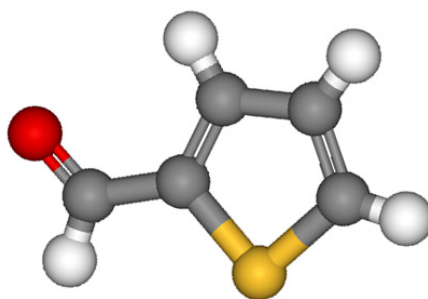
1. Introduction

Recently, many research groups have been studying structures of biomolecules adsorbed onto semiconductor surfaces in connection with the development of bio-related applications [1–5]. The Ge(100)- 2×1 reconstructed surface provides an excellent model of a semiconductor surface because the Ge(100)- 2×1 reconstructed surface displays zwitterionic character [6,7]. The zwitterionic dimers of the Ge(100)- 2×1 reconstructed surface act as Lewis bases and acids, which facilitate the adsorption of molecules with multifunctional groups [8–10].

Recent studies have shown that thiol groups react with Ge(100)- 2×1 reconstructed surface under photo-excitation with 510 nm light. This reaction is useful as a photo-induced thiol detector [11,12]. In terms of developing the other thiol detection applications, we have examined sulfur-containing compounds to understand how these molecules behave on the Ge(100)- 2×1 reconstructed surface at room temperature. 2-Thiophenecarboxaldehyde (TPCA) is a good test molecule as a detector because its molecular structure is relatively simple and sensitive.

Figure 1 shows the structure of a TPCA molecule containing a sulfur atom in an aromatic ring and an aldehyde functional group. Since Kurt Alder and Otto Diels first discovered the Diels–Alder reaction and were awarded the Nobel Prize in Chemistry in 1950, this reaction has been actively studied in a variety of chemical fields [13–18]. Contrary to what is known about the Diels–Alder reaction which occurs in aqueous state, TPCA is expected to form several cycloaddition structures under UHV conditions through reactions involving the TPCA diene and the dimer of the Ge(100)- 2×1 reconstructed surface atoms.

Figure 1. The structure of the TPCA molecule, which has a sulfur (yellow), an oxygen (red), five carbons (gray), and four hydrogens (white), respectively.



Previously, we reported that thiophene could adsorb onto the Ge(100)- 2×1 reconstructed surface. Hence, we expected that similar adsorption properties are possible for the TPCA molecule [19]. Thiophene adsorption proceeded via a Lewis acid–base reaction to form two distinct configurations on the Ge(100)- 2×1 reconstructed surface. At a low exposure (0.25 mL), thiophene formed one-dimensional (1-D) molecular chain structures in an S-dative bonded structure. Beyond exposure of 0.25 mL, the stable $[4 + 2]$ cycloaddition adducts were observed [20]. Here, we sought to compare the adsorption structures of thiophene and TPCA on the Ge(100)- 2×1 reconstructed surface using high-resolution photoemission spectroscopy (HRPES) and density functional theory (DFT) calculations.

NANO EXPRESS

Open Access

The aniline-to-azobenzene oxidation reaction on monolayer graphene or graphene oxide surfaces fabricated by benzoic acid

Myungjin Lee¹, Kijeong Kim² and Hangil Lee^{1*}

Abstract

The oxidation of aniline to azobenzene was conducted in the presence of either monolayer graphene (EG) or graphene-oxide-like surface, such as GOx, under ultra-high vacuum conditions maintaining a 365-nm UV light exposure to enhance the oxidation reaction. The surface-bound products were investigated using micro Raman spectroscopy, high-resolution photoemission spectroscopy, and work function measurements. The oxygen carriers present on the GOx surfaces, but not on the EG surfaces, acted as reaction reagents to facilitate the oxidation reaction from aniline to azobenzene. Increasing the aniline concentration at 300 K confirmed that the exchange ratio from the aniline to the azobenzene was enhanced, as determined by the intensity ratio between the aniline- and azobenzene-induced N 1 s core-level spectra. The work function changed dramatically as the aniline concentration increased, indicating that the aniline on the GOx surface conveyed *n*-type doping characteristics at a low coverage level. A higher aniline concentration increased the *p*-type doping character by increasing the azobenzene concentration on the GOx surface. A comparison of the oxidation reactivity of aniline molecules on the EG or GOx surfaces revealed the role of the oxygen carriers on the GOx surfaces in the context of catalytic oxidation.

Keywords: Oxidation reaction; GO surface; Raman spectroscopy; HRPES; Reaction reagent

Background

Among the various systems composed only of carbon 5atoms, graphene (a two-dimensional allotrope of carbon) presents a framework for understanding of the properties of all other carbon allotropes. The properties of graphene, including a high intrinsic mobility [1,2], a large theoretical specific surface area, and a high chemical stability, are potentially useful in applications ranging from chemical sensors to transistors [3-8]. Toward exploiting these unique properties of graphene, several research groups have attempted to fabricate large-scaled graphene oxide sheets [9-12]. Graphene oxide (GO) is a layered material consisting of hydrophilic oxygenated graphene oxide sheets bearing oxygen functional groups on their basal planes and edges [13]. It is a useful platform for fabricating functionalized graphene that can potentially confer improved mechanical, thermal, or electronic properties. The

numerous chemical functionalities on a GO surface are expected to readily lend themselves to further chemical functionalization. Graphene-based materials, therefore, show promise in a variety of technological applications. The use of GO surfaces as catalysts of synthetic transformations is a relatively new research area with outstanding potential. Current efforts are directed toward harnessing the oxygen carriers present on GO surfaces as heterogeneous catalysts [14-16].

In this study, we systematically compared and investigated the oxidation of aniline to form azobenzene on monolayer graphene (EG) or graphene-oxide-like (GOx) surfaces fabricated with benzoic acid. Moreover, we focus on examining the difference between EG and GOx surfaces in one substrate, simultaneously. Raman spectroscopy and high-resolution photoemission spectroscopy (HRPES) were used to characterize the surface-bound products. The carbonyl groups introduced onto the graphene surface upon oxidation by benzoic acid to GOx allowed aniline to react with the oxygen carriers. The oxidation of aniline proceed *via* a reaction between the aniline amine groups and the

* Correspondence: easyscan@sookmyung.ac.kr

¹Department of Chemistry, Sookmyung Women's University, Seoul 140-742, Republic of Korea

Full list of author information is available at the end of the article

oxygen groups on the GOx surface under ultra-high vacuum (UHV) conditions maintaining a 365-nm UV light exposure.

Generally, it is hard to distinguish the difference between EG and GOx surfaces in one substrate due to the large size of the HRPES beam. Hence, no previous systematic experimental studies have examined the oxidation of aniline on a GOx surface. However, this study is meaningful with regards to indicating this distinctive difference using the feature of micro Raman spectroscopy.

Methods

A Si-terminated 6H-SiC(0001) substrate (Cree Research, Durham, NC, USA) was used to fabricate EG. The substrate was degassed, annealed at 1,200 K under a Si flux (1 Å/min), and graphitized at temperatures up to 1,500 K (for 2 min) to produce a monolayer of graphene (EG). The annealing temperature was monitored using an infrared pyrometer (with an emissivity of 0.9).

A GOx surface was fabricated by exposing the EG surface to benzoic acid (Sigma Aldrich, purity, 97%, St. Louis, MO, USA). As shown in Figure 1, the EG surface structure was confirmed using LEED and C 1 *s* core-level spectroscopy (not shown here) and then immersed into a benzoic acid solution for 3 min. The surface was subsequently reintroduced into the UHV chamber.

Aniline (Sigma Aldrich, purity, 99.9%) was purified by turbo pumping to remove impurities prior to dosing onto the GOx surfaces. A direct doser, controlled by means of a variable leak valve, was used to dose the substrates. Raman spectra of the samples were collected using a home-built system equipped with an Ar⁺ ion laser (Spectra-Physics Stabilite 2017, Santa Clara, CA, USA) as an excitation source; a spectrometer (Horiba Jobin Yvon TRIAX 550, Kyoto, Japan), and a CCD detector (Horiba Jobin Yvon Symphony) cooled to 140 K. The wavelength of the incident excitation beam was 514.5 nm. HRPES experiments were performed at the 8A2 beamline at the Pohang

Accelerator Laboratory, which was equipped with an electron analyzer (SES100, Gamma Data Scienta, Uppsala, Sweden). The N 1 *s* core-level spectrum was obtained using photon energies of 460 eV. Secondary electron emission spectra (−20 V sample bias) and valence band spectra were measured at photon energies of 80 eV. The binding energies of the core-level spectra were determined with respect to the binding energies of the clean Au 4*f* core level and the valence band (Fermi energy) for the same photon energy. All spectra were recorded in the normal emission mode. The photoemission spectra were carefully analyzed using a standard nonlinear least-squares fitting procedure with Voigt functions [17].

Results and discussion

Raman spectroscopy, which is sensitive to the chemical functional groups on a surface, is a useful tool for comparing the properties of the EG and GOx surfaces. Optical microscopy images of the EG (a) and GOx (b) surfaces were acquired, and their corresponding Raman spectra at two positions (over a particle and over the bottom region) were collected, as shown in Figure 2. Figure 2a shows the optical microscopy image of the EG surface grown on a 6H-SiC (0001) substrate. The EG surface appeared clean, with a few small particles remaining (not oxide). The conditions of the surfaces were assessed by collecting the Raman spectra in a bottom region (marked (A)) and at a particle (marked (B)). A comparison of the D and G Raman bands revealed similar spectra that were characteristic of the EG surface. Note that the G band values (1,597.6 cm^{−1} and 1,597.9 cm^{−1}) were indistinguishable from the G band position of graphene. The ratio of the D and G band intensities, ID/IG, corresponded to the average value for graphene. The Raman D/G intensity ratios at both the bottom and small particle positions on the EG surface were 0.73, indicating that the surface properties at either position were typical of an EG surface [16].

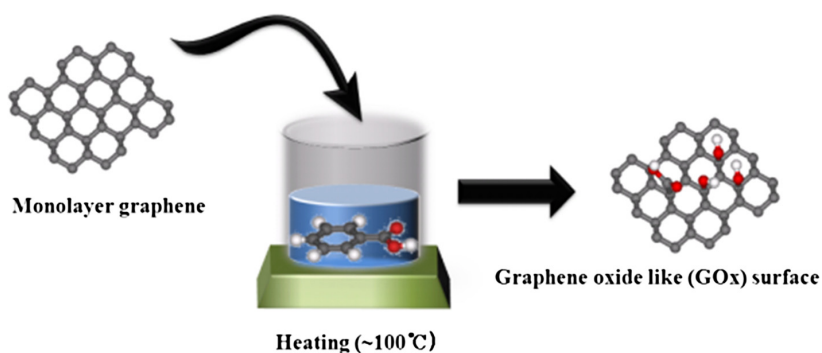


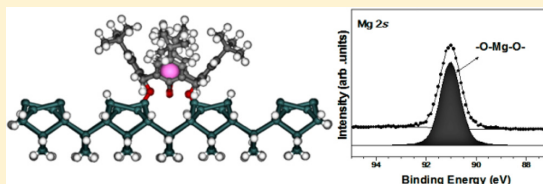
Figure 1 The method how fabricating graphene-oxide-like (GOx) surface. The scheme indicates that the fabrication of the GOx surfaces using benzoic acid.

Mg Capping Inside *p*-Tert-butylcalix[4]arene Adsorbed on a Ge(100) Surface

Minjeong Shin,[†] Myungjin Lee,[†] and Hangil Lee*

Department of Chemistry, Sookmyung Women's University, Seoul 140-742, Republic of Korea

ABSTRACT: The Mg capping inside *p*-tert-butylcalix[4]arene (*p*-TBCA) adsorbed onto a Ge(100) surface was investigated using high-resolution photoemission spectroscopy (HRPES) and their corresponding DFT calculation results. After 0.10 ML *p*-TBCA molecules had been adsorbed onto a Ge(100) surface, subsequent adsorption of a small amount of metallic Mg (~0.10 ML) resulted in the formation of a capped structure inside the preadsorbed *p*-TBCA molecules. The adsorption structures resulting from further deposition of Mg (~0.50 ML) onto the Ge(100) surface were monitored based on the surface charge state obtained from Mg 2s core level spectrum. We also clarified the evidence of the Mg capped inside *p*-TBCA adsorbed onto the Ge(100) surface as we monitored the work function change. Moreover, we confirmed that three different adsorption structures are experimentally favorable at room temperature through DFT calculation results.



INTRODUCTION

The interactions between functional groups in macromolecules and a variety of surfaces have been studied extensively toward improving our understanding of stable adsorption configurations between molecules and semiconductors or metal surfaces.^{1–5} Catalysis and device-related applications require an understanding of the role of functional groups at the molecular level when they adsorb onto the substrate.^{6–8} Organic macromolecules generally include two or more functional groups, such as hydroxyl groups or amine groups, and a variety of adsorption structures are possible for each molecular unit. The adsorption structures that form on a semiconductor surface, such as Ge(100) or Si(100) surfaces, may be differentiated by atomic and electronic properties of the surface dimer atoms.^{9–11}

p-Tert-butylcalix[4]arene (*p*-TBCA) molecules are characterized by a three-dimensional basket shape in which four hydroxyl groups interact via hydrogen bonding to stabilize a concave conformation (see Scheme 1).¹² A *p*-TBCA molecule can accelerate a chemical reaction within its cavity through a combination of local concentration effects and polar stabilization of the transition state.^{13–16} We sought to explore this property of *p*-TBCA molecules by asking whether deposited

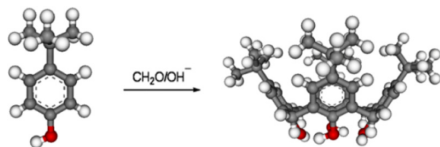
Mg atoms may be capped inside *p*-TBCA molecules adsorbed onto a Ge(100) surface. We additionally investigated whether additional Mg atoms could be directly positioned within the *p*-TBCA cavity. To this end, we examined evidence supporting Mg capping inside the *p*-TBCA molecule adsorbed onto a Ge(100) surface and characterized the electronic structures of Mg and *p*-TBCA adsorbed onto a Ge(100) surface as a function of the sample conditions using high-resolution photoemission spectroscopy (HRPES), work function measurements, and their corresponding DFT calculations.

To our knowledge, no systematic experimental studies have previously examined electronic and adsorption structures of Mg and *p*-TBCA adsorbed onto a Ge(100) surface under various conditions using HRPES and DFT calculations.

EXPERIMENTAL AND COMPUTATIONAL DETAILS

p-Tert-butylcalix[4]arene (*p*-TBCA), shown in Scheme 1, was synthesized by the condensation of *p*-tert-butylphenol with formaldehyde and sodium hydroxide (see Scheme 1).¹² The Ge(100) surface (*p*-type, $R = 0.10\text{--}0.39\ \Omega$) was cleaned by several cycles of sputtering with 1 keV Ar⁺ ions at 700 K for 20 min, followed by annealing at 900 K for 10 min. A solid doser was used to deposit the *p*-TBCA molecules. After warming to 345 K for 30 min, the chemical was deposited at a constant evaporation rate. Mg was evaporated by placing several Mg grains (99.99%) in a pyrolytic boron nitride (PBN) crucible. The Mg deposition flux was fixed at 0.05 ML/min (ML: 1 molecule per surface dimer), calibrated by a quartz thickness monitor. During the evaporation of the high-purity metals

Scheme 1. Synthesis of *p*-Tert-butylcalix[4]arene^a



^aEach red, gray, and white ball indicates oxygen, carbon, and hydrogen, respectively.

Received: July 30, 2013

Revised: October 7, 2013

Published: October 8, 2013

(0.9999), the pressure was increased by at most 2.5×10^{-10} Torr at 300 K.

High-resolution photoemission spectroscopy (HRPES) measurements were performed at the 10D beamline of the Pohang Accelerator Laboratory. The Ge 3d, C 1s, O 1s, and Mg 2s core-level spectra were obtained using a PHOIBOS 150 electron energy analyzer equipped with a two-dimensional charge-coupled device (2D CCD) detector (Specs GmbH) at photon energies of 80, 340, 590, and 160 eV, which enhanced the surface sensitivity. Secondary electron emission spectra (-20 V sample bias) were measured at a photon energy of 40 eV. The binding energies of the core-level spectra were determined with respect to the binding energies of the clean Au 4f core level and the valence band (Fermi energy) for the same photon energy. The photoemission spectra were carefully analyzed using a standard nonlinear least-squares fitting procedure with Voigt functions.¹⁷

All DFT adsorption energy calculations were performed using the JAGUAR 9.1 software package, which applied a hybrid density functional method and included Becke's three-parameter nonlocal exchange functional with the correlation functional of Lee–Yang–Parr (B3LYP).^{18,19} In these calculations, we considered twenty-dimer ($\text{Ge}_{161}\text{H}_{113}$) cluster models.

RESULTS AND DISCUSSION

The C 1s, O 1s, and Mg 2s core-level spectra were obtained to track the variation of the electronic and adsorption structures of Mg and *p*-TBCA adsorbed onto a Ge(100) surface varying with the sample conditions. Figure 1a–c shows the C 1s, O 1s, and

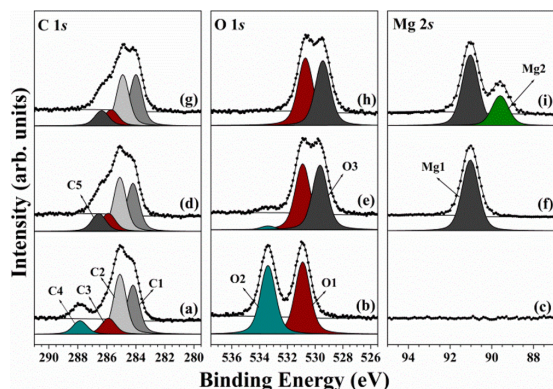


Figure 1. The C 1s, O 1s, and Mg 2s core-level spectra of Mg and *p*-TBCA adsorbed onto a Ge(100) surface under different conditions. Panels a–c were obtained after 0.10 ML *p*-TBCA deposited onto the Ge(100) surface. Panels d–f were obtained after 0.10 ML Mg deposition onto a 0.10 ML *p*-TBCA deposited onto the Ge(100) surface. Panels g–i were obtained after 0.50 ML Mg deposition onto 0.10 ML *p*-TBCA deposited onto the Ge(100) surface. The dots indicate experimental values and the solid lines represent the results of peak fitting.

Mg 2s core-level spectra obtained after 0.10 ML *p*-TBCA deposition onto a Ge(100) surface at 300 K. As shown in Figure 1a, the C 1s peaks exhibited four distinct bonding features. Considering the chemical inequivalency of the carbon atoms being included in the *p*-TBCA molecule, it would divide the molecule into three distinct features: tert-butyl-, phenyl-, or

hydroxyl groups (see Scheme 1). However, the observation of four peaks could be explained in terms of four surface-adsorbed structures formed through adsorption at two of the four hydroxyl groups in the *p*-TBCA molecule. Each C1, C2, C3, and C4 was assigned to the phenyl group (284.2 eV: C1), tert-butyl group (285.1 eV: C2), $-\text{C}-\text{O}-\text{Ge}$ (285.9 eV: C3), and $-\text{C}-\text{OH}$ (287.8 eV: C4), respectively. These results indicated that the *p*-TBCA molecules were well-adsorbed onto the Ge(100) surface through two hydroxyl groups included in the molecule.

The O 1s core-level spectrum was also acquired, as shown in Figure 1b. The chemical equivalency of the oxygen atoms in the *p*-TBCA molecule suggested that only one oxygen component was expected; however, as shown in Figure 1b, the O 1s peaks could be clearly resolved into two components designated as O1 and O2, which were assigned to $\text{O}-\text{Ge}$ (530.9 eV) and $-\text{C}-\text{OH}$ (533.4 eV), which were the similar values obtained at the serine and threonine adsorbed on the Ge(100) system, respectively.^{20,21} As in the context of the C 1s peak analysis, the hydroxyl and Ge bonded oxygen atoms were expected to yield different signals. The analysis of the O 1s peaks confirmed that the two hydroxyl groups in the *p*-TBCA molecule were adsorbed onto the Ge(100) surface in an OH dissociated bonded structure.

We next acquired three core level spectra after the deposition of 0.10 ML Mg onto the 0.10 ML *p*-TBCA adsorbed onto the Ge(100) surface, as shown in Figure 1d–f. The electronic structures characterized by these spectra directly indicated that the Mg atoms were capped inside the *p*-TBCA molecules adsorbed onto the Ge(100) surface. As mentioned, the focus of this study was to determine whether the deposited Mg atom could be capped inside the *p*-TBCA molecules adsorbed onto the Ge(100) surfaces and to identify the positioning of an additional Mg atom in the structure. The core level spectra permit resolution of the first question. As additional Mg atoms were deposited, the C 1s core level spectrum (Figure 1d) revealed that the nonbonded hydroxyl (C4) peak fully disappeared and a new peak grew at 286.5 eV (C5). These results indicated that the unadsorbed hydroxyl features were adsorbed onto another atom or molecule, although the stability of this adsorption structure was not clear (see Figure 4). Interestingly, the binding energy of the C 1s peak suggested that it originated from a $-\text{C}-\text{O}-\text{Mg}$ bonded feature. The analysis of the O 1s and Mg 2s core level spectra could be readily explained in terms of this variation in the adsorption structures as shown in Figure 1e,f. As expected, Figure 1e shows a large change in the bonding features of the oxygen atom. A comparison with Figure 1b showed that, under this condition, the hydroxyl-induced bonding feature disappeared and the metal-bonded oxygen feature appeared at 529.6 eV (marked as O3). The Mg 2s core level spectrum shown in Figure 1f confirmed the presence of an oxygen-bonded Mg peak ($-\text{O}-\text{Mg}-\text{O}-$) located at 91.0 eV (marked as Mg1). The binding energy of metallic Mg was located at 88.7 eV; thus, the value obtained indicated a range of Mg oxygen-related bonding features. An analysis of the three core level spectra obtained under this condition indicated that Mg atoms were adsorbed to the unadsorbed hydroxyl group. The electron charge state of Mg was expected to be +2 based on the binding energy obtained from the Mg 2s core level spectrum.

Finally, we addressed the second question about the positioning of additional Mg atoms. Mg atoms were further deposited up to a 0.50 ML coverage level. Under this condition,

Three Distinct Equilibrium States via Self-Assembly: Simple Access to a Supramolecular Ion-Controlled NAND Logic Gate

Dong Sub Kim,[†] Vincent M. Lynch,[†] Jung Su Park,^{*,§} and Jonathan L. Sessler^{*,†,‡}

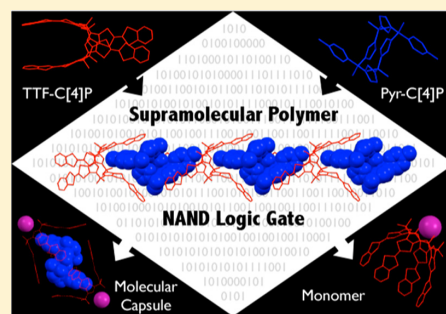
[†]Department of Chemistry and Biochemistry, The University of Texas at Austin, Austin, Texas 78712-1224, United States

[§]Department of Chemistry, Sookmyung Woman's University, Seoul 140-742, Korea

[‡]Department of Chemistry, Yonsei University, Seoul 120-749, Korea

S Supporting Information

ABSTRACT: During the past several decades, considerable effort has focused on self-assembled systems. However, most work has been directed toward understanding the equilibrium between two major chemical entities, namely the dissociated components and the corresponding associated complex. While there are quite a few examples of 'multi-responsive' materials, control over 'multistate' materials has proved difficult to achieve. Here, we report the formation and the interplay of a self-assembled calix[4]pyrrole array that exhibits three limiting forms, namely a 1:1 self-assembled oligomer, a 2:1 capsule, and the corresponding monomers. Interconversion between these states may be controlled by using the tetraethylammonium cation (TEA⁺) and/or iodide anion (I[−]) as chemical inputs. The combination of self-assembly and ion-based control may be used to create systems that display NAND logic behavior. The system outputs have been confirmed by a variety of analytic methods, including UV–vis and 2D ¹H DOSY, NOESY NMR spectroscopy, scanning electron microscopy, and single crystal X-ray diffraction analyses.



INTRODUCTION

The appeal of nanoscale computing and molecular informatics is driving efforts to store and read out information at the molecular level. Inspired by this challenge, considerable effort has been invested in developing molecular systems that allow access to more than two equilibrium states. This has led to such elegant systems as multistation mechanically linked rotaxanes,¹ coordination arrays,² and photochromic molecules.³ An appealing and labor saving alternative would be multimode systems generated via self-assembly. Such systems could consist of supramolecular ensembles that permit reversible and controllable access to multiple molecular states via an appropriate choice of external inputs. To date, self-assembly has been effectively used to create nanostructured materials,^{4,5} support dynamic combinatorial chemistry,⁶ and produce supramolecular polymers.^{7,8} For instance, synthetic supramolecular polymeric systems have been exploited to create a variety of functional materials including, but not limited to, self-healing,⁹ stimuli-responsive,¹⁰ and sensory materials.¹¹ However, this approach has yet to be used to produce self-assembled supramolecular polymers that display NAND-logic gate behavior. We believe that such systems may be obtained by creating self-assembled ensembles that allow access to three discrete states that may be interconverted via application of appropriate chemical inputs. Herein, we report a three-state system that can be prepared by simple mixing. Specifically, we detail the preparation of a self-assembled system based on two different calix[4]pyrrole (C[4]P) subunits that exists in three

limiting equilibrium states (monomer, capsule, and oligomer). This system undergoes ion-dependent conversion to three different molecular states as the result of controlled changes in the conformations of the constituent C[4]Ps. This gives rise to various spectroscopically distinct chemical products (output) based on the choice of the ion (input) and does so in accord with the terms of a NAND logic operation. (Note: A NAND gate is defined as a logic gate that produces a "false" response only if all inputs are true.)^{12,13} The present system differs from other NAND logic devices, both in terms of its ease of preparation and by virtue of the fact that the same output is produced from two of three underlying states, namely a supramolecular oligomeric structure and a molecular capsule. It also differs from other well-characterized multistate systems in that it does not rely on the use of either covalent or mechanical bonds.

Calix[4]pyrroles are attractive receptors for the generation of stimulus-controlled self-assembled structures, including as detailed below NAND logic devices. Calix[4]pyrroles are nonaromatic tetrapyrrolic macrocycles that typically undergo a dramatic conformational change from the more thermodynamically stable 1,3-alternate conformation to a bowl-like cone conformation upon the addition of coordinating anions, such as carboxylate, sulfate, phosphate, and halides.^{14–16} The C[4]Ps are also of interest because in their respective cone

Received: August 2, 2013

Published: September 5, 2013

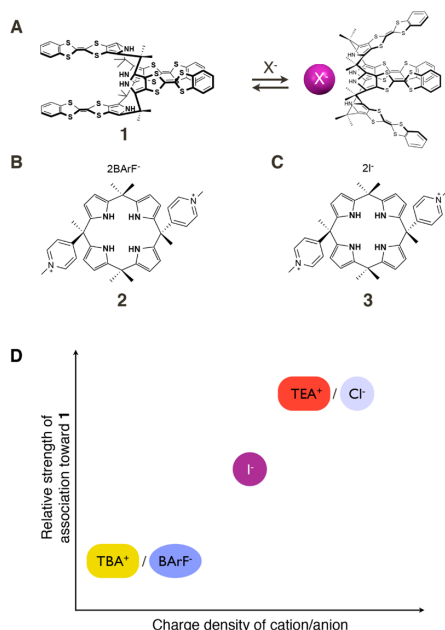


Figure 1. Chemical structures of three C[4]P species used in the present study. (A) The TTF-functionalized C[4]P (TTF-C[4]P, 1) acts as a switching unit through anion and/or cation recognition. (B) and (C) The two building blocks with which 1 is allowed to interact in a noncovalent fashion, namely, a bismethylpyridinium-functionalized C[4]P as the bis-BArF[−] (anions omitted for clarity) and the bis-iodide salts (species 2 and 3, respectively). (D) Schematic illustration of the relative binding affinity of relevant anions and cations for 1 plotted vs the charge density of the ion in question.

conformations (as produced in the presence of a strongly coordinating anion) they will bind moderately sized cations, such as Cs^+ and tetraethylammonium cation (TEA^+), but not larger species, such as the tetrabutylammonium cation (TBA^+).^{17,18} These features make the C[4]Ps attractive as building blocks for the creation of self-assembled materials. In particular, disparities in binding affinities, arising from differences in size and charge density, could be exploited to control the nature of C[4]P-derived self-assembled materials.

In a recent study, we showed the formation of supramolecular polymers from two different C[4]Ps, wherein the main attractive forces were hydrogen-bonding and charge-transfer interactions.¹¹ These self-assembled polymeric arrays were found to exhibit an environmental response, undergoing deaggregation when exposed to analytes such as trinitrobenzene (TNB) and chloride anion. In separate work, we demonstrated that electron transfer between an electron-rich TTF-annulated C[4]P and an electron-poor acceptor, such as bisimidazolium quinone (BIQ^{2+})¹⁹ or $\text{Li}@\text{C}_{60}$,²⁰ could be regulated via the addition of anions or cations. As a general rule, conformation change-driven switching is induced by the formation of pyrrole NH-anion hydrogen bonds, with smaller and more highly charge dense anions binding more tightly.¹⁴

Based on the above precedent, we considered it likely that a new supramolecular NAND logic device could be constructed via the use of two different C[4]Ps whose response to added anions and cations is slightly different. As detailed below, this objective was achieved by using a tetrathiafulvalene (TTF)-

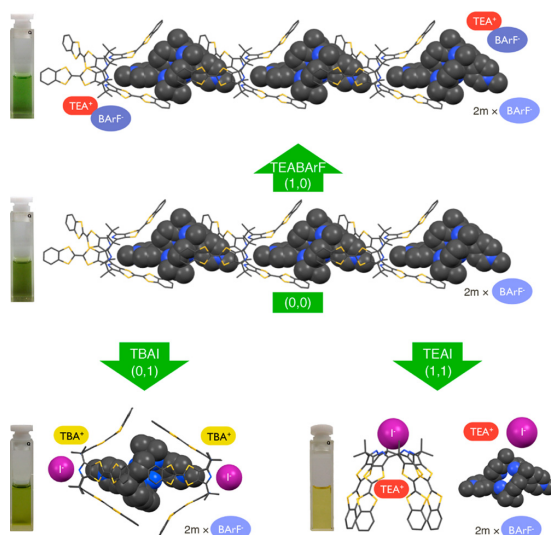


Figure 2. Schematic representation of the three limiting equilibrium states produced from 1 and 2 via treatment with appropriate ions. These states are illustrated starting from the self-assembled NAND logic gate (1·2)_n. The logic operations are triggered via the addition of various tetraalkylammonium salts (2 equiv each) as inputs. Also shown are the structures corresponding to the resulting outputs as observed in chloroform solution (where m is the number of equivalents of 2 present in solution).

functionalized C[4]P (TTF-C[4]P, 1) in conjunction with two appropriately chosen *trans*-bis(methylpyridinium)-*meso*-substituted C[4]P salts, namely Pyr-C[4]P 2 and Pyr-C[4]P 3. These two salts are comprised of the same C[4]P dication but differ in the nature of counteranions. Whereas 2 contains two noncoordinating tetrakis[bis(3,5-trifluoromethyl)phenyl]borate (BArF^-) anions,^{21–23} 3 bears two more strongly bound iodide counteranions. The structural changes induced by varying these anions and their salt forms give rise, in turn, to different colors in CDCl_3 solution, thus producing “readouts” that conform to NAND logic behavior. The various components of the present NAND system are shown in Figure 1. The underlying chemistry is summarized in Figure 2.

■ FORMING A SUPRAMOLECULAR OLIGOMER AND CREATING A FIRST EQUILIBRIUM STATE

When mixed in chloroform, 1 and 2 self-assemble to form a supramolecular oligomeric material, (1·2)_n, that is stabilized by donor–acceptor interactions between the electron-rich TTF moieties present in compound 1 and the electron-poor methylpyridinium units of compound 2. In the absence of a coordinating anion (i.e., using the BArF salt 2), each C[4]P exists in the 1,3-alterante conformation such that the electron-deficient pyridinium units are intercalated within the cavity generated by two electron-rich TTF moieties. Such behavior is consistent with what is normally observed for C[4]Ps and matches what was seen in a previous study wherein a bis-dinitrophenyl-functionalized C[4]P was found to be stabilized in the form of a sandwich-like complex in the presence of a second C[4]P that contains TTF subunits.¹¹

An immediate change in color accompanies the mixing of 1 and 2 (yellow to green) with corresponding differences in the

Pd(II)-catalyzed direct C–H acylation of *N*-Boc hydrazones with aldehydes: one-pot synthesis of 1,2-diacylbenzenes†

Cite this: *Org. Biomol. Chem.*, 2013, **11**, 7869

Satyasheel Sharma,^a Aejin Kim,^a Jihye Park,^a Mirim Kim,^a Jong Hwan Kwak,^a Young Hoon Jung,^a Jung Su Park^{*b} and In Su Kim^{*a}

Received 12th August 2013,
Accepted 27th September 2013

DOI: 10.1039/c3ob41644e

www.rsc.org/obc

A palladium(II)-catalyzed direct acylation of acetophenone *N*-Boc hydrazones with aldehydes via C–H bond activation is described. This protocol provides direct access to a range of 1,2-diacylbenzenes, which are useful precursors to construct biologically interesting and pharmaceutically important compounds.

Introduction

Aryl ketones are important structural scaffolds found in natural products, medicinally relevant molecules, and functional materials.¹ Notably, 1,2-diacylbenzenes are known to be crucial synthetic precursors due to the high prevalence of biologically active compounds and pharmaceuticals such as phthalazines, *N*-arylphthalimides, isobenzofurans, indanones, isoindoles, isoindolines, and isoquinolines.² Also, 1,2-diacylbenzenes are receiving increasing attention as fluorescence reagents for both qualitative and quantitative high-sensitivity analyses of amino acids and peptides.³ In particular, 1,2-diacylbenzenes were used in a fluorometric assay for biotinase due to their ability to react selectively with lysine.⁴ These facts have led to the development of an efficient method for the preparation of 1,2-diacylbenzenes.

General methods for the construction of 1,2-diacylbenzenes include oxidation of *N*-aroylhydrazones of *o*-hydroxyaryl ketone acylhydrazones with lead tetraacetate⁵ or hypervalent iodine reagents.⁶ However, from a synthetic point of view, all these reactions present intrinsic drawbacks, which include the use of toxic reagents, poor functional group tolerance, harsh reaction conditions, and the requirement of prefunctionalized starting materials. Therefore, it is highly desirable to develop atom-economical protocols that involve fewer synthetic steps

and readily available starting materials for the preparation of 1,2-diacylbenzenes.

Transition metal-catalyzed dehydrogenative cross-coupling reactions via selective C–H bond activation have emerged as a powerful tool to produce structurally diverse organic molecules, because such methods avoid preparation of preactivated starting materials and production of stoichiometric metallic waste.⁷ Thus, cross-coupling reactions via C–H bond activation can lead to an improved overall efficiency of the desired transformation. Recently, transition metal-catalyzed oxidative acylations of inactive sp² C–H bonds with aldehydes and alcohols as acyl sources have been described.⁸ Catalytic decarboxylative acylations of aromatic C–H bonds using α -oxocarboxylic acids as acyl surrogates were also reported.⁹ In addition, a palladium-catalyzed acylation of 2-arylpyridines and acetanilides via oxidation of benzylic C–H bonds in toluene derivatives was demonstrated.¹⁰ Notably, Yu^{8e} and Li^{8f} respectively disclosed efficient ways to prepare *ortho*-acylated acetophenone *O*-methyl oximes as the precursors for 1,2-diacylbenzenes via palladium- and rhodium-catalyzed oxidative coupling reactions between acetophenone *O*-methyl oximes and aldehydes. However, these approaches require an additional step to remove the oxime directing group for the preparation of 1,2-diacylbenzenes.

Hydrazones are common protection groups of the ketone moiety, and have become versatile cross-coupling partners in transition metal-catalyzed reactions.¹¹ However, hydrazones have been rarely used as directing groups of C–H bond activation protocols.¹² For example, Inamoto and Hiroya described a Pd-catalyzed C–H activation of *N*-tosylhydrazones followed by intramolecular amination to afford highly substituted indazoles (Fig. 1).^{12a} Xu and coworkers reported a Rh-catalyzed tandem C–H olefination and annulations between sulfonylhydrazones and alkenes to give 1,2-dihydrophthalazines.^{12b} Recently, Ge *et al.* disclosed a Cu-catalyzed dehydrogenative

^aSchool of Pharmacy, Sungkyunkwan University, Suwon 440-746, Republic of Korea. E-mail: insukim@skku.edu; Fax: +82 31 292 8800; Tel: +82 31 290 7788

^bDepartment of Chemistry, Sookmyung Women's University, Seoul 140-742, Republic of Korea. E-mail: jspark@sookmyung.ac.kr; Fax: +82 2 2077 7321; Tel: +82 2 2077 7549

†Electronic supplementary information (ESI) available: ¹H and ¹³C NMR copies of all products. See DOI: 10.1039/c3ob41644e

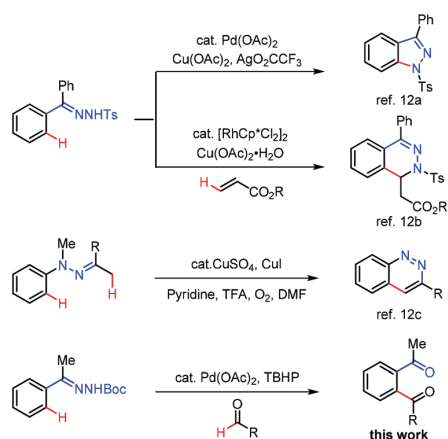


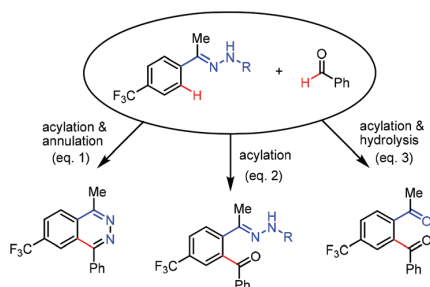
Fig. 1 Hydrazone directing groups in C–H bond activation.

intramolecular cyclization of *N*-methyl-*N*-phenylhydrazones to provide cinnolines.^{12c}

As part of an ongoing research program directed toward the development of catalytic acylation reactions of inactive C–H bonds, we became interested in developing an efficient synthetic route to 1,2-diacylbenzenes *via* C–H bond activation. Herein we present the tandem palladium-catalyzed *ortho*-acylation and deprotection of acetophenone *N*-Boc hydrazones with aldehydes under *tert*-butyl hydroperoxide (TBHP) as a convenient oxidant to afford 1,2-diacylbenzenes. To the best of our knowledge, this is the first report of direct catalytic acylation for the formation of 1,2-diacylbenzenes.

Results and discussion

Our initial investigation focused on the coupling of *p*-CF₃-acetophenone *N*-substituted hydrazones with benzaldehyde to afford 1,4-disubstituted phthalazines with interesting luminescent¹³ and anticancer activities¹⁴ by tandem acylation and intramolecular cyclization (Scheme 1, eqn (1)). After extensive screening of hydrazones with =NNH–Boc, =NNH–H, =NNH–Moc and =NNH–Ts directing groups, all reactions provide a mixture of *p*-CF₃-acetophenone, acylated hydrazones and 1,2-diacylbenzenes, instead of the phthalazine compounds



Scheme 1 Catalytic acylation strategy of hydrazone directing groups.

Table 1 Selected optimization of reaction conditions^a

Entry	Pd catalyst (mol%)	Oxidant (equiv.)	Solvent	Yield ^b (%)
1	Pd(OAc) ₂ (10)	TBHP (3)	DCE	58
2	Pd(OAc) ₂ (10)	(PhCOO) ₂ (3)	DCE	24
3	Pd(OAc) ₂ (10)	Ag ₂ O (3)	DCE	N.R.
4	Pd(OAc) ₂ (10)	(NH ₄) ₂ S ₂ O ₈ (3)	DCE	N.R.
5	Pd(TFA) ₂ (10)	TBHP (3)	DCE	28
6	PdCl ₂ (10)	TBHP (3)	DCE	22
7	Pd(OAc) ₂ (10)	TBHP (3)	Toluene	46
8	Pd(OAc) ₂ (10)	TBHP (3)	THF	24
9	Pd(OAc) ₂ (10)	TBHP (3)	MeCN	51
10	Pd(OAc) ₂ (10)	TBHP (3)	DMF	N.R.
11 ^c	Pd(OAc) ₂ (10)	TBHP (3)	DCE	50
12	Pd(OAc) ₂ (10)	TBHP (4)	DCE	63
13	Pd(OAc) ₂ (5)	TBHP (4)	DCE	48
14 ^d	Pd(OAc)₂ (10)	TBHP (4)	DCE	71
15 ^{d,e}	Pd(OAc) ₂ (10)	TBHP (4)	DCE	N.R.

^a Reaction conditions: **1a** (0.3 mmol), **2a** (0.6 mmol), Pd catalyst (quantity noted), oxidant (quantity noted), solvent (1 mL) at 70 °C for 10 h in pressure tubes. ^b Isolated yield by flash column chromatography. ^c AcOH (50 mol%) was added as an additive. ^d **2a** (0.9 mmol, 3 equiv.). ^e Room temperature, 20 h.

(Scheme 1, eqn (2) and (3)). Interestingly, *p*-CF₃-acetophenone *N*-Boc hydrazone (**1a**) was found to couple with 2 equiv. of benzaldehyde (**2a**) in the presence of 10 mol% of Pd(OAc)₂ and 3 equiv. of TBHP in DCE solvent at 70 °C for 10 h to afford 1,2-diacylbenzene **3a** in 58% yield (Table 1, entry 1). Thus, hydrazone derivative **1a** was chosen as an optimal substrate for the preparation of 1,2-diacylbenzene product.

Screening of oxidants showed that TBHP is superior to other oxidants such as (PhCOO)₂, Ag₂O and (NH₄)₂S₂O₈ (Table 1, entries 2–4). Further screening of the effect of palladium revealed that Pd(OAc)₂ was found to be most effective in this coupling reaction (Table 1, entries 5 and 6). After screening of solvents under otherwise identical conditions, DCE was found to be the most effective solvent in this coupling reaction, whereas other solvents such as toluene, THF, MeCN and DMF failed to facilitate high levels of conversion (Table 1, entries 7–10). Further study showed that AcOH additive displayed the decreased catalytic activity (Table 1, entry 11). Logically, it was thought that the formation of **3a** can be controlled by the amount of catalyst, oxidant and aldehyde (Table 1, entries 12–14). Indeed, the use of 3 equiv. of benzaldehyde (**2a**) and 4 equiv. of TBHP afforded our desired product **3a** in high yield (71%), as shown in entry 14. In addition, this coupling reaction did not proceed at room temperature, even for longer reaction time (Table 1, entry 15).

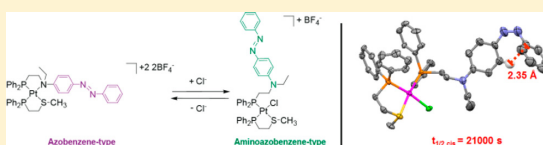
With the optimized reaction conditions in hand, the scope and limitation of the aldehyde were examined, as shown in Table 2. The coupling of *p*-CF₃-acetophenone *N*-Boc hydrazone (**1a**) and aldehydes **2b–2g** with electron-donating (OMe, OBn and Me) and electron-withdrawing groups (CF₃, F and Cl) at

Modulation of Electronics and Thermal Stabilities of Photochromic Phosphino–Aminoazobenzene Derivatives in Weak-Link Approach Coordination Complexes

Jung Su Park,^{†,§,||} Alejo M. Lifschitz,^{†,§} Ryan M. Young,^{†,‡} Jose Mendez-Arroyo,[†] Michael R. Wasielewski,^{†,‡} Charlotte L. Stern,^{†,‡} and Chad A. Mirkin^{*,†,‡}[†]Department of Chemistry and International Institute for Nanotechnology, Northwestern University, 2145 Sheridan Road, Evanston, Illinois 60208, United States[‡]Argonne-Northwestern Solar Energy Research (ANSER) Center, Northwestern University, Evanston, Illinois 60208, United States

S Supporting Information

ABSTRACT: A series of d⁸ transition-metal (Pt(II) and Pd(II)) coordination complexes incorporating phosphine-functionalized aminoazobenzene derivatives as hemilabile phosphino–amine (P,N) ligands were synthesized and studied as model weak-link approach (WLA) photoresponsive constructs. The optical and photochemical properties of these complexes were found to be highly influenced by various tunable parameters in WLA systems, which include type of metal, coordination mode, type of ancillary ligand, solvent, and outer-sphere counteranions. In dichloromethane, reversible chelation and partial displacement of the P,N coordinating moieties allow for toggling between aminoazobenzene- or pseudostilbene- and azobenzene-type derivatives. The reversible switching between electronic states of azobenzene can be controlled through either addition or extraction of chloride counterions and is readily visualized in the separation between π – π^* and n – π^* bands in the complexes' electronic spectra. In acetonitrile solution, the WLA variables inherent to semiopen complexes have a significant impact on the half-lives of the corresponding *cis* isomers, allowing one to tune their half-lives from 20 to 21000 s, while maintaining photoisomerization behaviors with visible light. Therefore, one can significantly increase the thermal stability of a *cis*-aminoazobenzene derivative to the extent that single crystals for X-ray diffraction analysis can be grown for the first time, uncovering an unprecedented edge-to-face arrangement of the phenyl rings in the *cis* isomer. Overall, the azobenzene-functionalized model complexes shed light on the design parameters relevant for photocontrolled WLA molecular switches, as well as offer new ways of tuning the properties of azobenzene-based, photoresponsive materials.



■ INTRODUCTION

Azobenzene and its analogues have long been scrutinized within the context of artificial photochromic systems capable of reversibly transforming between two isomeric forms possessing different electronic, structural, and functional properties.^{1,2} This photoactive synthon has thus been exploited in photoresponsive molecular switches^{3–6} in a variety of research areas ranging from industrial dyes^{7,8} to actuators,⁹ molecular machines,¹⁰ liquid crystals,¹¹ nonlinear optical devices,¹² catalysts,¹³ and biological applications.¹⁴ These extensive studies have led to the general categorization of azobenzene derivatives into three types, namely azobenzene (AB), aminoazobenzene (aAB), and pseudostilbene types (pAB), on the basis of the relative energetic order of their n – π^* and π – π^* transition bands.^{15,16} pAB- and aAB-type materials, obtained through attachments of electron-accepting and/or -donating moieties to the parent AB, display the desired red-shifted photoexcitation energies for isomerization in comparison to AB-type materials. However, both aAB and pAB exhibit significantly increased rates of thermal *cis* to *trans* isomerization, which may limit their use in applications that require longer-

lived *cis* isomers.^{15,16} Even though the AB family as a whole offers a wide span of photoexcitation energies and *cis* thermal stabilities,^{1–16} there is still a major challenge to developing new synthetic approaches that allow for forward- and back-isomerization of AB with visible light and thermal stabilities of the corresponding *cis* isomers within a wide range of time scales. Among several synthetic strategies employed to control the photochemical properties given above, azo-conjugated transition-metal complexes have been shown to provide new advanced molecular functions via combinations of the remarkable photoswitching properties of AB and the magnetic, electrochemical, or coordination properties of the transition-metal complexes.^{17–19}

We are particularly interested in exploiting the photochemistry of AB moieties in the context of the weak-link approach (WLA), a coordination chemistry based method for the construction of allosterically controlled supramolecular complexes.^{20–22} This approach allows for modulating the

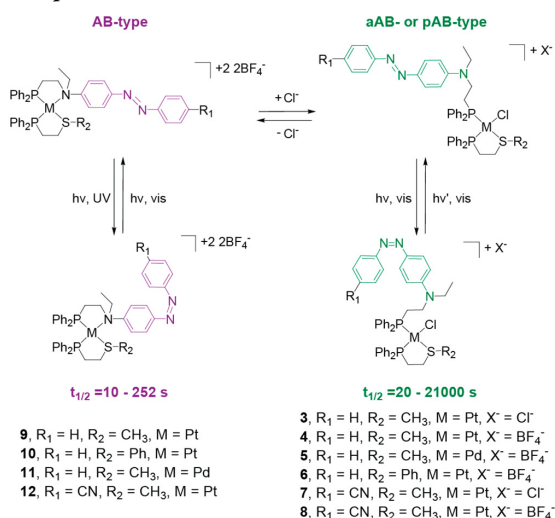
Received: July 12, 2013

Published: November 5, 2013

structure and properties of functional moieties embedded into the supramolecular architecture via coordination chemistry at distal sites, and it has been extensively employed in catalytic switches,^{23,24} chemical sensors,^{25,26} and signal amplification applications.^{27,28} We envisioned that AB could bestow WLA systems with multiple new functionalities, since its photoisomerization can allow for the toggling of properties intrinsic to the behavior of catalytic switches, such as reactive cavity sizes and substrate binding affinities.^{29,30} Furthermore, we hypothesized that AB could provide signaling for coordination changes that directly relate to the activity of a WLA system through the modulation of AB's electronics upon changes in coordination to the structural regulatory centers.

Toward this end, we have synthesized a series of model WLA d⁸ transition-metal complexes (3–12) that incorporate hemilabile phosphino–amine (P,N) ligands **1** and **2** modified with two representative AB-conjugated moieties (Scheme 1). In

Scheme 1. Modulation of Photoisomerization Excitation Energies and *cis* Thermal Stabilities of Model WLA Complexes^a



^aThe numbering of the compounds corresponds to the *trans* isomers unless otherwise noted.

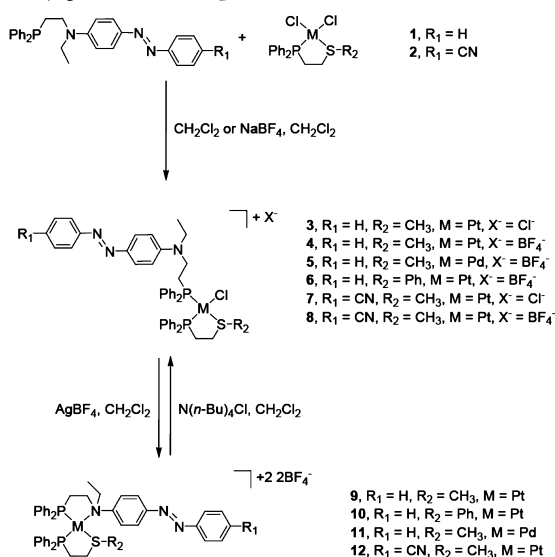
doing so, we have discovered that the WLA provides a promising new platform for tailoring AB's photochemical properties through changes in programmable WLA parameters such as type of metal, coordination mode, type of ancillary ligand, solvent, and outer-sphere counteranions. For instance, it was determined that the switching between coordination modes via abstraction or addition of chloride (closed 9–12 vs semiopen 3–8) induces dramatic electronic changes in parent AB chromophores, resulting in the toggling between two distinct types of AB derivatives in dichloromethane solution (AB vs aAB or pAB). While this toggling can be used, expectedly, to modulate the excitation energies for forward- and back-isomerization, we also determined that the WLA platform enables photochemical properties that are unique to the supramolecular ensemble. Most remarkably, we are able to generate semiopen complexes that retain the ability to be photoisomerized with visible light characteristic of aAB, while

also being able to modulate the half-lives of their *cis* conformers over a 3 order of magnitude range, reaching values typical of AB-type moieties. Such control over their half-lives allows one to significantly increase the lifetime of *cis* isomers of aAB derivatives to the extent that single crystals can be grown at ambient temperature and structural characterization in the solid state is, for the first time, possible. The model complexes studied herein thus represent general photoactive building blocks that can be incorporated into new WLA supramolecular switches with unparalleled photochemical properties.

RESULTS AND DISCUSSION

Synthesis. In many known cases, the azo group of AB can chelate to transition metals, thus hampering photoinduced *trans* to *cis* isomerization processes.³¹ In order to avoid such a situation, it was hypothesized that the attachment of a phosphine moiety to aAB derivatives would not only result in P,N-based hemilabile ligands for WLA model complexes but could also prohibit the coordination of the azo group to soft WLA metal centers. Under this assumption, two photoisomerizable ligands possessing a P,N donor set, **1** and **2**, were prepared (see Scheme 2 and the Supporting Information). These azobenzene derivatives can be classified into aAB type for **1** and pAB type for **2** according to the Rau classification.^{15,16}

Scheme 2. Synthetic Routes to Aminoazobenzene-Conjugated WLA Complexes



Using compounds **1** and **2**, a series of d⁸ transition-metal complexes were successfully synthesized in high yields via well-established standard WLA protocols (Scheme 2).^{32,33} The array of complexes as a whole introduces variables such as the nature of the metal cation (Pt(II) or Pd(II)), type of ancillary phosphino–thioether (P,S) ligand (P,S-Me or P,S-Ph) and inner- or outer-sphere anions (Cl[−] and BF₄[−]), which will be scrutinized later in terms of their effect on the photochemistry of the AB moieties. The model complexes were first isolated as semiopen complexes (3–8) in which the P,N-AB ligands are coordinated through the phosphorus to the transition-metal center and a second ancillary P,S ligand is chelated through

수학과

SELF-RECIPROCAL POLYNOMIALS WITH RELATED MAXIMAL ZEROS

JAEGUG BAE AND SEON-HONG KIM*

ABSTRACT. For each real number $n > 6$, we prove that there is a sequence $\{p_k(n, z)\}_{k=1}^{\infty}$ of fourth degree self-reciprocal polynomials such that the zeros of $p_k(n, z)$ are all simple and real, and every $p_{k+1}(n, z)$ has the largest (in modulus) zero $\alpha\beta$ where α and β are the first and the second largest (in modulus) zeros of $p_k(n, z)$, respectively. One such sequence is given by $p_k(n, z)$ so that

$$p_k(n, z) = z^4 - q_{k-1}(n)z^3 + (q_k(n) + 2)z^2 - q_{k-1}(n)z + 1,$$

where $q_0(n) = 1$ and other $q_k(n)$'s are polynomials in n defined by the severely nonlinear recurrence

$$4q_{2m-1}(n) = q_{2m-2}^2(n) - (4n+1) \prod_{j=0}^{m-2} q_{2j}^2(n),$$

$$4q_{2m}(n) = q_{2m-1}^2(n) - (n-2)(n-6) \prod_{j=0}^{m-2} q_{2j+1}^2(n)$$

for $m \geq 1$, with the usual empty product conventions, i.e., $\prod_{j=0}^{-1} b_j = 1$.

1. Introduction and statements of results

There are infinitely many sequences of monic integral polynomials

$$p_1(z), p_2(z), p_3(z), \dots$$

whose largest (in modulus) zero of $p_{k+1}(z)$ is $\alpha\beta$ where α and β are the first two largest (in modulus) zeros of $p_k(z)$. An example is taken when $p_1(z)$ is the minimal polynomial of a Salem number in which case we can take $p_k(z) = p_1(z)$ for all $k \geq 1$ because a Salem number is a real algebraic integer > 1 all of whose conjugates lie inside or on the unit circle, and at least one of these conjugates has modulus exactly 1. It is known that there are infinitely many Salem numbers. It does not seem obvious how to find such a sequence of

Received April 25, 2012.

2010 *Mathematics Subject Classification.* Primary 11B83; Secondary 30C15.

Key words and phrases. self-reciprocal polynomials, polynomials, sequences.

*This research was supported by the Sookmyung Women's University Research Grants 2012.

distinct polynomials each of which has the same degree. We exhibit here an explicit such sequence consisting of monic self-reciprocal polynomials of degree 4 with integer coefficients. For reference about self-reciprocal polynomial, see Chapter 7 of [1].

Consider the self-reciprocal polynomial

$$z^4 - z^3 - kz^2 - z + 1.$$

One may check that, for $0 \leq k \leq 4$, it has at least two zeros on the unit circle. But, for $k > 4$, no zeros lie on the unit circle. As a generalization of this, we define, for a real number $n > 6$,

$$p_1(n, z) = z^4 - z^3 - (n-2)z^2 - z + 1.$$

In this paper we show the following.

Theorem 1. *For each real number $n > 6$, there is a sequence $\{p_k(n, z)\}_{k=1}^{\infty}$ of fourth degree self-reciprocal polynomials such that the zeros of $p_k(n, z)$ are all simple and real, and every $p_{k+1}(n, z)$ has the largest (in modulus) zero $\alpha\beta$ where α and β are the first and the second largest (in modulus) zeros of $p_k(n, z)$, respectively. One such sequence is given by $p_k(n, z)$ so that*

$$p_k(n, z) = z^4 - q_{k-1}(n)z^3 + (q_k(n) + 2)z^2 - q_{k-1}(n)z + 1,$$

where $q_0(n) = 1$ and other $q_k(n)$'s are polynomials in n defined by the severely nonlinear recurrence

$$(1) \quad \begin{aligned} 4q_{2m-1}(n) &= q_{2m-2}^2(n) - (4n+1) \prod_{j=0}^{m-2} q_{2j}^2(n), \\ 4q_{2m}(n) &= q_{2m-1}^2(n) - (n-2)(n-6) \prod_{j=0}^{m-2} q_{2j+1}^2(n) \end{aligned}$$

for $m \geq 1$, with the usual empty product conventions, i.e., $\prod_{j=0}^{-1} b_j = 1$.

2. Proof of Theorem 1

We define

$$u_k(n) = \begin{cases} 4n+1 & \text{for } k \text{ odd,} \\ (n-2)(n-6) & \text{for } k \text{ even.} \end{cases}$$

Then one may write (1) as

$$(2) \quad 4q_k(n) = q_{k-1}^2(n) - u_k(n) \prod_{j=1}^{\lfloor \frac{k-1}{2} \rfloor} q_{k-2j-1}^2(n)$$

for $k \geq 1$. Note that $u_k(n) > 0$ for $n > 6$, and $u_k(n) = u_{k+2}(n)$ for any integer $k \geq 1$. For notational convenience, write q_k , u_k instead of $q_k(n)$, $u_k(n)$ when n is irrelevant to the context. The values of q_k for $0 \leq k \leq 5$ are as follows.

$$q_0 = 1,$$

TRANSLATION HYPERSURFACES WITH CONSTANT CURVATURE IN SPACE FORMS

KEOMKYO SEO

(Received August 26, 2011, revised November 24, 2011)

Abstract

We give a classification of the translation hypersurfaces with constant mean curvature or constant Gauss–Kronecker curvature in Euclidean space or Lorentz–Minkowski space. We also characterize the minimal translation hypersurfaces in the upper half-space model of hyperbolic space.

1. Introduction

In $\mathbb{R}^3$, a surface is called a *translation surface* if it is given by an immersion

$$X: U \subset \mathbb{R}^2 \rightarrow \mathbb{R}^3: (x, y) \mapsto (x, y, f(x) + g(y)),$$

where $z = f(x) + g(y)$ and f and g are smooth functions. One of the famous examples of minimal surfaces in 3-dimensional Euclidean space is a Scherk’s minimal translation surface. In fact, Scherk [10] showed in 1835 that except the planes, the only minimal translation surfaces are the surfaces given by

$$z = \frac{1}{c} \log \left| \frac{\cos cy}{\cos cx} \right|,$$

where c is a nonzero constant. This surface is called a Scherk’s minimal translation surface. In 1991, Dillen et al. [3] generalized this result to higher-dimensional Euclidean space. (See also [11].) A hypersurface $M \subset \mathbb{R}^{n+1}$ is called a *translation hypersurface* if M is a graph of a function

$$f: \mathbb{R}^n \rightarrow \mathbb{R}: (x_1, \dots, x_n) \mapsto f(x_1, \dots, x_n) = f_1(x_1) + \dots + f_n(x_n),$$

where f_i is a smooth function of one real variable for $i = 1, 2, \dots, n$. More precisely, they proved

Theorem ([3]). *Let M be a minimal translation hypersurface in $\mathbb{R}^{n+1}$. Then M is either a hyperplane or $M = \Sigma \times \mathbb{R}^{n-2}$, where Σ is a Scherk’s minimal translation surface in $\mathbb{R}^3$.*

Scherk's minimal translation surface in 3-dimensional Euclidean space $\mathbb{R}^3$ was generalized to translation surfaces with constant mean curvature or constant Gaussian curvature in $\mathbb{R}^3$ by Liu [7]. In particular, he proved

Theorem ([7]). *Let M be a translation surface with constant Gaussian curvature K in $\mathbb{R}^3$. Then M is congruent to a cylinder, and hence $K \equiv 0$.*

In Section 2 we generalize these previous results to translation hypersurfaces with constant mean curvature or constant Gauss–Kronecker curvature in Euclidean space and Lorentz–Minkowski space. In particular, we prove the following theorems.

Theorem 1.1. *Let M be a translation hypersurface with constant mean curvature H in $\mathbb{R}^{n+1}$. Then M is congruent to a cylinder $\Sigma \times \mathbb{R}^{n-2}$, where Σ is a constant mean curvature surface in $\mathbb{R}^3$. In particular, if $H = 0$, then M is either a hyperplane or $\Sigma \times \mathbb{R}^{n-2}$, where Σ is a Scherk's minimal translation surface in $\mathbb{R}^3$.*

Theorem 1.2. *Let M be a translation hypersurface with constant Gauss–Kronecker curvature GK in $\mathbb{R}^{n+1}$. Then M is congruent to a cylinder, and hence $GK \equiv 0$.*

One may ask the similar problems for translation hypersurfaces in the upper half-space model of hyperbolic space $\mathbb{H}^{n+1}$. Recently López [8] proved that there is no minimal translation surface of type I in $\mathbb{H}^3$. (See Section 3 for the definition of translation hypersurface of type I or type II in hyperbolic space.) In Section 3, we prove an analogue of López's result for higher-dimensional cases in hyperbolic space as follows:

Theorem 1.3. *There is no minimal translation hypersurface of type I in $\mathbb{H}^{n+1}$.*

Furthermore we characterize the minimal translation surfaces of type II in $\mathbb{H}^3$. (See Theorem 3.3.)

2. Translation hypersurface with constant curvature in Euclidean space and Lorentz–Minkowski space

Theorem 2.1. *Let M be a translation hypersurface with constant mean curvature H in $\mathbb{R}^{n+1}$. Then M is congruent to a cylinder $\Sigma \times \mathbb{R}^{n-2}$, where Σ is a constant mean curvature surface in $\mathbb{R}^3$. In particular, if $H = 0$, then M is either a hyperplane or $M = \Sigma \times \mathbb{R}^{n-2}$, where Σ is a Scherk's minimal translation surface in $\mathbb{R}^3$.*

Proof. Let a translation hypersurface M be an immersion given by

$$X: \mathbb{R}^n \rightarrow \mathbb{R}^{n+1}: (x_1, \dots, x_n) \mapsto (x_1, \dots, x_n, f(x_1, \dots, x_n)),$$



An optimal investment, consumption, leisure, and voluntary retirement problem with Cobb–Douglas utility: Dynamic programming approaches

Jung Lim Koo^a, Byung Lim Koo^a, Yong Hyun Shin^{b,*}

^a Wang Yanan Institute for Studies in Economics, Xiamen University, Xiamen, Fujian 361005, China

^b Department of Mathematics, Sookmyung Women's University, Seoul 140742, Republic of Korea

ARTICLE INFO

Article history:

Received 21 October 2012

Received in revised form 29 November 2012

Accepted 30 November 2012

Keywords:

Consumption and leisure
Voluntary retirement
Cobb–Douglas utility
Dynamic programming method
Portfolio selection

ABSTRACT

We consider an optimal consumption, leisure, investment, and voluntary retirement problem for an agent with a Cobb–Douglas utility function. Using dynamic programming, we derive closed form solutions for the value function and optimal strategies for consumption, leisure, investment, and retirement.

© 2012 Elsevier Ltd. All rights reserved.

1. Introduction

We consider an optimal consumption, leisure, and investment problem with voluntary retirement for an agent whose period utility function is a Cobb–Douglas utility function of consumption and leisure. In this model the agent can flexibly choose her leisure amount before retirement above a certain minimum labor requirement, and will receive labor income proportional to the amount of labor supplied. Upon retirement, the agent will enjoy full leisure, at the cost of forgoing all labor income. Using the dynamic programming method pioneered by Merton [1,2] and Karatzas et al. [3] we find closed form solutions to the value function and find the optimal consumption, leisure, and portfolio policies.

Barucci and Marazzina [4] consider this consumption, leisure, investment, and retirement problem in the case of stochastic labor income. Choi et al. [5] solve a similar problem for an agent who has constant elasticity of substitution (CES) period utility. Farhi and Panageas [6] also consider such a problem, where the choice of leisure is confined to only two values: l_1 while working and $\bar{l}$ after retirement. In all of these papers, the authors use the martingale method to solve their optimization problems (see also [7,8]). Shin [9] extends the results of Farhi and Panageas [6] by solving their problem using the dynamic programming method, and shows the equivalence of the solutions obtained through the martingale method and the dynamic programming method. Likewise, we provide a methodological contribution by solving our optimization problem using the dynamic programming method.

The work is organized as follows. Section 2 provides information on the financial market. Section 3 lays out and solves our optimization problem, with detailed proofs provided.

2. The financial market

In our continuous time financial market, we assume that there are two assets: a riskless asset $S_0(\cdot)$ and a risky asset $S_1(\cdot)$, which follow $dS_0(t)/S_0(t) = rdt$ and $dS_1(t)/S_1(t) = \mu dt + \sigma dB(t)$, respectively. The parameters $r > 0$,

* Corresponding author. Fax: +82 2 2077 7323.

E-mail addresses: koolimy@gmail.com (J.L. Koo), koolimy@hotmail.com (B.L. Koo), yhshin@sookmyung.ac.kr, yhshin@kias.re.kr (Y.H. Shin).

$\mu > r$ and σ are assumed to be constants, and $B(\cdot)$ is a standard Brownian motion on a given probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

The market price of risk is defined by $\theta := (\mu - r)/\sigma$. Let π_t be the dollar-amount invested in the risky asset $S_1(t)$ at time t , c_t be the consumption rate process at time t , l_t be the leisure rate process at time t , and τ be a $\mathcal{F}_t$ -stopping time considered as a voluntary retirement time from labor. π_t , c_t , and l_t are $\mathcal{F}_t$ -progressively measurable with $\int_0^t \pi_s^2 ds < \infty$, $\int_0^t c_s ds < \infty$, and $\int_0^t l_s ds < \infty$, for all $t \geq 0$ almost surely.

For before retirement, we consider l_t to be a control variable under the restriction $l_t \leq L < \bar{L}$. For after retirement, however, $l_t = \bar{L}$ is considered to be a constant. Let w be a constant wage rate, and let us define $w(\bar{L} - l_t)$ as labor income at time t . Thus the wealth process is given by

$$dX_t = [rX_t + \pi_t(\mu - r) - c_t + w(\bar{L} - l_t)] dt + \sigma \pi_t dB(t), \quad X_0 = x > -w\bar{L}/r,$$

where $w\bar{L}/r$ denotes the present value of the future labor income of the agent. The agent can consume and invest as long as $x > -w\bar{L}/r$.

3. The optimization problem

The agent in our model wants to maximize her expected utility

$$V(x) = \max_{(c, l, \pi, \tau) \in \mathcal{A}(x)} \mathbb{E} \left[\int_0^\infty e^{-\rho t} u(c_t, l_t) dt \right] = \max_{(c, l, \pi, \tau) \in \mathcal{A}(x)} \mathbb{E} \left[\int_0^\tau e^{-\rho t} u(c_t, l_t) dt + e^{-\rho \tau} U(X_\tau) \right], \quad (3.1)$$

where $\rho > 0$ is a subjective discount rate and $\mathcal{A}(x)$ is an admissible plan of quadruples (c, l, π, τ) such that

$$\mathbb{E} \left[\int_0^\tau e^{-\rho t} u^-(c_t, l_t) dt + e^{-\rho \tau} U^-(X_\tau) \right] < \infty,$$

where $u^- := \max(-u, 0)$. $u(c, l)$ is a Cobb–Douglas utility function defined by

$$u(c, l) := \frac{1}{\alpha} \cdot \frac{(c^\alpha l^{1-\alpha})^{1-\gamma}}{1-\gamma}, \quad 0 < \alpha < 1, \gamma > 0 \text{ and } \gamma \neq 1, \quad (3.2)$$

where γ is the agent's coefficient of relative risk aversion for two different goods c and l , and α is a constant parameter that measures the share of consumption's contribution to the agent's period utility. If we define $\gamma_1 := 1 - \alpha(1 - \gamma)$, then the Cobb–Douglas utility function (3.2) can be rewritten as

$$u(c, l) = \frac{c^{1-\gamma_1} l^{\gamma_1-\gamma}}{1-\gamma_1}.$$

The following assumption is needed to make our optimization problem well-defined and holds throughout the work without further comment.

Assumption 3.1.

$$K := r + \frac{\rho - r}{\gamma} + \frac{\gamma - 1}{2\gamma^2} \theta^2 > 0 \quad \text{and} \quad K_1 := r + \frac{\rho - r}{\gamma_1} + \frac{\gamma_1 - 1}{2\gamma_1^2} \theta^2 > 0.$$

Remark 3.1. The post-retirement value function $U(\cdot)$ in (3.1) is similar to that of the classical Merton's problem, and can be obtained as follows:

$$U(X_\tau) = \frac{\bar{L}^{\gamma_1-\gamma}}{K_1^{\gamma_1}(1-\gamma_1)} X_\tau^{1-\gamma_1}.$$

Remark 3.2. For later consideration, we introduce a quadratic equation

$$f(m) := \frac{1}{2} \theta^2 m^2 + \left(\rho - r + \frac{1}{2} \theta^2 \right) m - r = 0, \quad (3.3)$$

with two roots $m_+ > 0$ and $m_- < -1$.

The next theorem gives our main results.

OPTIMAL CONSUMPTION AND INVESTMENT PROBLEM WITH REGIME-SWITCHING AND CARA UTILITY

YONG HYUN SHIN*

ABSTRACT. We use the dynamic programming method to investigate the optimal consumption and investment problem with regime-switching. We derive the optimal solutions in closed-form with constant absolute risk aversion (CARA) utility.

1. Introduction

After the pioneer work of Merton [3, 4], portfolio optimization problem has been one of the most important and active area in mathematical finance. Recently, portfolio selection problems combined with regime-switching are widely considered. ([1, 6, 5])

In this paper we investigate the optimal consumption and investment problem with regime-switching under the dynamic programming framework of Karatzas *et al.* [2]. We use the constant absolute risk aversion (CARA) utility function to obtain the optimal solutions in closed-form.

2. The financial market

Let us define a standard Brownian motion B_t and a continuous-time two-state Markov chain ϵ_t on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We assume that B_t and ϵ_t are independent and the filtration $\{\mathcal{F}_t\}_{t \geq 0}$ is generated by both the Brownian motion B_t and the Markov chain ϵ_t .

We assume that two assets are traded in the financial market: One is a riskless asset with constant interest rate r and the other is a risky asset. We assume that there are two regimes 1, 2 in the market and regime i switches into regime j at the first jump time of an independent

Received September 06, 2012; Accepted November 09, 2012.

2010 Mathematics Subject Classification: Primary 91G10.

Key words and phrases: regime-switching, CARA utility, portfolio optimization.

Poisson process with intensity λ_i , for $i, j \in \{1, 2\}$. Here two states of regimes 1, 2 are economically considered as bull market and bear market. So, in regime $i \in \{1, 2\}$, the risky asset price process follows $dS_t/S_t = \mu_i dt + \sigma_i dB_t$. The market price of risk is defined by $\theta_i \triangleq (\mu_i - r)/\sigma_i$, $i = 1, 2$. Let π_t be the $\mathcal{F}_t$ -progressively measurable portfolio process, the dollar-amount of the agent's wealth invested in the risky asset at time t and c_t be the nonnegative $\mathcal{F}_t$ -progressively measurable consumption process at time t . We assume that they hold the following technical conditions:

$$\int_0^t c_s ds < \infty \quad \text{and} \quad \int_0^t \pi_s^2 ds < \infty, \quad \text{for all } t \geq 0, \text{ almost surely (a.s.).}$$

The agent's wealth process X_t at time t follows

$$dX_t = [rX_t + \pi_t(\mu_i - r) - c_t] dt + \sigma_i \pi_t dB_t, \quad X_0 = x > 0, \quad i = 1, 2.$$

3. The optimization problem

The agent's optimization problem with CARA utility $u(c) := -e^{-\gamma c}/\gamma$, is to maximize her expected utility

$$(3.1) \quad V_i(x) = \sup_{(c, \pi) \in \mathcal{A}(x)} \mathbb{E} \left[- \int_0^{\tau_i} \frac{e^{-\rho t - \gamma c_t}}{\gamma} dt + e^{-\rho \tau_i} V_j(X_{\tau_i}) \right],$$

where τ_i is the first jump time since the beginning of regime i , $\rho > 0$ is a subjective discount rate, $\gamma > 0$ is absolute risk aversion and $\mathcal{A}(x)$ is an admissible set of pairs (c, π) at x , where $i, j \in \{1, 2\}$ and $i \neq j$.

ASSUMPTION 3.1.

$$\rho - r + \frac{\theta_i^2}{2} > 0, \quad i \in \{1, 2\}.$$

Next theorem gives our main results.

THEOREM 3.2. *The value function is given by*

$$(3.2) \quad V_i(x) = -\frac{M_i}{\gamma r} e^{-\gamma r x},$$

where M_i and M_j are the solutions of the system of algebraic equations

$$(3.3) \quad \left(\rho + \lambda_i - r + \frac{\theta_i^2}{2} \right) M_i + r M_i \log M_i - \lambda_i M_j = 0,$$

통계학



IJRDM
41,9

688

Received 4 February 2012
 Revised 30 July 2012
 17 April 2013
 Accepted 1 May 2013

Online retailer reputation and consumer response: examining cross cultural differences

Jiyoung Kim and Kiseol Yang

Department of Merchandising, University of North Texas, Denton, Texas, USA, and

Bu Yong Kim

Department of Statistics, Sookmyung Women's University, Yonsan-Gu, Seoul, South Korea

Abstract

Purpose – The purpose of this study is to investigate cross cultural differences in the impact of online retailer reputation and retail quality on consumers' emotional and cognitive (i.e. perceived risk) reactions, which lead to purchase intention, based on stimulus-organism-response (S-O-R) model.

Design/methodology/approach – Two cultures, US (individualism, low uncertainty avoidance) and South Korea (collectivism, high uncertainty avoidance) were chosen for comparison for their contrast in cultural characteristics. Multiple group analysis in structural equation modeling (SEM) was employed in order to compare the structural model in two different cultures.

Findings – The results indicate that while the overall mechanism underlying the decision making process is similar for the two countries, differences are found in the relative importance of the factors determining consumers' cognitive and emotional reactions as well as their intention to purchase online.

Research implications/limitations – The majority of the respondents were female and their evaluation was mostly towards online apparel websites. It did not bias the result of this study as the two sample sets were comparable in their demographics and online shopping behavior, yet the result may be generalized to a limited extent.

Practical implications – The results suggest that retailing strategy may vary in response to cultural differences. In East Asian countries that share Confucian values, targeting consumer emotion through experiential cues may not yield as significant result as it may in the USA.

Originality/value – This study will add value to the current literature by examining the cultural difference in consumer psychological process and its consequent effect on purchase intention focusing on reputation as key external stimuli.

Keywords Cross-culture, Reputation, Website quality, Emotion, Perceived risk, United States of America, South Korea, Consumer behaviour, Shopping, Retailers

Paper type Research paper



1. Introduction

As online shopping has proliferated worldwide, differing consumer online behavior has been observed across countries. The differences in consumer behavioral responses in online shopping may be due to underlying cultural differences, as suggested by numerous studies that have investigated culture as a critical determinant explaining consumption behavior (e.g. Kailani and Kumar, 2011; Kim *et al.*, 2008; Liu *et al.*, 2008; Moon *et al.*, 2008; Tong, 2010). Even with increased wealth and convergence of technology across nations, cultural variables still act as a stable predictor of internet

consumption behavior in different countries (Goodrich and De Mooji, 2011). Further, culture generates different consumer response patterns towards presented online stimuli (e.g. Davis *et al.*, 2008), which is not surprising as culture consists of a set of values, norms, and symbols guiding individual behavior (Parsons and Shils, 1951). A number of studies on cross-cultural consumer behavior base their discussions on Hofstede's cultural dimensions and suggest that countries that differ in their cultural classification exhibit different behavioral propensities (e.g. Choi and Geitsfield, 2003; De Mooij and Hofstede, 2010; Goodrich and De Mooji, 2011; Javenpaa and Tractinsky, 1999; Mazaheri *et al.*, 2011; Steenkamp *et al.*, 1999).

This study focuses on cultural differences in the role of internal and external frame of reference as related to emotional, cognitive and behavioral reaction, based on Mehrabian and Russell's (1974) Stimulus-Organism-Response (S-O-R) framework. In the S-O-R framework, an organism is represented by affective and cognitive intermediary states and processes that mediate the relationship between the stimulus and response. Past studies found cross cultural differences in the relationship between website stimuli and consumer responses (e.g. Davis *et al.*, 2008) or the cross cultural differences in the effect of organismic response (i.e. emotion) on attitude and behavior (e.g. Mazaheri *et al.*, 2011). Nevertheless, there is to our knowledge no research that has examined how cultural difference determines the magnitude of the effect of internal and external stimuli on consumer emotional, cognitive, and behavioral responses. Therefore, this study aims to fill this void by proposing a comprehensive model that assesses the cross-cultural differences in the interrelationships among internal evaluative criteria (i.e. website quality), external frame of reference (i.e. reputation), and organismic reactions (i.e. emotion and perceived risk), which in turn impact behavioral response (i.e. purchase intentions).

In order to increase the understanding of the effects of different cultures on online consumer responses, this study selected the US and South Korea with a view to comparing their cultural characteristics and their impacts on online consumer behavioral responses. There are two main reasons for comparing Koreans and Americans. First, it is reported that the number of internet subscribers per 100 inhabitants of the two countries is ranked above the OECD average (OECD, 2011). In 2012, online sales in the US was increased by 13.9 percent to \$364.66 billion (eMarketer, 2013) and increased by 11.3 percent to about \$29.8 billion in South Korea (Statistics Korea, 2012), indicating that online retailing is prevalent in both countries. In addition, the two countries represent cultural characteristics that are different from one another (Park and Jun, 2003). In general, Americans are individualistic and have low uncertainty avoidance, whereas Koreans are traditionally collective and have high uncertainty avoidance (Hall, 1983). Hofstede (1980) reports high uncertainty avoidance for Koreans and Japanese, and low uncertainty avoidance for Americans. It is also noted that China, Japan, and Korea, countries that share a Confucian cultural background, exhibit higher levels of collectivism, whereas in North America, people desire "uniqueness" and wish to be independent of others (Triandis, 2000). The difference between individualism and collectivism is also noted in Lee and Green's (1991) study; this study reports that Koreans exhibit stronger collectivism in their decision making than Americans. Within the different cultural contexts of the two countries, this study aims to investigate the cross cultural differences in the effects of online retailer reputation and website quality on consumers' emotional and cognitive (i.e. perceived risk) reactions, which lead to purchase intention.

소지역 실업률의 패널추정을 위한 일반화커널추정방정식[†]

심주용¹ · 김영원² · 황창하³

¹인제대학교 데이터정보학과 · ²숙명여자대학교 통계학과 · ³단국대학교 응용통계학과

접수 2013년 6월 29일, 수정 2013년 8월 4일, 게재확정 2013년 8월 23일

요 약

오늘날 높은 실업률은 대부분의 국가에서 중요한 문제 중의 하나이다. 한편 소지역의 노동 관련 통계에 대한 요구가 지난 몇년간 급속도로 증가하였다. 그러나 대부분의 공식통계를 생산하기 위한 표본설계는 대영역의 통계를 생산할 목적으로 설계되기 때문에 소지역의 경우 배정되는 표본조사단위 수가 극히 적어 신뢰성 있는 통계 산출이 어렵다. 그리고 소지역 추정에 대한 대부분의 기존 연구들은 특정 시점에서의 추정에 국한 되어 왔다. 그러나 대부분의 공식통계들은 월, 분기 또는 연 단위로 측정되는 패널자료이기 때문에 이를 고려한 추정방법이 필요하다. 본 논문에서는 패널자료의 분석을 위해 유용하게 사용되고 있는 일반화추정방정식의 비모수적 버전인 일반화커널추정방정식을 도출하여 조사시점을 고려한 소지역 실업률의 추정에 활용하는 방안을 제안한다. 모의실험을 통하여 일반화커널추정방정식 방법, 일반화추정방정식 방법 및 일반화선형모형과 비교한다. 그리고 2005년 1월부터 12월까지 경상남도 및 울산광역시의 25개 시군구의 경제활동인구조사의 패널자료에 위에서 언급한 세 가지 방법을 적용하여 해당 소지역의 월별 실업률을 추정한다.

주요용어: 소지역 추정, 일반화선형모형, 일반화추정방정식, 일반화커널추정방정식, 커널기법, 패널자료, 패널추정.

1. 서론

통계청에 의해 매월 15일이 포함된 1주일 동안 전국 약 32,000 표본가구에 대해 실시되는 경제활동인구조사는 취업, 실업, 노동력 등과 같은 국민의 경제활동의 특성을 조사함으로써 거시경제 분석과 인력자원의 개발정책 수립에 필요한 기초자료를 제공한다는 측면에서 오늘날 매우 중요하다. 특히 실업률은 국가 경제와 관련된 주요지표이기 때문에 통계청에서는 특별시, 광역시 또는 도와 같은 대영역 (large domain) 단위의 실업률을 발표하고 있다. 그러나 지방자치제도가 정착되면서 시군구 등과 같은 소지역 (small area) 통계에 대한 관심과 요구가 증가하고 있다. 한편 국내외적으로 소지역 통계에 관한 관심이 높아지고 있으며 이와 관련하여 소지역 추정 (small area estimation)에 관한 많은 연구가 활발하게 진행되고 있다. 특히 미국, 캐나다, 이탈리아, 영국 등과 같은 통계 선진국에서는 분석대상에 따라 다양한 형태의 소지역 추정 관련 통계적 방법론에 대해 정부기관과 전문 학자들과의 공동연구들이 진행되고 있다. Pfeffermann (2013)과 Rao (2003)는 소지역 추정에 대한 최근까지의 많은 연구결과를 체

[†] 이 논문은 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임 (NRF-2011-0021125).

¹ (621-749) 경남 김해시 인제로 197, 인제대학교 데이터정보학과, 겸임교수.

² (140-742) 서울특별시 용산구 청파로47길 100, 숙명여자대학교 통계학과, 교수.

³ 교신저자: (448-701) 경기도 용인시 수지구 죽전로 152, 단국대학교 응용통계학과, 교수.

E-mail: chwang@dankook.ac.kr

계적으로 정리하여 소개하였다. 특히, 소지역 실업률의 추정과 관련된 연구결과들을 소개하면 Chung 등 (2003), Datta 등 (1999), Khoshgooyanfar와 Monazzah (2006), Kim과 Choi (2004), Pereira 등 (2013), Ugarte 등 (2009), Yeo 등 (2008), You 등 (2003) 등이 있다.

소지역 추정이란 용어는 다소 혼동을 주는 용어이다. 왜냐하면 소지역 추정에서 문제가 되는 것은 해당 소지역의 표본크기이지 소지역 자체의 크기가 아니기 때문이다. 또한 소지역이란 반드시 지리학적 지역만을 의미하는 것이 아니고 사회인구학적 집단 또는 산업형태에 따른 집단 등과 같은 다양한 영역(domain)을 나타내기도 한다. 경제활동인구조사를 포함한 대부분의 공식통계를 생산하기 위한 표본설계는 시·도를 중심으로 한 대영역의 통계를 생산할 목적으로 설계되기 때문에 소지역 추정에서는 계획에 없던 지역을 추정해야 하는 어려움이 있으며 가장 일반적으로 발생하는 어려움은 해당 소지역에 배정되는 표본조사단위수가 극히 적어 신뢰할 수 있는 통계를 산출하지 못한다는 것이다. 이러한 관점에서 대영역 기반의 표본설계 하에서 해당 소지역에 대한 행정자료나 센서스 자료를 보조정보로 이용하여 추정의 정확도를 높이는 소지역 추정법들의 필요성이 제기되고 있다.

일반적으로 소지역 추정법은 크게 직접추정법 (direct estimation), 간접추정법 (indirect estimation)과 모형기반추정법 (model-based estimation)으로 나눌 수 있다. 직접추정법은 조사된 자료 그 자체만을 이용하는 추정법이며 간접추정법은 해당 소지역의 행정자료 또는 센서스 자료와 인근 소지역에 관한 통계정보를 보조정보로 이용하는 추정법이다. 그리고 간접추정법에는 합성추정법 (synthetic estimation)과 복합추정법 (composite estimation)이 있다. 합성추정법은 추정하고자하는 소지역과 인근 소지역 또는 특성이 유사한 소지역들의 정보를 이용하여 관심변수의 추정값의 정도를 높이고자 하는 추정법이다. 직접추정량은 편향이 없는 추정량이지만 해당 소지역에 배정된 표본의 크기가 작은 경우에는 추정량의 분산이 커져서 신뢰성이 떨어지게 된다. 한편 합성추정량은 해당 소지역과 인근 유사 지역의 정보가 동질적이지 못할 경우 편향이 발생하는 문제점이 있다. 이런 문제점을 보완하기 위해 두 추정량의 가중평균을 사용하는 방법이 복합추정법이다. 모형기반추정법도 센서스 자료, 지역의 행정자료, 인근 소지역에 관한 통계정보 또는 이전의 연구결과 등과 같은 추가적인 정보들을 보조정보로 이용하기 때문에 보다 정밀한 추정이 가능하다. 모형기반추정법에서는 보조정보의 양이 많을수록 정확한 소지역 추정이 가능하므로 보조정보의 양을 증가시키는 방법이 연구되고 있다. 또한 소지역 간의 변동성을 오차구조에 반영하여 소지역 추정의 정확도를 높일 수 있다. 더욱이 이진 (binary), 범주형 또는 시계열 자료와 같은 다양한 자료에도 적용할 수 있을 뿐만 아니라 지수족 (exponential family)에 속하는 분포를 따르는 자료에 적용할 수 있는 일반화선형모형 (generalized linear model; GLM)과 일반화혼합선형모형 (generalized linear mixed model; GLMM)에 대한 연구가 Ghosh 등 (1998), Marker (1999), Noble 등 (2002) 등에 의해 활발하게 이루어지고 있다. 국내에서는 Yeo 등 (2008)이 패널자료의 분석을 위해 유용하게 사용되고 있는 일반화추정방정식 (generalized estimating equation; GEE)을 소지역 추정에 활용하였다.

대부분의 소지역 추정법들은 모수적 모형에 기초를 두고 있는데 최근 비모수적 방법이 소지역 추정에 적용되기 시작하였다. Jeong과 Shin (2012), Opsomer 등 (2008), Salvati 등 (2010), Shim과 Hwang (2012)이 비모수적 방법을 이용한 소지역 추정에 관하여 연구결과를 발표하였다. 지금까지의 비모수적 소지역 추정법들은 특정 시점에서의 추정에 국한되어 연구되었다. 본 논문에서는 통계적 학습이론에서 많이 활용되고 있는 커널기법을 GEE에 적용하여 비모수적 버전인 일반화커널추정방정식 (generalized kernel estimating equation; GKKE)을 도출하고 조사시점을 고려한 소지역 실업률 추정에 활용하는 방안을 제안하며 그 유효성을 살펴보고자 한다. 본 논문은 다음과 같이 구성되었다. 2절에서 GEE 방법을 간단히 설명하고, 3절에서 커널기법을 적용하여 GEE의 비모수적 버전인 GKKE를 유도하고 조사시점을 고려한 소지역 실업률 추정에 활용하는 방법을 살펴본다. 4절에서 실험연구를 통하여 제안된 방법의 우수성을 고찰하며, 5절에서 결론을 맺는다.

연구논문

2012년 총선 출구조사 예측오차 분석*

Analysis of Forecasting Error of the Exit Poll for the General Election of 2012 in Korea

곽은선^{a)} · 김지연^{b)} · 김영원^{c)}

Eun-Sun Kwak · Ji-Yeon Kim · Young-Won Kim

본 연구에서는 출구조사 예측오차의 발생 원인과 특성을 알아보기 위해 2012년에 치른 19대 총선 출구조사의 자료를 이용하여 출구조사 예측오차를 분석하였다. 예측 오차를 정리해 본 결과 새누리당에 대해서는 과소 예측이, 통합야당에 대해서는 과대 예측이 발생했으며, 시·도별로 편향 수준에 차이가 있었다. 예측오차를 투표소추출 오차와 실사오차로 구분해 보면 투표소추출오차에 비해 실사오차가 더 심각하였다. 또한 오차를 발생시킬 것으로 예상되는 변인들과 투표소추출오차 및 실사오차와의 상관분석을 통해 조사거절률이 실사오차에 유의한 양의 상관관계를 갖고 있으며, 표본 추출률을 증가시키면 투표소추출오차가 감소한다는 사실도 확인할 수 있었다. 한편 출구조사를 진행한 세 개 조사기관별로 오차를 비교해 본 결과 특히 실사오차에 있어서 조사기관간의 차이가 통계적으로 유의한 것으로 나타났다.

주제어: 출구조사 예측오차, 투표소추출오차, 실사오차, 조사거절률.

In this study, we investigate the data of the exit poll for the 19th general election of 2012 to analyze the causes and characteristics of exit poll forecasting error. The empirical studies show that systematic negative bias of candidates affiliated to Saenuri Party is serious in the specific regions. To explore the main causes and effects of the forecasting error, the total survey error is divided by

* 본 연구는 숙명여자대학교 2011학년도 교내연구비 지원에 의해 수행되었음.

a) 숙명여자대학교 통계학과 박사과정

b) 밀워드브라운미디어리서치 상무이사

c) 교신저자(corresponding author): 숙명여자대학교 통계학과 교수 김영원.

E-mail: ywkim@sm.ac.kr

sampling error of the polling sites and non-sampling error which comes from selecting voter and collecting responses from sampled voters in each polling sites. And we consider relationships among several factors, which are anticipated as main causes of error, and non-sampling error as well as sampling error. As a result of correlation analysis, we can figure out that the non-sampling error of the exit poll mainly come from non-response. Also we consider the house effect in exit poll.

Key words: exit poll, forecasting error, sampling error, non-sampling error

I. 서론

선거예측조사는 선거 전에 전화를 이용한 조사를 통해 후보자의 득표율을 예측하는 전화여론조사와 투표 당일 투표소에서 투표를 마치고 나오는 사람들을 조사해 선거결과를 예측하는 출구조사로 구분할 수 있다. 본 연구에서는 2012년에 치른 19대 총선의 출구조사 자료를 이용해 해당 출구조사에서 발생한 오차의 특징과 함께 그 오차의 발생 요인에 대해 분석해 보고자 한다.

통계학적 관점에서는 선거예측조사에서 발생한 오차가 어느 정도 수준인지와 함께 오차를 발생시킨 요인이 무엇인지, 그리고 그 결과를 바탕으로 향후 선거예측조사에서 보다 정확한 예측을 하기 위해서는 어떤 점을 개선해야 할 것인지에 대해 주로 관심을 갖게 된다. 여기서 선거예측조사의 정확성은 특정 후보의 지지율에 대한 예측값과 참값(개표결과)의 차이로 설명할 수 있다. 출구조사의 경우 투표 당일의 투표자를 대상으로 조사가 이루어지며, 투표가 끝나는 6시에 방송을 통해 예측 결과를 발표하게 된다. 이 때 각 후보자에 대한 예측 득표율과 개표가 끝난 후 집계되는 최종 득표율과의 비교를 통해 출구조사의 정확성을 평가하게 된다.

정치권에서는 각 정당이 몇 개의 의석을 차지하게 될 것인지에 대해서, 또는 후보자 개인 입장에서는 당락 여부 등이 관심대상이 될 수 있지만, 이준웅(2004)이 지적한 것처럼 당선자를 잘못 예측한 선거구 수를 기준으로 선거예측조사의 정확성을 단순히 평가하는 것은 과학적인 접근으로 볼 수 없다. 실제 출구조사는 상당히 복잡한 조사

Efficiency of Variance Estimators for Two-stage PPS Systematic Sampling

Young-won Kim^{a,1} · Yeny Kim^a · Hye-eun Han^a · Eun-sun Kwak^a

^aDepartment of Statistics, Sookmyung Women's University

(Received November 13, 2013; Revised December 4, 2013; Accepted December 4, 2013)

Abstract

In this paper, we investigate several variance estimators for pps systematic sampling. Unfortunately, there is no unbiased variance estimators for a systematic sample because systematic sampling can be regarded as a random selection of one cluster. This study provides guidance on which variance estimator may be more appropriate than others in several circumstances. We judge the efficiency of variance estimators for systematic sampling based on of their relative biases and relative mean square error. Also, we investigate variance estimation problems for two-stage systematic sampling applied for the Food Raw Material Consumption Survey and the Establishment Labor Force Survey simulation study, in order to consider the popular two-stage pps systematic sample design for establishment and household survey in Korea.

Keywords: Two-stage systematic sampling, variance estimator, pps systematic sampling, relative bias, mean square error.

1. 서론

계통추출(systematic sampling)은 단순확률추출(simple random sampling)보다 추출작업이 쉽고 효과적으로 활용하는 경우 단순확률추출보다 정도(precision)가 높은 추정결과를 얻을 수 있기 때문에 널리 사용되는 표본추출방법이다. 계통추출은 집락추출의 일종으로 전체 추출단위를 k (추출간격)개의 집락으로 나누고 그 중 하나의 집락을 추출하는 것으로 볼 수 있다. 따라서 계통추출은 실행이 간단한 반면, 하나의 계통표본으로는 추정량의 분산에 대한 불편추정량을 구하는 것은 이론적으로 불가능하다는 문제가 있다. 흔히 계통표본을 단순확률표본으로 간주하고 단순확률추출의 분산공식을 사용하여 계통표본의 분산을 추정한다. 하지만 이 경우 분산추정량은 편향을 가질 수 있을 뿐만 아니라 구간추정 등 통계적 추론을 하는데 문제를 발생시킨다.

비복원 불균등확률추출에서 가장 널리 사용되는 방법인 크기비례(probability proportional to size; pps) 계통추출 역시 분산에 대한 불편추정량을 구할 수 없고, 또한 pps 계통표본의 경우 비복원 불균등확률추출에서 널리 사용하는 Yates와 Grundy (1953)의 분산추정량을 사용하는 데 어려움이 있다.

본 논문의 2절에서는 계통추출된 하나의 표본을 가지고 편향이 적은 분산추정량을 구현하기 위한 기존 연구들을 정리하고 각 추정량들의 장·단점에 대해 간단히 논의했다. 그리고 실제 표본조사에서 많이 활

This Research was supported by the Sookmyung Women's University Research Grants 2011.

¹Corresponding author: Professor, Department of Statistics, Sookmyung Women's University, 52 Hyochangwon-gil, Yongsan-gu, Seoul 140-742, Korea. E-mail: ywkim@sm.ac.kr

용하는 2단계 불균등계통추출에서 분산 추정방법을 정리했다. 마지막으로 3절에서는 실제 사업체 및 가구 대상 조사 자료를 이용한 모의실험을 통해 2단계 pps 계통추출에서 각 분산추정량의 효율성을 비교했다.

2. 계통표본의 분산 추정

계통추출은 집락추출의 일종이라 할 수 있다. 다시 말하면 계통표본은 추출간격에 해당하는 k 개의 집락 중에서 하나의 집락을 추출하는 특수한 형태의 집락추출법으로 볼 수 있다. 그러므로 하나의 계통표본을 가지고 추정량의 분산에 대한 불편추정량을 구하는 것은 이론적으로 불가능하다. 본 장에서는 하나의 계통표본을 가지고 모평균이나 총계 추정량의 분산 추정에 대한 기존 연구 결과들을 우선 정리하고자 한다 (Wolter, 2007).

2.1. 균등확률계통표본의 분산 추정

균등확률계통표본은 추출간격이 k 일 때, k 개의 가능한 표본 중에서 $1/k$ 의 확률로 하나의 계통표본이 추출되는 것을 말하며, 계통표본의 모평균에 대한 분산을 추정하는 가장 간단한 방법은 단순확률추출의 분산추정량을 이용하는 것이며 표본평균에 대한 분산추정량은 다음과 같다.

$$v_1(\bar{y}) = (1 - f) \left(\frac{s^2}{n} \right), \quad (2.1)$$

여기서 $s^2 = \sum_{i=1}^n (y_i - \bar{y})^2 / (n - 1)$, $f = n/N = p^{-1}$ 이다. 모집단 단위들이 랜덤하게 정렬된 경우 계통추출의 분산추정량과 단순확률추출의 분산추정량은 근사적으로 같다.

모집단이 n 개의 층으로 구성되어 있고 각 층의 크기는 k (추출간격)라고 가정하면, 계통표본은 각 층에서 표본단위를 하나씩 추출한 층화표본이라고 할 수 있다. 분산을 추정하기 위하여 두 개의 층을 하나로 묶어서 모집단을 총 $n/2$ 개의 층으로 간주하면, 계통표본을 각 층에서 단위를 2개씩 추출한 것으로 보고 다음과 같이 분산을 추정할 수 있다 (Cochran, 1946; Park, 2000).

$$v_2(\bar{y}) = \frac{N - n}{Nn^2} \sum_{i=1}^{\frac{n}{2}} (y_{2i} - y_{2i-1})^2, \quad (2.2)$$

여기서 v_2 는 조사단위의 중복 사용을 허용하지 않은 것이고, 만약 조사단위를 중복해서 사용함으로써 분산추정량의 자유도를 증가시키는 경우 분산추정량은 아래와 같다.

$$v_3(\bar{y}) = (1 - f) \left(\frac{1}{n} \right) \sum_{i=1}^{n-1} \frac{(y_{i+1} - y_i)^2}{2(n - 1)}. \quad (2.3)$$

하나의 계통표본을 2개 이상의 크기가 같은 부차계통표본으로 나누어 표본평균들의 분산을 구함으로써 계통표본의 분산을 추정할 수도 있다. 크기가 n 인 계통표본을 p 개의 크기가 같은 부차계통표본으로 나누는 경우 α 번째($\alpha = 1, \dots, p$) 부차표본의 평균과 분산추정량은 다음과 같다. 참고로 본 논문의 각 분산추정량에 대한 사례분석 모의실험에서는 v_4 는 이용하지 않았음을 미리 밝혀둔다.

$$v_4(\bar{y}) = \frac{1}{p(p - 1)} \sum_{\alpha=1}^p (\bar{y}_\alpha - \bar{y})^2, \quad \text{여기서 } \bar{y}_\alpha = \frac{p}{n} \sum_{i=1}^{\frac{n}{p}} y_{p(i-1)+\alpha} \quad (2.4)$$

한편 어떤 경우에서나 최적인 계통추출에 대한 분산추정량은 존재하지 않는다. Wolter (2007) 등이 지적한 것처럼 v_1 은 선형추세나 층화효과가 강한 모집단에서는 효율성이 떨어진다는 한계를 갖고 있고,

Accuracy of linear approximation for fitted values in nonlinear regression[†]

Myung-Wook Kahng¹

¹Department of Statistics, Sookmyung Women's University

Received 10 December 2012, revised 3 January 2013, accepted 8 January 2013

Abstract

Bates and Watts (1981) have discussed the problems of reparameterizing nonlinear models in obtaining accurate linear approximation confidence regions for the parameters. A similar problem exists with computing confidence curves for fitted values or predictions. The statistical behavior of fitted values does not depend on the parameterization. Thus, as long as the intrinsic curvature is small, standard Wald intervals for fitted values are likely to be sufficient. Accuracy of linear approximation for fitted values is investigated using confidence curves.

Keywords: Confidence curves, fitted values, intrinsic curvature, likelihood confidence region, parameter-effects curvature, subset curvature, Wald confidence region.

1. Introduction

A standard paradigm for the analysis of nonlinear models is to assume that results for the linear model hold at least approximately for large enough sample sizes. Confidence regions for parameters of a normal nonlinear regression model are commonly constructed by using linear regression methods, replacing the solution locus with the tangent plane at the maximum likelihood estimate. Such linear regions are generally easier to construct and comprehend than corresponding likelihood regions. Likelihood regions, on the other hand, are not influenced by parameter-effects nonlinearity and generally have true coverage closer to the nominal level than do linear regions. Under suitable regularity conditions and with a sufficiently large sample size, linear and likelihood regions will be in good agreement, but in any particular problem the strength of this agreement is uncertain.

Bates and Watts (1980) propose measures of intrinsic and parameter-effects curvature for assessing the adequacy of the linear approximation. These ideas are extended and refined by Cook and Goldberg (1986). They develop measures for assessing the agreement between linear and likelihood regions for an arbitrary subset of parameters from a nonlinear model. The subset curvatures developed in this paper appear to be reliable indicators of the adequacy of linear confidence regions for most nonlinear models. This ability to deal with subsets greatly extends the usefulness of the Bates-Watts methodology.

[†] This research was supported by the Sookmyung Women's University Research Grants 2011.

¹ Professor, Department of Statistics, Sookmyung Women's University, Seoul 140-742, Korea.
E-mail: mwkahng@sookmyung.ac.kr

To ensure good agreement between the tangent plane and likelihood regions, the maximum curvature must be considerably smaller than the Bates-Watts guide. However, this criterion can be too stringent for certain parameter subsets if whole-parameter curvatures are used. By contrast, the subset curvature describes the shape of the likelihood region in the parameter subspace of interest. Thus, the subset curvature is more directly relevant to the linearization adequacy question.

Cook and Weisberg (1990) give the graphical alternative to likelihood and Wald confidence intervals for a component of the single parameter vector. They discussed that their methodology can obtain the confidence curves for the parameter subsets but it has limitation for a function of parameters. Kahng (2003) considers obtaining graphical summaries of uncertainty in estimates of a function of parameters in nonlinear models. This method overcomes the limitation.

In this paper we will explore the accuracy of linearization inference for the fitted values through confidence curves. Section 2 considers curvature measures and confidence curves in nonlinear regression model. Accuracy of the linear approximation for the fitted values is investigated using curvature measures and confidence curves in section 3. We provide an example in section 4. Section 5 contains concluding remarks.

2. Nonlinear regression model

The standard nonlinear regression model can be expressed as

$$y_i = f(\mathbf{x}_i, \boldsymbol{\theta}) + \epsilon_i, \quad i = 1, \dots, n$$

in which the i -th response y_i is related to the q -dimensional vector of known explanatory variables $\mathbf{x}_i$ through the known model function f , which depends on the p -dimensional unknown parameter $\boldsymbol{\theta} \in \Theta$, and ϵ_i is error. We assume that f is twice continuously differentiable in $\boldsymbol{\theta}$, and errors ϵ_i are i.i.d. normal random variables with mean 0 and variance σ^2 . In matrix notation we may write,

$$\mathbf{y} = \mathbf{f}(\mathbf{X}, \boldsymbol{\theta}) + \boldsymbol{\epsilon} \quad (2.1)$$

where $\mathbf{y}$ is an n -dimensional vector with elements $y_1, \dots, y_n$, $\mathbf{X}$ is an $n \times q$ matrix with rows $\mathbf{x}_1^T, \dots, \mathbf{x}_n^T$, $\boldsymbol{\epsilon}$ is an n -dimensional vector with elements $\epsilon_1, \dots, \epsilon_n$, and $\mathbf{f}(\mathbf{X}, \boldsymbol{\theta}) = (f(\mathbf{x}_1, \boldsymbol{\theta}), \dots, f(\mathbf{x}_n, \boldsymbol{\theta}))^T = (f_1(\boldsymbol{\theta}), \dots, f_n(\boldsymbol{\theta}))^T = \mathbf{f}(\boldsymbol{\theta})$. Given the response vector $\mathbf{y}$, the least squares estimate of $\boldsymbol{\theta}$ is denoted $\hat{\boldsymbol{\theta}}$, minimizes the residual sum of squares $S(\boldsymbol{\theta})$ and maximizes the likelihood function. The predicted response vector is $\hat{\mathbf{y}} = \mathbf{f}(\mathbf{X}, \hat{\boldsymbol{\theta}}) = \mathbf{f}(\hat{\boldsymbol{\theta}})$. A tangent plane approximation to the expectation surface at $\hat{\boldsymbol{\theta}}$ is used to make inferences about $\boldsymbol{\theta}$ through the derived linear model $\mathbf{f}(\boldsymbol{\theta}) = \mathbf{f}(\hat{\boldsymbol{\theta}}) + \hat{\mathbf{V}}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$ where $\mathbf{V} = \mathbf{V}(\boldsymbol{\theta}) = \partial \mathbf{f} / \partial \boldsymbol{\theta}^T$ is the $n \times p$ matrix and $\hat{\mathbf{V}} = \mathbf{V}(\hat{\boldsymbol{\theta}})$ is $\partial \mathbf{f} / \partial \boldsymbol{\theta}^T$ evaluated at $\hat{\boldsymbol{\theta}}$.

2.1. Accuracy in linear approximation

A standard likelihood confidence region for $\boldsymbol{\theta}$ can be written as

$$\left\{ \boldsymbol{\theta} \mid S(\boldsymbol{\theta}) - S(\hat{\boldsymbol{\theta}}) \leq ps^2 F_\alpha(p, n - p) \right\} \quad (2.2)$$

Log-density Ratio with Two Predictors in a Logistic Regression Model

Myung Wook Kahng^{a,1} · Jae Eun Yoon^a

^aDepartment of Statistics, Sookmyung Women's University

(Received November 27, 2012; Revised December 31, 2012; Accepted January 18, 2013)

Abstract

We present methods for studying the log-density ratio that enables the selection of the predictors and the form to be included in the logistic regression model. Under bivariate normal distributional assumptions, we investigate the form of the log-density ratio as a function of two predictors. If two covariance matrices are equal, then the crossproduct and quadratic terms are not needed. If the variables are uncorrelated, we do not need the crossproduct terms, but we still need the linear and quadratic terms. We also explore other conditions in which the crossproduct and quadratic terms are not needed in the logistic regression model.

Keywords: Binary regression, inverse regression, kernel mean function, log-density ratio, log-odds, logistic regression.

1. 서론

일반적으로 회귀분석에서 반응변수는 특정한 구간 안에 있는 값이라면 어느 값이라도 취할 수 있는 연속형 자료이다. 그러나 회귀분석을 수행하는데 있어서 반응변수가 범주형인 경우를 흔히 볼 수 있다. 이때에는 오차가 정규분포를 따른다는 가정을 할 수 없기 때문에 일반적인 정규선형모형(normal linear model)을 사용하는데 무리가 있다. 이것을 해결해 주는 방법 중 가장 일반적인 것이 로지스틱회귀모형(logistic regression model)이다.

Nelder와 Wedderburn (1972)이 제안한 일반화선형모형(generalized linear model)은 정규이론을 따르는 선형모형을 지수족(exponential family)과 연결함수(link function)를 이용하여 다음과 같은 두 가지 과정으로 일반화 될 수 있다. 첫째, 반응변수의 기댓값과 설명변수의 선형결합(linear predictor)을 연결시키는 연결함수를 설정한다. 둘째, 오차의 분포는 정규분포를 포함하는 지수족의 여러 가지 분포를 사용한다.

로지스틱회귀모형은 확률변수 y 가 시행횟수가 m 이고 성공확률이 θ 인 이항분포를 따르는 경우 y/m 를 반응변수로 하고 이것의 기댓값 θ 와 주어진 설명변수들의 선형결합 $\mathbf{x}'\boldsymbol{\beta}$ 를 로짓(logit) 연결함수로 이어주는 일반화선형모형의 한 형태이다.

This research was supported by the Sookmyung Women's University Research Grants 2011.

¹Corresponding author: Professor, Department of Statistics, Sookmyung Women's University, Yongsan-Gu, Seoul 142-742, Korea. E-mail: mwkahng@sookmyung.ac.kr

선형회귀모형에서 반응변수의 기댓값은 설명변수들의 선형결합 $\mathbf{x}'\boldsymbol{\beta}$ 이라고 가정한다. 로지스틱회귀모형에서도 역시 설명변수들의 선형결합이 이용된다. 하지만 설명변수의 선형결합만으로는 충분히 설명이 되지 못하고 설명변수의 변환된 형태 등의 추가적인 포함이 필요한 경우가 있다. 이러한 연구는 로그-밀도비(log-density ratio)를 근거로 하여 Kay와 Little (1987)에 의하여 시작되었고 설명변수가 하나인 경우 Scrucca (2003), Kahng과 Shin (2012)의 연구가 있다.

본 논문에서는 설명변수가 두 개일 때 설명변수의 선형결합만으로 반응변수를 충분히 설명할 수 있는지 아니면 추가적으로 설명변수의 변환된 요소가 필요한지를 알아보고자 한다. 2절에서는 성공-오즈(odds of success)에 로그를 취한 로그-오즈(log-odds)와 로그-밀도비에 대해 알아본다. 3절에서는 로지스틱회귀모형에서 설명변수가 두 개일 때 이변량 정규분포에 근거한 로그-밀도비를 알아보고 이를 이용하여 두 설명변수에 추가하여 이차항과 교차항이 필요한 조건을 알아본다. 4절에서는 실제 자료를 이용하여 3절에서 알아본 조건을 확인해본다.

2. 로지스틱회귀모형에서 로그-밀도비

성공횟수를 나타내는 확률변수 y 가 시행횟수가 m 이고 성공확률이 θ 인 이항분포를 따른다고 하자. 반응변수를 성공비율인 y/m 로 설명변수를 $\mathbf{x} = (x_1, \dots, x_p)'$ 로 하는 이항회귀(binomial regression)의 모형은 $E(y/m) = \theta(\mathbf{x}) = g(\mathbf{x}'\boldsymbol{\beta})$ 로 표현된다.

이항회귀모형은 일반화선형모형의 한 형태로 $g(\cdot)$ 는 커널평균함수(kernel mean function)이고 연결함수의 역함수이다. 커널평균함수로 로지스틱함수(logistic function)를 사용하는 다음의 일반적인 로지스틱회귀모형을 생각하자.

$$E\left(\frac{y}{m} \middle| \mathbf{x}\right) = \theta(\mathbf{x}) = \frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1 + \exp(\mathbf{x}'\boldsymbol{\beta})} = g(\mathbf{x}'\boldsymbol{\beta}). \quad (2.1)$$

모형 (2.1)은 로짓 연결함수를 통하여 다음과 같이 선형모형의 형태로 표현된다.

$$\text{logit}(\theta(\mathbf{x})) = \log\left(\frac{\theta(\mathbf{x})}{1 - \theta(\mathbf{x})}\right) = \mathbf{x}'\boldsymbol{\beta},$$

여기서 $\theta(\mathbf{x})/(1 - \theta(\mathbf{x}))$ 를 성공-오즈라 부른다. 이제 성공할 경우 “1”을, 실패할 경우 “0”을 가지는 반응변수 y 를 생각하자. y 의 조건부분포는 기댓값이 $E(y|\mathbf{x}) = P(y = 1|\mathbf{x}) = \theta(\mathbf{x})$ 가 되는 베르누이분포를 따른다. 이항회귀모형은 다음과 같이 쓸 수 있다.

$$E(y|\mathbf{x}) = g(\boldsymbol{\eta}'\mathbf{u}). \quad (2.2)$$

모형 (2.2)에서의 커널함수는 모형 (2.1)에서와 같은 $\mathbf{x}$ 의 선형결합인 $\mathbf{x}'\boldsymbol{\beta}$ 가 아닌 $\mathbf{u}$ 의 선형결합인 $\boldsymbol{\eta}'\mathbf{u}$ 의 함수이다. $\mathbf{u} = \mathbf{u}(\mathbf{x})$ 는 p 개의 설명변수 $\mathbf{x}$ 로부터 구한 벡터이다. 일반적으로 $\mathbf{u}$ 는 $\mathbf{x}$ 의 함수로 구성된다. 모형 (2.1)에서와 같이 로지스틱함수를 커널평균함수로 하면 모형 (2.2)는 다음과 같이 로지스틱회귀모형이 된다.

$$E(y|\mathbf{x}) = \theta(\mathbf{x}) = \frac{\exp(\boldsymbol{\eta}'\mathbf{u})}{1 + \exp(\boldsymbol{\eta}'\mathbf{u})} = g(\boldsymbol{\eta}'\mathbf{u}),$$

여기서 $\boldsymbol{\eta}'\mathbf{u}$ 에 대한 방정식을 풀면, 다음과 같이 쓸 수 있다.

$$\text{logit}(\theta(\mathbf{x})) = \log\left(\frac{\theta(\mathbf{x})}{1 - \theta(\mathbf{x})}\right) = \boldsymbol{\eta}'\mathbf{u}.$$



CENTRAL LIMIT THEOREMS FOR REDUCED U -STATISTICS UNDER DEPENDENCE AND THEIR USEFULNESS

TAE YOON KIM¹, JEONGCHEOL HA¹, SUN YOUNG HWANG², CHEOLYONG PARK¹
AND ZHI-MING LUO^{1,*}

Keimyung University and Sookmyung Women's University

Summary

A reduced U -statistic is a U -statistic with its summands drawn from a restricted but balanced set of pairs. In this article, central limit theorems are derived for reduced U -statistics under α -mixing, which significantly extends the work of Brown & Kildea in various aspects. It will be shown and illustrated that reduced U -statistics are quite useful in deriving test statistics in various nonparametric testing problems.

Key words: central limit theorem; reduced U -statistic; α -mixing; nonparametric test.

1. Introduction

Let $X_1, X_2, \dots$ be a sequence of stationary α -mixing (which will be defined at the end of this section) random variables (abbreviated as rvs). Consider one sample U -statistic of rvs

$$U_n = \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} \Psi(X_i, X_j), \quad (1)$$

where Ψ is a symmetric function. Limit theories for U -statistics in (1) have been obtained in the literature, and consider either non-degenerate or degenerate U -statistics. The U -statistic U_n is considered degenerate with respect to F if

$$\int_{-\infty}^{\infty} \Psi(x, y) dF(y) = 0,$$

for all x , where F is the marginal distribution of X . It is well known that under the independent identically distributed (iid) assumption, a central limit theorem (CLT) holds for the non-degenerate U -statistics while the limit distribution changes to a linear combination

*Author to whom correspondence should be addressed.

¹Department of Statistics, Keimyung University, Daegu 704-701, Korea.

e-mail: zhimingluo@kmu.ac.kr

²Department of Statistics, Sookmyung Women's University, Seoul 140-742, Korea.

Acknowledgements. We are very grateful to the associate editor and the referees for their valuable comments which improved the manuscript significantly. TY Kim's research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (KRF-2011-0015936 and KRF-2009-0073872). J Ha's research has been conducted by the Bisa Research Grant of Keimyung University in 2010 and supported in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2011-0014741). SY Hwang's work was supported by the grant of NRF-2012012872. Z Luo's work was supported by the Bisa Research Grant of Keimyung University in 2012.

of independent, centered χ^2 distributions for the degenerate U -statistics. For discussion of this refer to Gregory (1977), Hall (1979) and Weber (1981). Meanwhile, it has been proved for iid r.v.s that the CLT may hold for degenerate U -statistics if the U -statistic has a variable kernel Ψ depending on n (see Hall 1984). These results have been extended to dependent r.v.s, as in Eagleson (1979), Abdalimov (1988), Khashimov (1993), Min & Abdurakhmanov (2000), Dewan & Prakas Rao (2001), Huang & Zhang (2006), Dehling & Wendler (2010) and Kim, Luo & Kim (2011).

In this article, we are mainly concerned with reduced U -statistics, which are formally defined by Blom (1976) as

$$U_{nr} = \frac{\sum_{1 \leq |i-j| \leq \kappa_n} \Psi(X_i, X_j)}{N(\kappa_n)},$$

where $N(\kappa_n)$ is the number of distinct pairs satisfying $1 \leq |i-j| \leq \kappa_n$ and $1 \leq \kappa_n \leq n$. Brown & Kildea (1978) obtained a CLT for non-degenerate reduced U -statistics with a constant κ_n when the random variables are assumed iid and then discussed using the Hodges-Lehmann estimator as a major application. They introduced reduced U -statistics mainly for efficient computation of the corresponding U -statistics without losing efficiency. The main objective of the current article is twofold. First, we derive CLTs for reduced U -statistics under dependence, which significantly extends Brown & Kildea (1978) in various ways. See Theorems 1 and 2 with remarks following. Second, applications are made to various nonparametric testing problems, which reveals the potential usefulness of reduced U -statistics particularly under dependence. For application of Theorem 1 we propose enhanced nonparametric statistics for testing various properties of residuals and squared residuals (see Examples 1 and 2). Example 3, as an application of Theorem 2, studies a nonparametric regression test problem in which it is critical to remove a sufficient amount of data for handling correlated errors.

To extend the CLT from Brown & Kildea (1978) we modify U_{nr} to

$$U_{n1} = \frac{\sum_{1 \leq |i-j| \leq \kappa_n} \Psi_n(X_i, X_j)}{N(\kappa_n)},$$

which replaces the fixed function Ψ of U_{nr} by a sequence of symmetric functions $\{\Psi_n, n \geq 1\}$. To handle U_{n1} under dependence, we need the α -mixing coefficients for triangular array of r.v.s to be given as follows.

Let $\mathcal{M}_a^b$ denote the sigma field generated by X_i , $a \leq i \leq b$. Thus, $\{X_i\}_{i=1}^\infty$, with values in a measurable space $(S, \mathcal{T})$ is said to be α -mixing if

$$\alpha(\tau) = \sup_{A \in \mathcal{M}_1^i, B \in \mathcal{M}_{i+\tau}^\infty} |P(A \cap B) - P(A)P(B)| \rightarrow 0$$

as $\tau \rightarrow \infty$. Let $\{\zeta_{n,t} | \zeta_{n,t} = f_n(X_t) \text{ for } n \geq 1, 1 \leq t \leq n\}$ be a triangular array of $\{X_i\}_{i=1}^\infty$ for measurable functions f_n . For each $n \geq 2$ and $\tau = 1, \dots, n-1$, let the strong mixing coefficient of $\zeta_{n,1}, \dots, \zeta_{n,n}$ be defined as

$$\alpha_n(\tau) = \sup_{1 \leq t \leq n-\tau} \sup_{A \in \mathcal{M}_{n,1}^t, B \in \mathcal{M}_{n,t+\tau}^n} |P(A \cap B) - P(A)P(B)|,$$

where $\mathcal{M}_{n,1}^t = \sigma(\zeta_{n,1}, \dots, \zeta_{n,t}) \subset \mathcal{T}$ and $\mathcal{M}_{n,t+\tau}^n = \sigma(\zeta_{n,t+\tau}, \dots, \zeta_{n,n}) \subset \mathcal{T}$ are two sigma fields generated by $\{\zeta_{n,1}, \dots, \zeta_{n,t}\}$ and $\{\zeta_{n,t+\tau}, \dots, \zeta_{n,n}\}$, respectively. By convention, we set $\alpha_n(\tau) = 0$ for $\tau \geq n$.

A Graphical Improvement in Volatility Analysis for Financial Series

Jeong Won Lee^a · Jae Eun Yoon^a · Sun Young Hwang^{a,1}

^aDepartment of Statistics, Sookmyung Women's University

(Received August 12, 2013; Revised August 26, 2013; Accepted August 26, 2013)

Abstract

News Impact Curves(NIC) developed by Engle and Ng (1993) have been useful for graphically representing the volatilities arising from financial time series. Adding an improvement and refinement to the original NIC, this article proposes so called two dimensional NIC and principal component NIC. We illustrate the methodology via Kosdaq data.

Keywords: News Impact Curve, principal component, volatility.

1. 서론

KOSDAQ, KOSPI, 환율과 같은 금융시계열의 변동성(volatility, 조건부 분산, h_t 로 표현)을 살펴보면 수익률에 따라 값이 변동하는 성질을 볼 수 있으며 과거 시점에서의 수익률에 따라 그 변동성이 달라진다. Engle과 Ng (1993)에 의해 처음 소개된 News Impact Curve(NIC)는 $t - 2$ 이전 시점까지의 정보는 상수로 고정시키고, $t - 1$ 시점의 수익률과 그 다음 시점의 변동성인 h_t 간의 관계를 그래프로 표현한 것이다. 이는 $t - 1$ 시점에서의 수익률이 한 시점 후의 변동성에 얼마나 영향을 주는지 측정할 수 있게 한다.

본 논문에서는 기존의 NIC에서 $t - 2$ 시점의 수익률을 추가한 그래프인 이차원 NIC(2D-NIC)를 소개하고, $t - 1$ 시점에서부터 $t - 5$ 시점까지의 수익률의 선형결합인 주성분 분석(Principal Components Analysis; PCA) 방법을 통하여 새로운 인공축인 C_1 (제1주성분, the 1st principal component)으로 나타낸 그래프(PCA in NIC)를 소개하고자 한다. 주성분 분석은 서로 상관관계가 있는 변수들 사이의 복잡한 구조를 좀 더 간단히 이해하기 쉽게 설명하기 위하여 사용하는 분석기법이다. 이는 변수들의 선형결합을 통하여 변수들이 가지고 있는 전체정보를 최대한 설명할 수 있는 서로 독립적인 새로운 인공변수(artificial variable)들을 유도하여 해석하는 기법이며 (Seong, 1997, Chapter 4) 금융시계열에서는 차원축소 방법으로 널리 이용되고 있다. 시계열 변동성의 특성을 잘 파악하기 위해서는 과거 수익률과 변동성과의 관계가 어떻게 나타나는지 쉽게 볼 수 있어야 하므로 본 논문에서 제시하는 새로운 그래프 기법들은 시각적으로 보기 쉽게 나타내기 위한 실용적인 방법으로서 응용성이 있으리라 판단하며, 국내 금융시계열 자료인 Kosdaq에 적용해 보고자 한다.

¹Corresponding author: Professor, Department of Statistics, Sookmyung Women's University, Yongsan-Gu, Seoul 142-742, Korea. E-mail: shwang@sookmyung.ac.kr

2. GARCH모형과 T-GARCH모형

금융시계열의 변동성(volatility, 조건부 분산, h_t 로 표현)은 변동성 집중(volatility clustering)현상, 비대칭성 현상, 급첨(leptokurtic) 등이 있다. 이러한 성질들을 다른 일반 시계열 자료에서 잘 나타나지 않는 것으로 이 같은 특징을 잘 나타내는 모형으로 Engle (1982)은 ARCH(autoregressive conditional heteroscedastic)모형을 제안하였다. 이어서, Bollerslev (1986)는 ARCH모형의 확장으로서 GARCH모형을 다음과 같이 제안하였다.

$$\epsilon_t = \sqrt{h_t}e_t,$$

$$h_t = \alpha_0 + \sum_{i=1}^p \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^q \beta_j h_{t-j}$$

여기서, $\alpha_0 > 0, \alpha_i \geq 0, \beta_j \geq 0$ 을 의미하고, $\{\epsilon_t\}$ 는 평균이 0이고 분산이 1인 *i.i.d* 과정이다. 위의 모형을 우리는 GARCH(p, q)로 나타낸다. GARCH모형에서 오차항이 제곱의 형태로 나타나 있으므로 현재의 조건부 분산에 대하여 오차항의 부호는 영향을 미치지 않으며 현재의 조건부 분산은 대칭인 형태를 보인다. 이는 변동성에 대한 “leverage effect”를 효과적으로 나타내지 못한다는 단점이 있다. 여기서 “leverage effect”는 비대칭 변동성 효과로서 같은 양이라면 수익률의 하락이 상승보다 더 큰 변동성을 제공한다는 의미이다. Rabemananjara와 Zakoian (1993)은 변동성의 비대칭성을 반영하는 모형, T-GARCH(Threshold GARCH)모형을 제시하였다. T-GARCH(1, 1)모형은 다음과 같다.

$$\epsilon_t = \sqrt{h_t}e_t,$$

$$h_t = \alpha_0 + \alpha_{11} (\epsilon_{t-1}^+)^2 + \alpha_{12} (\epsilon_{t-1}^-)^2 + \beta_1 h_{t-1}$$

여기서, $\alpha_{11} \geq 0, \alpha_{12} \geq 0, \epsilon^+ = \max(\epsilon, 0), \epsilon^- = \max(-\epsilon, 0)$ 이다. 참고로 $\epsilon = \epsilon^+ - \epsilon^-$ 와 $|\epsilon| = \epsilon^+ + \epsilon^-$ 의 성질을 지니고 있다. T-GARCH모형에서 모수 α_{11} 와 α_{12} 가 다른 값을 가짐으로 비대칭성을 반영할 수 있게 된다. 이 모형은 좋은 정보와 나쁜 정보 각각에 따라 다르게 반응함으로써 “leverage effect”를 설명하기에 효과적이다. 여기서 만일 $\alpha_{11} = \alpha_{12}$ 가 된다면, 위의 T-GARCH(1, 1)모형은 GARCH(1, 1)모형과 같게 된다.

2.1. News Impact Curve

Engle과 Ng (1993)에 의해서 처음 소개된 News Impact Curve(NIC)는 $t - 2$ 시점 이하는 모두 상수로 변환한 후에 t 시점에서의 변동성과 $t - 1$ 시점에서의 수익률간의 관계를 쉽게 보기 위하여 그래프로 표현한 것이다. 여기서 NIC는 2차원 좌표 평면에서 수익률 ϵ_{t-1} 은 가로축에, 변동성 h_t 는 세로축에 나타낸다. 이는 비대칭적 효과(leverage effect)를 쉽게 확인하는데 적합한 도구이다. 좋은 정보는 NIC에서 $\epsilon_{t-1} > 0$ 부분에서 변동성의 기울기를, 나쁜 정보는 $\epsilon_{t-1} < 0$ 부분에서 변동성의 기울기를 봄으로써 이를 시각적으로 쉽게 확인 할 수 있다. 만약 “leverage effect”가 존재 한다면, $\epsilon_{t-1} < 0$ 부분에서의 기울기가 $\epsilon_{t-1} > 0$ 에서의 기울기 보다 더 클 것이다.

GARCH모형에서의 NIC 방정식은 다음과 같다.

$$h_t = \alpha_0 + \alpha_1 \epsilon_{t-1}^2 + \beta_1 \cdot \frac{\alpha_0}{1 - \alpha_1 - \beta_1}$$

$$= \alpha_1 \epsilon_{t-1}^2 + \alpha_0 + \frac{\alpha_0 \beta_1}{1 - \alpha_1 - \beta_1}.$$

Chapter 2

Martingale Estimating Functions for Stochastic Processes: A Review Toward a Unifying Tool

S. Y. Hwang and I. V. Basawa

SY Hwang is currently Head of the Department and the Director of Research Institute of Natural Sciences, Sookmyung Womens University.

IV Basawa is a Prof. Emeritus in the Department of Statistics at the University of Georgia.

2.1 Introduction

Various methods of estimation such as least squares, method of moments, maximum likelihood, pseudolikelihood, and quasilikelihood have been studied extensively in the literature. Historically, each of the estimating methods was developed individually to suit particular situations and at varying points of time. Large sample theory (covering consistency and limit distributions of the estimates) was also developed for each of the methods using diverse tools and limit theorems suited to the individual method. Most of the early work on estimation was devoted to independent observations. More recently, methods and theory of estimation (inference in general) have been extended to cover dependent observations in stochastic processes. See, for instance, Basawa and Prakasa Rao (1980a, b), Basawa and Koul (1988), and Basawa (1983, 2001). Martingale estimating functions provide a unified framework which covers various estimation methods under a single setting. See, among others, Godambe (1985); Bibby and Sorensen (1995); Wefelmeyer (1996); Basawa et al. (1997); and Heyde (1997). More recent research on large sample theory for estimating functions is focused on developing a unified approach to establish asymptotic optimality and large sample comparison of estimates obtained from estimating functions. Martingale limit theorems have proved useful when establishing large sample

S. Y. Hwang (✉)

Department of Statistics, Sookmyung Women's University, Seoul, Korea

I. V. Basawa

Department of Statistics, University of Georgia, Athens, GA, USA

properties of estimates for dependent observations. This paper is an overview of current research by the authors (Hwang and Basawa 2011b) on martingale estimating functions and asymptotic optimality of parameter estimates for stochastic processes.

Section 2.2 presents a general formulation of martingale estimating functions which includes conditional least squares, quasilikelihood, maximum likelihood, and pseudolikelihood, as special cases. As an illustration, various estimates for a general class of generalized autoregressive conditional heteroskedasticity (GARCH)-type processes are presented in a unified way. Asymptotic optimality for a certain class of martingale estimating functions (MEFs) for ergodic processes is established via the convolution theorems in Sect. 2.3. Applications to conditional linear autoregressive processes, GARCH-type processes, and bifurcating autoregressive processes are then presented as examples of ergodic processes. Section 2.4 covers the extension of results of Sect. 2.3 to the nonergodic case. Branching Markov processes and explosive autoregressive processes are discussed to illustrate the nonergodic case. Finally, Sect. 2.5 gives a brief summary of the results and some concluding remarks.

2.2 Martingale Estimating Functions: A Formulation

Let $\{X_t, t = 0, 1, \dots\}$ denote a discrete time stochastic process defined on a probability space. Suppose that the probability measure P_θ associated with $\{X_t\}$ is indexed by a $(k \times 1)$ vector parameter θ . Assume that θ takes values in Θ which is an open subset of the k -dimensional Euclidean space. It is noted that P_θ needs not be parametric (in the sense that θ determines the underlying distribution). Rather, most of the theory in the paper is applicable to semiparametric or even nonparametric cases with a restriction depending on the parameter θ , allowing additional (infinite dimensional) nuisance parameter. Based on a sample of size n observations $X_1, X_2, \dots, X_n$, we are concerned with estimating the parameter vector θ . Consider the following $(k \times 1)$ estimating function (EF) $U_n(\theta)$ given by

$$U_n(\theta) = \sum_{t=1}^n u_t(\theta) \quad (2.1)$$

where $\{u_t(\theta)\}$ is a sequence of martingale differences, i.e., $E(u_t(\theta)|F_{t-1}) = 0$. Here, F_t denote the σ -field generated by $X_t, X_{t-1}, \dots, X_1$. We shall refer to $U_n(\theta)$ as the MEF. Assume for the moment that $\{X_t\}$ is strictly stationary and ergodic. Nonergodic cases will be discussed separately in Sect. 2.4. Fix $\theta \in \Theta$ and the local neighborhood $N_\delta(\theta)$ of the radius $\delta > 0$ about θ is defined by

$$N_\delta(\theta) = \{\theta^* ; \sqrt{n} |\theta^* - \theta| < \delta\}. \quad (2.2)$$

where and throughout, the vector (or matrix) norm will be simply denoted by $|\cdot|$, viz., for any vector or matrix A , $|A|^2 = \text{tr}(A^T A) = \text{tr}(A A^T)$. Here, A^T is the transpose A . The neighborhood $N_\delta(\theta)$ is to be further specified in Sect. 2.4 for discussing “non-ergodic” cases. It is assumed that the $(k \times 1)$ vector $u_t(\theta)$ is differentiable (with

두 가지 유형의 보험청구가 있는 확산과정 리스크 모형의 파산확률[†]

원호정¹ · 최승경² · 이의용³

¹²³숙명여자대학교 통계학과

접수 2012년 10월 22일, 수정 2012년 11월 14일, 게재확정 2012년 12월 3일

요 약

본 논문에서는 잉여금이 양의 추세모수를 갖는 확산과정을 따라 움직이고, 두 가지 유형의 보험청구가 있는 리스크 모형을 소개한다. 두 유형의 보험청구 금액은 서로 독립이고, 각각 지수분포를 따른다고 가정한다. 유형 I의 보험청구는 잦은 빈도로 발생하지만 청구 금액은 적고, 유형 II의 보험청구는 상대적으로 드물게 발생하지만 청구 금액이 많다고 가정한다. 적미분 방정식을 세워 잉여금이 없어지는 파산확률을 구하고, 각 유형에 의한 파산확률과 확산과정에 의해 자연적으로 파산이 이루어지는 확률을 함께 구한다. 또한 예제를 통해 두 유형의 보험청구와 확산과정이 전체 파산확률에 미치는 영향을 수치적으로 비교 분석한다.

주요용어: 리스크 모형, 잉여금, 적미분 방정식, 파산확률, 확산과정.

1. 서론

본 논문에서는 잉여금이 양의 추세모수를 갖는 확산과정을 따라 움직이고, 두 가지 유형의 보험청구가 있는 리스크 모형이 소개된다. 보험 상품의 잉여금은 $u > 0$ 에서 시작해서 추세모수 $c > 0$ 와 확산모수 σ^2 을 갖는 확산과정을 따라 움직이며, 포아송 과정을 따라 발생하는 두 가지 유형의 보험청구에 의해 감소한다. 유형 I에 의한 보험청구는 발생률 λ_1 인 포아송 과정 $\{N_1(t), t \geq 0\}$ 을 따라 발생하고, i 번째 청구 금액인 Y_i 는 서로 독립이고 평균이 $\mu_1 > 0$ 인 지수분포 $G(y) = 1 - e^{-y/\mu_1}, y > 0$ 를 따른다고 가정한다. 반면에 유형 II에 의한 보험청구는 발생률 λ_2 ($\lambda_2 < \lambda_1$)인 또 다른 포아송 과정 $\{N_2(t), t \geq 0\}$ 을 따라 발생하고, j 번째 청구 금액인 Z_j 는 서로 독립이고 평균이 $\mu_2 > 0$ ($\mu_2 > \mu_1$)인 지수분포 $H(z) = 1 - e^{-z/\mu_2}, z > 0$ 를 따른다고 가정한다. 이때 두 포아송 과정 $\{N_1(t), t \geq 0\}$ 와 $\{N_2(t), t \geq 0\}$ 는 서로 독립이고, Y_i 들과 Z_j 들도 서로 독립이라고 가정한다. 잉여금의 움직임을 나타내는 확산과정의 추세모수는 $c = (1 + \theta)(\lambda_1\mu_1 + \lambda_2\mu_2)$ 로 표현될 수 있다. 여기서 $\lambda_1\mu_1 + \lambda_2\mu_2$ 는 두 가지 유형의 보험청구에 의해 단위시간당 빠져나가는 기대 금액이고, $\theta > 0$ 는 상대적인 보완 부가금이다. θ 로 인해 잉여금의 추세모수 c 는 보험금의 지급률인 $\lambda_1\mu_1 + \lambda_2\mu_2$ 보다 크게 된다. 보험 상품의 잉여금 과정을 수식으로 표현하면 다음과 같다.

$$U(t) = u - \sum_{i=1}^{N_1(t)} Y_i - \sum_{j=1}^{N_2(t)} Z_j + W(t).$$

[†] 본 연구는 숙명여자대학교 2011학년도 교내연구비 지원에 의해 수행되었음.

¹ (140-742) 서울특별시 용산구 청파동 2가, 숙명여자대학교 통계학과, 석사.

² (140-742) 서울특별시 용산구 청파동 2가, 숙명여자대학교 통계학과, 박사.

³ 교신저자: (140-742) 서울특별시 용산구 청파동 2가, 숙명여자대학교 통계학과, 교수.

E-mail : eylee@sookmyung.ac.kr

여기서, $\{W(t), t \geq 0\}$ 은 추세모수가 $c > 0$ 이고, 확산 모수가 σ^2 인 브라운 운동 (Brownian motion)이다. 즉, $W(t)$ 는 정규분포 $N(ct, \sigma^2 t)$ 를 따른다.

잉여금 과정 $U(t)$ 의 표본경로 (sample path)를 그리면 Figure 1.1과 같고, 여기서는 잉여금이 확산과정에 의해 자연적으로 파산되는 경우를 나타내고 있다.

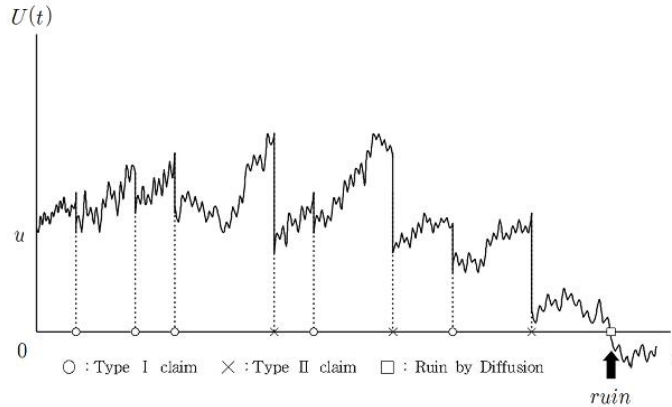


Figure 1.1 A sample path of surplus process $U(t)$

다양한 리스크 모형에서 잉여금의 파산확률에 대한 연구는 많은 연구자들에 의해 수행되어왔다. 한 가지 유형의 보험청구가 있는 고전적인 리스크 모형에서의 파산확률에 대한 연구 결과는 Klugman 등 (2004)에 잘 요약되어 있다. Dufresne와 Gerber (1991)는 잉여금이 확산과정을 따라 움직이고, 한 가지 유형의 보험청구가 있는 리스크 모형에서 잉여금의 파산확률을 연구하였다. 잉여금이 확산과정을 따라 움직인다는 가정은 향후 잉여금 움직임의 불확실성을 반영한 것으로, 이후 많은 연구자들의 연구 대상이 되었다.

예를 들어, Wang과 Wu (2000)는 Dufresne와 Gerber (1991)의 모형과 같은 리스크 모형에서 파산이 있기까지 잉여금의 최대치에 관련된 확률적 성질을 구하고, 파산 시 잉여금 수준의 분포에 대한 연구를 하였고, Zhang과 Wang (2003)은 파산까지의 시간, 파산 직전의 잉여금 수준, 파산 시 결손액의 결합확률분포를 구하였다. 또, Tsai (2009)는 확산과정이 있는 리스크 모형에서 보험청구의 빈도와 크기를 바꿔가면서 이들이 파산확률에 미치는 영향을 알아보았고, Tsai와 Lu (2010)는 확산과정이 있는 리스크 모형에서 파산확률의 바운드를 연구하였다. 하지만 이들 연구에서는 한 가지 유형의 보험청구만 고려되었다.

최근 들어, 두 가지 유형의 보험청구가 있는 리스크 모형이 연구되었다. Chan 등 (2003)은 두 가지 잉여금 과정이 있는 이 차원 리스크 모형을 제시하고, 이 차원 파산확률이 만족하는 적미분 방정식을 유도하였다. 하지만 명확한 해를 구하지는 못했다. Guo 등 (2007)도 두 개의 잉여금 과정을 고려하고, 이들의 합으로 표현되는 잉여금 과정의 파산확률을 연구하였다. Li와 Garrido (2005)와 Lv 등 (2010)은 두 가지 유형의 보험청구가 있는 리스크 모형을 좀 더 일반화하여, 한 유형은 포아송 과정을 따라 발생하고, 다른 유형은 재생과정을 따라 발생하는 경우에 전체 잉여금 과정의 파산확률을 연구하였다. Li와 Garrido (2005)는 파산확률의 라플라스 변환을 유도하였고, Lv 등 (2010)은 파산확률의 지수적 바운드를 구하였다.

두 가지 유형의 보험청구가 있는 리스크 모형의 예는 자동차 보험을 들 수 있다. 자동차 보험에서 보험청구는 대인보상의 유무에 따라 두 가지 유형으로 나눌 수 있다. 대물보상만 있는 보험청구는 빈도는

Transient and Stationary Analyses of the Surplus in a Risk Model

Eon Young Cho^a, Seung Kyoung Choi^a, Eui Yong Lee^{1, a}

^aDepartment of Statistics, Sookmyung Women's University

Abstract

The surplus process in a risk model is stochastically analyzed. We obtain the characteristic function of the level of the surplus at a finite time, by establishing and solving an integro-differential equation for the distribution function of the surplus. The characteristic function of the stationary distribution of the surplus is also obtained by assuming that an investment of the surplus is made to other business when the surplus reaches a sufficient level. As a consequence, we obtain the first and second moments of the surplus both at a finite time and in an infinite horizon (in the long-run).

Keywords: Risk model, surplus process, characteristic function, integro-differential equation, stationary distribution.

1. Introduction

In this paper, we consider a classical risk model in which the surplus at time $t > 0$ is given by

$$U(t) = u + ct - \sum_{i=1}^{N(t)} Y_i,$$

where, $u = U(0)$, c is the premium rate, $N(t)$ is the number of claims by time t , and Y_i is the amount of the i th claim. $\{N(t), t > 0\}$ is assumed to be a Poisson process of rate $\lambda > 0$ and Y_i 's are assumed to follow, independently and identically, distribution function G with mean $\mu > 0$. The premium rate c is usually assumed to be larger than $\lambda\mu$ which is the expected total amount of claims per unit time.

This risk model has been studied by many authors by assuming that a ruin occurs when the surplus becomes negative. They have obtained the ruin probability of the surplus and some related characteristics of the risk model. The core result on the ruin probability is well summarized in Klugman *et al.* (2004). The first passage time of the surplus to a certain level was introduced by Gerber (1990), thereafter, Gerber and Shiu (1997) obtained the joint distribution of the time of ruin, the surplus immediately before ruin and the deficit at ruin. Dickson and Willmot (2005) calculated the density of the time to ruin by an inversion of its Laplace transform.

However, until now, most works have been concentrated on the ruin probability of the surplus and its related characteristics. In this paper, we assume that the surplus process continues though it becomes negative and analyze stochastically the level of the surplus in the risk model. In Section 2, we derive the characteristic function of the level of the surplus at finite time $t > 0$, by establishing a partial

This research was supported by the Sookmyung Women's University Research Grants 2012.

¹ Corresponding author: Professor, Department of Statistics, Sookmyung Women's University, Cheongpa-ro 47-gil 100, Yongsan-Gu, Seoul 140-742, Korea. E-mail: eylee@sookmyung.ac.kr

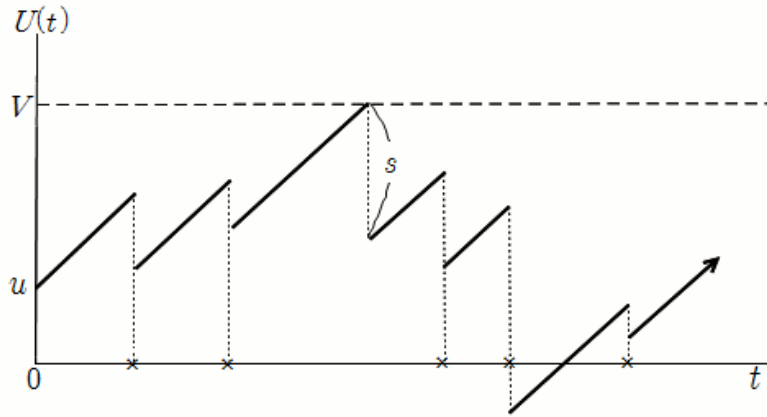


Figure 1: A sample path of the surplus process with investment

integro-differential equation for the distribution function of the surplus and solving the equation for it. As a consequence, we obtain the first and second moments of the surplus at finite time $t > 0$.

Since the premium rate is larger than the loss rate $\lambda\mu$, the surplus process goes eventually to infinity in the classical risk model. To analyze stochastically the level of the surplus in an infinite horizon, we assume that an investment of the surplus to other business is made by an amount s ($0 < s < V$), if the level of the surplus reaches a sufficient level $V > u$. This is a very practical assumption, since the large stock of the surplus increases the opportunity cost. The concept of the investment of the surplus was considered in Jeong *et al.* (2009) and Jeong and Lee (2010). They studied some optimal policies related to managing the surplus in the risk model.

The surplus process in the risk model with investments is illustrated in Figure 1.

In Section 3, we obtain the characteristic function of the stationary distribution of the surplus in the modified risk model by establishing and solving an ordinary integro-differential equation which does not depend on time t . By differentiating the characteristic function, we obtain the first and second moments of the surplus in the long-run (in an infinite horizon).

2. Transient Analysis of the Surplus

Let $F(x, t)$ be the distribution function of $U(t)$, the surplus at time $t > 0$, that is,

$$F(x, t) = P\{U(t) \leq x\}, \quad \text{for } -\infty < x < \infty.$$

Conditioning on whether a claim arrives in a small interval $[t, t + \Delta t]$, we can have the following relations between $U(t)$ and $U(t + \Delta t)$:

- (i) If no claims arrive, then

$$U(t + \Delta t) = U(t) + c\Delta t.$$

- (ii) If a claim arrives, then

$$U(t + \Delta t) = U(t) + c\Delta t - Y.$$

Hence, we can obtain the following equation for $F(x, t)$:

$$F(x, t + \Delta t) = [1 - \lambda\Delta t + o(\Delta t)] F(x - c\Delta t, t) + [\lambda\Delta t + o(\Delta t)] P\{U(t) - Y \leq x - c\Delta t\} + o(\Delta t).$$

STUDY PROTOCOL

Open Access

Yukmijihwang-tang for the treatment of xerostomia in the elderly: study protocol for a randomized, double-blind, placebo-controlled, two-center trial

Gajin Han¹, Jae-Woo Park², Seok-Jae Ko¹, Jihee Son², Jongki Seon¹, Juyeon Kim¹, Seulki Kim¹, Inkwon Yeo³, Bongha Ryu² and Jinsung Kim^{2*}

Abstract

Background: Xerostomia, a subjective sense of dry mouth, is not generally regarded a disease despite its high prevalence among the elderly, and therefore continues to impair affected patients' quality of life. In traditional Korean medicine, '*Yin-Deficiency*' has been implicated in the pathogenesis of xerostomia among the elderly. *Yukmijihwang-tang* is a famous herbal prescription used to relieve '*Yin-Deficiency*', and reportedly has antioxidant effects; therefore, it is postulated that *Yukmijihwang-tang* can be used to treat xerostomia in the elderly. However, to our knowledge, no clinical trial has been conducted on the effects of *Yukmijihwang-tang* on xerostomia. Thus, we designed a randomized clinical trial to investigate the effects and safety of *Yukmijihwang-tang* on xerostomia in the elderly. In addition, we will clarify the aforementioned assumption that '*Yin-Deficiency*' is the major cause of xerostomia in the elderly by identifying a correlation between xerostomia and '*Yin-Deficiency*'.

Methods/Design: This randomized, double-blind, placebo-controlled trial will be carried out at two centers: Kyung Hee University Korean Medicine Hospital and Kyung Hee University Hospital at Gangdong. We will recruit 96 subjects aged 60-80 years who have experienced xerostomia for 3 months prior to participation. Subjects who present with score >40 on the visual analogue scale for xerostomia and unstimulated salivary flow rate under 0.3mL/min will be included and the randomization will be carried out by an independent statistician by using a random number creation program. The subjects and all researchers except the statistician will be blinded to the group assignment. *Yukmijihwang-tang* or placebo will be administered to each group for 8 weeks. The primary outcome is change in the scores for the visual analogue scale for xerostomia and the dry mouth symptom questionnaire from 0 to 8 weeks.

Discussion: It will be assessed whether *Yukmijihwang-tang* can be used as a new herbal treatment for xerostomia in the elderly by demonstrating its therapeutic effects in a well-designed clinical trial.

Trial registration: ClinicalTrials.gov Identifier: NCT01579877

Keywords: Xerostomia, *Yukmijihwang-tang*, *Yin-Deficiency*, Visual analogue scale

* Correspondence: oridoc@khu.ac.kr

²Department of Internal Medicine, College of Korean Medicine, Kyung Hee University, Seoul, Republic of Korea

Full list of author information is available at the end of the article

Background

Xerostomia, which is a subjective sense of oral dryness [1], is a fairly common condition among the elderly [2]. It is estimated that approximately 30% of the population aged 65 years and older have xerostomia [3], and several studies have reported a higher prevalence in the elderly than in the younger population [4-7].

The causes of xerostomia are long-term uses of xerostomia-inducing medications for chronic diseases and systemic diseases (for example, Sjögren's syndrome, diabetes mellitus, Alzheimer's disease, and Parkinson's disease), head and neck radiation therapy, aging, heavy smoking, and stress [8]. In the elderly, the long-term use of medications and aging are regarded as major causes of xerostomia [9]. However, recent xerostomia studies mostly focused on Sjögren's syndrome or postradiotherapy-induced xerostomia, which are related to salivary gland damage, and only a few studies have assessed xerostomia in elderly subjects who did not exhibit obvious structural damage of the salivary glands [10,11].

Xerostomia is not generally accepted as a disease despite its high prevalence in the elderly; therefore, this condition continues to impair the affected patient's quality of life (QoL) [7,12]. For example, xerostomia in the elderly is known to cause dental problems and changes in chewing, swallowing, digestion, and taste [12]. Therefore, many xerostomia patients turn to solutions such as frequent water intake, sour foods, and alternative therapies (herbs or nutritional supplements) [13]. However, it is unclear whether these behaviors or therapies provide relief from xerostomia symptoms.

In traditional Korean medicine (TKM), 'Yin-Deficiency' is regarded as the main pathologic cause of xerostomia in the elderly [14]. Generally, the balance of *Yin* and *Yang* is important for maintaining health and treating diseases in TKM [14]. Particularly, *Yin* in 'Yin-Deficiency' refers to general body fluids, including blood, and *Jing* (mostly translated as 'essence of body') [15]. Therefore, 'Yin-Deficiency' is closely correlated with a shortage of body fluids, which results in various symptoms (dry mouth, dry eyes, or constipation) [14]. TKM has many herbal prescriptions for treating 'Yin-Deficiency' [10,16].

Yukmijihwang-tang (YMJ); *Liu-wei-di-huang-tang* in traditional Chinese medicine; *Lokumijio-to* in Kampo Medicine) is described in the ancient Chinese literature, 'Key to Therapeutics of Children's Diseases' [17], and is composed of six herbs. It has long been used for enriching the fluid-humor of the body in TKM, and it can treat *Deficiency Fire Flaming Upward* due to 'Yin-Deficiency'. Recently, many studies have described the pharmacological effects of YMJ, such as protection against renal ischemia/reperfusion in rats [18], memory enhancement in rats [19], and antioxidant and antidiabetic effects in early diabetic nephropathy patients [20]. In xerostomia-related aspects, YMJ is a

famous herbal formula for body enrichment, which includes relieving 'Yin-Deficiency' [21], because of its demonstrated antioxidant and free radical scavenging activities in animal studies [18,22]. Thus, it is postulated that YMJ can improve xerostomia in the elderly by enriching the *Qi* of the body and tonifying *Yin*.

However, to our knowledge, no clinical trial has examined the effects of YMJ on xerostomia. Therefore, we designed a randomized clinical trial to assess the effects and safety of the use of YMJ on xerostomia in the elderly. To evaluate any changes in xerostomia symptoms, the visual analogue scale (VAS) for xerostomia and the salivary flow rate will be analyzed.

Methods/Design

Objectives

The aims of this study are to:

- (1) obtain clinical evidence regarding the safety and effects of YMJ in elderly individuals with xerostomia;
- (2) identify a correlation between xerostomia in the elderly and the 'Yin-deficiency' state.

Hypothesis

We hypothesize that 8 weeks of YMJ administration will improve xerostomia in the elderly and will not induce severe side effects.

Design

This is an academic research study that will be conducted as a randomized, placebo-controlled, double-blind, two-center trial at Kyung Hee University Korean Medicine Hospital and Kyung Hee University Hospital at Gangdong in Seoul, Korea.

This clinical trial will consist of an 8-week oral administration of YMJ with two visits at 4-week intervals and a 2-week follow-up period. Before enrollment, all subjects will go through a 7-day run-in period.

After randomization, all subjects will be divided into two groups: the YMJ and placebo groups. Each group will be provided with YMJ or a placebo for 8 weeks and the subjects will be required to take 3 g of YMJ or placebo three times a day. During the 8-week administration period, subjects will be prohibited from taking any kind of xerostomia-relieving drug such as pilocarpine or cevimeline. Primary and secondary outcomes will be measured at baseline and at 4 weeks, 8 weeks, and 10 weeks after randomization.

The flow of the entire trial is shown in Figure 1.

Sample size calculation

The formula for estimating the sample size is as follows:

$$n_t = n_c = \left\{ (Z_{\alpha/2} + Z_{\beta})^2 \sigma^2 (\lambda + 1) / \lambda \right\} / (\mu_c - \mu_t)^2$$

Research Article

Availability of Tongue Diagnosis System for Assessing Tongue Coating Thickness in Patients with Functional Dyspepsia

Juyeon Kim,¹ Jiyoung Son,² Seungwon Jang,² Dong-Hyun Nam,³ Gajin Han,² Inkwon Yeo,⁴ Seok-Jae Ko,² Jae-Woo Park,² Bongha Ryu,² and Jinsung Kim²

¹ Department of Clinical Korean Medicine, Graduate School, Kyung Hee University, Seoul 130-701, Republic of Korea

² Department of Gastroenterology, College of Korean Medicine, Kyung Hee University, Seoul 130-702, Republic of Korea

³ Department of Biofunctional Medicine and Diagnosis, College of Korean Medicine, Sangji University, Wonju 220-702, Republic of Korea

⁴ Department of Statistics, Sookmyung Women's University, Seoul 140-742, Republic of Korea

Correspondence should be addressed to Jinsung Kim; oridoc@khu.ac.kr

Received 20 June 2013; Accepted 15 August 2013

Academic Editor: Raheleh Khorsan

Copyright © 2013 Juyeon Kim et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Tongue diagnosis is an important procedure in traditional Korean medicine (TKM). In particular, tongue coating thickness (TCT) is deemed to show the progression of the disease. However, conventional tongue diagnosis has limitations because of various external factors. Therefore, it is necessary to investigate the availability of tongue diagnosis system (TDS) in the assessment of TCT. This study has been designed as a prospective clinical trial involving 60 patients with functional dyspepsia. Tongue images will be obtained by TDS twice with a 30 min interval. The system will measure the percentage of TCT and classify it as either no coating, thin coating, or thick coating according to the existing diagnostic criteria. After finishing the collection of 60 patients' tongue images, TCT on the images will be simultaneously evaluated by the conventional method to establish the gold standard for assessing TCT by 5 well-trained clinicians. The evaluation will be repeated by the same clinicians after 2 weeks, but the order of the images will be changed. This trial is expected to provide clinical evidence for the availability of TDS as a diagnostic tool and to contribute to the standardization of the diagnosis system used in TKM. This trial is registered with ClinicalTrials.gov NCT01864837.

1. Introduction

Tongue diagnosis is an important diagnostic procedure that is used in traditional Korean medicine (TKM) to examine the physiological function and pathological changes of the internal organs. It involves the examination of the body of the tongue (its color, shape, moisture, and movement) and the coating on the tongue (its color, thickness, distribution, and characteristics at the root) [1, 2]. The coating is believed to reflect the condition of the stomach as well as the nature and the site of the pathogenic factors. In particular, tongue-coating thickness (TCT) is considered to reflect the progression of a pathological condition [3]. A conventional method used in tongue diagnosis involves direct examination of a protruded tongue by a practitioner, so the result is not only

affected by his or her subjectivity but also by environmental factors such as light [4]. Moreover, the available diagnostic criteria are insufficient and vague.

To overcome the above limitations, a number of tongue diagnosis systems (TDSs) have been developed in recent years [5]. In general, TDS has a number of components including image capturing and storage, color correction, tongue segmentation, and image analysis. The processes of image capturing and storage are hardware resources, while those of color correction and tongue segmentation are image-preprocessing steps in tongue diagnosis. In the process of image analysis, the tongue is classified according to selected diagnostic parameters [5, 6]. In this study, tongues will be classified based on the measurement of TCT into those with

no coating, thin coating, or thick coating according to the criteria developed by Kim et al. [7].

Functional dyspepsia (FD) is chronic or recurrent dyspepsia without evidence of organic disease [8]. The prevalence of FD is from 11.5% to 14.7% in the general population in Western countries [9], while in Republic of Korea, 25% of the population is affected as shown by a recent epidemiological survey [10]. We decided to include patients with FD in our study, because according to TKM, FD is one of the common stomach diseases that are well reflected by various features of tongue coating [11, 12].

For the objective and standardized tongue diagnosis system in TKM, it is necessary not only to develop an objective diagnostic tool but also to propose a new study protocol for clinical trial. We aimed to investigate the agreement between the results provided by TDS and those obtained by the conventional method in the assessment of TCT, which constitutes a part of tongue diagnosis in TKM, for evaluating on the availability of TDS in practice.

2. Materials and Methods

2.1. Study Setting and Participants. This study has been designed as a prospective clinical trial on the availability of TDS as a new diagnostic tool. Because there are no data available in the literature that would indicate the effective sample size to investigate the availability of TDS, we estimated the approximate sample size based on the number of outpatients with FD presenting at the Korean Medicine Hospital of the Kyung Hee University in Seoul, Korea.

The study will include a total of 60 patients aged 20 years and older who meet the Rome III criteria for FD [8]. Patients complaining of dyspepsia within the previous 3 months and reporting the onset of symptoms at least 6 months prior to enrollment will be included in the study. They should have 1 or more of the following symptoms: bothersome postprandial fullness, early satiation, and epigastric pain and burning, without evidence of structural disease (on upper endoscopic examination) that could explain the symptoms.

The exclusion criteria are as follows: pregnancy, mental disorders, severe systemic organ diseases such as cancer, infectious or other serious consumptive diseases, and geographic tongue. Patients unable to open their mouths or protrude their tongues; those unable to read, write, hear, or see; and those unable to sign a written informed consent form will also be excluded.

The study will be conducted in accordance with the standards of the International Committee on Harmonization of Good Clinical Practice and the revised version of the Declaration of Helsinki. The protocol of the trial has been approved by the institutional review board of the Korean Medicine Hospital of the Kyung Hee University (IRB no. KOMCIRB-2013-01). A written informed consent will be obtained from all patients before enrollment.

All patients will be required to visit the hospital twice during the study period. At the first visit, the written informed

consent will be obtained, and patients will be screened for eligibility criteria. All eligible patients will be then instructed to avoid mouth rinsing as well as food and liquid intake for 4 hours before the examination with the TDS at the next visit.

Tongue images will be obtained with the TDS under constant conditions at the second visit. Then, the percentage of tongue coating will be measured using a TDS analysis software, and TCT will be classified with the existing diagnostic criteria [7]. To assess the reproducibility of the TDS, the images will be taken again by the same method after 30 min. After two examinations of TDS, the weight of the tongue coating (mg) will be measured quantitatively to determine the correlation with the percentage of tongue coating.

After we complete the collection of 60 patients' tongue images, the images will be simultaneously evaluated by 5 clinicians with 5 years or more of clinical experience to establish the gold standard for the assessment of TCT. The evaluation will be repeated after 2 weeks to analyze intra- and interrater reliability of the clinicians. Finally, we will evaluate the agreement between the results provided by TDS and those obtained by clinicians. The flow of the trial is shown in Figure 1.

2.2. Outcomes

2.2.1. Primary Outcome. TDS will measure the percentage of tongue coating and classify TCT as either no coating, thin coating, or thick coating according to the existing diagnostic criteria [7]. At the same time, TCT shown on the images will be assessed by clinicians. We will then evaluate the agreement between the results of the TDS and the final diagnosis established by the clinicians.

2.2.2. Secondary Outcome. To evaluate the performance of TDS, we will assess its reproducibility by comparing the 2 measurements of TCT in each patient and analyze the correlation between the percentage of coating measured with the TDS and the weight of coating obtained by tongue scraping [13]. Additionally, the intra- and interrater reliability will be analyzed to ensure the validity of the results obtained by 5 clinicians in the assessment of TCT.

2.3. Components of TDS. The TDS consists of image acquisition system, LED illuminator, case, and analysis software. It is equipped with a vision camera (HVR-2130CPA, Hyvision System, Korea) and a mount lens (H2Z0414C-MP, Hyvision System, Korea). The camera can take color images of SXGA resolution (1280 × 1024 pixels) and automatically adjust white balance, exposure time, and gain settings. To evenly illuminate the dorsal surface of the tongue, 12 white LED lamps ($\varnothing$ 5 mm; first peak wavelength of 470 nm; second peak wavelength of 580 nm; color temperature of 6800 K) are arranged around the camera and thin double diffusion plates are placed in front of the lamps. The case is equipped with a part in contact with the face and a screw for adjusting the visual axis of the camera in order to take the tongue image

Research Article

The Effects of *Banha-sasim-tang* on Dyspeptic Symptoms and Gastric Motility in Cases of Functional Dyspepsia: A Randomized, Double-Blind, Placebo-Controlled, and Two-Center Trial

Jae-Woo Park,¹ Seok-Jae Ko,¹ Gajin Han,¹ Inkwon Yeo,² Bongha Ryu,¹ and Jinsung Kim¹

¹ College of Korean Medicine, Kyung Hee University, 26 Kyungheedaero, Dongdaemun-gu, Seoul 130-701, Republic of Korea

² College of Sookmyung Women's University, Cheongpa-dong 2ga, Yongsan-gu, Seoul 140-132, Republic of Korea

Correspondence should be addressed to Jinsung Kim; oridoc@khu.ac.kr

Received 7 April 2013; Revised 14 May 2013; Accepted 19 May 2013

Academic Editor: Chang Gue Son

Copyright © 2013 Jae-Woo Park et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Introduction. Functional dyspepsia (FD) is highly prevalent, and no standard treatments exist for this condition. Herbal prescriptions are widely used to treat FD. In traditional Korean medicine, *Banha-sasim-tang* (BST) is a famous herbal prescription for dyspepsia. This study aimed to evaluate the efficacy of BST and to examine the relationship between gastric slow waves and dyspeptic symptoms. **Materials and Methods.** In total, 100 FD patients were recruited; BST or placebo was administered for 6 weeks. The gastrointestinal symptom scale, FD-related quality of life scale, and frequency or power variables regarding gastric slow waves were measured at 0, 6, and 14 weeks. **Results.** There were no significant differences in the overall dyspeptic symptoms or quality of life between the BST and placebo groups. However, early satiety was significantly improved in the BST group ($P = 0.009$, at 6 weeks by intention-to-treat analysis). Abnormal gastric dysrhythmias and power ratios were also significantly improved by BST. **Conclusion.** BST had no significant effects on FD. However, early satiety appeared to improve after BST administration. Electrogastrography may be a useful technique for assessing changes in gastric motility dysfunction after interventions for FD. Further investigation focused on specific symptoms or subtypes of FD is required.

1. Introduction

Functional dyspepsia (FD) is chronic or recurrent dyspepsia without evidence of organic disease [1]. In western countries, the prevalence of FD is 11.5–14.7% in the general population [2]. In South Korea, an epidemiologic survey showed that 25% of the population was afflicted by FD [3].

At present, the causes of FD are believed to be multifactorial, and pathogenesis-based treatments, including administration of acid secretion inhibitors and prokinetics and eradication of *Helicobacter pylori*, have been used without success [4–6]. Therefore, many FD patients seek alternative and more effective treatments such as herbal prescriptions, nutritional supplements, acupuncture, or moxibustion [1, 7–9].

Banha-sasim-tang (BST; *Hange-shashin-to* in Kampo medicine; *Ban xia xie xin tang* in traditional Chinese medicine) is one of the most famous herbal prescriptions in the

old herbal prescription literature, where it is referred to as “*Shan-han-za-bing-lin* (傷寒雜病論)” [10]. In traditional Korean medicine, BST, which consists of 7 herbs, has been used to treat “gastric stuffiness” [10], a condition similar to dyspepsia. Recently, some studies have reported that BST has anti-inflammatory effects on the gastric mucosa and affects gastric function by influencing gut hormones or plasma peptides [11–13]. A recent clinical study investigating modified BST, conducted in China, demonstrated beneficial effects on only a subgroup of FD patients [14]. BST can be easily obtained without a prescription in Korea. Despite many previous experimental and clinical research studies, including a randomized controlled trial (RCT) in China [14], more reliable evidence about BST as a therapeutic alternative for FD are needed. However, few relevant RCTs have examined the effects of BST on FD or its mechanisms. The aim of

the present study was to investigate the effect of BST on dyspeptic symptoms and the quality of life in FD patients. To assess the changes in gastric motility dysfunction caused by BST, as a mechanical study, cutaneous multichannel electrogastrography (EGG), which detects gastric myoelectric activity (GMA), was used. The relationship between EGG frequency variables and FD symptoms was analyzed before and after oral administration of BST.

2. Materials and Methods

2.1. Study Design. This study was conducted as a randomized, placebo-controlled, double-blind, and 2-center trial at the Oriental Hospital, Kyung Hee University Medical Center, and at the Kyung Hee University Hospital at Gangdong in Seoul, Korea. The protocol of the trial was approved by the ethics review boards of both hospitals. The permission numbers were KOMC IRB 2009-05 for the Oriental Hospital and KHNMC-OH-IRB 2009-001 for the Kyung Hee University Hospital. Written informed consent was obtained from each participant before enrollment.

2.2. Study Subjects. Patients aged 19–75 and meeting the Rome III criteria for FD [15] were recruited from the clinics at both hospitals and from responders to local advertisements. In this trial, FD patients were classified into 1 of 3 subtypes: PDS, EPS, or mixed type: (1) meal-induced postprandial distress syndrome (PDS), characterized by postprandial fullness and early satiety; (2) epigastric pain syndrome (EPS), characterized by epigastric pain and burning; (3) PDS mixed with EPS [15]. The severity of dyspepsia was assessed using a validated gastrointestinal symptom (GIS) scale, which consists of 10 dyspeptic symptom subscales and is used to measure the severity of individual dyspeptic symptoms [16]. The existence of “moderate” severity on at least 3 subscales of the GIS resulted in the patient meeting the study’s inclusion criteria [1]. However, patients with medical histories of peptic ulcers, reflux diseases, previous abdominal surgeries, mental diseases, such as major depression, predominant irritable bowel syndrome, severe organ diseases, or continuous administration of analgesic agents were excluded as were lactating or pregnant women. In addition, patients who were using antibiotics, proton-pump inhibitors, bismuth salts, prokinetic agents, and herbal prescriptions were excluded, as previously described [17].

2.3. Randomization and Blinding. Randomization was performed by an independent clinical research coordinator (CRC). The randomization document with the subject’s basic information was transmitted by facsimile to the independent statistician (IKY) without a confirmed randomization number. The statistician determined the randomization number based upon the allocation sequence generated in advance by a random number program. Subsequently, the statistician returned the randomization document with the confirmed number to the CRC. The random allocation ratio for the 2 sites was 1:1. The randomization process was assured by

the authorized contract research organization (CRO; Marinert, Seoul, Korea), as previously described [17].

During the study, the investigators did not contact the CRC, the clinical pharmacist, or the statistician. In particular, the CRC was separated from all researchers to ensure that the researchers did not have any influence on enrollment or randomization. The statistician received the randomization document with the number and returned it to CRC in order. Thus, the statistician’s contact with other investigators was avoided. This blinding procedure was also verified by the authorized CRO.

Thus far, no standard therapy has been established for FD; therefore, the placebo used in this study did not have any active components. As experimental agents, Bansasin granules (Hanpoong Pharm & Food, Jeonju, Korea) in the current study are not artificial chemical therapeutics; a relevant placebo with an identical appearance, color, and flavor was required for effective blinding. After several attempts, a placebo that could not be distinguished from the real Bansasin granule by 6 healthy persons was successfully produced. Placebo was made of starch, lactose, brown caramel food coloring (BF2481, SaeRom BNF Co., Korea), and flavor (BF24781, SaeRom BNF Co., Korea) similar to color, flavor, and scent with Bansasin granule. Samples of the placebo were also kept by Hanpoong Pharm & Food. Additionally, at the end of the study, all subjects were asked whether the experimental agents that they had been given were real or placebo in order to evaluate the success of blinding.

2.4. Interventions. The intervention of the current study, BST (1/3 pack of herbal medicine (貼): pinelliae tuber (the rhizome of *Pinellia ternata* (Thunb.) Breit., family Araceae), 1.67 g; scutellariae radix (the root of *Scutellaria baicalensis* Georgi, family Labiatae), 1.00 g; ginseng radix (the root of *Panax ginseng* C.A. Meyer, family Araliaceae), 1.00 g; glycyrrhizae radix (the root of *Glycyrrhiza uralensis* Fisch., family Leguminosae), 1.00 g; zizyphi fructus (the fruit of *Zizyphus jujuba* Mill. var. *inermis* Rehder, family Rhamnaceae), 1.00 g; zingiberis rhizoma (the rhizome of *Zingiberis officinale* Roscoe, family Zingiberaceae), 0.83 g; coptidis rhizoma (the rhizome of *Coptis chinensis* Franch., family Ranunculaceae), 0.33 g) was extracted (0.91 g) with boiled water and mixed with starch (1.57 g) and lactose (0.52 g) then given the Bansasin granule (3 g). The manufacture was processed according to Korean Good Manufacturing Practice and permitted and regulated by the Korean Food & Drug Administration. The standard chemical components in the Bansasin granule (3 g) are berberin (11.6 mg), glycyrrhizin acid (25.0 mg), and baicalin (100.0 mg). Sample specimens were kept at the laboratory of Hanpoong Pharm & Food. Generally, 3 g of Bansasin granules 3 times per day is the recommended adult dosage for dyspeptic symptoms, including nausea, vomiting, diarrhea, abdominal pain, or anorexia [17].

This clinical trial consisted of a 6-week oral administration of BST or placebo (3 g, 3 times/day) and a 2-month follow-up period. Before randomization, all participants underwent a 7-day washout phase for elimination of any traces of previous medications. During the administration phase, subjects were prohibited from taking any

Prediction of the Number of Food Poisoning Occurrences by Microbes

In-Kwon Yeo^{a,1}

^aDepartment of Statistics, Sookmyung Women's University

(Received August 19, 2013; Revised October 15, 2013; Accepted October 28, 2013)

Abstract

This paper proposes a method to predict the number of foodborne disease outbreaks by microbes. The weekly data of food poisoning occurrences by microbes in Korea contain many zero-valued observations and have dependency between outbreaks. In order to model both phenomena, the number of food poisonings is predicted by an autoregressive model and the probabilities of food poisoning occurrences by microbes (given the total of food poisonings) are estimated by the baseline category logit model. The predicted number of foodborne disease outbreaks by a microbe is obtained by multiplying the predicted number of foodborne disease outbreaks and the estimated probability of the food poisoning by the corresponding microbe. The mean squared error and the mean absolute value error are evaluated to compare the performances of the proposed method and the zero-inflated model.

Keywords: Autoregressive model, baseline category logit model, zero-inflated negative binomial regression.

1. 서론

식품의약품안전처 조사에 의하면 우리나라에서 발생하는 식중독 중 60% 정도는 세균이나 바이러스로 인한 것으로 알려져 있으며 주요 원인균으로는 노로바이러스, 병원성 대장균, 살모넬라, 장염비브리오균, 황색포도상구균 등이 있다. 이들 균의 활성도는 온도나 습도와 같은 기후변수에 영향을 받는 것으로 알려져 있는데 이와 관련된 연구가 Bentham과 Langford (1995), Patrick 등 (2004), Magny 등 (2008)과 Fleuly 등 (2006)에 의해 수행되었으며 우리나라의 경우 Choi 등 (2008), Jung 등 (2012), Yeo (2012)에 의해 연구되었다.

Jung 등 (2012)의 분석에 의하면 원인균별 식중독 발생건수는 각기 다른 기후변수에 영향을 받고 있으며 대부분의 식중독균은 기온이 올라갈수록 활성도가 높아지는데 반하여 노로바이러스에 의한 식중독은 기온이 떨어지면 오히려 증가하는 추세를 가지는 것으로 나타났다. 이것은 원인균별로 기후변수의 영향력이 서로 상쇄되어 전체 식중독 건수를 예측하는데 기후변수의 설명력을 떨어뜨리는 요인으로 작용할 수 있다는 것을 의미한다. Yeo (2012)의 비교분석에 의하면 기후변수를 설명변수로 설정한 포아송회귀모형으로 예측했을 때보다 이전 식중독 발생건수를 설명변수로 가지는 자기회귀모형에 의한 예측력이

This research was supported by the research fund of Sookmyung Women's University (2012).

¹Professor, Department of Statistics, Sookmyung Women's University, Chongpa-dong 2ga, Yongsan-gu, Seoul 140-742, Korea. E-mail: inkwon@sm.ac.kr

훨씬 우수한 것으로 나타났다. 실제로 우리나라의 경우 단체급식이나 식당에서의 식중독 발생비율이 상대적으로 높다. 이는 위생 및 보건시설의 인프라 구축이나 행정적 조치 등에 영향을 받을 수 있으며 식중독이 발생했을 때 확산을 막을 조치를 취하지 않으면 전염성이 높은 식중독 경우 지속적으로 증가하는 경향이 있다.

식중독 예방 및 처치는 원인균에 따라 다르게 적용해야 하기 때문에 원인균별로 식중독 발생여부나 건수에 대한 예측은 매우 중요하다. 원인균별 식중독 발생건수에 대한 Jung 등 (2012)의 연구에서는 각 원인균별로 주요 변수를 파악하는데 도움이 되었지만 모형에 시계열적 요소를 포함하지 않고 있어 식중독 발생건수를 예측 하는데 한계가 있었다.

이 논문에서는 기후변수를 설명변수로 설정한 다범주 로짓모형으로 원인균별로 식중독 발생 비율을 추정하고 자기회귀모형을 이용하여 예측된 전체 식중독 발생건수를 곱해 원인균별로 식중독 발생 건수를 예측한다. 이에 대한 타당성을 확인해 보기 위해 Jung 등 (2012)에서 사용했던 영과잉 음이향회귀모형과 비교한다. 이를 위해 2003년 5월부터 2010년 4월까지 전국에서 신고된 주별 전체 식중독 발생건수와 원인균별 식중독 발생건수, 해당 주의 전국 60개 기상관측소에서 관측된 평균기온, 평균습도, 평균일조량 자료를 이용하여 모형을 추정하고 예측값의 평균제곱오차와 평균절대값오차를 통해 비교하고자 한다.

2. 원인균별 식중독 예측 방법론

먼저 기상자료가 식중독발생에 어떻게 영향을 주는지 알아보기 위해 기상자료의 추이와 식중독발생건수에 대한 시계열 그림을 그려보았다. Figure 2.1에서 평균기온, 습도, 일조량은 부분적으로 차이는 있지만 계절주기를 가지는 것을 볼 수 있다. 이에 반해 식중독 발생건수는 기상자료와는 상당히 다른 패턴을 보이고 있다. 이것은 일반화선형모형과 같은 모형으로 기상자료와 식중독 발생건수의 관계를 유도하면 좋은 결과를 얻을 수 없다는 것을 보여준다. Yeo (2012)는 교차상관관계를 통해 기후변수가 식중독 발생건수 간에 시차를 두고 영향을 주는 것을 확인했으나 기후변수와 이전 식중독 발생건수가 설명변수인 포아송회귀모형은 분석모형으로 적절하지 않은 것을 실증분석을 통해 보였다.

식중독 발생은 기후요인에 영향을 받지만 전염성이 있고 식재료관리와 조리법에 영향을 많이 받기 때문에 어떻게 대처하는가가 발생건수에 더 중요한 요인이 될 수 있다. 또한 Jung 등 (2012)에서 언급한 것처럼 원인균별로 기후변수에 대한 활성도가 다르고 어떤 경우에는 역으로 반응하는 경우가 있어 전체 식중독 발생건수를 예측하는데 어려움이 발생한다. 원인균별로 따로 분석할 경우 0의 값을 가지는 관측값이 상대적으로 많아져 영과잉모형을 사용해야 한다. 영과잉모형을 사용했던 Jung 등 (2012)의 분석결과를 보면 원인균별 식중독 발생건수의 예측에서 여전히 만족할 만한 결과를 보이고 있지 않다.

Yeo (2012)의 비교분석에서는 포아송회귀모형보다 자기회귀모형이 전체 식중독발생건수를 예측하는데 더 좋은 예측력을 가지는 것으로 나타났다. 이 논문에서는 자기회귀모형으로 전체 식중독발생건수를 예측하고 다범주로짓모형을 통해 해당시점에서 각 원인균별로 식중독이 발생할 확률을 추정하여 예측된 전체 식중독발생건수에 각각의 확률을 곱해 원인균별 식중독 발생건수를 계산하는 방법을 제안한다. 이론적 내용과 자세한 예측 과정은 다음과 같다.

식중독 원인을 p 개로 분류할 수 있고 시점 t 에서 j -번째 원인에 의해 발생한 식중독 건수를 Y_{jt} 라고 하자. 이 시점에서 발생한 전체 식중독 건수를 N_t 라고 하고 각각의 식중독 발생은 하나의 원인에 의해서만 발생한다고 가정하면 $N_t = \sum_{i=1}^p Y_{it}$ 가 된다. 각 원인균에 의한 식중독 발생건수는 서로 독립이고 평균이 μ_{jt} 인 포아송 분포를 따른다고 가정한다. 그러면 N_t 는 평균이 $\mu_t = \sum_{i=1}^p \mu_{it}$ 인 포아송 분포를 따르게 된다.

Research Article

Effect of Korean Herbal Medicine Combined with a Probiotic Mixture on Diarrhea-Dominant Irritable Bowel Syndrome: A Double-Blind, Randomized, Placebo-Controlled Trial

Seok-Jae Ko,¹ Gajin Han,¹ Seul-Ki Kim,¹ Jae-Gu Seo,² Won-Seok Chung,² Bongha Ryu,¹ Jinsung Kim,¹ Inkwon Yeo,³ Beom-Joon Lee,⁴ Jin-Moo Lee,¹ and Jae-Woo Park¹

¹ College of Korean Medicine, Kyung Hee University, 26 Kyungheedaero, Dongdaemun-gu, Seoul 130-701, Republic of Korea

² Cell Biotech, Co., Ltd., Gaegok-Ri, Wolgot-myeon, Gyeonggi-do, Gimpo-si 415-871, Republic of Korea

³ College of Sookmyung Women's University, Cheongpa-dong 2-ga, Yongsan-gu, Seoul 140-742, Republic of Korea

⁴ Kangnam Korean Hospital, Kyung Hee University, 26 Kyungheedaero, Dongdaemun-gu, Seoul 130-701, Republic of Korea

Correspondence should be addressed to Jae-Woo Park; pjw2907@khu.ac.kr

Received 18 April 2013; Revised 4 November 2013; Accepted 15 November 2013

Academic Editor: Kenji Watanabe

Copyright © 2013 Seok-Jae Ko et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Introduction. Although combination therapy with herbal medicine and probiotics is gaining popularity for controlling diarrhea-dominant irritable bowel syndrome (D-IBS) symptoms, few studies have investigated its clinical effects. **Materials and Methods.** Fifty-three patients with D-IBS were randomly allocated into 1 of the following 4 groups: herbal medicine (*Gwakhyangjeonggisang*; GJS) plus probiotics (Duolac7S; DUO), GJS plus placebo DUO, placebo GJS plus DUO, and placebo GJS plus placebo DUO. The study period consisted of a 2-week run-in, 8 weeks of administration, and 2 weeks of follow-up. The primary outcomes were weekly adequate relief (AR) of overall IBS symptoms and the proportion of responders (PR) during the administration period. The secondary outcomes included individual IBS symptoms, stool assessment, and quality of life. Changes of intestinal microbiota and intestinal permeability were also analyzed. **Results and Discussion.** Weekly AR was not different among the 4 groups throughout the treatment period. However, the 3 treatment groups exhibited significant improvements in PR compared to the findings in the placebo group. In the intestinal microbiota assessment, herbal medicine and probiotics synergistically increased beneficial bacteria counts. **Conclusion.** Combination therapy with herbal medicine and probiotics appears to relieve overall IBS symptoms by synergistically increasing beneficial intestinal microbe counts.

1. Introduction

Irritable bowel syndrome (IBS) is a functional gastrointestinal disorder characterized by abdominal pain, discomfort, and bowel disturbances without any structural abnormality [1]. IBS may cause significant inconvenience to patients, impair their social functioning, and deteriorate their quality of life [2]. The involvement of various factors in the pathophysiology of IBS makes treatment more difficult [3]. Factors such as imbalances of intestinal microbiota and increased intestinal permeability have been identified as important elements in the pathophysiology of IBS [4, 5]. Therefore,

therapeutic approaches aimed at resolving disturbances in the intestinal microbiota and maintaining mucosal barrier homeostasis can be helpful in the treatment of IBS. However, due to dissatisfying results with conventional IBS treatments, complementary therapies including herbal medicine and probiotics are becoming attractive options for many patients [6].

Herbal medicines have long been used in Asian countries due to their safety and having only a few side effects. *Gwakhyangjeonggisang* (GJS; *Kkako-shoki-san* in Kampo Medicine; *Huoxiang-zhengqi-san* in Traditional Chinese Medicine), a study agent in this trial, has been found to relieve abdominal pain, diarrhea, and vomiting as an

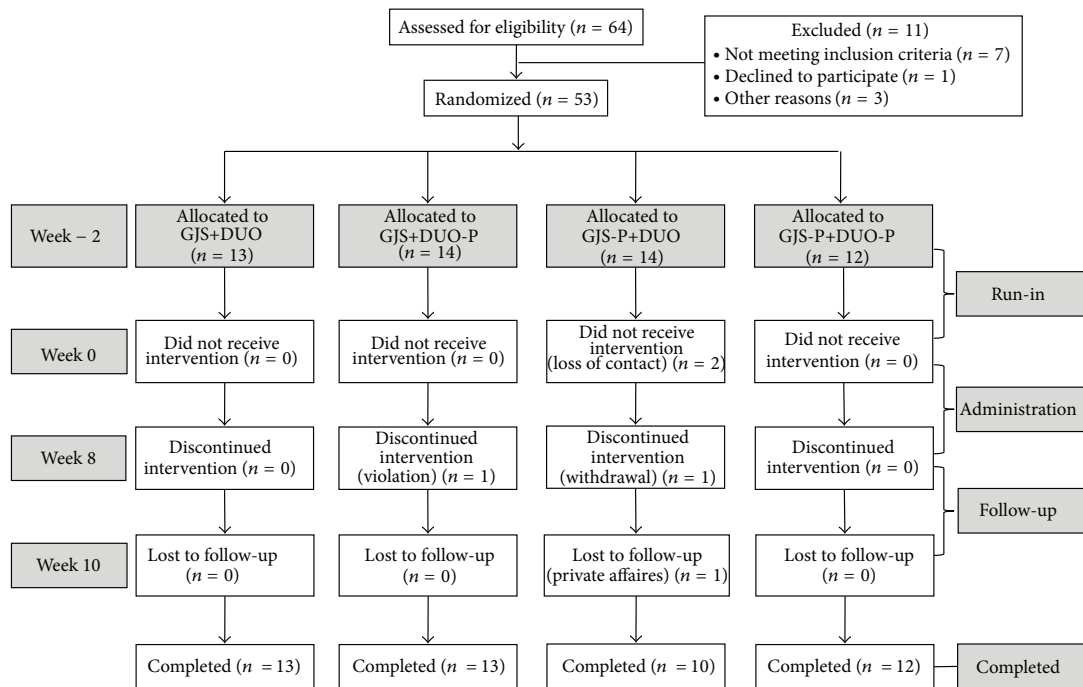


FIGURE 1: Flow chart of trial. GJS: *Gwakhyangjeonggis*, DUO: Duolac7S, GJS-P: placebo of *Gwakhyangjeonggis*, DUO-P: placebo of Duolac7S.

over-the-counter or prescribed medicine [7–9]. However, the clinical evidence supporting the efficacy of herbal medicine against IBS is weak [10]. Duolac7S (DUO), a probiotics mixture, has been also reported to have beneficial effects on IBS in a previous study [11]. Recently, the simultaneous administration of herbal medicines and probiotics has become a popular treatment for IBS in Korea [12]. However, no clinical studies have investigated the effect of combination therapy with herbal medicine and probiotics on IBS. In the present study, we evaluated and compared the effect of GJS combined with a multistrain probiotic mixture (DUO) on diarrhea-dominant IBS (D-IBS) symptoms to that of a placebo. In addition, intestinal permeability was assessed, and common strains of bacteria in the intestine were quantified using denaturing gradient gel electrophoresis (DGGE) and reverse transcription-polymerase chain reaction (RT-PCR) for mechanism analysis.

2. Materials and Methods

2.1. Participants. Sixty-four participants who met the criteria of D-IBS based on Rome III [1] were recruited at Kyung Hee University Hospital at Gangdong in Seoul. The inclusion and exclusion criteria for this study are shown in the protocol paper [13].

2.2. Study Protocol. The current study was conducted as a double-blind, placebo-controlled trial with 64 participants being randomly allocated to 1 of the following 4 groups: (1) the

real GJS and real DUO group (GJS+DUO), (2) the real GJS and placebo DUO group (GJS+DUO-P), (3) the placebo GJS and real DUO group (GJS-P+DUO), and (4) the placebo GJS and placebo DUO group (GJS-P+DUO-P). The participants completed a 2-week run-in (weeks –2 to 0), 8 weeks of administration (weeks 0–8), and a 2-week follow-up period (weeks 8–10). During the administration period, participants were to take 1 pack of GJS or its placebo 3 times a day (2 h after each meal) and 1 capsule of Duolac7S or its placebo 2 times a day (2 h after breakfast and dinner). Randomization was performed after an independent statistician screened participants by using random allocation numbers from a random number creation program. The investigator, clinical research coordinator (CRC), clinical pharmacist, and participants were blinded to randomization until the end of the study. We calculated the sample size based on previous similar studies [14, 15], as the study agent had never been studied, and determined that 64 participants were necessary because a sample size of 48 was regarded as the minimal number of participants needed for clinical significance, assuming a 25% dropout rate. This sample size provided 80% power to demonstrate the superiority of the study agents to placebos. The flow of the entire trial is described in Figure 1.

Informed consent was properly acquired prior to the trial at week –2. Participants were required to record their daily symptoms, responses to medication, and potential adverse effects. Safety was examined by blood testing at the end of the administration period (week 8). Compliance was calculated by determining the amount of medication

An Empirical Characteristic Function Approach to Selecting a Transformation to Symmetry

In-Kwon Yeo and Richard A. Johnson

Abstract Somewhat surprisingly, the empirical characteristic function can provide the basis for selecting a transformation to achieve near symmetry. In this paper, we propose to estimate the transformation parameter by minimizing a weighted squared distance between the empirical characteristic function of transformed data and the characteristic function of a symmetric distribution. Asymptotic properties are established when a random sample is selected from an unknown distribution. We also propose the method to determine the influence functions for vector-valued estimators based on M -equations taking the form of U -statistics. According to our influence function calculations, these estimates are less sensitive to a few outliers than the maximum likelihood estimators when the target distribution is normal. A small Monte Carlo simulation shows transforming data by our method lead to more symmetry than those by the maximum likelihood method when the population has heavy tails.

1 Introduction

Many statistical techniques are based on assumption about the form of population distribution. The validity of those results may depend on the assumed conditions being satisfied. When the observed data seriously violate these assumptions, transformation of data can improve the agreement with the assumption about underlying distribution. Since an objective way of determining a transformation was introduced

In-Kwon Yeo

Department of Statistics, Sookmyung women's University, Seoul 140-742, Korea, e-mail: inkwon@sm.ac.kr

Richard A. Johnson

Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706, USA, e-mail: rich@stat.wisc.edu

by Box and Cox (1964), transformation of data has widely used in applied statistics as well as theoretical statistics.

Generally, a main goal of transforming data is to enhance the normality and homoscedasticity of data. Box and Cox (1964) discussed estimating transformation parameter by the maximum likelihood approach and by a Bayesian method. It is well-known that, under the normality assumption, the maximum likelihood estimator of the Box-Cox transformation parameter is very sensitive to outliers, see Andrews (1971). Carroll (1980) proposed a robust method for selecting a power transformation to achieve approximate normality in a linear model.

Robust techniques sometimes require symmetry rather than normality of data. Hinkley (1975) and Taylor (1985) suggested methods for estimating the transformation parameter in the Box-Cox transformation when the goal is to obtain approximate symmetry. Yeo and Johnson (2001) and Yeo (2001) introduced an M -estimator which is obtained by minimizing the integrated square of the imaginary part of the empirical characteristic function of Yeo-Johnson transformed data.

Many authors including Koutrouvelis (1980), Koutrouvelis and Kellermeier (1980), Fan (1997), Klar and Meintanis (2005), and Jimenez-Gamero et al. (2009) have proposed the goodness-of-fit test statistics based on measuring differences between the empirical characteristic function and the characteristic function in the null hypothesis.

Our estimators are obtained by minimizing a squared distance between the empirical characteristic function of the transformed data and the target characteristic function. Specifically, we minimize the integral of the squared modulus of the difference of the two characteristic functions multiplied by a weight function. This estimation procedure for a vector-valued parameter, can be viewed as solving estimating equations based on a U -statistic, see Lee (1990). In order to calculate the influence function, we take the approach of calculating this function in terms of the asymptotically equivalent statistic that is a sum of independent and identically distributed terms. The resulting expressions show that our procedure is more robust than the maximum likelihood procedure under normality.

2 Estimation

Let $\psi(x, \lambda)$ be a general class of transformations which are indexed by the transformation parameter λ . Examples include the families introduced by Box and Cox (1964), John and Draper (1980), Burbidge, Magee, and Robb (1988), and Yeo and Johnson (2000). Based on calculations of the relative skewness by van Zwet (1964), the Box-Cox transformation and the Yeo-Johnson transformation can improve the symmetry of data and be applied to skewed data. By contrast, the modulus transformation by John and Draper (1980) and the inverse hyperbolic sine transformation by Johnson (1949) and Burbidge, Magee, and Robb (1988) are useful to reduce the kurtosis of heavy-tailed data. Hence, we focus on the Box-Cox transformation, for $x > 0$,

Nonparametric Two-Sample Tests of Longitudinal Data with Termination Events

YANG-JIN KIM¹, CHUNG MO NAM²,
AND JINHEUM KIM³

¹Department of Statistics, Sookmyung Women's University, Seoul, Korea

²Department of Preventive Medicine, Yonsei University
College of Medicine, Seoul, Korea

³Department of Applied Statistics, University of Suwon,
Gyeonggi-Do, Korea

In longitudinal data, observations of response variables occur at fixed or random time points, and can be stopped by a termination event. When comparing longitudinal data for two groups, such irregular observation behavior must be considered to yield suitable results. In this article, we propose the use of nonparametric tests based on the difference between weighted cumulative mean functions for comparing two mean functions with an adjustment for difference in the timing of termination events. We also derive the asymptotic null distributions of the test statistics and examine their small sample properties through simulations. We apply our method to data from a study of liver cirrhosis.

Keywords Liver cirrhosis data; Marked point processes; Termination event times; Weighted cumulative mean function.

Mathematics Subject Classification Primary 60G55; Secondary 60G46.

1. Introduction

Longitudinal datasets involve the repeated measurement of response variables over time. As parametric models, linear mixed effects models and the marginal models have been applied for the analysis of longitudinal data (Diggle et al., 1994). Nonparametric models using techniques such as kernel, spline, and local spline smoothing methods have also been suggested for investigating a pattern of change in the response variable over time (Wu and Zhang, 2006). When collecting longitudinal datasets, response measurements are sometimes stopped due to a terminal event such as death or censoring.

Received August 31, 2010; Accepted May 20, 2011.

Address correspondence to Jinheum Kim, Department of Applied Statistics, University of Suwon, San 2-2, Wau-Ri Bongdam-Myeon, Gyeonggi-Do 445-743, Korea; E-mail: jinhkim@suwon.ac.kr

For the analysis of longitudinal data with terminal events, Robins et al. (1995) proposed a class of inverse probability of censoring weighted estimators for parameters that define the mean response. Their approach used an extension of the generalized estimating equations proposed by Liang and Zeger (1986). Rotnitzky et al. (1998) extended this concept to more general observation schemes. In the random effects joint model, Wulfsohn and Tsiatis (1997) combined a linear Gaussian model for longitudinal responses and a proportional hazards model for the terminal event times. Xu and Zeger (2001) proposed a similar model, but their model was adapted to the setting of generalized linear models. Henderson et al. (2000, 2002) developed a more general version of the random effects joint model. However, all of these methods assume that response variables are measured at fixed or defined times. Sometimes response variables are measured at irregular times, and for these situations the methods above cannot be applied. Therefore, we consider a measurement process that accounts for random measurement times and the occurrence of terminal events. The aim of this article is to compare the mean functions of two groups with different measurement process and different terminal event process. We define a weighted cumulative mean function to include information of a terminal event distribution, and derive its asymptotic properties using general martingale techniques. Finally, we propose test statistics based on the difference between weighted cumulative mean functions of two groups.

The remainder of this article is organized as follows. Sec. 2 reviews marked point processes which simultaneously incorporate measurement process and mark process (response variable). Sec. 3 suggests one-sample nonparametric estimation of the weighted cumulative mean function and extends to two-sample test statistics for comparing two mean functions with an adjustment for different terminal event time distributions. In Sec. 4, we report the results of a simulation study to evaluate the suggested method and the application of this method to a liver cirrhosis dataset. We conclude with a discussion in Sec. 5.

2. Marked Point Processes

We introduce the notation and a general review of marked point processes. For the subject i , let $T_{i1} < \dots < T_{iL_i}$ be random observation times and $Z_i(T_{i1}), \dots, Z_i(T_{iL_i})$ the response variables measured at the respective observation times $T_{i1}, \dots, T_{iL_i}$, where $i = 1, \dots, n$ and L_i is a random last observation number. This constitutes a marked point process characterized by the jump time and the associated mark. The marked point process generates a sequence of pairs $\{(Z_{ij}, T_{ij}) | i = 1, \dots, n; j = 1, \dots, L_i\}$, where $Z_{ij} = Z_i(T_{ij})$. Let $(R, \mathcal{B})$ denote a mark space, where $\mathcal{B}$ denotes the Borel σ -field on $R = (-\infty, \infty)$. For each $A \in \mathcal{B}$, the counting process is defined as

$$N_i^*(A, t) = \sum_{j=1}^{L_i} I(Z_{ij} \in A) I(T_{ij} \leq t),$$

which counts the number of observation times with associated marks in A until time t , where $I(\cdot)$ is the indicator function. As a special case, the simple counting process with no marks is defined with $A = R$, that is, $N_i^*(t) = N_i^*(R, t)$. The marked point processes can also be identified with the counting measure $p_i^*(ds \times dz)$ defined by

$$p_i^*((0, t] \times A) = N_i^*(A, t), \quad A \in \mathcal{B}.$$

Statistical Analysis of Bivariate Recurrent Event Data with Incomplete Observation Gaps

Yang-Jin Kim^{1,a}

^aDepartment of Statistics, Sookmyung Women's University

Abstract

Subjects can experience two types of recurrent events in a longitudinal study. In addition, there may exist intermittent dropouts that results in repeated observation gaps during which no recurrent events are observed. Therefore, these periods are regarded as non-risk status. In this paper, we consider a special case where information on the observation gap is incomplete, that is, the termination time of observation gap is not available while the starting time is known. For a statistical inference, incomplete termination time is incorporated in terms of interval-censored data and estimated with two approaches. A shared frailty effect is also employed for the association between two recurrent events. An EM algorithm is applied to recover unknown termination times as well as frailty effect. We apply the suggested method to young drivers' convictions data with several suspensions.

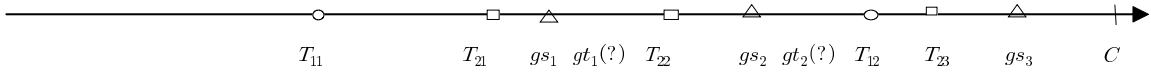
Keywords: Bivariate recurrent event data, frailty effect, observation gap, piecewise constant.

1. Introduction

A bivariate recurrent event arises when two different types of events repeatedly occur. Like other types of multivariate failure times, a frailty effect approach (Cook *et al.*, 2010) and a marginal approach (Cai and Schaubel, 2004) are applied to analyze bivariate recurrent event data. In this paper, we consider a special situation occurring in a longitudinal study where a subject is assumed to be followed until the termination of study. However, some subjects can leave the study and then return to the study at a later time. Such an intermittent missing is denoted as the observation gaps (Zhao and Sun, 2006) and is regarded as a non-risk status at statistical analysis. Furthermore, risk status becomes unclear when such observation gaps result in an incomplete form. As a motivation of this paper, Young Traffic Offence Program (YTOP) data included 40 subjects who experienced at least one suspension and the suspended subjects were banned from driving cars during certain periods. Once the suspension began, subjects dropped out of the risk state and came back to the risk state after completing the suspension. Thus, the observation gap was determined by both start and termination of suspensions and the duration of observation gap depends on the terminating time of suspension. With a complete information about observation gap, Therneau and Hamilton (1997) remarked on discontinuous intervals at risk in recurrent event data and utilized a counting process technique. For similar problems, Duchateau *et al.* (2003) applied a gap time scale to analyze asthma data with the non-risk period. However, the YTOP data provides only the starting time of the suspension while the terminating time is not available. Instead, the terminating time was placed somewhere between the starting time and the recurrent event time occurring subsequently after suspension. To estimate covariate effects on bivariate recurrent event, a frailty is incorporated and an EM algorithm is applied to recover two kinds of incomplete

This Research was supported by the Sookmyung Women's University Research Grants 2012.

¹ Assistant Professor, Department of Statistics, Sookmyung Women's University, Chungpa-Dong, Yonnsan-Gu, Seoul 140-742, Korea. E-mail: yjin@sookmyung.ac.kr



- : recurrent event of type1
- : recurrent event of type2
- △ : suspension starting time

Figure 1: Interval censored termination time of suspension with bivariate recurrent events

data: frailty effects and termination times. The reminder of this paper is organized as follows. Section 2 describes models and notations and Section 3 develops an estimation procedure. In Section 4, the suggested method is applied to the YTOP dataset. Concluding remarks appear in Section 5.

2. Model

Consider n subjects with two different types of recurrent events. Denote T_{1ij} and $T_{2ij'}$ as the j^{th} and j'^{th} recurrent event time of two types of events of subject i , where $i = 1, 2, \dots, n$, $j = 1, \dots, n_{1i}$ and $j' = 1, \dots, n_{2i}$. Define $N_{1i}(t) = \sum_j I(T_{1ij} \leq t)$ as a counting process recording the cumulative number of the first event of subject i with $dN_{1i}(t) = 1$ indicating subject i has experienced the first event at time t . Similar definitions are applied to $N_{2i}(t)$ and $dN_{2i}(t)$ for the second recurrent event. Assume that two events follow proportional hazard rates with covariates vectors, $x_{i1}(t)$ and $x_{i2}(t)$, respectively. Given shared frailty v_i and covariates vectors, the intensity functions are

$$\begin{aligned}\lambda_{1i}(t|x_i, v_i) &= Y_i(t) \lambda_{01}(t) \exp(\beta'_1 x_{i1}(t) + v_i), \\ \lambda_{2i}(t|x_i, v_i) &= Y_i(t) \lambda_{02}(t) \exp(\beta'_2 x_{i2}(t) + v_i),\end{aligned}$$

where $Y_i(t) = I(t \leq C_i)$ with a censoring variable C_i indicates at risk process and is a predictable process. $\lambda_{01}(t)$ and $\lambda_{02}(t)$ are arbitrary baseline intensity functions with cumulative functions $\Lambda_{01}(t)$ and $\Lambda_{02}(t)$, respectively. In addition, β_1 and β_2 are regression coefficients of covariate vectors for the first and second recurrent events, respectively. For modeling a correlation between recurrent events, a shared frailty, v_i is assigned and assumed to be independent and identically distributed with $N(0, \sigma^2)$. In order to incorporate observation gaps, let $G_i = \{g_{i1}, \dots, g_{im_i}\}$ be the set of suspensions for subject i and every subject has a different suspension number, $(0 \leq m_i < \infty)$. Each g_{ik} consists of two time points under suspension, $g_{ik} = [gs_{ik}, gt_{ik}]$, where gs_{ik} and gt_{ik} are the starting time and terminating time of the k^{th} suspension of subject i , respectively. Then the risk indicator is redefined as

$$Y_i^*(t) = I(t \leq C_i, \text{ and } t \notin g_{ik} = [gs_{ik}, gt_{ik}], k = 1, \dots, m_i).$$

An incompleteness of G_i makes it impossible to use Y_i^* directly because of unknown terminating times. Figure 1 shows two types of censoring for termination times of observation gaps. gt_1 and gt_2 are the first and second terminating times and are interval-censored with $[gs_1, T_{22})$ and $[gs_2, T_{12})$, respectively. For the third observation gap, gt_3 is right-censored, $gt_3 \in [gs_3, \infty)$.

For the easy implementation of computation, a piecewise constant baseline rate function is applied. Lawless and Zhan (1998), Liu *et al.* (2004) showed that the pieces defined with appropriately selected cutpoints that perform well at several situations. Let $0 = a_{r0} < a_{r1} < a_{r2} < \dots < a_{rq_r} = \tau_r$ be the cutpoints of the r^{th} ($r = 1, 2$) recurrent event, where τ_r is the largest event time of the r^{th} recurrent

Keywords: renal cell carcinoma; survival; erythrocyte sedimentation rate; anaemia; prognostic biomarkers

Erythrocyte sedimentation rate and anaemia are independent predictors of survival in patients with clear cell renal cell carcinoma

Y Choi^{1,2}, B Park³, K Kim⁴, BC Jeong³, SI Seo³, SS Jeon³, HY Choi³, JE Lee^{1,2} and HM Lee³

¹Women's Health Research Institute, Sookmyung Women's University, Seoul, Republic of Korea; ²Department of Food and Nutrition, Sookmyung Women's University, Seoul, Republic of Korea; ³Department of Urology, Samsung Medical Centre, Sungkyunkwan University School of Medicine, 50 Irwon-dong, Gangnam-gu, Seoul 135-710, Republic of Korea and ⁴Department of Statistics, Sookmyung Women's University, Seoul, Republic of Korea

Background: The 1997 international consensus conference on renal cell cancer (RCC) prognosis suggested erythrocyte sedimentation rate (ESR), alkaline phosphatase (ALP), and anaemia as prognostic biomarkers, but most studies reviewed were limited by small sample sizes.

Methods: The Cox proportional hazards model was used to evaluate whether ESR, ALP, haemoglobin (Hb), and haematocrit (Hct) could predict survival outcomes in 1307 patients with clear cell RCC (ccRCC) who underwent nephrectomy during 1994–2008.

Results: During a median follow-up of 43 months, we found that the patients with preoperative high levels of ESR, had a 2.10-fold (95% confidence interval (CI): 1.21–3.67) greater risk of dying from RCC compared with patients with low levels (normal range). Patients with preoperative anaemia, assessed by Hb and Hct, had a 3.11-fold (95% CI: 1.17–8.25) and 6.20-fold (95% CI: 2.30–16.72) greater risk of dying from other illnesses, respectively, compared with patients without anaemia. ALP levels were not associated with ccRCC patients' survival. These associations for ESR and anaemia were more pronounced in patients with body mass index (BMI) <25 compared with patients with BMI ≥25 kg m⁻².

Conclusion: Preoperative high ESR, but not ALP, was a significant predictor for cancer-specific survival among ccRCC patients. Anaemia increases the risk of death from other illness.

The age-standardised kidney cancer incidence rates per 100 000 population have risen from 3.0 in 1999 to 5.1 in 2009 in Korea (Jung *et al*, 2012). During a similar time period, the 5-year relative survival rates have also increased from 62% during 1993–1995 to 77.1% during 2005–2009 (Jung *et al*, 2012). Of the renal cell carcinoma (RCC) subtypes, clear cell RCC (ccRCC) is the most common, representing 70% of all RCCs (Cheville *et al*, 2003), and is characterised by the loss of the short arm of chromosome 3p (Presti *et al*, 1991).

Several biomarkers for RCC patients were proposed to be associated with RCC prognosis in the 1997 international consensus conference, convened by the American Joint Committee on Cancer

(AJCC) and the Union Internationale Contre le Cancer (UICC); such factors included erythrocyte sedimentation rate (ESR), anaemia (defined by haemoglobin; Hb), serum calcium, and alkaline phosphatase (ALP) (Srigley *et al*, 1997). Anaemia has been identified in 35–52% of RCC patients and is associated with more advanced disease and worse survival (Kim *et al*, 2003; Magera *et al*, 2008; Jensen *et al*, 2009). The ESR is a non-specific measure of inflammatory response (Brigden, 1999) and RCC patients with high-stage and high-grade tumour had high levels of ESR (Sengupta *et al*, 2006). A recent meta-analysis has shown that elevated ESR levels were associated with poor overall survival (OS), cancer-specific survival (CSS), and recurrence-free survival (RFS)

*Correspondence: Dr HM Lee, E-mail: besthtml@medimail.co.kr

Received 9 August 2012; revised 31 October 2012; accepted 21 November 2012; published online 8 January 2013

© 2013 Cancer Research UK. All rights reserved 0007–0920/13

(Wu *et al*, 2011). Some studies, but not all (Magera *et al*, 2008), have observed that elevated ALP levels, a possible sign of paraneoplastic syndrome of RCC patients, were associated with poor RCC survival (Lee *et al*, 2006; Kume *et al*, 2011).

Most studies published to date on RCC prognosis included small sample sizes or did not take into account distinct histological subtypes or stages of RCC. We therefore evaluated whether the preoperative high levels of ESR, ALP, or anaemia, which are routinely measured parameters, could predict worse survival outcomes of ccRCC patients treated with radical or partial nephrectomy in a clinical-based cohort study containing large numbers of patients and information about histological subtypes and stages of RCC.

MATERIALS AND METHODS

Study patients. This study has been described elsewhere (Choi *et al*, 2012). We included 1307 patients who underwent initial nephrectomy for ccRCC at the Samsung Medical Center, Seoul, Republic of Korea, between October 1994 and December 2008. Patients were excluded if they had missing information on the main exposures of interest, ESR ($n = 81$) or ALP variables ($n = 1$), leaving 1226 and 1306 patients, respectively. In all, 1307 patients were included for analyses of Hb and haematocrit (Hct).

The primary end points were OS, CSS, and other-cause survival (OCS), defined as the elapsed time between the date of surgery and either the date of death from any cause, from RCC, and from causes other than RCC, respectively, or 31 December 2008. This study was approved by an institutional review board at the Samsung Medical Centre.

Data collection. All patients received a clinical and pathological review. Data were manually abstracted from the electronic medical records using a standardised data collection form. At diagnosis, patients were asked about their smoking and alcohol habits, and whether they ever had RCC-related symptoms or weight loss compared with their usual weight. Height and weight were directly measured. All patients had a clinical abdominal computed tomography scan before surgery for the status of regional lymph nodes near the kidney and the presence of distant metastases. Tumour staging was confirmed by microscopic examination of surgically resected specimens of RCC. The size and location of tumours were measured at the time of macroscopic examination. The histological subtypes were classified as: clear cell, papillary, chromophobe, collecting duct RCC, and unclassified cell carcinoma (Kovacs *et al*, 1997). The tumour grade was based on the Fuhrman nuclear grading system (Fuhrman *et al*, 1982). The American Society of Anaesthesiologists (ASA) score was assessed by anaesthesiologists.

Cause of death. Death information of all patients was collected based on our institution's vital records data, which incorporated both in-hospital death data at our institution and data from the Korean National Statistical Office (KNSO). If the cases where the death occurred out of our institution, the patient's medical records were electronically linked to death certificates via unique 13-digit personal identification number in the KNSO, the system of which complete ascertainment of cause for deaths reached to 96.2% (Korean National Statistical Office, 2010). Causes of death were ascertained and verified by a medical record review by physicians, the patient's death certificate, or autopsy report, and coded according to the Korean Standard Classification of Causes of Death (Korean National Statistical Office, 1995), which was based on the International Classification of Disease for Oncology 10th edition (World Health Organization, 1992).

Laboratory assays. Venous blood samples were collected as part of routine clinical procedures before surgery at our institution and laboratory parameters for this study were abstracted from medical records. The ESR was measured using the quantitative capillary photometry method (TEST-1 system, Alifax, Italy) with the inter- and intra-assay coefficient of variation (CV) for ESR was $\leq 10\%$. Haemoglobin was measured using the cyanide-free haemoglobin spectrophotometry method and Hct was measured using electric impedance (SYSEM XE-2100, Kobe, Japan) with the inter- and intra-assay CVs of $\leq 3\%$ for Hb and Hct. ALP was measured using the International Federation of Clinical Chemistry method (Roche modular DP-210, Mannheim, Germany) with the inter- and intra-assay CVs of $\leq 1.4\%$.

To define cutoff points for high levels (abnormal laboratory values) of ESR and ALP, the gender-specific cutoff values from *a priori* study were referred to: ESR—males > 22 mm per hour, females > 29 mm per hour (Magera *et al*, 2008); the cutoff point for ALP was $> 115 \text{ U l}^{-1}$ for both sexes (Motzer *et al*, 1999). The anaemia was defined using the following cutoff points: Hb—males $< 13.5 \text{ g dl}^{-1}$, females $< 12 \text{ g dl}^{-1}$; and Hct—males $< 41\%$, females $< 36\%$ (Magera *et al*, 2008). We also analysed the data with reference values used at our institution as follows: ESR—males 0–22 mm per hour; females 0–27 mm per hour; ALP—males 53–128 U l^{-1} ; females 42–98 U l^{-1} ; Hb—males 13.6–17.4 g dl^{-1} ; females 11.2–14.8 g dl^{-1} ; Hct—males 40.4–51.3%; females 31.8–43.8%.

Statistical analysis. Because ccRCC represents a distinct clinical entity in relation to pathogenesis and genetic alterations from other histological subtypes of RCC (Presti *et al*, 1991; Cheville *et al*, 2003), we restricted the analysis to only patients with ccRCC in the main analysis. The Kaplan–Meier model was used to estimate survival rates, and the Cox proportional hazard regression model was used to estimate multivariate-adjusted HRs and 95% confidence intervals (CIs).

We conducted three separate multivariate analyses. The first model was stratified by age at diagnosis (years) and sex. The second model was further adjusted for body mass index (BMI) (continuous), smoking (never or ever), weight loss (yes or no), tumour stage (pT1, pT2, or pT3–4), nephrectomy type (radical or partial), Fuhrman grade (G1–2 or G3–4), symptom presence (yes or no), tumour size (< 5 or ≥ 5 cm), and ASA class (< 2 or ≥ 2). The third model additionally simultaneously adjusted for the other all exposures of interest (ESR, ALP, and anaemia) in addition to the covariates listed above. Adjustment for alcohol did not appreciably change the results; therefore, we did not include this in the analyses. We used the false discovery rate (FDR) adjustment < 0.05 as the significance threshold in each of the tests to account for multiple comparisons (Benjamini and Hochberg, 1995). The study also evaluated whether tumour stage (pT–2 or pT3–4) or preoperative BMI (< 25 or $\geq 25 \text{ kg m}^{-2}$) (Steering Committee of the Western Pacific Region of the World Health Organization, the International Association for the Study of Obesity, the International Obesity Task Force, 2000), modified the associations. In the sensitivity analyses, patients who had metastases ($n = 78$) or those who died due to any cause within the first 2-year follow-up period after surgery ($n = 79$) were excluded to reduce the potential effects of reverse causality. All *P*-values reported are two sided, with $P < 0.05$ used to denote statistical significance. The analyses were performed using the STATA (version 11; Stata Corp., College Station, TX, USA).

RESULTS

Patients' characteristics prior to surgery. During a median follow-up of 43 months, 165 of 1307 patients (12.6%) died from

Exploring *Trans*-acting regulators of gene expression associated with metabolic syndrome: a coupled application of factor analysis and linkage analysis

Kyee-Zu Kim · Jin-Young Min · Kyunga Kim ·
Joohon Sung · Sung-Il Cho

Received: 25 April 2011 / Accepted: 6 August 2012 / Published online: 5 February 2013
© The Genetics Society of Korea 2013

Abstract Metabolic syndrome is an established public health problem because it increases the risk of cardiovascular disease. Although many genes contribute to the etiology of metabolic syndrome, the effect size of each gene is much weaker than predicted by heritability. Exploring the regulation of gene expression levels can bridge this gap. Here, we aimed to explore genomic loci for a *trans*-regulator of gene expression associated with metabolic syndrome via factor analysis and linkage analysis. Forty-nine gene-expression quantitative traits (eQTs) associated with metabolic syndrome were selected and clustered into three latent eQTs by factor analysis. These included insulin-related, lipid-related, and glucose-related latent eQTs. Linkage analyses were performed for 49 original eQTs and 3 latent eQTs. Two original eQTs, adipose differentiation-related protein and IRS2, showed the highest LOD scores (3.09 and 3.08, respectively). We observed that the insulin-related and the lipid-related latent eQTs had stronger linkage evidence than the original eQTs, suggesting the presence of common gene regulators captured as latent eQTs. Several single nucleotide polymorphism markers

located on the peaks were co-localized with specific genes. Five of them were related to the neuronal system, which plays a role in the metabolic syndrome pathway. The findings related those eQTs could contribute to our understanding regarding the genetic pathway for insulin and lipid metabolism, particularly the regulation of gene expression.

Keywords Expression quantitative trait · Gene regulation · Mixed application

Introduction

Metabolic syndrome represents a cluster of endocrine-metabolic disorders, including hypertension, dyslipidemia, glucose intolerance, and abdominal obesity (Reaven 2002). In recent years, metabolic syndrome has been a growing public health problem, since it is known to increase the risk of cardiovascular disease (Ford et al. 2002; Park et al. 2003, 2007). Insulin resistance is believed to be an

K.-Z. Kim
Institute of Health & Environment, Graduate School of Public Health, Institute on Aging, Seoul National University, Seoul, South Korea
e-mail: kyeezu@gmail.com

J.-Y. Min
Institute of Health & Environment, Graduate School of Public Health, Seoul National University, Seoul, South Korea
e-mail: yaemin00@snu.ac.kr

K. Kim
Department of Statistics, Sookmyung Women's University, Seoul, South Korea
e-mail: kyunga@sookmyung.ac.kr

J. Sung · S.-I. Cho (✉)
Graduate School of Public Health, Seoul National University, Seoul, South Korea
e-mail: scho@snu.ac.kr; persontime@hotmail.com

J. Sung
e-mail: jsung@snu.ac.kr

J. Sung · S.-I. Cho
Institute of Health & Environment, Graduate School of Public Health, Seoul National University, Seoul, South Korea

underlying risk factor for these endocrine disorders, (Reaven 1988) and its pathologic and epidemiologic characteristics, including genetic components, are continually being investigated.

Clarifying the genetic components of metabolic syndrome and other complex diseases has been a challenge. To unravel the genetic mechanisms underlying complex diseases, various concepts and methods have been applied, e.g., gene-networking (Wessels et al. 2001), candidate genes (Tabor et al. 2002), or gene-regulator approaches (Hasty et al. 2001; de Jong 2002). Genetic studies of metabolic syndrome are mainly conducted using the candidate gene approach, and numerous successes have been made in identifying the genetic components of metabolic syndrome (Groop 2000; Mercado et al. 2002; Stern 2003; Horenstein and Shuldiner 2004; Wang et al. 2004; Pollex and Hegele 2006; Sale et al. 2006; Vasseur et al. 2006; Yang and Chuang 2006; Ma et al. 2007; Bonnet et al. 2008; Yamada et al. 2008). The mechanisms underlying the interrelationships among related genes, however, are poorly understood.

Many recent studies have recognized the utility of expression quantitative traits (eQTs) for unraveling the genetic components of human diseases. These studies have focused on the identification of genetic components underlying eQTs, called expression quantitative trait loci (eQTL) (Rockman and Kruglyak 2006; Frazer et al. 2009). The eQTL represent *cis*- and *trans*-acting regulators of gene expression processes, and may act as a common gene regulation mechanism (Morley et al. 2004).

Although many genes may contribute to the complex disease pathology, the impact of an individual gene might not be sufficiently great for it to be defined as a “real candidate gene” (Cookson et al. 2009). In this case, investigating the association between the disease and a group of genes that can be clustered using certain data reduction methods, such as factor analysis, would be more effective. Explanatory factor analysis is a statistical method that has been used to examine the validity of observed variables with the fewer latent variables (Pett et al. 2003; Wang et al. 2009). Factor analyses have also been recommended for reducing high dimensionality when data contain large-scale variables, such as microarray data (Quackenbush 2001). Using a factor analysis, observed variables can be clustered with fewer unobserved variables (i.e., factors or latent variables) that represent the properties of the observed variables. In this study, we applied factor analysis to characterize and cluster genes expression associated with insulin resistance, also called original expression quantitative traits (eQTs). By utilizing these resulting factors (i.e., latent eQTs), we investigated the regulatory mechanisms of genes within similar pathways. While each original eQT can have more than one function,

and its biological pathway may be largely unknown, each latent eQT may represent a characterized group of original eQTs. Examples of latent eQTs include insulin resistance-related components, such as insulin, lipid, and glucose. This property of characterized grouping would contribute to our understanding of gene expression interactions among eQTs that have similar pathways.

In this study, we aimed to explore genomic loci for *trans*-regulators of gene expression associated with metabolic syndrome using factor analysis and linkage analysis. We hypothesized that unknown latent eQTs can *trans*-regulate target genes associated with various endocrine functions. The latent variables may include metabolic syndrome-related factors. Several eQTs, including both original and latent eQTs, may play a role in the transcriptional regulation network. Genetic regulators were examined via the comparison of original and latent eQTs with the linkage peak loci or high LOD scores at single nucleotide polymorphism (SNP) markers on 22 autosomal chromosomes.

We present the clustering of eQTs and their properties related to metabolic syndrome in the “**Materials and methods**” section. The methods used to explore *trans*-acting gene regulators and regulation loci are also described. In the “**Results**” section, locations of linkage peak and eQT LOD scores are presented, along with the resulting latent variables from the explanatory factor analysis. Furthermore, the section includes results on how latent variables work as gene regulators among metabolic syndrome-related eQTs. Finally, we discuss the genetic and biological implications of our results in the “**Discussion**” section.

Materials and methods

Data and materials

Data were from the 15th Genetic Analysis Workshop (GAW15), Problem 1. Fourteen of three-generation Centre d’Etude du Polymorphisme Humain (CEPH) Utah families with 194 individuals were included in the dataset (Morley et al. 2004; Cheung and Spielman 2007). The phenotypes comprise the 3554 genes’ expression level in lymphoblastic cells. Among those 3554 eQTs, 49 eQTs were selected, known to participate in metabolic syndrome-related traits, such as hypertension, dyslipidemia, central obesity, and glucose intolerance, based on information from previous research, review papers, the NCBI gene database, and the Gene Ontology database. Genes known to influence endocrine-metabolic functions including lipid, insulin, adipose tissue, and glucose mechanism were recruited. Table 1 presented the characteristics of the 49 eQTs.

RESEARCH

Open Access

Identification of multiple gene-gene interactions for ordinal phenotypes

Kyunga Kim¹, Min-Seok Kwon², Sohee Oh³, Taesung Park^{2,3*}

From Second Annual Translational Bioinformatics Conference (TBC 2012)
Jeju Island, Korea. 13-16 October 2012

Abstract

Background: Multifactor dimensionality reduction (MDR) is a powerful method for analysis of gene-gene interactions and has been successfully applied to many genetic studies of complex diseases. However, the main application of MDR has been limited to binary traits, while traits having ordinal features are commonly observed in many genetic studies (e.g., obesity classification - normal, pre-obese, mild obese and severe obese).

Methods: We propose ordinal MDR (OMDR) to facilitate gene-gene interaction analysis for ordinal traits. As an alternative to balanced accuracy, the use of tau-b, a common ordinal association measure, was suggested to evaluate interactions. Also, we generalized cross-validation consistency (GVCV) to identify multiple best interactions. GVCV can be practically useful for analyzing complex traits, especially in large-scale genetic studies.

Results and conclusions: In simulations, OMDR showed fairly good performance in terms of power, predictability and selection stability and outperformed MDR. For demonstration, we used a real data of body mass index (BMI) and scanned 1~4-way interactions of obesity ordinal and binary traits of BMI via OMDR and MDR, respectively. In real data analysis, more interactions were identified for ordinal trait than binary traits. On average, the commonly identified interactions showed higher predictability for ordinal trait than binary traits. The proposed OMDR and GVCV were implemented in a C/C++ program, executables of which are freely available for Linux, Windows and MacOS upon request for non-commercial research institutions.

Background

Because most complex biological phenotypes are often affected by multiple genes and environmental factors, the investigation of gene-gene and gene-environment interactions can be essential in understanding the genetic architecture of complex traits [1]. It has been pointed out that focusing only on marginal effects of individual genes may result in low power and a low replication rate in genetic association studies of complex traits [2,3].

Many different methods have been proposed to analyze gene-gene interactions in genetic association studies [4,5], and can be categorized to methods based on regression modeling [6-9], pattern recognition [10,11], and data reduction [12-14]. Recently, machine learning approaches,

such as random forest [15], support vector machine [16] and ensemble learning [17], were applied to gene-gene interaction analysis.

While each method has its own advantages and disadvantages, the multifactor dimensionality reduction (MDR) method, a data-reduction approach, is known to have the advantages in examining high-order interactions and detecting interactions without main effects [13,18-20], and has been widely applied to detect gene-gene interactions in many common diseases (see the related literature available on <http://epistasis.org>). In addition, because the mode of genetic inheritance of a common complex trait is usually unknown a priori, MDR can be more useful to study a complex trait in that it does not require any assumption on genetic model. Since the MDR method was first introduced, it has been extended in many directions. Examples include family data [21], covariate adjustment and quantitative traits [22], the quantitative measure

* Correspondence: tspark@stats.snu.ac.kr

²Interdisciplinary Program in Bioinformatics, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul, South Korea

Full list of author information is available at the end of the article

of multi-locus genotype risk [23], and the selection of a parsimonious genetic model [24]. However, the applicability of existing MDR approaches is still restricted mainly to binary traits.

In the MDR analysis for binary traits, multi-locus genotype combinations of a set of genetic variables/markers (e.g., single nucleotide polymorphisms or SNPs) are induced to two levels (e.g., high risk and low risk) of a new binary variable, called an MDR classifier. The induction is conducted via assessing odds of two phenotypic classes for each genotype combination. Among MDR classifiers representing specific marker sets, the single best MDR classifier is selected by evaluating their classification performances, such as cross-validation consistency (CVC). As a result, the corresponding set of genetic markers is identified as having the strongest association with a trait of interest.

While MDR was introduced for binary traits, there is no existing approach that is applicable to ordinal categorical traits. In many genetic association studies, examples of traits having ordinal features are commonly available, such as the obesity classification based on body mass index (e.g., normal, pre-obese, mild obese and severe obese), the diabetes diagnosis based on glucose level (e.g., normal, impaired glucose tolerance and diabetes) and the severity classification of metabolic syndrome. The current application of MDR to these ordinal traits requires to dichotomization of traits by combining several categories, which results in the loss of ordinal information and powers.

In this study, we propose an ordinal MDR (OMDR) approach that enables one to analyze a joint effect of multiple genetic variables on an ordinal categorical trait. The proposed OMDR generates a classifier for each set of genetic markers in the form of a categorical variable with ordinal levels. The performance of each OMDR classifier is evaluated to select the best OMDR classifiers. For performance evaluation, we suggest the use of common ordinal association measures, such as tau-b [25], which test for the trend of directional association between two ordinal variables. By using the ordinal association measures, the performance of OMDR classifier can be evaluated by the degree of tendency of positive association between the observed categories of an ordinal trait and the estimated categories by OMDR.

In addition, we propose a way to report multiple candidates of gene-gene interactions in OMDR as well as MDR analyses. The original MDR approach reports only a single best candidate. This feature can be impractical and/or unreasonable when causal gene-gene interactions are searched for complex traits, especially in a genome-wide scale. Because genome-wide association studies with up to ~1 million SNPs became common, there is a growing need for more efficient criterion to report multiple candidates of gene-gene interactions in the MDR analysis.

Thus, we propose a new evaluation measure, generalized cross-validation consistency (GCVC), according to which one can report multiple best gene-gene interactions associated with the ordinal trait. Specifically, a pre-specified number (K) of the best classifiers are selected via this GCVC.

Simulations are conducted to investigate performance of the proposed new OMDR method and GCVC. We apply the proposed method to an ordinal obesity trait for body mass index (i.e., normal, pre-obese, mild obese and severe obese) of Age-Related Eye Disease Study data [26].

Methods

Overall procedure of OMDR

The OMDR procedure is same as the MDR procedure for binary traits, and consists of multiple steps. First, the dataset is partitioned into L (usually equal-size) subsets for L -fold cross-validation (CV). For example, $L = 10$ hereafter. Out of 10 subsets, one subset is taken as an independent testing dataset, and the remaining nine subsets are assigned to a training dataset. As a result, a total of 10 CV datasets are generated. Second, all possible OMDR classifiers are constructed for the corresponding combinations of m SNPs, and the K best ones are selected based on classification performance on a training data for each CV set (see the following two sections for details). Third, the best OMDR classifiers are chosen over all CV sets for the fixed m . The predictability of the selected OMDR classifiers is evaluated via the average value of the evaluation measure with a testing dataset over all the 10 CVs. In addition, the selection strength of a particular OMDR classifier is suggested via $GCVC^K$ which is the number of times the classifier is identified as one of the K best classifiers across all the CVs. The best OMDR classifiers across the CVs are chosen if having the maximum predictability and maximum $GCVC^K$. Finally, the overall best OMDR classifiers are selected based on the predictability and $GCVC^K$ among the best ones for various values of m , which result from the previous steps. For additional details, refer to the original MDR procedure described in literature [13,27,28].

OMDR classifier construction

Let 1, 2, ..., J be classes for an ordinal phenotype of interest. For example, 'low blood pressure (BP)', 'normal' and 'high BP' classes can be viewed as classes 1, 2 and 3, respectively, in the analysis of the BP classification trait. Note that $J = 2$ for a binary trait (e.g., classes 1 and 2 respectively for control and case).

Suppose that an m -way interaction is under consideration. For the corresponding m SNPs, let n_{ij} be the number of individuals with the i th multi-locus genotype and let n_{+j} be the total number of individuals in phenotypic

Functional Implication of β -Carotene Hydroxylases in Soybean Nodulation^{1[C][W][OA]}

Yun-Kyoung Kim², Sunghan Kim², Ji-Hyun Um, Kyunga Kim, Sun-Kang Choi, Byung-Hun Um, Suk-Woo Kang, Jee-Woong Kim, Shinichi Takaichi, Seok-Bo Song, Choon-Hwan Lee, Ho-Seung Kim, Ki Woo Kim, Kyoung Hee Nam, Suk-Ha Lee, Yul-Ho Kim, Hyang-Mi Park, Sun-Hwa Ha, Desh Pal S. Verma, and Choong-Il Cheon*

Department of Biological Science (Y.-K.K., S.K., J.-H.U., K.H.N. C.-I.C.) and Department of Statistics (K.K.), Sookmyung Women's University, Seoul 140-742, Korea; Gangneung Science Industry Foundation, Gangneung 210-340, Korea (S.-K.C.); Natural Products Research Center, KIST Gangneung Institute, Gangneung 210-340, Korea (B.-H.U., S.-W.K.); Electron Microscopy Laboratory, Dental Research Institute (J.-W.K.), and School of Plant Science (S.-H.L.), Seoul National University, Seoul 151-742, Korea; Department of Biology, Nippon Medical School, Nakahara, Kawasaki 113-8602, Japan (S.T.); Department of Functional Crops, National Institute of Crop Science, Milyang 441-857, Korea (S.-B.S.); Department of Molecular Biology, Pusan National University, Busan 609-735, Korea (C.-H.L., H.-S.K.); School of Ecological and Environmental Systems, Kyungpook National University, Sangju 702-701, Korea (K.W.K.); National Institute of Crop Science, Suwon 441-857, Korea (Y.-H.K., H.-M.P.); Graduate School of Biotechnology and Crop Biotech Institute, Kyung Hee University, Yongin 446-701, Korea (S.-H.H.); and Biotechnology Center, Ohio State University, Columbus, Ohio 43210 (D.P.S.V.)

Legume-*Rhizobium* spp. symbiosis requires signaling between the symbiotic partners and differential expression of plant genes during nodule development. Previously, we cloned a gene encoding a putative β -carotene hydroxylase (*GmBCH1*) from soybean (*Glycine max*) whose expression increased during nodulation with *Bradyrhizobium japonicum*. In this work, we extended our study to three *GmBCHs* to examine their possible role(s) in nodule development, as they were additionally identified as nodule specific, along with the completion of the soybean genome. In situ hybridization revealed the expression of three *GmBCHs* (*GmBCH1*, *GmBCH2*, and *GmBCH3*) in the infected cells of root nodules, and their enzymatic activities were confirmed by functional assays in *Escherichia coli*. Localization of *GmBCHs* by transfecting Arabidopsis (*Arabidopsis thaliana*) protoplasts with green fluorescent protein fusions and by electron microscopic immunogold detection in soybean nodules indicated that *GmBCH2* and *GmBCH3* were present in plastids, while *GmBCH1* appeared to be cytosolic. RNA interference of the *GmBCHs* severely impaired nitrogen fixation as well as nodule development. Surprisingly, we failed to detect zeaxanthin, a product of *GmBCH*, or any other carotenoids in nodules. Therefore, we examined the possibility that most of the carotenoids in nodules are converted or cleaved to other compounds. We detected the expression of some carotenoid cleavage dioxygenases (*GmCCDs*) in wild-type nodules and also a reduced amount of zeaxanthin in *GmCCD8*-expressing *E. coli*, suggesting cleavage of the carotenoid. In view of these findings, we propose that carotenoids such as zeaxanthin synthesized in root nodules are cleaved by *GmCCDs*, and we discuss the possible roles of the carotenoid cleavage products in nodulation.

Legume-*Rhizobium* spp. symbiosis results in the formation of the root nodule, in which rhizobia fix atmospheric nitrogen. Nodule development requires

diverse events, such as Nod factor synthesis in the rhizobia, perception of the Nod factor on plant roots by receptor-like kinases, endocytosis of rhizobia into plant cells, and so on (Stacey et al., 2006; Oldroyd et al., 2011; Singh and Parniske, 2012). Sequential expression of numerous plant genes occurs during nodulation, contributing to different stages including nitrogen fixation. Arbuscular mycorrhizal (AM) symbiosis exhibits many similarities to the nodulation process (Oldroyd et al., 2009). For example, *SymRK*, the receptor-like kinase gene, is required for both rhizobial and AM symbioses (Stracke et al., 2002). Similarly, the signal transduction pathways following perception are also in part the same, and the genes common to the two pathways have been referred to as the common *symbiosis* (*SYM*) genes (Kistner et al., 2005). These similarities may reflect common mechanisms for host plant cells to respond to symbionts, although the commonality is not globally defined yet.

¹ This work was supported by the Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Education, Science, and Technology (grant no. 2-1205-0037) and by the Next-Generation BioGreen 21 Program (grant no. PJ009076012013), Rural Development Administration, Republic of Korea.

² These authors contributed equally to the article.

* Corresponding author; ccheon@sookmyung.ac.kr.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantphysiol.org) is: Choong-Il Cheon (ccheon@sookmyung.ac.kr).

[C] Some figures in this article are displayed in color online but in black and white in the print edition.

[W] The online version of this article contains Web-only data.

[OA] Open Access articles can be viewed online without a subscription.

www.plantphysiol.org/cgi/doi/10.1104/pp.113.215020

Plant carotenoids are mostly C₄₀ tetraterpenoid pigments with a series of double bonds (DellaPenna and Pogson, 2006; Lu and Li, 2008). They play essential roles in photosynthesis. The phytohormone abscisic acid (ABA) is synthesized from xanthophylls, oxygenated derivatives of carotenoids. The beneficial effects of carotenoids for human disease prevention and health promotion are well established and are based on their antioxidant activities (Kopsell and Kopsell, 2006; Rao and Rao, 2007; von Lintig, 2010). Metabolic engineering approaches have produced crop plants with enhanced carotenoid contents and improved nutritional value (Giuliano et al., 2008). For example, enhancement of β -carotene, provitamin A, by engineering the carotenoid biosynthetic pathway resulted in the development of cv Golden rice (*Oryza sativa*; Ye et al., 2000; Paine et al., 2005; Ha et al., 2010).

The initial step of carotenoid biosynthesis is the production of phytoene by the enzyme phytoene synthase (Fig. 1; DellaPenna and Pogson, 2006; Cazzonelli and Pogson, 2010). The subsequent activities of desaturases, isomerase, and cyclase convert phytoene into lycopene and further into β -carotene. Xanthophyll synthesis begins with the action of β -carotene hydroxylase (BCH) on β -carotene, producing initially β -cryptoxanthin and thereafter zeaxanthin (Kim et al., 2009). Overexpression of BCH has been found to confer tolerance to light stress (Davison et al., 2002). The subsequent steps catalyzed by zeaxanthin epoxidase (ZEP) and neoxanthin synthase lead to the synthesis of ABA (Takaichi and Mimuro, 1998).

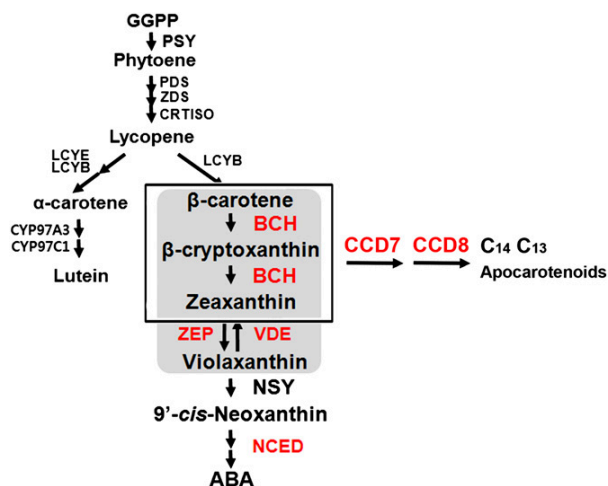


Figure 1. The biosynthetic pathway of carotenoids in plants. GGPP, Geranylgeranyl diphosphate; PSY, phytoene synthase; PDS, phytoene desaturase; ZDS, ζ -carotene desaturase; CRTISO, carotene isomerase; LCYB, lycopene β -cyclase; CYP97A3 and CYP97C1, cytochrome P450 enzymes; NSY, neoxanthin synthase; LCYE, lycopene ϵ -cyclase; CRTR-E, ϵ -carotene hydroxylase. Enzymes in red were examined in this study.

Various carotenoid cleavage dioxygenases (CCDs) catalyze the formation of apocarotenoids with functions as hormones, flavors, and pigments (Auldrige et al., 2006b; Strack and Fester, 2006; Tsuchiya and McCourt, 2009; Walter et al., 2010). Recently, CCD7 and CCD8 were shown to control the synthesis of strigolactones, newly discovered hormones that inhibit shoot branching (Gomez-Roldan et al., 2008; Umehara et al., 2008; Vogel et al., 2010; Ruyter-Spira et al., 2013). In addition, carotenoid cleavage products have been discovered in plant roots colonized by AM fungi (Strack and Fester, 2006). During AM symbiosis, roots synthesize apocarotenoids at the same time as activating plant genes for carotenoid metabolism. Although RNA interference (RNAi)-mediated inhibition of apocarotenoid synthesis suggests that apocarotenoids are functionally significant (Snowden et al., 2005; Floss et al., 2008), their role in AM symbiosis is unknown.

In a search for genes differentially induced during soybean (*Glycine max*)-*Rhizobium* spp. symbiosis, several antioxidant genes, including a gene encoding a putative BCH, were identified. In this report, we describe genes (*GmBCHs*) encoding a putative BCH whose expression increased in soybean root nodules. Therefore, the biochemical activities of BCHs were investigated. RNAi inhibition of *GmBCH* expression interfered with nitrogen fixation as well as nodule development. Subsequent analysis of the expression and biochemical activities of *GmCCDs* in root nodules led us to hypothesize that *GmCCD8* could be involved in the synthesis of apocarotenoids from zeaxanthin in these nodules.

RESULTS

Expression of Genes Encoding Putative BCHs in Soybean

We isolated from root nodules of soybean a complementary DNA (cDNA) with strong homology to BCH (*GmBCH1*), whose expression was higher in nodules than in roots (Fig. 2A; Lee et al., 2005). In a BLAST search with the *GmBCH1* sequence against the soybean genome (<http://www.phytozome.net/soybean.php>; Schmutz et al., 2010), we found several open reading frames encoding products with amino acid sequences highly homologous to that of *GmBCH1*; they were designated *GmBCH2* to *GmBCH5* (GenBank accession numbers are as follows: *GmBCH1* [AY575953], *GmBCH2* [BT093388], *GmBCH3* [BT098487], *GmBCH4* [JF970190], *GmBCH5* [JF970191]). The open reading frames of *GmBCH1* and *GmBCH2* are very similar to functionally confirmed BCHs (e.g. they have 74% and 75% sequence identity, respectively, to the BCH from coffee [*Coffea arabica*], CaCRTR-B; Simkin et al., 2008; Supplemental Fig. S1A). They have divergent N-terminal regions, like most previously reported BCHs, but carry four His-containing motifs, in which the spacing of the His residues was conserved (e.g.

Prediction of type 2 diabetes in women with a history of gestational diabetes using a genetic risk score

Soo Heon Kwak · Sung Hee Choi · Kyunga Kim · Hye Seung Jung · Young Min Cho ·
Soo Lim · Nam H. Cho · Seong Yeon Kim · Kyong Soo Park · Hak C. Jang

Received: 10 July 2013 / Accepted: 4 September 2013 / Published online: 22 September 2013
© Springer-Verlag Berlin Heidelberg 2013

Abstract

Aims/hypothesis Women with a history of gestational diabetes mellitus (GDM) are at increased risk of future development of type 2 diabetes. Recently, over 65 genetic variants have been confirmed to be associated with diabetes. We investigated whether this genetic information could improve the prediction of future diabetes in women with GDM.

Methods This was a prospective cohort study consisting of 395 women with GDM who were followed annually with an OGTT.

Electronic supplementary material The online version of this article (doi:10.1007/s00125-013-3059-x) contains peer-reviewed but unedited supplementary material, which is available to authorised users.

S. H. Kwak · H. S. Jung · Y. M. Cho · S. Y. Kim · K. S. Park (✉)
Department of Internal Medicine, Seoul National University Hospital,
101 Daehak-Ro, Chongno-Gu, Seoul 110-744, South Korea
e-mail: kspark@snu.ac.kr

S. H. Choi · Y. M. Cho · S. Y. Kim · K. S. Park · H. C. Jang
Department of Internal Medicine, Seoul National University College of
Medicine, 103 Daehak-Ro, Chongno-Gu, Seoul 110-799, South Korea

S. H. Choi · S. Lim · H. C. Jang (✉)
Department of Internal Medicine, Seoul National University Bundang
Hospital, 82 Gumi-Ro, 173 Beon-Gil, Bundang-Gu, Seongnam-Si,
Gyeonggi-Do 463-707, South Korea
e-mail: janghak@snu.ac.kr

K. Kim
Department of Statistics, Sookmyung Women's University,
Seoul, South Korea

N. H. Cho
Department of Preventive Medicine, Ajou University
School of Medicine, Suwon, South Korea

K. S. Park
World Class University Department of Molecular Medicine and
Biopharmaceutical Sciences, Graduate School of Convergence Science
and Technology, Seoul National University, 103 Daehak-Ro,
Chongno-Gu, Seoul 110-799, South Korea

A weighted genetic risk score (wGRS), consisting of 48 variants, was assessed for improving discrimination (C statistic) and risk reclassification (continuous net reclassification improvement [NRI] index) when added to clinical risk factors.

Results Among the 395 women with GDM, 116 (29.4%) developed diabetes during a median follow-up period of 45 months. Women with GDM who went on to develop diabetes had a significantly higher wGRS than those who did not (9.36 ± 0.92 vs 8.78 ± 1.07 ; $p < 1.56 \times 10^{-7}$). In a complex clinical model adjusted for age, prepregnancy BMI, family history of diabetes, blood pressure, fasting glucose and fasting insulin concentration, the C statistic marginally improved from 0.741 without the wGRS to 0.775 with the wGRS ($p = 0.015$). The addition of the wGRS to the clinical model resulted in a modest improvement in reclassification (continuous NRI 0.430 [95% CI 0.218, 0.642]; $p = 7.0 \times 10^{-5}$).
Conclusions/interpretation In women with GDM, who are at high risk of diabetes, the wGRS was significantly associated with the future development of diabetes. Furthermore, it improved prediction over clinical risk factors.

Keywords Genetic risk score · Gestational diabetes · Risk prediction · Type 2 diabetes

Abbreviations

GDM	Gestational diabetes mellitus
GRS	Genetic risk score
GWA	Genome-wide association
IGT	Impaired glucose tolerance
IQR	Interquartile range
IR	Insulin resistance
IS	Insulin secretion
NGT	Normal glucose tolerance
NRI	Net reclassification improvement
SNP	Single nucleotide polymorphism
uGRS	Unweighted genetic risk score
wGRS	Weighted genetic risk score

Introduction

Women with gestational diabetes mellitus (GDM) have an increased risk of developing type 2 diabetes in the future. It has been reported that women with GDM have a 7.43-fold increased risk of developing type 2 diabetes compared with those who had a normoglycaemic pregnancy [1]. We have recently reported that as many as 50% of Korean women with GDM progress to diabetes within 8.1 years after parturition [2]. The risk factors associated with progression to type 2 diabetes include age, a family history of diabetes, degree of hyperglycaemia and decreased beta cell function during pregnancy, and prepregnancy obesity [2–5]. We have shown in our previous reports that certain genetic variants are also associated with the progression to diabetes [2, 6]. Because women with GDM are at substantially higher risk of developing type 2 diabetes, it is important to accurately predict the future risk of diabetes in these women.

Lifestyle intervention and medications, including metformin and troglitazone, have been shown to prevent or delay the development of type 2 diabetes in women with GDM [7, 8]. In the Troglitazone in Prevention of Diabetes study, troglitazone treatment was associated with a 55% reduction in diabetes compared with placebo in women previously diagnosed with GDM [7]. In addition, in the subgroup analysis of the Diabetes Prevention Program, both lifestyle intervention and metformin treatment reduced the risk of diabetes by approximately 50% compared with placebo in women who were reported to have GDM [8]. Because these findings indicate that it is possible to prevent, or at least delay, the development of diabetes in women with GDM, it is important to have an accurate prediction model to identify high-risk subgroups among such women.

During the past 5 years, the advent of genome-wide association (GWA) studies has rapidly increased the number of genetic variants associated with type 2 diabetes. At least 65 genetic variants have been associated with the risk of diabetes at a genome-wide significance level [9]. These common variants, however, confer only a small relative risk and explain only a limited fraction of the heritability of diabetes [10]. Furthermore, most recent studies have shown that genetic information in the form of a genetic risk score (GRS) did not improve the current diabetes prediction model, which is based on well-validated clinical risk factors [11–14]. Most of these studies, however, examined the general population with an average risk of diabetes, and they consisted of relatively old participants. We hypothesised that genetic information could improve the prediction of type 2 diabetes in women with GDM, because these women are at particularly high risk at a relatively young age. In this study, we developed a GRS consisting of 48 confirmed common genetic variants and tested whether the addition of this GRS in a clinical risk model could improve our prediction of diabetes in women with GDM.

Methods

Study participants This study was based on a prospective cohort of 395 women with GDM who were followed for more than 1 year postpartum. The diagnosis and follow-up protocols of the women are described in previous reports [2, 15]. Briefly, all pregnant women were screened with a 50 g 1 h glucose challenge test, and women who tested positive (cut-off value ≥ 7.2 mmol/l) underwent a 100 g OGTT.

The diagnosis of GDM was made if two or more of the following criteria were met during the 100 g OGTT, according to the criteria of the Third International Workshop-Conference on Gestational Diabetes Mellitus: fasting glucose ≥ 5.8 mmol/l, 1 h glucose ≥ 10.6 mmol/l, 2 h glucose ≥ 9.2 mmol/l, 3 h glucose ≥ 8.1 mmol/l [16]. Insulin levels were also measured during the 100 g OGTT. All women with GDM were tested 2 months postpartum and annually afterwards with a 75 g OGTT at each visit. We classified participants as having normal glucose tolerance (NGT), impaired glucose tolerance (IGT) or type 2 diabetes according to American Diabetes Association criteria [17]. Anthropometric measures were investigated while participants were wearing light clothing and no shoes. BMI was calculated as the body weight (in kilograms) of a patient divided by the square of that patient's height (in metres). Blood pressures were measured using standard protocols; mean arterial pressure was calculated as $1/3(\text{systolic pressure}) + 2/3(\text{diastolic pressure})$. We retrieved an obstetric history for each patient, including the gestational week at diagnosis of GDM and the requirement for insulin therapy during pregnancy. Participants who had any first-degree relatives with diabetes were considered to have a positive family history of diabetes.

The institutional review board of the Clinical Research Institute at Seoul National University Hospital approved the study protocol (H-1205-130-411), and written informed consent was obtained from each participant. All clinical investigations were conducted according to the principles expressed in the Declaration of Helsinki as revised in 2008.

Biochemical measurements Plasma glucose concentration was determined by the glucose oxidase method using a YSI 2300 STAT Glucose and Lactate Analyzer (YSI, Yellow Springs, OH, USA). Insulin concentration was measured by a human-specific radioimmunoassay kit (Linco Research, St Charles, MO, USA). The AUC of the glucose and insulin concentrations was calculated using trapezoidal methods. The 1 h $\Delta I/\Delta G$ was derived as $(\text{insulin}_{1\text{h}} - \text{insulin}_0)/(\text{glucose}_{1\text{h}} - \text{glucose}_0)$. Insulin sensitivity was estimated by the Matsuda index as $10,000/\sqrt{[(\text{fasting glucose}) \times (\text{fasting insulin}) \times (\text{mean glucose}) \times (\text{mean insulin})]}$ [18]. Pancreatic beta cell function was evaluated with the insulin secretion (IS)/insulin resistance

Quality of Fit Measures in the Framework of Quantile Regression

HOHSUK NOH, ANOUAR EL GHOUGH and INGRID VAN KEILEGOM

Institut de Statistique, Université catholique de Louvain

ABSTRACT. In regression experiments, to learn about the strength of the relationship between a covariate vector and a dependent variable, we propose a ‘coefficient of determination’ based on the quantiles. Such a coefficient is a ‘local’ measure in the sense that the strength is measured at a prespecified quantile level. Once estimated, it can be used, for example, to measure the relative importance of a subset of covariates in the quantile regression context. Related to this coefficient, we also propose a new ‘local’ lack-of-fit measure of a given parametric model. We provide some asymptotic results of the proposed measures and carry out a Monte Carlo simulation study to illustrate their use and performance in practice.

Key words: coefficient of determination, inadequacy index, lack-of-fit measure, local polynomial estimation, non-parametric regression, quantile regression

1. Introduction

Given a random variable Y , the response and a random vector $\underline{X} = (X_1, \dots, X_d)^T$ with $d \geq 1$, the covariate, the amount of variability in Y that can be explained by $\underline{X}$, via an (unknown) link function, is a fundamental question in statistics. Typically, the link is given as the conditional mean function $m(\underline{X}) = E(Y | \underline{X})$, but it may be any function of interest like the conditional median or, more generally, a conditional quantile function. When the mean regression function m is of interest, Doksum & Samarov (1995) proposed to use

$$\eta^2 = \frac{\text{Var}(m(\underline{X}))}{\text{Var}(Y)} = 1 - \frac{E |Y - m(\underline{X})|^2}{E |Y - E(Y)|^2}$$

as a measure of the explanatory power of $\underline{X}$. This is nothing but a non-parametric version of the well-known coefficient of determination R^2 , which is largely used within the linear regression framework as an assessment of the prediction power of a given linear model.

While η^2 can address the question ‘is $\underline{X}$ (or a subset of it) important to understand the variation in Y ’, it cannot answer some important questions like ‘does $\underline{X}$ exert any significant effect on the tail levels of Y ?’ or ‘is the effect of $\underline{X}$ as important on the tail levels of Y as on its central level?’ To study such a question one should renounce the mean and adopt a function that gives a more comprehensive picture of the effect of $\underline{X}$ on Y like the conditional quantile. As we will show, by using the latter we are able to figure out the exact effect of $\underline{X}$ at a specific quantile level q of Y . To be clear, let us first consider the case of the median, i.e. $q = 0.5$. An L_1 natural analogue of η^2 is given by $1 - E |Y - m_{0.5}(\underline{X})| / E |Y - \xi_{0.5}|$, where $m_{0.5}(\underline{X})$ is the conditional median of Y given $\underline{X}$ and $\xi_{0.5}$ is the median of Y . An obvious generalization of such a measure is given by

$$R^1(q) = 1 - \frac{E[\rho(Y - m(\underline{X}))]}{E[\rho(Y - \xi_q)]},$$

where, for a fixed $0 < q < 1$, $m(\underline{X}) \equiv m_q(\underline{X}) = \arg \min_{\theta} E[\rho_q(Y - \theta) | \underline{X}]$ is the (unique) conditional q th quantile of Y given $\underline{X}$, $\rho \equiv \rho_q(y) = (2q - 1)y + |y|$ is the well-known check function and $\xi_q = \arg \min_{\theta} E[\rho_q(Y - \theta)]$ is the (unique) q th quantile of Y . While η^2 only assesses the

explanatory power of $\underline{X}$ on Y in the conditional mean, $R^1(q)$ enables us to get a more complete picture of the explanatory strength of $\underline{X}$ on Y by varying q . A nice practical illustration of the importance of considering different values of q when studying the association between $\underline{X}$ and Y can be found in Koenker & Machado (1999), who already proposed $R^1(q)$ when $m(\cdot)$ is linear and cited an empirical study from Chamberlain (1994) on the US labour union wage premium.

Surprisingly, even in the linear model, the $R^1(q)$ has never been studied in the literature either from a theoretical or from a practical point of view. The only exception that we have found is the work of McKean & Sievers (1987), who proposed $R^1(0.5)$ as a robust alternative to R^2 in the case of a parametric linear model. Assuming a fixed design, they proposed a parametric version of the estimator of $R^1(0.5)$ and proved its consistency.

This motivates us to consider $R^1(q)$ in a more general context, where (1) no parametric restriction is made on the form of m , (2) $\underline{X}$ is allowed to be a random vector of any given dimension $d \geq 1$ and (3) the quantile level q can be any value within $(0, 1)$. Under relatively weak assumptions, we propose a consistent estimator of $R^1(q)$, for which we obtain a Bahadur representation that allows us to prove its asymptotic normality. Note that by allowing m to be a constant it can be seen that $0 \leq R^1(q) \leq 1$. $R^1(q)=0$ corresponds to the case when $m(\underline{X})=\xi_q$ with probability one, i.e. no variability is captured by $\underline{X}$, while $R^1(q)=1$ corresponds to the case when $m(\underline{X})=Y$ with probability one, i.e. all the variability in Y at its q th quantile level is captured by $\underline{X}$ via the quantile function m . In other words, $R^1(q)$ exactly quantifies the relative gain of introducing the covariate $\underline{X}$ in estimating the q th quantile of Y . It can be used to robustly reduce the dimensionality of $\underline{X}$ by keeping only the significant components of it or to discriminate between different combinations of fixed numbers of covariates with the objective of selecting the best combination with the highest explanatory power.

Another related and important problem of common interest is to assess the discrepancy of a given parametric model, say $m(\theta, \underline{X})$, for estimating a certain conditional quantile of Y . The objective is to measure the inevitable loss of information that can be attributed to (and only to) the parametric restriction. From similar motivations as for $R^1(q)$, we define

$$\Phi(q) = 1 - \frac{E[\rho(Y - m(\underline{X}))]}{E[\rho(Y - m(\theta^*, \underline{X}))]},$$

where $\theta^* \equiv \theta^*(q)$ is a ‘pseudo-true’ parameter, i.e. an argument that minimizes $E[\rho(Y - m(\theta, \underline{X}))]$ with respect to all θ in a set Θ , or equivalently $m(\theta^*, \cdot)$ is the best approximation to the true regression function m that can be found within the parametric family $\{m(\theta, \cdot)\}_{\theta \in \Theta}$. The numerator of $\Phi(q)$, i.e. $E[\rho(Y - m(\theta^*, \underline{X}))] - E[\rho(Y - m(\underline{X}))]$, is a non-negative scalar that represents the amount of unexplained ‘variation’ which is measured by the check function ρ , due to the fact that one uses the parametric model $m(\theta^*, \cdot)$ instead of the true quantile regression curve m . Thus, it follows that $\Phi(q)$ is the fraction of the parametric residual variation that can be completely attributed to the lack of fit of the parametric quantile function $m(\theta, \cdot)$. In the following, $\Phi(q)$ will be shortly called the q -inadequacy index of $m(\theta, \cdot)$. Like $R^1(q)$, the q -inadequacy index is a local measure. We use the word ‘local’ to emphasize that it quantifies the quality (or to be more correct the poorness) of $m(\theta, \cdot)$ for a given fixed quantile level q . Allowing $m(\theta, \cdot)$ to be constant, one can easily check that $0 \leq \Phi(q) \leq R^1(q) \leq 1$. Unlike $R^1(q)$, the case $\Phi(q)=0$ corresponds now to the best case in which $m(\theta^*, \underline{X})$ coincides almost surely with the true q -quantile curve $m(\underline{X})$. The case $\Phi(q)=R^1(q)$ represents the worst case when the parametric model fails to capture any variation in the data that could have been captured if $m(\underline{X})$ was used. In such case $m(\theta^*, \underline{X})=\xi_q$, with probability one. In the mean regression framework, this type of inadequacy index was already proposed by Doksum & Samarov



Assessing model adequacy in possibly misspecified quantile regression

Hohsuk Noh ^{*}, Anouar El Ghouch, Ingrid Van Keilegom

Institut de Statistique, Biostatistique et Sciences Actuarielles, Université Catholique de Louvain, 20 Voie du Roman Pays, Louvain-la-Neuve 1348, Belgium

ARTICLE INFO

Article history:

Received 5 January 2012

Received in revised form 20 July 2012

Accepted 21 July 2012

Available online 27 July 2012

Keywords:

Coefficient of determination

Conditional quantiles

Lack-of-fit

Linear model

Prediction quality

ABSTRACT

Possibly misspecified linear quantile regression models are considered. A measure for assessing the combined effect of several covariates on a certain conditional quantile function is proposed. The measure is based on an adaptation to quantile regression of the famous coefficient of determination originally proposed for mean regression, and compares a 'reduced' model to a 'full' model, both of which can be misspecified. An estimator of this measure is proposed and its asymptotic distribution is investigated both in the non-degenerate and the degenerate case. The finite sample performance of the estimator is studied through a number of simulation experiments. The proposed measure is also applied to a data set on body fat measures.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

Quantile regression has emerged as an attractive alternative to the classical mean regression based on the quadratic loss function. Since it was introduced by [Koenker and Bassett \(1978\)](#) as a robust (to outliers) and flexible (about error distribution) linear regression method, quantile regression has received considerable interest in both theoretical and applied statistics (see [Koenker, 2005](#) and references therein). As this method has widened its applications to many domains like economics, biology, ecology and finance, it becomes very attractive for practitioners to develop a simple and effective assessment measure of goodness-of-fit for quantile regression in the spirit of the well-known R^2 : the coefficient of determination for linear mean regression.

To this end, when it is assumed that the true conditional quantile function is linear with respect to the given covariates, [Koenker and Machado \(1999\)](#) proposed a measure of model adequacy in quantile regression that aims at measuring prediction quality of a certain set of covariates. However, theoretical properties of its estimator have not been studied except for the work of [Mckean and Sievers \(1987\)](#), who showed the consistency of the estimator for the conditional median. Further, no results exist when the linear model is misspecified. The consideration of misspecification deserves attention because it is rare in reality that such a linear model assumption is correct. Yet practitioners often prefer to use a linear model for reasons of parsimony or interpretability even when they are not sure about the correct specification.

Motivated by these observations, we revisit the measure proposed by [Koenker and Machado \(1999\)](#) and reinterpret it as a way to compare two nested linear quantile models regardless of whether the corresponding true conditional quantile functions are linear or not. This kind of reinterpretation is made possible because of the result in [Kim and White \(2003\)](#) and [Angrist et al. \(2006\)](#), who showed the consistency for certain "pseudo-true" parameter values and the asymptotic normality of the quantile estimator when the model is misspecified. Further, we provide an asymptotic representation of the estimator of the proposed measure, which implies its consistency and asymptotic normality. This result is meaningful in that such a representation is not known (to the best of our knowledge) even for the case of a correctly specified linear model. Although

^{*} Corresponding author. Tel.: +32 10 479403; fax: +32 10 473032.

E-mail address: word5810@gmail.com (H. Noh).

URL: <http://perso.uclouvain.be/ingrid.vankeilegom/DataProgramCSDA.zip> (I. Van Keilegom).

our main goal in this paper is to propose a simple assessment measure of quantile regression under misspecification, it is possible to use the asymptotic results for a specification test as is shown in Section 4.

The rest of this paper is organized as follows. In Section 2 we introduce our measure and provide some insight into it using a few examples. We provide asymptotic results about the estimator of the proposed measure in Section 3 and describe statistical inference based on it in Section 4. In Section 5 we present some Monte Carlo evidence of the developed theory, whereas the analysis of data on body fat measures is given in Section 6. All the theoretical proofs are deferred to the Appendix.

2. Model adequacy measure under misspecification

To introduce our measure, we bring in some notations. Define the check loss function for a fixed quantile level $q \in (0, 1)$ as $\rho_q(u) = 2u(q - I(u < 0))$. Let Y be a one-dimensional dependent variable and $\mathbf{X} = (\mathbf{X}_0^\top, \mathbf{X}_1^\top)^\top$ be a random covariate vector of dimension $d_0 + d_1$, with $d_0, d_1 \geq 1$. The first (and only the first) element of $\mathbf{X}_0$ is 1. Then, our measure $\zeta(q)$ for assessing the effect of $\mathbf{X}_1$ under possible misspecification of a linear q -th quantile regression model is defined as

$$\zeta(q) = 1 - \frac{E[\rho_q(Y - \mathbf{X}^\top \boldsymbol{\beta}_q^*)]}{E[\rho_q(Y - \mathbf{X}_0^\top \boldsymbol{\beta}_{0,q}^*)]}, \quad (1)$$

where $\boldsymbol{\beta}_{0,q}^*$ and $\boldsymbol{\beta}_q^*$ are pseudo true parameters in the sense that they are assumed to be the unique minimizers of $E[\rho_q(Y - \mathbf{X}_0^\top \mathbf{b}_0)]$ and $E[\rho_q(Y - \mathbf{X}^\top \mathbf{b})]$ with respect to $\mathbf{b}_0$ and $\mathbf{b}$, respectively. Equivalently, we can say that they are the best approximations to the true quantile regression function that can be found within the two given families of linear models. Neither of the two linear models are supposed to be correct, and they are both possibly subject to model misspecification. In terms of the check loss distance, $E[\rho_q(Y - \mathbf{X}^\top \boldsymbol{\beta}_q^*)]$ represents the amount of variation of Y that cannot be explained through a ‘full’ but possibly incorrect linear model in $\mathbf{X}$, and $E[\rho_q(Y - \mathbf{X}_0^\top \boldsymbol{\beta}_{0,q}^*)]$ is the variation of Y that cannot be explained through the reduced linear model, which is also possibly incorrect. Consequently, $\zeta(q)$ is nothing but the relative loss of explained variation in terms of the check distance that can be attributed to the lack-of-fit of the reduced q th quantile linear model compared to the full one. From the definition, it is clear that $0 \leq \zeta(q) \leq 1$. $\zeta(q) = 0$ is equivalent to saying that $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_0^*, \mathbf{0})$, i.e. no information is lost when considering only the restricted linear model. Note that unlike R^2 , which is a global measure, several $\zeta(q)$ ’s (for different values of q) are able to show a more complete picture of the effect of $\mathbf{X}_1$ both in the center and the tails. In other words, the proposed $\zeta(q)$ is a local measure of goodness of fit.

As noted by a referee, the proposed measure $\zeta(q)$ is only informative about the prediction qualities of the full and reduced models for the outcome under the asymmetric loss function, but it is not directly informative about the approximation properties of these models to the conditional quantile function under a precise loss function. To judge the approximation quality one might consider measuring a distance between $Q_q(Y|X)$ and $X^\top \boldsymbol{\beta}^*$ and compare it with the distance between $Q_q(Y|X_0)$ and $X_0^\top \boldsymbol{\beta}_0^*$. However, such a comparison is problematic because the target functions, $Q_q(Y|X)$ and $Q_q(Y|X_0)$, are different and further we need to use nonparametric methods to estimate those quantities. A possible solution to this problem could be to use partial quantile regression as explained in Angrist et al. (2006) or to consider a different type of coefficient as proposed in Noh et al. (in press).

Before moving into theoretical analysis of $\zeta(q)$ and its estimator, first we will present two interesting applications of $\zeta(q)$ when $d_0 = 1$ (coefficient of determination) and $d_1 = 1$ (covariate importance). We use the following model for illustration of $\zeta(q)$:

$$Y_i = 0.5 + X_{1i} - 2X_{2i} + \beta_3 X_{3i} + \nu g(X_{1i}) + \sigma \varepsilon_i, \quad i = 1, \dots, n, \quad (2)$$

where $g(X_{1i}) = 1 - \cos(\pi X_{1i}/2)$, $\sigma = 0.4$ and the $\mathbf{X}_i = (X_{1i}, X_{2i}, X_{3i})^\top$ ’s are i.i.d. and generated from a truncated multivariate normal distribution with the constraints $0 \leq X_{ki} \leq 1$, $k = 1, 2, 3$. The errors ε_i ’s are independent standard normal variables and are independent of the $\mathbf{X}_i$ ’s. The details of this model are given in Example 3 of Section 5. Note that ν controls the degree of the linear quantile regression model. Because it is difficult to calculate the value of $\boldsymbol{\beta}_q^*$ and $\boldsymbol{\beta}_{0,q}^*$ analytically, we use a Monte Carlo approximation for the computation of $\zeta(q)$, using a large sample of size 500,000.

2.1. Coefficient of determination ($d_0 = 1$)

If $d_0 = 1$, then $\boldsymbol{\beta}_{0,q}^*$ becomes $\xi_q = \arg \min_b E[\rho_q(Y - b)]$, which is the marginal q th quantile of Y , and then $\zeta(q)$ becomes

$$\zeta(q) = 1 - \frac{E[\rho_q(Y - \mathbf{X}^\top \boldsymbol{\beta}_q^*)]}{E[\rho_q(Y - \xi_q)]} \quad (3)$$

which we will call $R^*(q)$ hereafter. $R^*(q)$ is the quantile analogue of the well known Pearson’s correlation ratio $\eta^2 = 1 - E[Y - \mathbf{X}^\top \boldsymbol{\beta}^*]^2 / E[Y - E(Y)]^2$, i.e. the ‘theoretical’ R^2 for the linear mean regression model $E(Y|\mathbf{X}) = \mathbf{X}^\top \boldsymbol{\beta}^*$. For this reason, when the underlying distribution is asymmetric or in the presence of outliers, $R^*(0.5)$ could be used as a robust alternative to R^2 . Like η^2 , $R^*(q)$ lies in $[0, 1]$. $R^*(q) = 0$ corresponds to the case when $\mathbf{X}^\top \boldsymbol{\beta}_q^* = \xi_q$ with probability one, i.e. all components of $\boldsymbol{\beta}_q^*$ vanish except the first one, which coincides with ξ_q . In that case, no variability is captured by $\mathbf{X}$ via a linear

Copula-Based Regression Estimation and Inference

Hohsuk NOH, Anouar El GHOUGH, and Taoufik BOUEZMARNI

We investigate a new approach to estimating a regression function based on copulas. The main idea behind this approach is to write the regression function in terms of a copula and marginal distributions. Once the copula and the marginal distributions are estimated, we use the plug-in method to construct our new estimator. Because various methods are available in the literature for estimating both a copula and a distribution, this idea provides a rich and flexible family of regression estimators. We provide some asymptotic results related to this copula-based regression modeling when the copula is estimated via profile likelihood and the marginals are estimated nonparametrically. We also study the finite sample performance of the estimator and illustrate its usefulness by analyzing data from air pollution studies.

KEY WORDS: Dependence modeling; Profile likelihood; Semiparametric regression; Vine copula.

1. INTRODUCTION

Let $\mathbf{X} = (X_1, \dots, X_d)^\top$ be a random vector of dimension $d \geq 1$ and Y be a random variable with continuous cumulative distribution functions (cdf's) $F_1, \dots, F_d$ and F_0 , respectively. Y is our response variable and $\mathbf{X}$ is our set of covariates. We denote the density of X_j by f_j and that of Y by f_0 . For a given $\mathbf{x} = (x_1, \dots, x_d)^\top$, we will use $\mathbf{F}(\mathbf{x})$ as a shortcut for $(F_1(x_1), \dots, F_d(x_d))$. From the inspiring work of Sklar (1959), the cdf of $(Y, \mathbf{X})$ evaluated at $(y, \mathbf{x})$ can be expressed as $C(F_0(y), \mathbf{F}(\mathbf{x}))$, where C is the copula distribution of $(Y, \mathbf{X})$, which is the function from $[0, 1]^{d+1}$ to $[0, 1]$ defined by $C(u_0, u_1, \dots, u_d) = P(U_0 \leq u_0, U_1 \leq u_1, \dots, U_d \leq u_d)$, where $U_0 = F_0(Y)$ and $U_j = F_j(X_j)$, $j = 1, \dots, d$. This is nothing but a joint distribution function with margins that are uniform over the unit interval $[0, 1]$. Copulas enable us to model the dependence between variables separately from their marginal distributions and by specifying a copula we summarize all the dependencies between margins. See Nelsen (2006) for more about this subject. From the definition of a copula function, the conditional density of Y given $\mathbf{X} = \mathbf{x}$ is given by

$$f_0(y) \frac{c(F_0(y), \mathbf{F}(\mathbf{x}))}{c_{\mathbf{X}}(\mathbf{F}(\mathbf{x}))},$$

where $c(u_0, \mathbf{u}) \equiv c(u_0, u_1, \dots, u_d) = \partial^{d+1} C(u_0, u_1, \dots, u_d) / \partial u_0 \partial u_1 \dots \partial u_d$ is the copula density corresponding to C and $c_{\mathbf{X}}(\mathbf{u}) \equiv c_{\mathbf{X}}(u_1, \dots, u_d) = \partial^d C(1, u_1, \dots, u_d) / \partial u_1 \dots \partial u_d$ is the copula density of $\mathbf{X}$. Obviously, the conditional mean, $m(\mathbf{x})$,

of Y given $\mathbf{X} = \mathbf{x}$ can be written as

$$m(\mathbf{x}) = \mathbb{E}(Y w(F_0(Y), \mathbf{F}(\mathbf{x}))) = \frac{e(\mathbf{F}(\mathbf{x}))}{c_{\mathbf{X}}(\mathbf{F}(\mathbf{x}))}, \quad (1)$$

where $w(u_0, \mathbf{u}) = c(u_0, \mathbf{u}) / c_{\mathbf{X}}(\mathbf{u})$ and

$$e(\mathbf{u}) = \mathbb{E}(Y c(F_0(Y), \mathbf{u})) = \int_0^1 F_0^{-1}(u_0) c(u_0, \mathbf{u}) du_0. \quad (2)$$

The equality (1) shows that, given the marginals, one can obtain the mean regression function relating Y to $\mathbf{X}$ directly from the copula density, or equivalently the copula distribution of $(Y, \mathbf{X}^\top)^\top$. It also implies that the conditional mean is just a weighted mean with weights induced by the unknown "conditional" copula function w defined above. This relation is not new and has already been applied in Sungur (2005), Leong and Valdez (2005), and Crane and Van Der Hoeck (2008) to compute the mean regression function corresponding to several well-known copula families (Gaussian, t , Farlie–Gumbel–Morgenstern (FGM), Iterated FGM, Archimedean, etc.) with single ($d = 1$) and multiple covariate(s). To illustrate the idea, we briefly cite two examples:

- If the copula density of (Y, X_1) belongs to the FGM family with a parameter θ , that is, $c(u_0, u_1) = 1 + \theta(1 - 2u_0)(1 - 2u_1)$, then we have

$$m(x_1) = \mathbb{E}(Y) + \theta(2F_1(x_1) - 1) \int F_0(y)(1 - F_0(y)) dy. \quad (3)$$

A similar formula holds for the multiple covariate case; see Leong and Valdez (2005).

- If the copula of $(Y, \mathbf{X}^\top)^\top$ is Gaussian with correlation matrix

$$\Sigma_{Y, \mathbf{X}} = \begin{bmatrix} 1 & \boldsymbol{\rho}^\top \\ \boldsymbol{\rho} & \Sigma_{\mathbf{X}} \end{bmatrix},$$

then we have

$$m(\mathbf{x}) = \mathbb{E}\left[F_0^{-1}\left(\Phi(\mathbf{u}^\top \Sigma_{\mathbf{X}}^{-1} \boldsymbol{\rho} + \sqrt{1 - \boldsymbol{\rho}^\top \Sigma_{\mathbf{X}}^{-1} \boldsymbol{\rho}} Z)\right)\right], \quad (4)$$

where $\mathbf{u} = (\Phi^{-1}(F_1(x_1)), \dots, \Phi^{-1}(F_d(x_d)))^\top$, $Z \sim \mathcal{N}(0, 1)$, and Φ is the cdf of a standard normal distribution.

Hohsuk Noh is postdoctoral researcher, Institut de Statistique, Biostatistique et Sciences Actuarielles, Université Catholique de Louvain, Louvain la Neuve 1348, Belgium (E-mail: word5810@gmail.com). Anouar El Ghouh is Professor, Institut de Statistique, Biostatistique et Sciences Actuarielles, Université Catholique de Louvain, Louvain la Neuve 1348, Belgium (E-mail: anouar.elghouch@uclouvain.be). Taoufik Bouezmarni is Assistant Professor, Département de Mathématiques, Université de Sherbrooke, Sherbrooke, Quebec J1K 2R1, Canada (E-mail: taoufik.bouezmarni@usherbrooke.ca). H. Noh's research is supported by IAP research network P6/03 of the Belgian Government (Belgian Science Policy) and the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013)/ERC Grant agreement no. 203650. A. El Ghouh's research is supported by IAP research network P6/03 of the Belgian Government (Belgian Science Policy), and the contract "Projet d'Actions de Recherche Concertées" (ARC) 11/16-039 of the "Communauté Française de Belgique," granted by the "Académie Universitaire Louvain." The authors thank two anonymous referees, the associate editor, and the coeditor for their valuable suggestions that have significantly improved the article.

Note that in the single covariate case, we have $c_X(\mathbf{u}) \equiv c_{X_1}(u_1) = 1$ for all $u_1 \in [0, 1]$. In such a case, the weight function w coincides with the copula density c and (1) reduces to $m(x_1) = e(F_1(x_1)) = \mathbb{E}(Yc(F_0(Y), F_1(x_1)))$. Also, if the covariates are mutually independent, then $c_X(\mathbf{u}) = 1$ and $m(\mathbf{x})$ coincides with $e(\mathbf{F}(\mathbf{x}))$. In other words, $e(\mathbf{F}(\mathbf{x}))$, the numerator of $m(\mathbf{x})$ in (1), is the mean regression function of Y given $\mathbf{X}$ assuming independence between the covariates or, equivalently, assuming that the conditional density of $Y|\mathbf{X}$ is $f_0(y)c(F_0(y), \mathbf{F}(\mathbf{x}))$. Thus, in term of copulas, the mean regression function is the ratio of a numerator that only captures the mean dependence between Y and $\mathbf{X}$ and a denominator that captures the dependence within $\mathbf{X}$.

The equality (1) can also be used as an estimating equation. In fact, if $\hat{w}$, $\hat{F}_0$, and $\hat{F}_j$ are any given estimators for w , F_0 , and F_j , respectively, then m can obviously be estimated by

$$\hat{m}(\mathbf{x}) = \int_{-\infty}^{\infty} y \hat{w}(\hat{F}_0(y), \hat{\mathbf{F}}(\mathbf{x})) d\hat{F}_0(y), \quad (5)$$

where $\hat{\mathbf{F}}(\mathbf{x}) = (\hat{F}_1(x_1), \dots, \hat{F}_d(x_d))^\top$. To the best of our knowledge, such an approach has never been proposed or investigated in the literature, neither in the single nor in the multiple covariate case. To estimate w , one needs an estimator for the copula densities c and c_X . The copula density c_X can be obtained from c by integration. In fact,

$$c_X(\mathbf{u}) = \int_{-\infty}^{\infty} f_0(y)c(F_0(y), \mathbf{u})dy = \int_0^1 c(u_0, \mathbf{u})du_0. \quad (6)$$

Therefore, given an estimator $\hat{c}$ for c , one can easily estimate c_X using the plug-in method and then estimate m by (5).

Since, in the literature, there are many different methods available for estimating a copula and a cdf, $\hat{m}(\mathbf{x})$ defines a large new class of interesting estimators. Depending on the method of estimating the components in (5), $\hat{m}(\mathbf{x})$ can be a nonparametric or a semiparametric or a fully parametric estimator. For example, using nonparametric estimators for c , F_0 , and F_j , $j = 1, \dots, d$, leads to a fully nonparametric estimator. Nonparametric methods for estimating c include kernel-smoothing estimators (Gijbels and Mielniczuk 1990; Charpentier, Fermanian, and Scaillet 2006; Chen and Huang 2007; Omelka, Gijbels, and Veraverbeke 2009), Bernstein estimator (Bouezmarni, Rombouts, and Taamouti 2010), and spline estimator (Kauermann et al. in press) to name a few. In spite of the great flexibility, nonparametric methods are typically affected by the curse of dimensionality and come with the difficult problem of selecting a good smoothing parameter. On the other hand, imposing a parametric structure on both the copula and marginal distributions can lead to severely biased and inconsistent (fully parametric) estimators in case of misspecification. For these reasons and to avoid such problems as much as possible, we consider here a semiparametric approach where the copula is modeled parametrically but the marginal distributions are modeled nonparametrically. As will be shown in the next sections, the proposed method has many interesting properties both from a theoretical and a practical point of view. In particular, the asymptotic properties are easy to obtain, the numerical calculations can be done directly using existing software packages and, unlike many semiparametric methods, no iteration procedure is needed

to guarantee consistency. Also, the asymptotic variance can be estimated without any extra complications.

The plan of the article is as follows. Section 2 presents the general theoretical framework of the method with the necessary notations and assumptions. In Section 3, we identify the asymptotic representation of the proposed estimator in the single and multiple covariate cases. From this representation, we establish the asymptotic distribution of the estimator. Further, we propose and study an estimator of its asymptotic variance. In Section 4, we investigate theoretical properties of the estimator under misspecification. In Section 5, we present some numerical simulations designed to confirm the theoretical results and to compare the performance of our estimators with that of some competitors. Finally, in Section 6, we analyze data from air pollution studies to illustrate the usefulness of the proposed estimator. All the proofs appear in the Appendix.

2. THEORETICAL BACKGROUND

Let $\mathbf{X}_i = (X_{1,i}, \dots, X_{d,i})^\top$ and let $(Y_i, \mathbf{X}_i)$, $i = 1, \dots, n$, be an independent and identically distributed (iid) sample of n observations generated from the distribution of $(Y, \mathbf{X})$. Clearly, the shape and the performance of our estimator $\hat{m}$ in (5) will depend heavily on the methods of estimation for c , F_0 , and F_j . In this work, F_0 is estimated by a rescaled empirical distribution function:

$$\hat{F}_{0,n}(y) = \frac{1}{n+1} \sum_{i=1}^n I(Y_i \leq y).$$

Estimating the other cdf's F_j , $j = 1, \dots, d$, can also be done in the same way via $\hat{F}_{j,n}$. However, this results in a nonsmooth estimate $\hat{m}(\mathbf{x})$ as is illustrated in Figure 1, where we show the resulting estimator using $\hat{F}_{1,n}$ in the univariate case. To get a more visually attractive regression curve or surface, one should smooth the empirical cdf's $\hat{F}_{1,n}, \dots, \hat{F}_{d,n}$. The simple way to do that is to use a kernel-smoothing method. Let $k(\cdot)$ be a function, which is a symmetric probability density function and $h \equiv h_n \rightarrow 0$ be a bandwidth parameter. Then, a kernel-smoothing estimator of F_j is given by

$$\tilde{F}_{j,n}(x) = \frac{1}{n} \sum_{i=1}^n K\left(\frac{x - X_{j,i}}{h}\right),$$

where $K(x) = \int_{-\infty}^x k(t)dt$. The methods of estimating the marginals $F_1, \dots, F_d$ do not bring out any difference in asymptotic behavior of the estimator $\hat{m}(\mathbf{x})$ as long as they satisfy the following assumption:

Assumption A. For a given point of interest $\mathbf{x} \in \mathbb{R}^d$,

$$\tilde{F}_{j,n}(x_j) = n^{-1} \sum_{i=1}^n I(X_{j,i} \leq x_j) + o_p(n^{-1/2}), \quad j = 1, \dots, d.$$

Note that the kernel-smoothing estimator $\tilde{F}_{j,n}$ satisfies Assumption A if $nh^4 \rightarrow 0$. Evidently, this assumption also holds true for the rescaled empirical cdf $\hat{F}_{j,n}$.

We examined the effect of the methods of estimating the marginal distributions on the regression fit in an extensive simulation study and came to the conclusion that both estimators ($\hat{F}_{j,n}$ and $\tilde{F}_{j,n}$) show more or less similar performance. Here,

컴퓨터과학과

다중 인프라 기반의 공력 설계 교육을 위한 과학 워크플로우 통합 환경

김서영^{*}, 강혜정^{**}, 김윤희^{***}, 김종암^{****}

요 약

전 세계적으로 다양한 응용과학 분야의 연구자들은 그들의 연구 개발에 필수적인 고성능 컴퓨팅 자원의 확보와 복잡한 수치 해석 기법 개발을 위해 막대한 연구를 수행해 왔다. 특히 항공 우주 분야에서는 공력 최적 설계를 위해 소요되는 시간과 비용을 상당 부분 줄이기 위해 진보적인 수치기법을 개발하고 컴퓨팅 기술의 발전에 의존해왔으나, 여전히 1회의 실험에 막대한 비용 지출과 수개월의 소요 기간을 감수하고 있는 실정이다. 본 논문에서는 항공 우주 분야 연구자들의 연구 개발 편의성을 도모하고자 다양한 컴퓨팅 자원 인프라를 제공하는 통합 공학 교육·실험 환경을 소개하고 그 우수성을 보인다. 다양한 컴퓨팅 인프라 구조로의 연결을 통해 산재되어 있는 다수의 컴퓨팅 자원 활용이 가능하므로 다수의 교육 대상자 및 연구자들에게 장소에 제한 없는 실험 시도를 가능케 함으로써 연구 개발의 복잡성을 줄이고 생산성을 높일 수 있다. 또한 통합 환경을 교육에 활용하여 교육 효율성을 극대화시킬 수 있다.

An Integrated Scientific Workflow Environment over Multiple Infrastructures for Engineering Education of Aerodynamics

Seoyoung Kim[†], Hyejeong Kang^{**}, Yoonhee Kim^{***}, Chongam Kim^{****}

ABSTRACT

All around the world, numerous scientists have been carried out researches of e-Science to improve performance of computations and accessibility of their experimental flexibilities for a long times. However, they still have been in difficulty securing high-performance computing facilities. In case of Aerodynamics, for example, a single experiment costs a tremendous amount of budget and requires a span of more than 6 months even though researchers have been developed diverse improved mathematical methods as well as relied on advanced computing technologies to reduce runtime and costs. In this paper, we proposed a multiple infrastructure-based scientific workflow environments for engineering education in fields of design optimization of aircraft and demonstrated the superiority. Since it offers diverse kind of computing resources, it can offer elastic resources regardless of the number of tasks for experiments and limitations of spaces. Also, it can improve education efficiency by using this environment to engineering education.

Key words: Multiple Infrastructure(다중 인프라), Scientific Workflow(과학 워크플로우), e-Science(e-사이언스), Science Cloud(사이언스 클라우드), Aerospace(항공 우주), Engineering education (공학 교육)

※ 교신저자(Corresponding Author): 김윤희, 주소: 서울 시 용산구 청파로47길 100 (청파동 2가) 숙명여자대학교 새싹관 402호 (140-742), 전화: 02) 710-9664, FAX: 02) 710-9664, E-mail: yulan@sookmyung.ac.kr
접수일: 2012년 11월 5일, 수정일: 2012년 11월 23일
완료일: 2012년 11월 30일

[†] 준회원, 한국과학기술정보연구원
(E-mail: sssyy77@kisti.re.kr)

^{****} 준회원, 숙명여자대학교 컴퓨터과학부
(E-mail: hjkang@sookmyung.ac.kr)

^{****} 정회원, 숙명여자대학교 컴퓨터과학부

^{****} 정회원, 서울대학교 기계항공공학부
(E-mail: chongam@snu.ac.kr)

※ 본 연구는 숙명여자대학교 2012학년도 교내연구비 지원에 의해 수행되었음.

1. 서 론

전 세계적으로 e-사이언스 환경을 더욱 유연하고 효율적으로 구축하기 위한 연구가 활발하게 진행되고 있다[1]. e-사이언스[2]는 국내외에 위치한 연구자, 연구 장비, 연구 정보 등의 연구 자원을 사이버 공간에서 공동 활용하여 연구를 수행하는 차세대 연구 활동으로, 특히 복잡한 형태의 계산 중심 응용을 주로 다루는 전산 유체 역학(Computational Fluid Dynamics: CFD[3])분야에 대한 컴퓨터 과학 분야의 협력이 국내외로 진행되고 있는 실정이다[4]. 이러한 CFD분야에서 최적 설계 지원을 위해서는 고성능 컴퓨팅 자원 확보뿐만 아니라, 전문 연구가를 위한 실험에 필수적인 과학 워크플로우 설계 도구 역시 함께 요구된다. 현재 전문 연구가를 위해 고수준의 연구 환경을 지원하면서 동시에 고성능 계산 자원의 접근을 허용하는 e-사이언스 환경에 대한 연구는 미미하여 CFD를 다루는 다양한 응용과학 분야 중 항공우주 응용을 대상으로 하는 e-사이언스 프로토타입을 개발하였다[5]. 기존 연구[5,6]의 경우, 그리드 환경[7]으로의 자원 확장으로 대규모의 고성능 컴퓨팅 자원 활용이 가능했으나 자원을 공유하여 사용하는 만큼 특정시간대에 집중적으로 다수의 작업 수행 요구가 발생하면 자원 사용을 위한 긴 대기 시간을 초래하는 문제가 발생하였다. 본 논문에서는 자원 요구가 늘어날 경우, 가상화 기술을 바탕으로 동적인 자원 확장이 가능한 클라우드를 다중 인프라에 포함하여 실험 수행이 가능하도록 확장한다. 즉, 로컬 클러스터와 그리드 컴퓨팅 환경뿐만 아니라 동적인 자원 확장이 가능한 클라우드 환경까지 모두 접근 가능한 웹 기반의 워크플로우 설계 서비스를 제안한다.

본 논문의 2장에서는 관련 연구에 대해 논한 후, 본 논문에서 프로토타입으로 한 대상 응용에 대한 특징 및 실행 속성을 3장에서 간단하게 알아본다. 4장에서는 제안한 포털 시스템의 전체 시스템 구조에 대한 설명과 성능 검증 실험을 논하며, 그 후에는 이를 기반으로 구현된 포털을 소개한 뒤, 마지막으로 결론 및 향후 연구에 대해 논하겠다.

2. 관련 연구

과학 분야에서는 방대한 데이터의 관리와 분석을

자동화하기 위해서 과학 워크플로우를 이용해왔다. GridFlow[8], P-GRADE[9]와 같은 과학 워크플로우 시스템의 경우는 대규모의 그리드 자원을 기반으로 수행된다.

GridFlow는 그리드 환경에서 워크플로우 관리 서비스를 제공하는 시스템이다. 그리드 자원 관리 시스템을 통해 워크플로우의 관리 서비스를 제공하여 시뮬레이션 실행과 모니터링 기능을 가능하도록 해준다. 하지만 이 시스템은 응용 형태로 제공되어 별도의 설치 과정 및 컴퓨팅 환경과의 연결을 위한 추가 설정이 필요하므로 연구자에게 불편함을 준다는 단점이 존재한다. P-GRADE 포털은 다양한 그리드(EGEE, UK, NGS, GIN VO, OSG, TeraGrid 등) 자원에 접근하여 큰 단위의 작업들을 생성, 실행, 관리할 수 있는 서비스를 제공한다. 그리드 자원 관리를 위한 미들웨어를 제공하며, 워크플로우 작업을 변경하지 않고 다른 그리드 자원으로 쉽게 마이그레이션하는 기능도 제공한다. 대용량의 작업 수행에 다양한 그리드 자원을 지원한다는 점에서 상호운용성을 제공하지만, 단순한 작업의 반복 실행만을 수행하기 때문에 복잡한 수치에 대한 연산 수행에는 한계가 있다.

최근 그리드의 한계를 극복하기 위해 이기종 분산 컴퓨팅 자원에 속하는 계산 그리드와 클라우드를 통합·구축하여 과학 워크플로우 수행을 지원하는 연구가 활발하게 진행되고 있다. VGrADS(Virtual Grid Application Development Software)프로젝트[10]는 고성능 컴퓨팅 수행을 위해 기존 그리드 환경과 공용 클라우드 자원인 아마존 EC2를 동시에 이용 가능하도록 컴퓨팅 자원을 확장하였다. ASKALON[11]의 경우, 기존 응용 개발 및 실행 환경을 확장하여 그리드와 클라우드 환경을 동시에 활용할 수 있는 응용개발환경 구축을 목표로 두 가지 환경을 통합한 서비스를 연구하고 있다. 사용자는 UML 그래픽 도구를 이용하여 워크플로우를 쉽게 설계할 수 있으며, 'Enactment engine'이 워크플로우에 최적인 자원을 스케줄링하는 역할을 담당한다. Nimrod툴[12]는 High Throughput Computing(HTC) 중심 응용 수행을 지원하는 미들웨어 통합 실행 프레임워크로, 로컬 클러스터 또는 대규모 그리드 자원 상에서 응용 수행이 가능하도록 그리드 미들웨어 및 인터페이스를 구축하고 더불어 효율적인 실행을 위한 스케줄링 정책

An Allocation and Provisioning Model of Science Cloud for High Throughput Computing Applications

Seoyoung Kim
National Institute of
Supercomputing and
Networking, KISTI
Daejeon, Republic of Korea
ssyyy77@kisti.re.kr

Jik-Soo Kim
National Institute of
Supercomputing and
Networking, KISTI
Daejeon, Republic of Korea
jiksoo.kim@kisti.re.kr

Soonwook Hwang
National Institute of
Supercomputing and
Networking, KISTI
Daejeon, Republic of Korea
hwang@kisti.re.kr

Yoonhee Kim^{*}
Dept. of Computer Science
Sookmyung Women's Univ.
Seoul, Republic of Korea
yulan@sookmyung.ac.kr

ABSTRACT

Recent cloud computing enables numerous scientists to earn advantages by serving on-demand and elastic resources whenever they desire computing resources. This science cloud paradigm has been actively developed and investigated to satisfy requirements of the scientists such as performance, feasibility and so on. However, effective allocation and provisioning virtual machines on clouds are still considered as a challenging issue in scientists using high throughput computing, since it determines whether they can earn benefits from economy of scale in clouds or not. Moreover, allocating the “right” provisioned cloud resources on an optimal data center is very important as performance can vary widely depending on where and under what circumstances it actually runs. In these reasons, it is required that an appropriate and suitable model for science cloud to support increasing scientists and computations.

In this paper, we present an allocation and provisioning model of science cloud, especially for high throughput computing applications. In this model, we utilize job traces where statistical method is applied to pick the most influential features for improving application performance. With the feature, the system determines where VM is deployed (allocation) and which instance type is proper (provisioning). An adaptive evaluation step which is subsequent to the job execution enables our model to adapt to dynamical computing environments. We show performance achievements as comparing the proposed model with other policies through experiments. Finally, we expect that improvement

on performance as well as reduction of cost from resource consumption through our model.

Categories and Subject Descriptors

C.2.4 [Computer-Communication Networks]: Distributed Systems

General Terms

Resource Management

Keywords

Science Cloud, High Throughput Computing, Job Profiling, Cloud Provisioning, PCA(Principal Components Analysis)

1. INTRODUCTION

Cloud computing nowadays enables numerous scientists to earn advantages by serving on-demand and elastic resources whenever they desire computing resources. This science cloud paradigm has been actively developed and investigated to satisfy requirements of the scientists such as performance, feasibility and so on. However, effective allocation and provisioning virtual machines on clouds are still considered as a challenging issue in scientists using high throughput computing, since it determines whether they can earn benefits from economy of scale in clouds or not. Moreover, allocating the “right” cloud resources(i.e., virtual machine type, cloud service site) on an optimal data center is very important as performance can vary widely depending on where and under what circumstances it actually runs. In these reasons, it is required that an appropriate and suitable model for science cloud to support increasing scientists and computations. Fortunately, job history which potentially involves status of resources where jobs are executed as well as properties of applications can lead us to devise performance-optimized provisioning scheme with meaningful factors through an appropriate processing of the traces.

In this paper, we present a **PHAP**(ProfilHistorical factor based Allocation and Pro- visioning model) which is

^{*}Corresponding Author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CAC '13, August 05 - 09 Miami, FL, USA
Copyright 2013 ACM 978-1-4503-2172-3/13/08 ...\$15.00.

an allocation and provisioning model of science cloud, especially for high throughput computing applications. In this model, we utilize job traces where statistical method is applied to pick the most influential features for improving application performance. With the feature, the system determines where VM is deployed (allocation) and which instance type is proper (provisioning). An adaptive evaluation step which is subsequent to the job execution enables our model to adapt to dynamical computing environments. We show performance achievements as comparing the proposed model with other policies through experiments. Finally, we expect that improvement on performance as well as reduction of cost from resource consumption through our model.

The rest of this paper is structured as follows. Section 2 discusses related work and Section 3 presents introduction of the proposed system model where application model will be introduced. We discuss allocation and provisioning model in details in Section 4, while Section 5 discusses experiment and its results. Finally, we conclude this paper and discuss future work in Section 6.

2. RELATED WORK

As growing attentions to clouds in several science communities, a lot of efforts to provide optimized and facilitated virtual environments on science clouds have been made. It is associated with that almost science applications typically involve long-term computer executions and huge dataset processing and large-scale computations requiring the availability of abundant computing infrastructures. Therefore, it is important to identify requirements of users as well as applications and to allocate adequate quantities of resources. Here, we are going to introduce several related works focused on two aspects; 1) cloud provisioning and resource allocating issue 2) utilization of job traces.

Lei Wang et al.[11] had investigated cloud modeling for scientific applications and evaluated their model using two kinds of workloads-MTC(Many Task Computing) and HTC (High Throughput Computing). The model adopted a policy which is based on the ratio of waiting jobs to total available VM to lease VM.

Another work, [12] proposed a controlling system which allocates appropriate resources through monitoring and analyzing current workloads of applications. In [12], a virtual server is operated on a group of physical machines and each server is responsible for particular application on the heterogeneous machines. The system, in addition, predicts future workload through analysing resource utility and real-time requirements of applications and schedules using the predicted information. In this way, diverse physical machines can be maintained in optimal status.

However, the above two studies mainly focused on the proper resource management and utilization without considering performance issue.

On the other hand, there also exist some researches using job histories. One of them [6] had proposed for predicting application runtime and queue wait time. In [6], it had exploited genetic algorithm to search similar jobs among job histories. This study predicts two metrics by mining historical workloads and the two are correspond to job execution time and wait time in queue. It firstly searches the most similar job among the job histories using genetic algorithm considering both characteristics of applications and resources. With the discovered one, it predicts two metrics (application

runtime and queue wait time) using instance-based learning techniques. However, the main focus of the [6] do not lie on resource managements, but only runtime predictions.

Meanwhile, Bhuvan et al. [10] also exploited application profiling in shared resources to offer guaranteed performances as well. Their method for analysis, on the other hand, is different with ours since they consider only an aspect among various features.

An allocation and provisioning model we present focuses on the followings: First, selection of a cloud site which serves optimal performance without bursty workloads and results in efficient resource allocation. Secondly, a proper virtual machine provisioning which contributes to satisfy requirements of user and application. To determine the above two, our model utilizes three representative factors through profiling job traces. It is performed adaptably by evaluating profiled results and enables to adapt to current status of resources regardless of failure or overloading on systems.

3. SYSTEM MODEL

3.1 Target Application Model

Our model mainly focused on two kinds of workload types; High Throughput Computing (High Throughput Computing) and Many Task Computing (MTC) are generally found in most of workloads from various scientific applications. These kinds of workloads which are also called as ‘Bag-of Tasks(BoT)’ have several characteristics in common like the followings:

- Massively Paralleled
- Independent in each task (minor dependent tasks)
- Adoption of ‘throughput’ as a main metric(over a fixed period of time)

Assume that a job(j_i) is submitted by some user has n number of tasks($t_{x+1}, t_{x+2}, \dots, t_{x+n}$, here x is arbitrary job id) by parameter sweeping. The running time of each task is associated with overall total makespan of a job(j_i). ‘throughput’ indicates the number of completed tasks in a fixed period of time and so we can induce a throughput of j_i as the following(Eq. 1).

$$throughput(j_i) = \frac{n}{makespan(j_i)} \quad (1)$$

To increase throughput of j_i , we have to reduce makespan of j_i where all of n number of tasks should be completed since n is fixed by user when he submits the job. However, it is a hard work to predict makespan even though duration of one task is given, since each of them can have dynamic execution time by varied parameter value with independent operations. Hence, it is essentially required to recognize properties of the submitted job(i.e. tasks).

3.2 System Model

Figure 1 shows an abstract architecture where proposed PHAP(Profiling Historical factor based Allocation and Provisioning model) model is applied. It consists of five layers in total and main layer is the third layer named as ‘Profiling & Provisioning Layer’. Once users desire to compute their application, make their requests through the first layer denoted as ‘User request Layer’. The made requests in the first layer are sent to the next layer; ‘Job Management

■ 2013년도 학생논문 경진대회 수상작

하이브리드 클라우드에서 과학 응용을 위한 정책 기반의 적응형 작업 스케줄링

(Adaptive Policy-based Task Scheduling for Scientific
Applications in Hybrid Cloud)

고 정 인 [†] 강 혜 정 [†] 김 윤 희 ^{††}
(Jung In Koh) (Hyejeong Kang) (Yoonhee Kim)

요 약 방대한 계산 처리를 필요로 하는 과학 응용은 그리드를 포함한 다양한 계산 자원을 수년간 이용해왔다. 하지만 기존의 계산자원만으로 과학 응용의 가변적인 작업 부하 및 동적 자원 요구를 충족시키기에는 어려움이 있기 때문에 필요에 따라 자원을 빌려 쓸 수 있는 클라우드 자원으로의 확장이 요구된다. 또한 클라우드 자원 이용을 통해 발생하는 비용, 성능 오버헤드 등의 고려를 통해 응용의 특성 및 서비스 수준 협약(Service Level Agreement)을 보장하는 동적인 자원 할당이 가능한 작업 스케줄링 기법이 요구된다. 본 논문에서는 하이브리드 클라우드 자원의 이용 효율을 향상시키고, 주어진 데드라인 안에서 응용 수행을 마칠 수 있도록 수행 중 필요에 따른 작업 재배치를 위해 온디맨드로 다중 인프라 자원 할당이 가능한 정책 기반의 적응형 작업 스케줄링 기법을 제안하고자 한다. 실험을 통하여 정책 조건을 만족하면서 주어진 데드라인 내에서 다중 인프라 자원을 충분히 활용하여 작업을 수행하는 결과를 보임으로써 제안하는 기법의 우수성을 검증하였다.

키워드: 적응형 스케줄링, 하이브리드 클라우드, 정책 기반 스케줄링, 계산 과학 응용

Abstract As many scientific workflows require high computing power to fulfill their goals in a reasonable time period, diverse resources such as Grids have been considered for several years. However, the existing computing resources are difficult to meet dynamic resource demands caused by variable workload of recent scientific applications. As Cloud offers on-demand resources for computation-intensive applications which require large-scale computing power in certain time period, hybrid cloud capable of scaling out multi-infrastructure more easily satisfies the need. In a hybrid cloud, task scheduling is one of difficult problems to provide good matching between diverse resources and the requirement of application characteristics. Particularly, a scheduling policy, which satisfies user's SLA (Service Level Agreement) such as performance or cost, can be the first priority to achieve the goal of application execution. In this paper, we propose a task scheduling service that provides on-demand multiple infrastructures by applying policies to execute an application in a given deadline. The goal of the task scheduling service is maximizing the utility of multiple infrastructures and providing an adaptive scheduling which can change the initial scheduling plan during execution to get better results. We have evaluated the performance of our scheduling service. The result shows that with deadline constraints the service can enhance the utilization of various infrastructures under policies.

Keywords: Adaptive scheduling, hybrid cloud, policy-based scheduling, scientific application

· 본 연구는 숙명여자대학교 2013학년도 교내연구비 지원에 의해서 수행되었음

[†] 학생회원 : 숙명여자대학교 컴퓨터과학과
jungin@sookmyung.ac.kr
hjkang@sookmyung.ac.kr

^{††} 종신회원 : 숙명여자대학교 컴퓨터과학과 교수
yulan@sookmyung.ac.kr
(Corresponding author)

논문접수 : 2013년 5월 22일

심사완료 : 2013년 9월 9일

Copyright©2013 한국정보과학회 : 개인 목적이거나 교육 목적인 경우, 이 저작물의 전체 또는 일부에 대한 복사본 혹은 디지털 사본의 제작을 허가합니다. 이 때, 사본은 상업적 수단으로 사용할 수 없으며 첫 페이지에 본 문구와 출처를 반드시 명시해야 합니다. 이 외의 목적으로 복제, 배포, 출판, 전송 등 모든 유형의 사용행위를 하는 경우에 대하여는 사전에 허가를 얻고 비용을 지불해야 합니다.

정보과학회논문지 : 컴퓨팅의 실제 및 레터 제19권 제11호(2013.11)

1. 서론

정보화 사회가 고도화됨에 따라 고성능 컴퓨터 등의 각종 IT계산자원의 수요가 급격히 증가하면서 더욱 높은 계산능력을 가진 자원을 확보하려는 노력이 증가하고 있다. 그 중에서 그리드 컴퓨팅 기술은 지역적으로 분산된 이기종 계산 자원과 대용량 저장 장치, 다양한 과학 기술 연구 장비 등을 통합하여 과학기술 및 정보 서비스를 제공하기 때문에 많은 계산과학 분야에서 이용되고 있다. 하지만 현대 응용의 동적인 작업 부하로 인해 응용 수행에 필요한 계산 자원의 수요가 가변적이고 이에 따라 필요한 양을 유동적으로 제공해야 할 필요가 있다. 그리드 자원의 경우 적절한 양을 탄력적으로 서비스하기에는 어려움이 따른다. 반면, IT분야에서 새로운 패러다임으로 떠오르는 클라우드 컴퓨팅은 각종 컴퓨팅 자원들을 가상화 기술로 통합하여 필요한 양만큼 이용 가능한 IT 환경을 제공하는 기술로서 기존 계산 자원들과는 달리 자원 사용의 유연성을 제공한다. 그러나 단순히 클라우드 자원만을 사용하게 될 경우 기존 자원의 낭비가 발생할 수 있으며, 가상화 기술을 기반으로 하하는 클라우드 컴퓨팅의 특성 상, 성능면에서 추가적인 오버헤드가 발생할 수 있다. 따라서 각 분산 자원의 서로 다른 특성을 파악하고 높은 계산 성능을 위해 다양한 분산 컴퓨팅 자원을 포함하는 하이브리드 클라우드에 관한 효율적인 자원 관리 및 작업 스케줄링 기법이 반드시 필요하다. 특히, Amazon EC2[1], Windows Azure[2], Rackspace[3]와 같은 상용 클라우드 컴퓨팅의 경우 사용시간에 비례하여 비용이 증가하기 때문에 자원 사용에 따른 비용 발생 여부 및 비용 발생량이 핵심적인 고려사항이 될 수 있다.

본 논문에서는 슈퍼컴퓨팅, 그리드, 클라우드 컴퓨팅과 같은 분산 컴퓨팅 인프라들을 기반으로 하는 이기종 자원을 효율적으로 활용하기 위한 적응형 작업 스케줄링 기법을 제안한다. 또한 다양한 이기종 자원으로 인해 사용자의 응용 수행 요구가 복잡해지면서 사용자와 제공자간의 서비스 수준 협약(SLA: Service Level Agreement)도 복잡해졌다. 이러한 SLA를 충족시키기 위해서 계산 과학 응용의 특성과 자원의 속성을 고려한 성능 및 비용 정책을 수립하고, 정책을 수용할 수 있는 테드라인 기반의 동적 작업 스케줄링 기법에 대한 연구를 수행하고자 한다. 전산 유체 역학 분야(CFD; Computational Fluid Dynamics)[4]의 응용을 대상으로 제안한 기법의 성능 측정 및 비교/분석을 위해서 스케줄링 기법에 대한 실험을 수행하여 제안하는 기법의 우수성을 검증하고자 한다.

본 논문의 구성은 다음과 같다. 1장의 서론에 이어 2장에서는 관련 연구들을 살펴보고, 3장에서는 본 논문에서

제안하는 스케줄링 서비스를 적응형 스케줄링 방식과 정책 기반의 스케줄링 방식으로 설명한다. 4장에서는 실험 환경 및 다양한 실험 결과에 대해 살펴본다. 마지막으로 5장에서 결론을 도출한다.

2. 관련 연구

여러 가지 자원들을 복합적으로 이용할 경우 다양한 정책들을 고려하여 작업들을 효율적으로 클라우드 자원에 스케줄링 하기 위한 연구가 활발하게 진행되고 있다.

Bossche R, et al.[5]은 그리드와 클라우드 자원을 통합한 하이브리드 클라우드 환경에서 테드라인 내에 계산 비용 및 데이터 전송 비용을 고려하여 가장 효율적인 자원을 선택할 수 있도록 하는 요소들을 확률적 성능 평가 모델을 통해 구하고 수식으로 표현하였다. 응용은 BoT (Bag of Tasks) 워크플로우 모델로 정의하였고 자원은 GoGrid[6]와 Rackspace, Amazon EC2를 실제 속성(인스턴스 가격, 네트워크 등)에 따라 정의하여 시뮬레이션을 통해 비용 효율적인 스케줄링 시 고려해야 할 인자들에 대해 정의하였다. 그러나 이 연구에서 제안한 기법을 활용하기 위해서는 충분한 이력 데이터가 쌓여야 의미가 있으며, 해당 논문에서도 이력으로 쓸 데이터가 부족하여 확률적 모델을 통한 성능 평가를 수행하였다.

HCOC[7]는 워크플로우 응용의 테드라인을 고려한 자원 대역 및 수행 비용 최적화 알고리즘을 제안한다. 테드라인을 워크플로우의 CP(Critical Path)로 설정한다. CP는 최고 사양의 자원 이용 시 응용의 수행 시간을 의미하며 이상적인 테드라인 내에 작업 수행을 보장하기 위한 동적 스케줄링이 가능하다. 그러나 BoT 속성을 가진 응용에 대해서는 테드라인을 CP로 설정하기에는 무리가 있다.

VGrADS[8]는 시간적 요소에 대한 고려가 필수적으로 수반되는 응용의 수행을 위해 스케줄링 정책 및 인프라 구축을 목표로 진행되는 프로젝트이다. Tera-Grid[9], Amazon EC2, 로컬 클러스터와 Eucalyptus[10](사설 클라우드)로 구성된 그리드와 클라우드 기술의 협동 사용으로 사이언스 워크플로우의 수행 성능을 균형 있게 하며 안정성, 비용 고려 방식을 계획하는 새로운 워크플로우를 가능하게 하는 것을 목표로 한다. 하지만 Fault-tolerance를 지원하는 가상 자원 프로비저닝만을 제공하며, 그 밖의 정책(테드라인, 비용 등)을 고려한 프로비저닝 연구는 여전히 진행 중이고 공개된 연구 결과는 존재하지 않는다.

Aneka[11] 프로젝트에서는 테드라인을 우선순위로 고려하는 프로비저닝을 연구한다. 이 시스템은 스케줄러 서비스와 프로비저닝 서비스로 이루어져 있으며, 스케줄러 서비스가 작업들의 평균 실행 시간과 테드라인을 고려해서 자원의 양을 결정하면 프로비저닝 서비스가 해당 자원을 물리적으로 할당하고, 스케줄러 서비스에서 불필

과학 계산 응용 실행을 위한 하이브리드 클라우드에서의 SLA기반 VM 오토-스케일링 기법

(A SLA-based VM Auto-Scaling Method in Hybrid Cloud Computing for Scientific Computational Applications)

강 혜 정 ^{*}
(Hyejeong Kang)

고 정 인 ^{*}
(Jung In Koh)

김 윤 희 ^{**}
(Yoonhee Kim)

요 약 클라우드 컴퓨팅은 동적인 자원 사용을 가능하게 함으로써 계산 과학 응용 수행에 활발히 이용되어 왔다. 계산 과학 분야에서 클라우드 컴퓨팅 환경을 이용하는 데 있어 중요한 이슈 중 하나는 응용 수행에 실제 필요한 양만큼 자원을 할당하여 사용하는 것이다. 실제 필요한 양만큼의 자원을 사용함으로써 자원 사용비용을 줄이기 위함이다. 그러나 가변적인 작업 부하를 갖는 현대 응용의 특성상 이는 매우 어려운 작업이다. 기존 상용 클라우드들은 규칙 기반의 메커니즘을 이용하여 동적인 자원할당을 시도하고 있으나 대부분의 기법들은 단순히 자원의 성능 지표만을 기준으로 삼음으로써 응용의 Service Level Agreement(SLA; e.g. 데드라인, 비용 등)는 고려하지 않는다. 본 논문에서는 사실/상용 클라우드로 구성된 하이브리드 클라우드 환경에서 가변적인 자원 요구에 따라 자원을 할당하고 SLA 위반을 최소화하는 오토-스케일링 기법을 제안한다. 실험을 통해 주어진 데드라인 내에서 동적인 자원 활용을 통해 작업을 수행하는 결과를 보임으로써 제안하는 오토-스케일링 기법의 우수성을 검증한다.

키워드: 오토-스케일링, 하이브리드 클라우드 컴퓨팅, SLA, 다중 정책

Abstract Cloud computing facilitates on-demand resource usage. Scientists can perform large-scale scientific computational experiments over cloud environment. Executing experiments in a timely manner with the proper number of resources is important for the scientists. Scientists can save the cost by using resources actually needed. However, providing resources for processing dynamic workload of modern applications have difficulty. Rule-based and schedule-based mechanisms have been tried to allocate resources dynamically, but most of the auto-scaling methods just simply support performance metric such as CPU utilization but rarely are aware of Service Level Agreements (SLA) including execution deadline or cost. In this paper, we propose an auto-scaling method that automatically allocates resources depending on variable resource requirements in hybrid clouds satisfying a user's requirements on SLA. We have evaluated the performance of our auto-scaling method. The result shows that the method can perform a scientific application with the proper number of resources in deadline constraints.

Keywords: auto-scaling, hybrid cloud computing, SLA, multi-policies

· 이 논문은 2013년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. NRF-2013R1A1A3007866)

· 이 논문은 2013 한국컴퓨터종합학술대회에서 '하이브리드 클라우드 환경에서의 SLA기반 오토-스케일링 기법'의 제목으로 발표된 논문을 확장한 것임

^{*} 학생회원 : 숙명여자대학교 컴퓨터과학부
hjkang@sookmyung.ac.kr
jungin@sookmyung.ac.kr

^{**} 종신회원 : 숙명여자대학교 컴퓨터과학부 교수
yulan@sookmyung.ac.kr
(Corresponding author)

논문접수 : 2013년 8월 19일

심사완료 : 2013년 10월 11일

Copyright©2013 한국정보과학회 : 개인 목적이나 교육 목적의 경우, 이 저작물의 전체 또는 일부에 대한 복사본 혹은 디지털 사본의 제작을 허가합니다. 이 때, 사본은 상업적 수단으로 사용할 수 없으며 첫 페이지에 본 문구와 출처를 반드시 명시해야 합니다. 이 외의 목적으로 복제, 배포, 출판, 전송 등 모든 유형의 사용행위를 하는 경우에 대하여는 사전에 허가를 받고 비용을 지불해야 합니다.

정보과학회논문지 : 시스템 및 이론 제40권 제6호(2013.12)

1. 서 론

최근 계산 과학 분야에서 대용량 컴퓨팅 자원의 요구를 능동적으로 만족시키기 위해, 자원을 필요한 만큼 빌려 쓸 수 있는 클라우드 기술을 적용하여 과학 클라우드(Science Cloud)를 구축하는 연구가 세계 여러 곳에서 진행되고 있다. 빅데이터 처리 응용이나 계산 과학 응용(예, 과학 워크플로우)은 고성능 컴퓨팅(HPC: High Performance Computing) 또는 다처리 컴퓨팅(HTC: High Throughput Computing)을 요구하는 대규모 사이즈 응용(MTC: Many Task Computing)으로 정의할 수 있다. 이런 응용은 고성능의 자원을 장시간 활용해야 하며 안정적이고 지속적인 자원의 공급이 응용 실행의 필수이다. 이는 엑사스케일(Exascale) 컴퓨팅, 페타스케일(Petascale) 컴퓨팅의 연구로 이어지고 있으며 이런 환경에서 성공적인 작업 수행을 위한 실행관리, 자원 관리 등을 제공하는 계산 문제 풀이 환경(Computational Problem Solving Environment)에 대한 연구가 더욱 절실하다.

이런 문제 풀이 환경에 필요한 클라우드 자원을 효율적으로 활용하기 위한 방법으로 온-디맨드 자원 가상화 특성을 활용한 오토-스케일링 기법을 고려할 수 있다. 오토-스케일링 기법은 응용의 수행 중에 작업 수행을 위한 가상 자원의 수를 줄이거나 늘리는 방법으로 AWS(Amazon Web Service)[1]의 “Auto-scaling” 서비스가 대표적인 예이다. AWS나 Scalr[2]와 같은 오토-스케일링의 서비스들은 사용자가 정의한 규칙(예, CPU 이용률이 60% 이상일 때, 가상머신 2개 추가 할당)을 기반으로 자원의 규모를 결정한다. 단순한 방법으로 오토-스케일링을 가능케 하는 장점이 있지만 현대 응용의 가변적인 워크로드로 인한 동적인 자원 요구 패턴을 만족하지 못하는 문제점이 있다. 특히, 응용의 자원 요구가 충족되지 않을 경우 응용의 데드라인 위반, 수행 실패의 증가와 같은 SLA(Service Level Agreement) 위반으로 이어질 수 있는 심각한 문제점이 존재한다.

본 논문에서는 하이브리드 클라우드 컴퓨팅 환경에서 자원을 효율적으로 활용하기 위한 오토-스케일링 기법을 제안한다. 제안하는 오토-스케일링 기법은 데드라인, 비용/성능 정책과 같은 SLA를 충족한다. 또한 계산 집약, I/O 집약, 데이터 집약 등 응용의 특성과 함께 상용 클라우드의 인스턴스 타입을 고려함으로써 비용 효율적인 오토-스케일링이 가능하다.

본 논문의 구성은 다음과 같다. 1장의 서론에 이어 2장에서는 관련 연구들을 살펴보고, 3장에서는 본 논문에서 제안하는 SLA기반 오토-스케일링 기법을 적용한 프레임워크 구조에 대하여 설명한다. 4장에서는 SLA 기반 오토-스케일링 기법에 대해 설명하고 5장에서는 실

험에 대해 살펴본다. 마지막으로 6장에서 결론을 맺는다.

2. 관련 연구

클라우드 컴퓨팅은 가상화 기술을 통해 자원의 무한한 제공이 가능하고, 자원 규모의 확장과 축소가 용이하다. 이로 인해 효율적인 자원 관리 방법으로 오토-스케일링 기법에 대한 연구가 활발히 이루어지고 있다.

대표적인 오토-스케일링 기술로는 AWS[1]의 “Auto-scaling” 서비스가 존재함을 앞서 언급하였다. 이 서비스는 사용자가 정의한 규칙을 통해 자원의 규모를 확장하고 축소하는 규칙 기반의 자원 할당 자동화 기법으로 CPU 사용량, 디스크 사용량과 같은 하드웨어의 성능과 관련된 수치를 기준으로 한다. 기준이 되는 성능치에 상한 값과 하한 값을 정하고 그 값을 기준으로 미리 정해진 수만큼 가상머신을 추가하거나 제거하여 자원의 규모를 확장하거나 축소시키는 방법이다. 비슷한 서비스로 Windows Azure[3]에서 사용되는 Paraleap[4]과 Scalr[2], RightScale[5]이 존재한다. 규칙 기반의 스케일링 기법은 비교적 단순한 방법으로 동적인 자원 할당이 가능하지만 복잡한 수행 패턴을 갖는 작업에 대해서 실제로 필요한 만큼 가상머신이 추가/제거되지 않을 수 있다. 이러한 문제는 작업 수행 데드라인이나 자원 사용 비용과 같은 SLA를 갖는 작업들에서는 SLA 위반이라는 큰 문제를 야기할 수 있다.

작업 수행 데드라인이나 자원 사용 비용 등을 고려한 오토-스케일링 기법 연구로는 [6,7]이 존재한다. [6]에서는 사용자가 요청한 작업의 데드라인 내에서 작업 수행이 가능하도록 오토-스케일링을 수행한다. 이 연구에서는 작업의 데드라인뿐만 아니라 자원 할당에 따른 비용까지 고려하였다. 그러나 상용 클라우드를 단독으로 사용하였을 경우의 자원 할당만을 고려하였다는 제한사항이 존재한다. [7]은 자원 사용 비용을 최소화하기 위해 물리머신(Physical Machine)의 사용량과 자원 사용료 간에 트레이드-오프가 최적인 물리머신 사용량을 찾아 해당하는 만큼 가상머신을 스케일링하는 기법을 제안한 연구이다. 가상머신의 개수를 추가/제거하여 자원 규모를 확장하고 축소하는 수평적(Horizontal) 스케일링과 가상머신의 하드웨어 사양을 조절하는 수직적(Vertical) 스케일링 기법을 혼합 활용한다. 그러나 이 연구에서도 수평적 스케일링 시에는 고정된 개수의 가상머신만을 추가/제거하여 복잡한 수행 패턴의 작업에 대하여 동적인 자원 할당이 가능하다고 하기엔 부족하다.

본 논문에서는 하이브리드 클라우드 환경에서 데드라인과 자원 사용 비용과 같은 SLA를 고려하여 자원 할당이 가능한 오토-스케일링 기법을 제안한다. 지속적인 모니터링을 통하여 자원 할당이 이루어지며 자원 할당

하둡을 이용한 소셜네트워킹의 TV광고효과 분석 시스템 설계

A Design of Analysis System on TV Advertising Effect of Social Networking Using Hadoop

허 서 연¹ 김 윤 희*
Seoyeon Hur Yoonhee Kim

요 약

빅데이터가 화두가 되면서, 그 대표적인 예인 SNS를 이용한 서비스 개발도 활기를 띠고 있다. SNS는 기존 매체와는 다르게 실시간으로 의견을주고받는 하나의 장으로 확장되었고, 다양하고 많은 개인들의 의견을 분석하고자 하는 서비스들도 늘어나고 있다. 한편, 매체가 다양화되면서, TV광고계에서도 광고에 대한 의견의 확보와 분석에 새로운 접근방법이 필요해졌다. 이에 본 연구에서는 TV광고의 효과를 트위터 데이터를 기반으로 분석하며 특히 하둡을 이용하여 트위터 데이터와 같은 빅데이터를 저장 및 분석하도록 하는 LiveAD라는 시스템을 설계 및 구축하여, 트위터를 대상으로 TV광고 분석을 빠르게 수행할 수 있음을 보여주었다.

☞ 주제어 : 하둡, 광고효과분석, TV광고, 소셜 네트워크 서비스, 트위터

ABSTRACT

As 'Big data' has been one of challenging issues, development of new services using Social Network Service (SNS) which is its typical example became active. SNS has developed as a media where everyone communicates at real time and the number of SNS opinion analyzing services is increasing. Meanwhile, new approach to acquire and analyze twitter data becomes necessary in TV advertisement system. This paper proposes LiveAD system, which store and analyze big data such as twitter data as well as analyze TV advertising effect based on twitter data. As a proof of concept, the proposed system has been implemented collecting and analyzing twitter data using Hadoop. The result of collected information over the system increases the chance of analyzing TV advertising effect on twitter in real-time.

☞ keyword : Hadoop, Analysis on Advertising Effect, TV Advertisement, Social Network Service, Twitter

1. 서 론

SNS 시장이 커져가면서 이를 활용하기 위한 방안에 대한 논의 역시 활발해지고 있다. 대표적인 SNS 매체 중 하나인 트위터는 단문으로 된 의견을 공개하거나 공유하는 특성과 스마트폰의 보급으로 모바일 환경에서의 접근이 가능해지면서, 새로운 정보와 의견을 빠르게 전하는 하나의 매체로 자리잡았다. 이미 트윗트렌드와 같이 트위터를 기반으로한 분석 서비스들이 등장하며 트위터의 인기와 활용도를 알렸다. SNS의 대표적인 특징은 신속성, 개인성과 정보 개방성이다 [1]. 즉, 이러한 특성으로 인해

개개인의 의견이 공개되어 트윗은 제품 • 기업 선호도 및 트렌드 분석에 중요한 자료가 된다. 트윗 각각은 140자의 적은 양이다. 하지만 트윗은 실시간으로 전세계적으로 생성된다는 점에서 그 양이 많고 다루기 어려워 빅데이터라고 할 수 있다. 실제 7주년을 맞아 트위터에서 발표한 자료에 따르면 하루에 전송되는 트윗 수가 4억 건이 넘는다고 발표했다 [2]. 빅데이터는 기존의 방식으로서는 다루기 어렵고 처리하는데도 효율이 떨어져, 트윗을 처리하는데 기존 방법이 아닌 새로운 방식이 필요해졌다.

한편, 광고는 기업과 소비자를 이어주는 창이다. 그 중에서도 우리가 가장 쉽게 접하는 광고는 바로 TV에서 프로그램 사이사이에 나오는 CF영상들이다. TV는 드라마, 뉴스, 예능 등 다양한 콘텐츠를 제공하며, 보급률과 신뢰도가 높다. 이와 같은 매체 특성에 의해, TV광고는 많은 시청자들에게 노출될 수 있어 그 효과가 좋으나, 반대로 노출 범위가 넓기 때문에 매체의 특성상 그 반응을 즉각 측정하기 어렵다. 그럼에도 불구하고, 방영 후 반응을

¹ Dept. of Computer Science, Sookmyung Women's Univ., Seoul, 140-742, Korea

* Corresponding author (yulan@sm.ac.kr)

[Received 1 August 2013, Reviewed 4 August 2013, Accepted 31 November 2013]

☆ 2013년도 정부(미래창조과학부)재원으로 한국과학창의재단이 공개계우수연구(HCR) 지원과제로 수행하였음.

조사하는 일은 매우 중요하다. TV광고는 비용이 크며, 기대효과가 클 뿐만 아니라 광고계의 빠른 흐름을 파악하는 것은 보다 효과적인 광고 제작을 가능케 하기 때문이다. 위와 같은 이유로, 이전부터 광고 효과를 분석하는 다양한 기법들이 시도되었다 [3]. 특히 설문 기법이 주로 쓰이고 있으며 이를 위해 자료를 수집하고, 분석하는 절차를 자동화 아닌 수동으로 처리하고 있다.

본 논문에서는 트위터의 장점을 이용해 광고 효과 분석을 더욱 효율적으로 시도하기 위해 하둠을 이용하여 빠르게 트위터를 분석하는 시스템에 대해 설명한다. 위에서 언급했듯이, 트위터는 사람들의 의견에 대한 많은 데이터를 가지고 있기 때문에, 이 데이터를 광고 효과 분석에 이용할 수 있다. 하지만 트위터의 데이터양은 너무 많아 기존 방법으로는 처리하기 어렵고, 기존 TV광고 분석 기법은 시간과 비용이 상당히 소요된다는 단점이 있다. 이러한 문제와 빅데이터 분석에 걸리는 시간을 효과적으로 단축하기 위해, 본 논문에서는 하둠을 이용하였다. 이전에도 트위터 서비스는 존재해왔으나, 상용화된 기존 트위터 서비스는 트윗 검색, 트위터 내 트렌드 분석 등 트위터 내에 한정된 경우가 많았다. 그러나 본 연구는 트위터 자체가 아닌 트위터와 광고를 결합하여 분석하는데 초점이 맞추어져 있다. 본 연구의 목적은 광고와 트위터 데이터를 수집하며, 수집한 데이터를 토대로 광고의 효과를 트위터 상에의 언급도로 측정하는 것이다. 본 논문에서는 광고와 트위터의 데이터를 수집하고, 트위터에서의 시간별 광고의 언급도를 측정하는 시스템인 LiveAD에 대한 설계와 구축에 대해 기술한다. 본 논문의 구성은 다음과 같다. 2장에서는 관련 연구에 대해 설명하며, 3장에서는 시스템 설계에 대해, 4장에서는 시스템 구현에 대해 설명한다. 5장에서는 분석 결과를 기술하며, 6장에서는 결론을 맺는다.

2. 관련 연구

광고 효과의 측정이란 광고에 의해 광고목적이 어느 정도 달성되었는가를 판정하기 위해 광고내용이 전달된 후 수용자에게 미친 영향을 과학적인 방법으로 측정하는 것이다[4]. 방송 후, 방송광고에 대한 광고 효과를 측정하는 기존 방식은 정성적 접근과 정량적 접근, 두 가지 방식이 있다. 먼저 정성적 접근방식은 일일 후 회상조사(Day-After Recall, DAR), 테스트 마켓(Test Market), 단일 원천 추적조사(Single-source Tracking Test)가 있다 [3]. 이

중에서 일반적으로 쓰이는 방법은 일일 후 회상조사로 버크 테스트(Burke Test)라고도 불린다. 광고가 방영되고 하루가 지난 다음 집행된 프로그램을 시청한 사람들을 대상으로 전화면접을 실시한다. 응답자가 광고를 기억할 경우와 그렇지 않은 경우를 나누어 조사한다. 그러나 버크 테스트는 응답자의 불확실한 기억 또는 설문자의 도움에 의존하기 때문에, 응답 결과가 정확하지 않다는 단점이 존재한다[5]. 또한 조사기간이 오래 걸리고 비용 또한 크다는 문제가 존재한다.

TV 방송 광고의 효과를 측정하는 정량적인 수단은 지표를 이용하는 것이다. 여기에는 시청률에 기반한 CPRP와 GRPs라는 두 가지 지표가 있다. CPRP는 Cost Per Rating Point의 약자로, 1%의 시청률을 얻기 위해 투입된 광고비의 규모를 의미한다. CPRP는 광고비를 시청률로 나눈 값으로 광고의 효율성을 측정하는 지표이다. 한편, GRPs는 Gross Rating Points의 약자로 특정 기간의 시청률의 합이다 [6]. 그러나 두 가지 지표 모두 시청률에 기반하여 측정하므로, 현재의 다매체 다채널 환경에 맞지 않는다는 문제가 있다 [7].

한편 트위터를 하둠을 이용하여 분석하는 시스템에 대한 시도가 있었다. 그 연구에서 하둠을 이용하여 트위터 메시지를 분석하는 것을 제안하였으나 시스템의 설계에 그치고 활용효과를 언급하지 않았다 [8].

3. 설 계

본 연구에서 제안한 LiveAD는 그림 1과 같이 하둠을 이용하여 데이터를 수집하여 분산저장 및 분산처리 후 웹을 통해 보여주는 형식으로 구성된다. 전체 시스템은 하둠을 이용하여 광고데이터와 트위터를 매일 자동 수집하는 데이터 크롤러, 특정 광고를 언급하는 트위터를 분석해주는 광고 효과 분석기와 사용자에게 분석 결과를 보여주는 웹 인터페이스로 나눌 수 있다. 각 기능의 자세한 설명은 다음과 같다.

시스템은 크게 데이터 크롤러(Data Crawler), 하둠 클러스터(Hadoop Cluster) 그리고 웹 서버(Web Server)로 구성된다. 데이터 크롤러는 트위터와 TV광고 데이터를 수집하며 자세한 내용은 3.1절에서 설명한다. 하둠 클러스터는 데이터베이스와 트윗 분석기(Twit Analyzer)로 구성된다. 데이터베이스에는 수집된 트위터와 TV광고 데이터가 저장되며, 트윗 분석기는 광고 효과를 맵과 리듀스를 통해 분석한다. 자세한 분석 내용은 3.2절에서 소개한다.

워크플로우 응용을 위한 에너지 효율적인 가상 자원 할당 기법 (Energy-Efficient Provisioning of Virtual Machines for Workflow Applications)

이준영^o, 강혜정, 김윤희*

숙명여자대학교 컴퓨터과학부

{pinekiss07, hjkang, yulan}@sm.ac.kr

요 약

최근 클라우드 환경에서 가상화 기술을 통한 자원의 동적인 확장에 대한 연구가 활발히 이루어지고 있으며, 특히 이런 클라우드 컴퓨팅의 확장성을 통해 사용자의 요구 사항을 충족시키는 자원 할당 기법에 대한 연구가 다양하게 진행되고 있다. 더 나아가 전 세계적으로 지구 온난화에 대한 관심이 증대하면서, 높은 컴퓨팅 수준을 보장하면서 전력 소모량을 절감하는 그런 컴퓨팅이 새로운 트렌드로 떠오르고 있다. 본 논문에서는 저전력 기술 중 **Dynamic Voltage and Frequency Scaling (DVFS)** 기술을 적용하여 전력 소모량을 고려하는 저전력 가상 자원 할당 기법에 대하여 제안한다. 기존 연구의 경우, 사용자의 요구 사항을 충족하지 못하거나, 혹은 연구 대상이 되는 응용 유형이 **Bag of tasks**에 한정되어있었다. 하지만, 본 논문에서 제안하는 프로비저닝 기법은 클라우드 컴퓨팅 환경에서 상호간 의존성이 존재하는 서브 태스크들로 구성된 워크플로우 응용을 연구 대상으로 하며, 이는 사용자의 데드라인 요구 조건을 위반하지 않으면서 **Auto-scaling** 기술을 통해 사용자의 비용을 축소하고, 전력 소모량 최소화를 통해 플랫폼 관리자의 비용을 감소한다.

Keywords: Power-aware, Energy-efficient, Cloud, Workflow, DVFS

1. 서론

클라우드 컴퓨팅의 출현으로, 대규모 가상화 컴퓨팅과 데이터 센터는 IT 산업에서 중요한 요소로 자리 잡게 되었다. 이로 인해 좀 더 높은 수준의 컴퓨팅이 가능하게 되었지만, 데이터 센터의 증가로 전력 소모량이 급증하였다. 데이터 센터의 컴퓨팅 머신들이 대기상태에 있더라도 소규모 도시가 소비하는 전력량에 달하는 많은 양의 전력을 필요로 하고, 뿐만 아니라 서버들을 적정 온도로 유지하기 위해서는 마찬가지로 비슷한 양의 전력으로 쿨링 시스템을 가동해야 한다. 이러한 요인들은 CO2 발생량을 증가시키고 현재 이슈가 되고 있는 지구 온난화를 가속화하게 된다. 그래서 고성능 컴퓨팅의 새로운 트렌드인 그런 컴퓨팅으로 이러한 문제를 해결하고자 한다. 그런 컴퓨팅은 높은 성능 수준 달성과 전력 소모량 절감을 통하여 전체 시스템의 효율을 최대화한다. 최근 활발히 연구되고 있는 저전력 기술 중 가장 대표적인 것으로는 **Dynamic Voltage and Frequency Scaling (DVFS)** 기술이 있다[1][2]. DVFS 기술은 시스템이나 시스템에 연결된 장치들이 동작하는 동안에는 소비 전력을 줄일 수 없는 기술의 한계를 극복하기 위해 제안된 것으로 머신의 동작 속도를 줄이는 대신 소비 전력에서 이득을 취하는 방식의 기술이다. DVFS는 CPU에서 소비되는 에너지 E 가 주파수 f 에 비례한다는 점을 이용하여 CPU clock frequency를 줄여 소비 전력을 줄인다.

Auto-scaling 기법은 클라우드 컴퓨팅의 특성을 활용하여 수행 중인 작업 및 자원의 지속적인 모니터링을 통해 자원 규모를 확장/축소한다. 이 기법을 기반으로 한 자원 프로비저닝에 대한 여러 가지 연구가 활발하게 진행되고 있는데, 사용자의 데드라인 요구 조건을 만족하면서 사용자 비용의 최소화까지 고려하는 스케줄링 기법에 대한 연구도 존재한다[3][4]. 더 나아가, 전력을 고려하는 가상 자원 할당에 대한 연구도 존재한다[5][6][7]. 특히, 선행 연구[7]에서는 클라우드 컴퓨팅 환경에서 서브 태스크간 의존성이 존재하지 않는 **Bag of tasks** 어플리케이션을 대상으로 자원 할당 자동화 기법을 이용한 동적 부하 관리에 대한 연구를 진행하였다. 이를 통해 사용자의 요구 조건을 충족하면서 클라우드 플랫폼 관리자의 비용을 절감하였지만, 서브 태스크들 간의 의존성이 없는 **Bag of tasks** 어플리케이션을 대상으로 하였으므로 연구 적용 대상에 대

* 교신저자(Corresponding Author): 김윤희

이 논문은 2013 년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. NRF-2013R1A1A3007866)

한 일반화가 부족하다고 할 수 있다.

본 논문에서는 선행 연구를 확장하여 클라우드 컴퓨팅 환경에서 서버 태스크간 의존성이 존재하는 워크플로우 어플리케이션을 대상으로 하여 자원 할당 자동화 기법을 이용하여 작업의 데드라인을 보장하는 동적인 부하 관리 기법을 제안한다. 사용자의 데드라인 요구 조건을 만족함과 동시에 DVFS 기술을 적용한 저전력 할당 기법을 통하여 플랫폼 관리자의 비용도 절감한다. 그리고 워크플로우 어플리케이션을 대상으로 연구를 진행하여 기존의 선행 연구보다 다양한 형태의 어플리케이션에 대해 기법을 적용할 수 있도록 하였다.

본 논문의 2 장에서는 관련 연구에 대해 다루고, 3 장에서 저전력 가상 자원 동적 할당 기법을 설명한다. 마지막 장에서는 결론을 맺는다.

2. 관련연구

클라우드 컴퓨팅 환경에서 가상화 기술의 도입으로 자원의 무한한 제공이 가능해지면서 무한한 자원들을 동적으로 할당할 수 있는 기법에 대한 연구가 활발하게 진행되고 있다. AWS(Amazon Web Service)[8]의 Auto-Scaling 기술은 자원 할당 자동화 서비스의 대표적인 예로, 사용자가 정의한 규칙을 기반으로 자원의 규모를 결정한다. 자원 규모의 확장과 축소는 하드웨어의 성능과 관련된 수치의 상한 값과 하한 값을 미리 정해 놓고 그 수만큼 가상 머신(Virtual Machine: VM)을 추가하거나 제거하는 방식으로 이루어진다. 그러나 이 방식은 데드라인과 같은 사용자의 요구사항을 고려하지 않는다는 단점이 있다.

Auto-scaling 기술을 적용하여 데드라인과 사용자 비용을 고려한 자원 할당 기법인 [3]연구는 이러한 단점을 보완한 방식으로, 작업의 데드라인 및 실제 클라우드 컴퓨팅 환경의 요소들을 고려해 사용자 비용 절감을 가능하게 한다. 하지만 클라우드 플랫폼 관리자의 비용은 고려하지 않았다. 이를 보완한 형태의 [5]연구는 작업의 데드라인을 고려하면서 전력 소모량까지 동시에 고려한 자원 할당 기법을 제안하였다. 이 기법은 사용자의 비용을 낮추고 데드라인 요구사항을 지켜 사용자 만족도를 높이는 것은 물론, 전력 사용량 감축을 통한 관리자 측면에서의 비용 또한 줄여준다. 이를 위해 가상 머신이 물리 머신(Physical Machine: PM)에 프로비저닝 될 때 발생하는 전력 소모량을 고려하고 가상 머신의 가격이 같을 경우 전력 소모량이 적은 자원을 할당해준다. 이 기법은 태스크의 요구 성능에 따라 가상 머신을 제공하기 때문에 사용자에게 불필요한 비용부담을 줄 수 있다는 단점이 있다. [5]연구의 단점을 보완하는 [7]연구는 클라우드 컴퓨팅 환경의 실제 제약 요소들과 전력 소모량을 함께 고려하여 사용자와 플랫폼 관리자, 둘 모두의 비용을 최소화하는 자원 할당 기법을 제안하였다. 그러나 이 기법은 태스크들간 상호 의존성이 존재하지 않는 Bag of tasks 어플리케이션을 대상으로 하였기 때문에, 주로 워크플로우 형태의 과학 어플리케이션에 기법을 적용할 수 없다는 한계점이 있다. 그러므로 이 연구는 기법을 적용할 수 있는 어플리케이션이 한정되어 있으므로 적용 대상의 일반화가 부족하다고 할 수 있다. [9]연구는 [7]연구의 단점인 Bag of tasks 어플리케이션으로 제한을 두지 않고 서버 태스크들 사이에 상호 의존성이 존재하는 워크플로우 형태의 어플리케이션을 대상으로 스케줄링 기법을 제안한다. 특히, [9]연구에서는 본 논문에서 활용한 static PCH(Path Clustering Heuristic) 기법을 소개하고 이에 따른 자원 스케줄링 기법을 제안한다. 하지만, 클라우드 환경이 아닌 그리드 환경을 기반으로 하였다. 이를 보완한 [4]연구는 [9]연구와 달리 클라우드 환경에서 워크플로우 어플리케이션을 위한 자동화 자동 할당 기법을 제안하지만, 사용자의 비용만 고려할 뿐, 전력 소모량을 고려하지 않았다.

본 논문에서는 앞서 언급한 관련연구들의 단점인 적용 대상의 일반화 부족을 보완하여 사용자의 데드라인 요구사항을 만족하면서 사용자와 클라우드 플랫폼 관리자, 모두의 비용을 축소할 수 있는 워크플로우 형태의 어플리케이션을 대상으로 한 저전력 자원 할당 기법을 제안한다.

3. 저전력 동적 가상 자원 할당 기법

3.1 서비스 구조

A SLA driven VM Auto-Scaling Method in Hybrid Cloud Environment

Hyejeong Kang Jung-in Koh Yoonhee Kim
Dept. of Computer Science
Sookmyung Women's University
Seoul, Korea
{hjkang, jungin, yulan}@sookmyung.ac.kr

Jaegyoon Hahm
Korea Institute of Science and Technology Information
Daejeon, Korea
jaehahm@kisti.re.kr

Abstract—The advent of Science Clouds enables scientists to facilitate large-scale scientific computational experiments over cloud environment besides specialized supercomputers in diverse science domains. Cloud computing service elicits efficiency on on-demand resource usage and timely execution at any given time depending on experimental requirements. Hybrid clouds, composing of private and public clouds, even extend research opportunities on resource selection for further complicated experiments but increase the needs of dynamic resource management to maximize its utilization. At existing public cloud providers for commercial use, rule-based and schedule-based mechanisms have been tried for automatic resource allocation to provide resources for processing dynamic workload of modern applications. However, most of the auto-scaling methods just simply support performance metric such as CPU utilization but rarely are aware of Service Level Agreements (SLA) including execution deadline or cost. In this paper, we propose an auto-scaling method that automatically allocates resources depending on variable resource requirements in hybrid clouds satisfying a user's requirements on SLA. We present experimental results which show that the proposed auto-scaling can minimize SLA violations and acceptable cost if needed.

Keywords—*auto-scaling, hybrid cloud computing, SLA, multi-policies;*

I. INTRODUCTION

The advent of Science Clouds enables scientists to facilitate large-scale scientific computational experiments over cloud environment besides specialized supercomputers in diverse science domains. Cloud computing enables applications to exploit on-demand and scalable resources. Computational scientific applications (e.g. scientific workflow) and big data processes can be classified as many task computing (MTC) requiring high performance computing (HPC) or high throughput computing (HTC). It is essential for such applications to supply high performance resources in a long term and guarantee stable executions of applications. In these days, as the needs of high performance computing increase, it leads the advent of exa-scale computing or peta-scale computing. Hence, it is getting important to study computational problem solving environment which provides the management of job executions or resources in the large scale of computation. Auto-scaling with virtualization enables efficient utilization of cloud resources for the computational problem solving environment. Technically, auto-scaling can

dynamically increase or decrease the number of VMs (Virtual Machine) during execution of an application and a notable example is “Auto-scaling” of AWS (Amazon Web Service) [1]. Auto-scaling services like AWS or Scalr [2] are based on user-defined rules (e.g. when the utilization of CPU is more than 60%, add 2 VMs) when deciding scale of resources. Although the rule-based auto-scaling methods are simple, it is difficult for them to satisfy dynamic patterns of resource requirements caused by variable workloads of modern applications. Especially, if resource requirements of an application cannot be fulfilled, this would result in SLA (Service Level Agreement) [3] violation: deadline violation, growth of execution failure and so on.

In this paper, we propose an auto-scaling method to provide efficient resource utilization in a hybrid cloud computing environment. The proposed auto-scaling meets SLAs such as deadline, cost-oriented or performance-oriented policies. It is also available to auto-scale in economic way by considering both characteristics of applications (computation-intensive, data-intensive or I/O-intensive) and instance types of cloud resources for commercial use (public cloud). Based on Amazon EC2 instance types, in case of that two different instances (e.g. High-CPU type and Standard type) have same CPU resources, CPU-intensive type has lower price with the same execution time on computation-intensive jobs. The main assumptions that are implied in this paper are that (1) jobs of an application have no dependencies; these jobs are able to do parallel process for grand scale job from their Bag-of-Tasks type [4], and that (2) jobs, once started, should not be moved to a different virtual resource, so only queued jobs are adaptively scheduled during the auto-scaling phases.

The continuation of the paper is structured with the sections as follows; we introduce some related works in Section 2, and then present architecture of our auto-scaling framework. Section 4 explains auto-scaling algorithm, and Section 5 contains contents about a scenario of using auto-scaling. After that, experiment and simulation results are discussed in Section 6, while conclusion and future work are in final section.

II. RELATED WORK

Cloud computing offers unlimited resource by virtualization technology and facilitates extension and

reduction of resources. Auto-scaling issues are currently being discussed and studied as effective resource management.

"Auto-scaling" of AWS, as mentioned before, is one of existing representative auto-scaling methods. As a rule-based auto-scaling method which elastically expands or contracts resources to follow user-defined metrics, this service configures metrics related to hardware performance as criteria: CPU utilization, hard disk utilization, etc. Through setting upper bound and lower bound of a metric, VMs can be created or destroyed based upon the values. Similarly, there exist Paraleap [5] for Windows Azure [6], Scalr [2], and RightScale [7]. While rule-based auto-scaling methods are simple enough to allocate resources dynamically, it may not be possible to satisfy an amount of resources actually needed, in the case of applications which have a complex workload, then it could cause a severe problem called SLA violation (deadline, cost, and etc.).

References [8], [9] are the studies of auto-scaling considering deadline of applications or cost for resource usage. Reference [8] guarantees execution of an application within its deadline using their auto-scaling method. They respect cost as well in the auto-scaling method. However, they use only public cloud, which means that the scope of resource is limited to demonstrate performance of their auto-scaling method in a general way. In order to minimize expenses for resource usage, [9] proposes an auto-scaling method mapping optimal physical machines to resource requirements. It is combined with horizontal scaling which adds or removes VMs and vertical scaling which increases or decreases a VM size. In spite of the combination of auto-scaling methods, it is still inadequate for fulfilling resource requirements of dynamic workloads because it fixes the number of VMs created or deleted in the horizontal scaling.

In this paper, we propose an auto-scaling method which enables resource allocation considering SLA such as deadline and expenses in hybrid cloud environment. Using continuous monitoring service, it is possible to calculate the numbers of requisite VMs so that VMs can be properly added or removed whenever the need varies.

III. SERVICE ARCHITECTURE OF AUTO-SCALING FRAMEWORK

In this section, we discuss architecture of services. Fig. 1 related to the whole architecture which supports auto-scaling in hybrid cloud environment. In order to decide how many VMs will be created or destroyed on a dynamic basis, the framework needs a SLA description from a user. An application is submitted with the SLA and queued before they are processed while the SLA goes to 'Job Metadata' module.

Four major services form auto-scaling framework as illustrated in Fig. 1. Each service has sub-modules to perform specific work for auto-scaling.

'Metadata Mgmt. Service' (MMS) maintains information of three different categories and mainly stores static information. One of the categories is 'Job Metadata' module which contains length, execution history, definition (e.g. input, output), and SLA descriptions of applications. Another is 'Resource Metadata' module. This module reserves resource

specifications and the number of physical machines. The last module in MMS is 'VM Catalog'. 'VM catalog' manages specifications and templates of VMs in private cloud, with specifications and cost of instance types in public cloud.

'Auto-Scaling Service' (ASS) is core of our framework. Dynamic workload of modern applications leads service providers to have difficulties of predications on execution of jobs and their resource requirements. Because of that, all of processes of 'Auto-Scaling Service' are conducted on a regular basis, so that ASS allocates actually required resources at any given time. This service consists of three modules: 'Scheduling', 'SLA Monitoring' and 'Run-time Scaling'. 'Scheduling' module allocates jobs to VMs which fit resource requirements of jobs under the selected policy. We offer two kinds of policies. One is cost-oriented policy which is aware of resource utilization cost of public cloud environment. Through this policy, users can expense reasonable costs, while executing applications within a deadline. The other is performance-oriented policy which is aiming to finish application as fast as possible (i.e. this policy has a short deadline). If a user chooses the performance policy, a deadline which must be bigger than execution time of Critical Path (CP) and minimum performance requirement, the user requests, will be included in SLA. In the performance policy, jobs are allocated on VMs which have much more resources than minimum performance requirement. 'SLA Monitoring' module is responsible for estimation of performance and cost. Performance can be estimated by comparing the biggest EFT (Estimated Finish Time) of the last job on a VM with the deadline in SLA whether the EFT would exceed the deadline or not. 'Run-time Scaling' module decides the number of VMs to create or destroy for application execution and then gets the decisions across to 'Dynamic Resource Mgmt. Service' and 'Job Execution Service'. Scale-out which expands scale of resources through adding extra VMs is fulfilled when a resource allocation is executing, whereas scale-in is performed at going after the resource allocation.

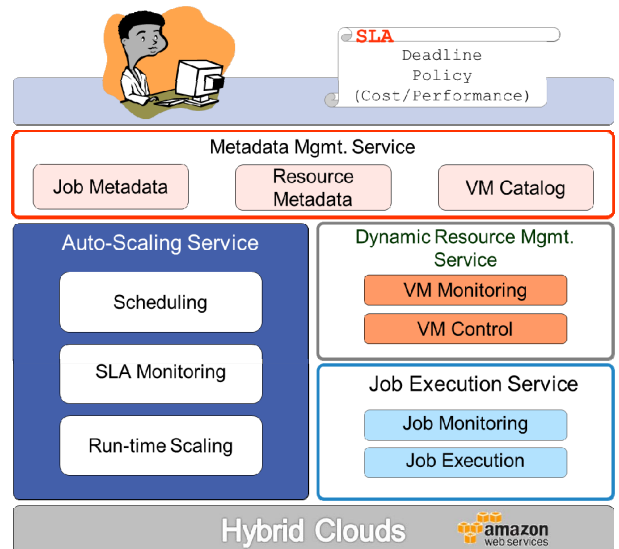


Fig. 1. Service Architecture

협업필터링의 희박성 문제 개선을 위한 통합 유사도 모델

유 석 종*

Integrated Preference Similarity Algorithm for Improving Sparsity Problem in Collaborative Filtering

Seok-Jong Yu*

본 논문은 2011년도 숙명여자대학교 교내연구비 지원에 의해 수행되었음.

요 약

협업필터링은 유사 사용자들 간의 집단지성에 기반한 사회적 추천 알고리즘으로 전자상거래 시스템에 도입되면서 상업적으로 활용되어 왔다. 반면, 평가행렬의 희박성, 확장성, 신규 추천 등의 한계가 있으며 이에 대한 개선 연구가 지속적으로 이루어지고 있다. 그 예로 복합 추천 알고리즘은 주로 사용자 프로파일 정보 또는 아이템 속성 등 내용정보와 협업필터링을 결합하여 단점을 보완하고자 하였다. 기존 복합 알고리즘 중 사용자 프로파일 정보, 아이템 속성정보, 아이템 평가행렬을 모두 통합하는 연구는 없었으며, 본 연구에서는 특히 사용자 간의 공통평가 아이템 수가 적어 추천이 어려운 희박성 문제를 개선하기 위하여 통합 유사도 알고리즘을 제안하고 기존의 방법과의 성능을 비교한다.

Abstract

Collaborative filtering is a social recommendation algorithm based on collective intelligence of similar users, and has been actively introduced to commercial electronic business systems. However, collaborative filtering has limitations such as sparsity of item rating matrix, scalability, and cold start, and it has been improved continuously to overcome these fundamental problems. As a result, hybrid recommendation algorithm was suggested to compromise the weaknesses of content-based recommendation and collaborative filtering. Existing hybrid algorithms have been focused on user-profile or item attribute with collaborative filtering, and there has been no attempt to integrate them together. Therefore, to improve the sparsity problem, this study suggests a new hybrid recommendation algorithm particularly for the case of sparse matrix by introducing the integrated similarity model, and also presents comparison results with conventional recommendation methods.

Keywords

collaborative filtering, content-based recommendation, hybrid recommendation

* 숙명여자대학교 컴퓨터과학부

· 제1저자(First Author): 유석종(sjyu@sm.ac.kr)

· 접수일: 2013년 04월 19일, 수정일: 1차- 2013년 05월 06일, 2차- 2013년 06월 12일, 게재확정일: 2013년 07월 15일

I. 서 론

인터넷 상에서 거래되고 있는 영화, 책, 앱 등의 상품은 일반적으로 그 종류와 수가 대규모이며 신상품으로 인해 지속적으로 그 수가 증가하는 경향이 있다. 따라서 인터넷 쇼핑몰 시스템에서는 효과적인 검색기능은 매우 중요하며, 상품을 고객에게 효과적으로 노출시키기 위하여 상품 추천 기능을 채택하고 있다. 추천 시스템은 사용하는 추천 알고리즘에 의하여 성능이 좌우되며, 내용기반 추천 방법(Content-based)과 협업필터링(Collaborative Filtering), 그리고 복합(Hybrid) 추천 알고리즘으로 구분된다. 내용기반 추천은 사용자 프로필의 내용과 상품의 속성정보를 비교하여 관심이 있을 만한 상품들을 추천해 주는 방법이다. 이 방법은 속성정보가 충분하지 않을 경우 추천 성능이 떨어지고, 비교 대상 아이템 수가 늘어날수록 연산 시간이 늘어나는 문제점이 있다. 협업필터링은 아이템에 대한 고객의 선호도를 나타내는 평가 행렬(Rating Matrix)에 기반한 추천 방법이다[1]-[2]. 협업필터링은 아이템에 대한 사용자의 평가행렬을 기반으로 미경험 아이템에 대한 선호도를 예측하는 추천 방법이다. 평가한 아이템의 수가 적어 추천이 어려울 경우 사용자 프로필과 아이템 속성 정보를 이용하는 내용기반 방법과 결합하여 적용하는 복합 추천 방법이 제안되었다. 이 방법에는 협업필터링 추천 결과와 내용기반 방법에 의한 추천 결과를 병합하는 방법이 있으며, 사용자 클러스터 또는 아이템 클러스터 상에서 협업필터링을 적용하는 방법이 있다. 기존 협업필터링은 평가행렬을 통해 사용자간 선호도 유사도(Preference Similarity)를 측정하는 방법을 사용한다. 본 연구에서는 평가행렬뿐만 아니라 사용자 프로필, 아이템의 속성정보별 유사도를 결합하는 통합 유사도 개념을 제안하고 특히 두 사용자간 공통 평가 아이템의 수가 희박한 경우 추천 성능에 미치는 영향을 비교하고자 한다.

II. 관련연구

2.1 내용기반 추천

내용기반 추천 알고리즘은 고객의 프로필정보, 아이템의 분류와 같은 속성 정보를 분석하여 관심도가 일치하는 아이템을 추천하는 방법이다. 고객들을 나이, 성별, 직업 등으로 분류하여 동일 그룹의 사용자들이 많이 구매한 아이템을 추천해주는 방법과, 기존 구매 아이템이나 조회 횟수가 높은 아이템과 유사한 아이템을 추천해주는 방법을 사용한다[2]. 이 방법은 사용자가 이미 알고 있고 있거나 이와 유사한 아이템들이 주로 추천되는 포트폴리오 현상이 발생하는 단점이 있다[3].

2.2 협업필터링

협업필터링[3]-[5]은 아이템 평가기록을 바탕으로 선호도가 유사한 사용자 그룹(이웃)을 탐색하여 미경험 아이템에 대한 목표 사용자의 선호도를 예측하여 이중 상위 아이템들을 추천해 주는 방법이다. 협업필터링의 추천의 질은 사용자-아이템 평가행렬에 좌우되며 평가한 아이템의 수가 많을수록 예측 정확도도 향상된다. 내용기반 추천방법과는 달리 선호도가 유사한 이웃에 기반하고 있기 때문에 다양한 장르에 걸친 아이템의 추천이 가능하다. 그러나 구조적으로 초기평가, 확장성, 희박성의 문제점을 포함한다[6]. 첫째로, 새로운 사용자나 평가된 적이 없는 아이템에 대해서 추천하기 어려운 특성이 있으며, 둘째로 희박성 문제는 전체 상품 수에 비해 평가된 상품의 수가 극히 적을 경우 유사 사용자 탐색이 어려우며 추천의 질도 떨어진다[2][4][7]. 확장성은 사용자와 아이템의 수가 증가할수록 유사 이웃 탐색에 필요한 연산 시간이 비례하여 증가하여 실시간 추천을 어렵게 한다.

2.3 kNN 탐색과 클러스터링

k-Nearest Neighbor(kNN) 탐색은 목표 사용자와 전체 사용자간의 선호 유사도를 계산한 후 유사도 상위 k명을 이웃으로 선정하는 방법이다[7]. k가 너무 크면 추천 정확도나 성능을 떨어뜨릴 수 있으며, k가 너무 작으면 유사도가 높지 않은 사용자들의 아이템에 대해서는 예측할 수 없는 문제점이 발생



Skyline queries on keyword-matched data

Hyunsik Choi^a, HaRim Jung^a, Ki Yong Lee^b, Yon Dohn Chung^{a,*}

^a Department of Computer Science and Engineering, College of Information and Communication, Korea University, Seoul 136-713, Republic of Korea

^b Department of Computer Science, Sookmyung Women's University, Seoul, Republic of Korea

ARTICLE INFO

Article history:

Received 28 July 2009

Received in revised form 9 January 2012

Accepted 30 January 2012

Available online 10 February 2012

Keywords:

Information technology and system

Database management

Query processing

Spatial database

Textual database

ABSTRACT

Given a set of d -dimensional tuples with textual descriptions, a keyword-matched skyline query retrieves a skyline computed from tuples whose textual descriptions contain all query words. For example, suppose a customer prefers cars with low mileage and low price, and finds a car equipped with 'air bag' and 'sunroof' in an online shop. In such a case, a keyword-matched skyline query is highly recommended. Although there are many applications for this type of query, to date there have not been any studies on the keyword-matched skyline queries. In this paper, we define a keyword-matched skyline query and propose an efficient and progressive algorithm, named Keyword-Matched Skyline search (KMS). KMS utilizes the IR^2 -tree as an index structure. To retrieve a keyword-matched skyline, it performs nearest neighbor search in a branch and bound manner. While traversing the IR^2 -tree, KMS effectively prunes unqualified nodes by means of both spatial and textual information of nodes. To demonstrate the efficiency of KMS, we conducted extensive experiments in various settings. The experimental results show that KMS is very efficient in terms of computational cost and I/O cost.

© 2012 Elsevier Inc. All rights reserved.

1. Introduction

Recently, skyline queries have emerged as an important query for various applications involving multi-criteria decision making. In terms of retrieving the best tuples, skyline queries are very useful. Given a set of d -dimensional tuples, a skyline contains a set of tuples which are not dominated by other tuples. A tuple tp is said to dominate another tuple tp' when tp is not worse than tp' on all dimensions and tp is better than tp' on at least one dimension.

Skyline queries, unlike general exemplary queries involving one criteria (e.g., top- k and nearest neighbors), can retrieve numerous results according to data distribution and cardinality. However, many results are meaningless to users, that is the user has to search for desired information again. Therefore, skyline queries should consider various user preferences in order to enable users to find interesting information effectively. In this paper, we use keywords as new preference of skyline queries.

Initially, we assume that each d -dimensional tuple is additionally described by a textual attribute. There have been many internet applications in which an enormous volume of data is represented by textual descriptions.

In this paper, we integrate both skyline queries and keyword searches in order to provide an improved means to search for desired information. We introduce a keyword match as a preference predicate, and we propose skyline queries on

* Corresponding author.

E-mail addresses: hyunsikchoi@korea.ac.kr (H. Choi), harim3826@korea.ac.kr (H. Jung), kiyong.lee@gmail.com (K.Y. Lee), ydchung@korea.ac.kr (Y.D. Chung).

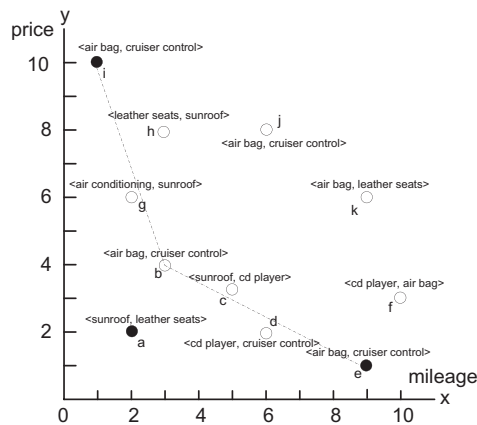


Fig. 1. An example of keyword skyline.

keyword-matched data. In terms of representing human's cogitation well, the word is one of the most useful means to describe users' preferences. Suppose that a dataset consists of d -dimensional tuples with textual information, and a query tries to retrieve skyline tuples whose textual descriptions contain specific query keywords. We call such queries *keyword-matched skyline queries*.

For simplicity, we assume a textual attribute is a set of keywords such as tags in *Web 2.0* [23]. However, it can be replaced with a narrative description without any loss of generality.

Example 1. A customer wants to find a used car on a shopping web site. He/she prefers cars with both lower mileage and lower price and desires a car equipped with air bags for safety.

Example 2. Through an online hotel search service provided by a travel agency, a traveler is looking for a hotel which is cheap and close to the beach. In addition, he/she needs wireless internet and a baby sitter service.

In such cases, a keyword-matched skyline query is highly recommended. Example 1 is shown in Fig. 1, where 2-dimensional points denote used cars in an online shop. A used car (a point) has two attributes: mileage (x -axis) and price (y -axis). The keywords enclosed by angle brackets (i.e., $\langle \rangle$) denote features (e.g., air bag, cruiser control, sunroof, etc.) with which cars are equipped. As we mentioned above, a car tp dominates tp' if tp is not larger¹ than another car tp' in all axes and tp is smaller than tp' in at least one axis.

In this example, a keyword-matched skyline query returns a skyline computed from tuples whose textual descriptions contain the keyword 'air bag'. That is, keyword-matched skyline tuples for 'air bag' are $\{b, e, i\}$ denoted as a dashed line in Fig. 1. Note that the result of the keyword-matched skyline is different to that of the original skyline (i.e., the original skyline contains the points $\{a, e, i\}$ denoted as filled circles, and the points $\{e, i\}$ belong to both the keyword-matched skyline and the original skyline). Although there are many applications for keyword-matched skyline queries, to date there have been no efficient algorithms known.

Intuitively, this problem can be solved by the combination of a keyword search algorithm and a skyline algorithm. The combination approach first retrieves the keyword-matched tuples from the dataset, and then computes skyline tuples among the keyword-matched tuples. This approach results in correct answers by definition 4 as will be discussed in Section 3.1. With respect to a skyline computation, several index-based algorithms (e.g., *nearest neighbor search* and *branch and bound skyline search*) [15,24] show the best performance among existing skyline algorithms. However, they are not applicable to the combination approach since index structures cannot be instantly constructed from the keyword-matched tuples. Therefore, we can consider non-index-based skyline algorithms [2] (i.e., *block-nested-loop* and *divide-and-conquer*) as a part of skyline computation in keyword-matched skyline, but we exclude a divide-and-conquer approach due to its scalability problem [3].

As a straightforward approach, we introduce an *INverted-index-based Keyword Skyline algorithm* (INKS) which combines an inverted index [9,33,25] and a block-nested-loop algorithm. Although INKS shows reasonable performance, there are two disadvantages. First, as we mentioned above, the skyline computation of INKS cannot be helped by any index structure because index structures cannot be instantly constructed from keyword-matched tuples retrieved by using an in-

¹ Without loss of generality, we assume that lower values are better.



Efficient processing of multiple continuous skyline queries over a data stream

Yu Won Lee^a, Ki Yong Lee^b, Myoung Ho Kim^{a,*}

^a KAIST, Daejeon, Republic of Korea

^b Sookmyung Women's University, Seoul, Republic of Korea

ARTICLE INFO

Article history:

Received 8 November 2011

Received in revised form 15 September 2012

Accepted 25 September 2012

Available online 4 October 2012

Keywords:

Multiple skyline query

Data stream

Continuous query

ABSTRACT

Given a set of data objects, the skyline query returns the objects that are not dominated by others. Although skyline computation has been studied extensively for static data, there has been relatively less work on data streams. Recently, a few methods have been proposed to process a single continuous skyline query over a data stream. However, efficient techniques that can handle multiple skyline queries have not been much considered. In this paper, we propose a new method, called *FAST*, for processing multiple continuous skyline queries over a data stream. *FAST* uses a filtering technique that can early discard an object that will not be a member of any future skyline of continuous queries, and uses a discriminant that can efficiently determine which objects in memory are skyline objects for which queries. We present that the proposed method *FAST* can compute skylines of multiple continuous queries very efficiently. Through extensive experiments, we show the high performance and great scalability of the proposed method.

© 2012 Elsevier Inc. All rights reserved.

1. Introduction

Skyline queries are widely used in applications that support multi-criteria decision making and analysis. Given a set of d -dimensional data objects, the skyline query returns the objects that are not dominated by any other objects. An object o is said to *dominate* another object o' if o is not worse than o' on all dimensions and o is strictly better than o' on at least one dimension. Fig. 1a shows an example of skyline, where each 2-dimensional point represents a stock record with two attributes: *risk* (x -axis) and *commission cost* (y -axis). We assume that o is better than o' on a certain dimension if o is smaller than o' on that dimension. In Fig. 1a, the skyline is $\{i, h, b, d\}$ because they are not dominated by any other stock. They are the stocks of our interest in general because a stock that is not in the skyline is dominated by at least one of the stocks in the skyline. For example, stock e has a higher risk and a higher cost than d . On the other hand, b and d are incomparable since they do not dominate each other in all dimensions.

Due to the recent growth of applications supporting multi-preference analysis and decision making, skyline computation has received considerable attention [18,5,6,9,11,33,16,27,21]. However, most of these works assume static data, i.e., all the input data are known in advance before processing. On the other hand, there has been relatively less work on the continuous skyline query over a data stream, where data arrive at the system continuously over time. A query over a data stream mostly implies a continuous query. Thus, in the rest of the paper, we will often omit the word “continuous”, e.g., queries instead of continuous queries, if there is no ambiguity.

* Corresponding author.

E-mail addresses: ywlee@dbserver.kaist.ac.kr (Y.W. Lee), kiyonglee@sookmyung.ac.kr (K.Y. Lee), mhkim@dbserver.kaist.ac.kr (M.H. Kim).

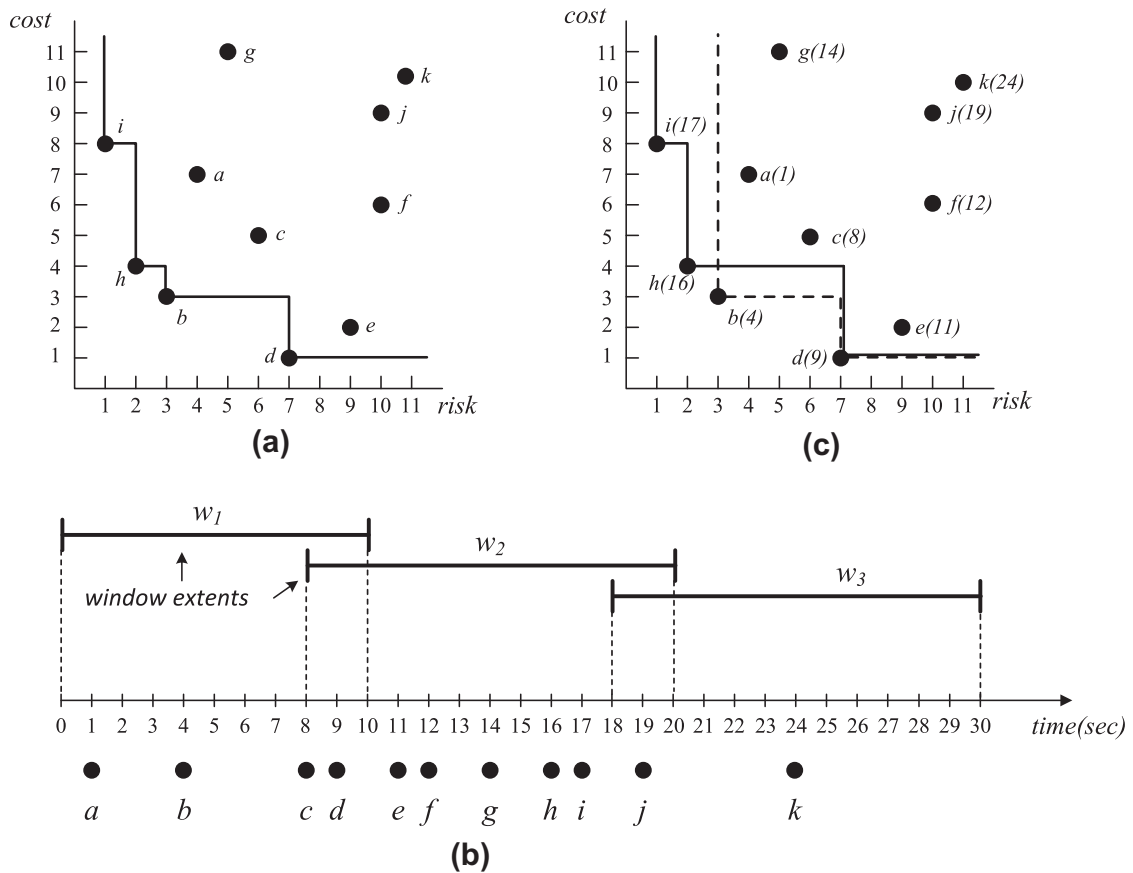


Fig. 1. Skyline examples. (a) A skyline on a static data set. (b) A data stream and window extents by the sliding window with RANGE = 12 and SLIDE = 10. (c) Two skylines on two window extents w_1 and w_2 .

Since a data stream is often unbounded, a query over a data stream is generally specified with a *sliding window*. Note that a sliding window can be time-based or tuple-based. We in this paper assume the time-based sliding window. A time-based sliding window, shortly a sliding window in this paper, decomposes a data stream into substreams of equal-sized time intervals, which are called *window extents* [23]. A sliding window is defined by two parameters, namely, *RANGE* and *SLIDE*. *RANGE* specifies the length of the window extent and *SLIDE* specifies the step by which the window extent moves. The query is then evaluated for each window extent. For example, consider a skyline query defined over a data stream of stock records. Suppose that the skyline query over a data stream is specified by a sliding window with *RANGE* = 12 s and *SLIDE* = 10 s as in Fig. 1b, where the data stream is decomposed into window extents $w_1, w_2, \dots$, by the sliding window. Fig. 1c shows two skylines for window extents w_1 and w_2 . In the figure, the number beside each point represents the arrival time of a stock record. At time 10 s, the query is evaluated for the window extent w_1 , whose result is $\{b, d\}$ (the dashed line in Fig. 1c). At time 20 s, the query is evaluated for the window extent w_2 , whose result is $\{i, h, d\}$ (the solid line in Fig. 1c).

Recently, a few methods have been proposed to process a single continuous skyline query over a data stream [25,34,26]. These methods maintain two buffers in memory, namely, B_{sky} and B_{cand} . B_{sky} stores objects that are currently in the skyline, and B_{cand} stores objects that are not currently in the skyline but may appear in the skyline in the future after some object in B_{sky} expires. A change in the skyline can occur only when a new object arrives or some object in B_{sky} expires. Let a new object o_{new} arrive. If o_{new} is dominated by some object in B_{sky} , o_{new} is added to B_{cand} . Otherwise, o_{new} is added to B_{sky} and the objects in either B_{sky} or B_{cand} that are dominated by o_{new} are removed. When an object in B_{sky} expires, new skyline objects are found, if exist, from B_{cand} and then added to B_{sky} .

Although these methods can process a single continuous skyline query efficiently, to the best of our knowledge, there has been no work that addresses the processing of multiple continuous skyline queries. In real-world applications, many queries can run concurrently, where each query has different *RANGE* and *SLIDE*. For the example of stock records, stock traders who run algorithmic trading systems may request a skyline query that updates the result every 5 min over the past 8 min (i.e., *RANGE* = 8 min and *SLIDE* = 5 min) to the trading server, while another user may run a skyline query that updates the result every 10 min over the past 15 min (i.e., *RANGE* = 15 min and *SLIDE* = 10 min) depending on the specific financial models they

전자상거래 데이터의 실시간 분석을 위한 데이터 스트림과 다수 릴레이션 간의 효율적인 연속 조인 처리 기법

Efficient Processing of Continuous Join Queries between a Data Stream and Multiple Relations for Real-Time Analysis of E-Commerce Data

김해리(Haeri Kim)*, 이기용(Ki Yong Lee)**

초 록

최근 들어 전자상거래 데이터의 실시간 공급이 가능해지면서, 전자상거래 데이터를 실시간으로 분석하고자 하는 요구가 급증하고 있다. 이를 위해서는 전자상거래 데이터 스트림과 디스크에 저장된 대규모 릴레이션 간의 연속 조인 질의를 효율적으로 처리하는 것이 매우 중요하다. 본 논문에서는 전자상거래 데이터 스트림과 디스크에 저장된 다수 릴레이션 간의 효율적인 연속 조인 질의 기법을 제안한다. 제안 방법은 기존 방법에 비해 서비스율을 크게 향상시키는 한편, 메모리 사용량을 크게 줄인다. 분석과 다양한 실험을 통해, 제안 방법은 기존 방법에 비해 서비스율과 메모리 사용량에서 더 효율적임을 보인다.

ABSTRACT

Recently, as real-time availability of e-commerce data becomes possible, the requirement of real-time analysis of e-commerce increases significantly. In the real-time analysis of e-commerce data, it is very important to efficiently process continuous join queries between an e-commerce data stream and disk-based large relations. In this paper, we propose an efficient method for processing a continuous join query between an e-commerce data stream and multiple disk-based relations. The proposed method improves the service rate significantly, while reducing the amount of required memory substantially. Through analysis and various experiments, we show the efficiency of the proposed method compared with the previous one in terms of service rate and memory usage.

키워드 : 연속 조인 질의, 데이터 스트림, 실시간 분석
Continuous Join Query, Data Streams, Real-Time Analysis

* M.S. Candidate, Division of Computer Science, Sookmyung Women's University

** Corresponding Author, Assistant Professor, Division of Computer Science, Sookmyung Women's University
(E-mail : kiyonglee@sookmyung.ac)

2013년 07월 17일 접수, 2013년 08월 05일 심사완료 후 2013년 08월 09일 게재확정.

1. 서 론

최근 들어 전자상거래 데이터의 실시간 공급이 가능해지면서, 전자상거래 데이터를 실시간으로 분석하고자 하는 요구가 급증하고 있다. 전자상거래 데이터를 실시간으로 분석하기 위해서는 실시간으로 끊임없이 유입되는 전자상거래 데이터 스트림(data stream)과 디스크에 저장된 대규모 릴레이션(relation) 간의 연속 조인 질의(continuous join query)를 효율적으로 처리하는 것이 매우 중요하다[1, 2, 3].

실시간으로 유입되는 전자상거래 데이터 스트림을 S 라 하고, 디스크에 저장된 대규모 릴레이션을 R 이라 할 때, S 와 R 간의 연속 조인 질의(continuous join query)는 $S \bowtie R$ 로 표현된다. 그 사용 예로서, 어떤 시각(timestamp)에 어떤 제품(productID)이 판매되었을 때 그를 나타내는 튜플(tuple) (timestamp, productID)가 데이터 스트림 S 에 실시간으로 유입된다고 하고, 디스크에 저장된 릴레이션 R 의 각 튜플(productID, price)이 각 제품(productID)의 가격(price)을 나타낸다고 할 때, 연속 조인 질의 $S \bowtie R$ 의 각 결과 튜플(timestamp, productID, price)는 실시간으로 판매되는 각 제품의 가격을 나타낸다.

이러한 연속 조인 질의 $S \bowtie R$ 을 처리하기 위해, 전통적인 중첩 루프 조인(nested loop join)[4]이나 색인 기반 조인(index-based join)[4]을 사용하면, S 에 도착하는 매 튜플마다 그와 조인되는 R 의 튜플을 찾기 위해 디스크를 여러 번 접근해야 하므로 성능이 매우 저하된다. 이를 극복하기 위해 최근에 메시 조인(mesh join)이라는 기법이 제안되었다[5, 6]. 메시 조인은 S 에 도착하는 w 개의 튜플

당 디스크를 한 번만 접근함으로써 $S \bowtie R$ 의 처리 속도를 크게 향상시켰다.

하지만 최근 들어 전자상거래 데이터 스트림 S 와 디스크에 저장된 다수의 릴레이션 $R_1, R_2, \dots, R_N$ 간의 연속 다중(multi-way) 조인 $S \bowtie R_1 \bowtie \dots \bowtie R_N$ 에 대한 처리가 점차 중요해지고 있다[2]. 예를 들어, 어떤 시각(timestamp)에 어떤 제품(productID)가 어떤 매장(storeID)에서 판매되었을 때 그를 나타내는 튜플(timestamp, productID, storeID)이 데이터 스트림 S 에 실시간으로 유입된다고 하자. 디스크에 저장된 릴레이션 R_1 의 각 튜플(productID, price)은 각 제품(productID)의 가격(price)을 나타내고, 디스크에 저장된 다른 릴레이션 R_2 의 각 튜플(storeID, location)은 각 매장(storeID)의 위치(location)를 나타낸다고 할 때, 연속 조인 질의 $S \bowtie R_1 \bowtie R_2$ 의 각 결과 튜플(timestamp, productID, price, storeID, location)은 실시간으로 판매되는 각 제품의 가격과 판매 위치를 나타낸다.

그러나 연속 다중 조인 $S \bowtie R_1 \bowtie \dots \bowtie R_N$ 을 처리하기 위해 기존의 메시 조인을 단순히 적용하면 문제점이 발생한다. 제 2.1절에서 다시 자세히 설명하겠지만, 메시 조인은 각 릴레이션 $R_1, R_2, \dots, R_N$ 을 고정된 크기의 블록(block)으로 나누고, 각 릴레이션에서 한 블록씩을 메모리에 적재한다. 각 릴레이션 $R_i (i = 1, 2, \dots, N)$ 의 블록 수를 $B(R_i)$ 라 하자. 메시 조인은 S 에 w 개의 튜플이 새로 도착할 때마다, 메모리에 적재된 $R_1, R_2, \dots, R_N$ 의 블록들 중 하나를 그가 속한 릴레이션의 그 다음 블록과 교체한다. 그리고 S 에 가장 최근에 도착한 $w \cdot B(R_1) \cdot B(R_2) \cdot \dots \cdot$

웹소켓을 이용한 실시간 커뮤니케이션 서비스 시스템 구현 *

김나영, 신다영, 최혜진, 심준호, 이기용

숙명여자대학교 컴퓨터과학부

e-mail : na07love@naver.com, shindanim@nate.com, chjlove6@naver.com, jshim@sookmyung.ac.kr,
kiyonglee@sookmyung.ac.kr

Implementation of a Real-Time Communication Service System Using WebSocket

Nayoung Kim, Dayoung Shin, Hyejin Choe, Junho Shim, Ki Yong Lee
Division of Computer Science, Sookmyung Women's University

요 약

본 논문은 HTML5 의 새로운 표준인 웹소켓(WebSocket)을 이용하여 구현한 실시간 커뮤니케이션 시스템을 소개한다. 본 논문에서 구현한 시스템은 PC 의 브라우저와 모바일 디바이스 모두를 통해 접근이 가능하다. 본 논문에서 구현한 실시간 커뮤니케이션 서비스 시스템은 이벤트 기반의 비동기 방식을 채택함으로써, 클라이언트가 리로드(reload)하지 않더라도 필요한 정보를 빠르게 제공한다. PC 의 브라우저가 접근하는 부분은 Node.js 를 사용하여 구현하였으며, 모바일 디바이스가 접근하는 부분은 PhoneGap 플랫폼을 사용하여 구현하였다. 본 논문에서는 시스템의 전체 구조와 관련 기술을 설명하고, 구체적인 구현 코드의 일부를 제시하며, 추후 확장 가능성과 비전을 제시한다.

1. 서론

커뮤니케이션 시스템이란 사용자 간 의사소통을 위해 사용되는 시스템을 말한다. 이러한 시스템을 통해 사용자들은 언제든지 다른 사람과의 의견을 주고받을 수 있다. 페이스북이나 트위터와 같은 소셜 네트워크 서비스도 커뮤니케이션 시스템의 일종으로 볼 수 있다. 본 논문에서는 여러 커뮤니케이션 시스템 중 특정 대학의 학과를 위해 구축된 커뮤니케이션 시스템을 소개한다. 이 시스템에서 요구되는 기능으로 실시간 양방향 커뮤니케이션 제공은 필수적이다. 또한 다수의 동시 접속자들에게 원활한 서비스를 제공해야 한다.

본 논문에서는 이를 위해 웹소켓(WebSocket)을 이용하여 구현한 실시간 커뮤니케이션 시스템을 소개한다. 웹소켓은 HTML5 의 새로운 표준으로서, 웹을 통한 양방향 통신을 가능하도록 하는 기술이다 [1]. 웹소켓은 웹에서 소켓을 구현함으로써 진정한 서버 푸시가 가능하다. 본 논문에서 구현한 시스템은 PC 의 브라우저와 모바일 디바이스 모두를 통해 접근이 가능하다. PC 의 브라우저가 접근하는 부분은 서버 개발 플랫폼인 Node.js 를 사용하여 구현하였다.

Node.js 는 웹 소켓을 모듈로 제공하기 때문에 사용이 쉬우며, 이벤트 기반 I/O 방식을 사용하기 때문에 서비스를 신속하게 제공할 수 있다는 장점이 있다. 한편 모바일 디바이스가 접근하는 부분은 하이브리드 애플리케이션 개발 플랫폼인 PhoneGap[2]을 이용하여 구현하였다. PhoneGap 의 장점은 JavaScript 와 같은 스크립트 언어를 사용하여 카메라와 같은 모바일 디바이스를 제어할 수 있다는 것이다.

본 논문에서 구현한 커뮤니케이션 시스템은 각 클라이언트가 리로드(reload)를 하지 않더라도 사용자 간의 의사 전달이 가능하고, PC 와 모바일 디바이스 등 다양한 디바이스에서 접근할 수 있다는 점에서 기존 시스템과 차별화된다.

본 논문의 구성은 다음과 같다. 2 장에서는 관련 기술을 소개하며, 3 장에서는 시스템 구조와 구현 코드에 대해 설명한다. 마지막으로 4 장에서는 관련 기술과 관련된 비전 및 추후 연구 방향을 제시한다

2. 관련 기술

2.1 WebSocket

*이 논문은 2013 년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2012R1A1A1001269)

웹소켓은 클라이언트와 서버 간에 전이중(full-duplex) 양방향의 단일 소켓 연결을 제공한다. 웹소켓을 이용하면 단일 HTTP 요청만으로 클라이언트와 서버가 연결될 수 있으며, 클라이언트에서 서버로의 요청/응답과 서버에서 클라이언트로의 요청/응답이 같은 연결을 재사용한다. 웹소켓을 이용하면 지연이 줄어든다. 핸드셰이킹이 한 번 이루어지면 하나의 연결로 메시지를 주고받을 수 있으며, 주기적으로 요청을 보내지 않아도 되기 때문이다 [1].

웹소켓 기능을 사용하기 위해, 본 논문에서는 다음 절에서 설명하는 Node.js 에서 제공하는 웹소켓 관련 API 를 이용한다.

2.2 Node.js

Node.js 는 2009 년 Ryan Dahl 에 의해 개발된 소프트웨어 플랫폼이다. JavaScript 기반의 언어를 사용하여 서버를 개발할 수 있도록 해준다. 그 동안 JavaScript 는 client-side 개발을 위한 언어로, 주로 웹 브라우저 콘텐츠를 만들기 위해 활용되었다. 하지만 JavaScript 는 실행속도에 있어 C 나 C++와 같은 언어에 비해 느리다는 단점을 가지고 있다. 그렇지만 JavaScript 의 간단하고 활용도가 높다는 장점 때문에 다른 분야에 적용하려는 노력 또한 계속되었다. 2008 년 구글의 Chrome 웹 브라우저에 V8 엔진을 탑재하면서 JavaScript 의 느린 단점을 극복할 수 있는 계기가 마련되었다. 이를 이용해 Ryan Dahl 은 V8 JavaScript 엔진을 사용해 동작하는 Node.js 를 개발하였다.

Node.js 의 가장 큰 특징은 이벤트 기반의 비동기 I/O 를 제공한다는 점이다. 이 방식은 단일 스레드(thread)로 동작한다. 다중 스레드를 사용하는 기존의 서버 엔진과는 다르게 단일 스레드를 사용함으로써, 스레드 개발에 필요한 자원을 아낄 수 있다. 또한 기존의 서버 엔진은 메모리를 많이 차지해 사용자의 요청이 몰릴 시 엔진의 성능을 저하시키는 단점이 있는 반면, 이벤트 기반 I/O 방식은 요청이 증가하더라도 서버 성능의 저하를 줄일 수 있다는 장점이 있다.

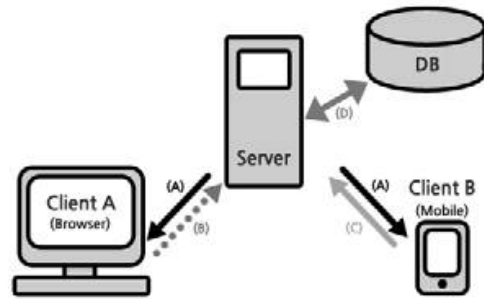
또한 Node.js 는 기본 모듈이나 객체에 더해 외부에서 제작된 모듈을 포함시키는 것이 가능하다는 장점을 가지고 있다. 이는 개발에 필요한 자원을 크게 줄일 수 있으며, 웹소켓 등 다양한 외부 모듈을 사용하여 시스템을 간편하게 구현할 수 있도록 한다.

3. 실시간 커뮤니케이션 시스템의 설계 및 구현

3.1. 시스템 구조

클라이언트가 처음 시스템에 접속하면 (그림 1)의 (A)와 같이 서버에서 해당 클라이언트에게

데이터를 전송한다. 만약 PC 의 브라우저를 통해 접근한 클라이언트인 Client A 가 (B)와 같이 서버에 요청을 보내면, 서버는 (D)와 같이 DB 에 접근하여 정보를 처리하고, 처리 결과를 (A)로 다시 전송한다. 모바일 디바이스를 통해 접근한 클라이언트인 Client B 도 마찬가지로 (C)와 같이 서버에 요청을 보내면, 서버는 (D)와 같이 DB 에 접근하여 정보를 처리하고, 처리 결과를 (A)로 다시 전송한다.



(그림 1) 시스템 구조

3.2. Node.js 를 통한 웹 소켓의 구현

Node.js 에서는 외부 모듈인 socket.io 를 통해 웹소켓 기능을 사용할 수 있다. 외부 모듈을 사용하기 위해서는 해당 모듈을 사용 전에 설치하는 과정이 필요하다. NPM(Node Packaged Modules)은 외부 모듈의 다운로드 및 설치에 사용되는 시스템으로서, 명령 창에서 실행할 수 있다. 다음은 socket.io 모듈을 다운로드 받아 설치하는 NPM 명령어를 나타낸다.

```
> npm install socket.io
```

socket.io 모듈이 설치되어 사용할 준비가 되면, Node.js 에서 다음과 같은 코드로 웹소켓 기능을 사용할 수 있다.

```
var io = require('socket.io').listen(server); ①
io.sockets.on('input_reserve', function(data) { ②
  var Rsq1 = 'SELECT * FROM cabinet ';
  Rsq1 += 'WHERE cab_user = ';
  Rsq1 += 'W'+server_user+W';
  client.query(Rsq1, function(err, rows) { ③
    if (err)
      console.log(err);
    else
      for (i in rows)
```

공개 데이터의 프라이버시 안전성: 진료정보 통계자료 HIRA-NPS 2011 사례 분석*

김수형*, 정연돈**, 이기용***

*고려대학교 IT융합학과

**고려대학교 컴퓨터·통신공학부

***숙명여자대학교 컴퓨터과학부

e-mail:{soohyung, ydchung}@korea.ac.kr, kiyonglee@sookmyung.ac.kr

The Privacy Safety of Public Data: A Case Study on Medical Statistics HIRA-NPS 2011

Soohyung Kim*, Yon Dohn Chung**, Ki Yong Lee***

*Dept. of IT Convergence, Korea University

**Dept. of Computer Science and Engineering, Korea University

***Division of Computer Science, Sookmyung Women's University

요 약

개인정보가 포함된 데이터가 많은 기관에서 다양한 목적을 위해 배포되고 있다. 이러한 공개 데이터는 프라이버시 문제를 야기할 수 있기 때문에, 배포에 앞서 항상 데이터에 대한 프라이버시 보호가 고려되어야 한다. 그러나 현재 배포되는 많은 데이터는 충분하지 못한 프라이버시 보호 과정을 거쳐 배포되고 있다. 이 논문에서는 개인정보를 포함하는 데이터의 프라이버시 안전성을 분석한다. 이를 위해 우리는 건강보험심사평가원에서 배포한 2011년 진료정보 통계자료(HIRA-NPS)를 실험에 사용한다. 분석을 위해 기존에 널리 쓰이는 프라이버시 보호 모델 k -익명성(k -anonymity)과 l -다양성(l -diversity)을 차용하여 안전성 판단의 척도를 정의한다. 또한 실제 데이터에 이 척도를 적용하여 프라이버시 안전성을 측정하고, 그 결과가 갖는 의미를 분석한다.

1. 서론

개인정보가 포함된 데이터는 현재 마케팅, 사회 현상 분석, 연구 등의 다양한 분야에서 활용되고 있다. 애플의 음성인식 서비스 시리(Siri)는 사용자와 스마트폰의 대화 기록을 분석하여 서비스 품질을 높이는데 활용한다. 대표적인 검색 엔진 구글은 사용자의 검색 기록, 방문한 페이지 등을 수집하여 검색 엔진의 성능 향상에 활용한다. 이외에도 개인의 의료 정보, 위치 정보, 소셜 네트워크 정보 등이 여러 기관에서 배포되어 다양한 용도로 이용되고 있다.

개인정보를 포함하는 데이터의 활용은 항상 프라이버시 유출 가능성을 수반한다. 만약 데이터에 포함된 개인이 식별된다면, 그 개인의 민감한 정보가 노출될 수 있다. 이러한 프라이버시 문제를 방지하기 위하여, 데이터를 배포하기 이전에 적절한 프라이버시 보호 기법의 적용이 필요하다[1,2].

그러나 데이터를 배포하는 많은 기관은 단순히 개인을 구분하는 식별자(주민등록번호, 학번, 사번 등)만을 제거하여 데이터를 배포하는 등의 충분하지 못한 프라이버시 보호 과정을 따르고 있다. 실제로, 미국의 비디오 스트리밍 대역 업체 Netflix사는 추천 시스템의 공모를 위해 개인정

보가 담긴 데이터를 공개하였으나, 일부 개인의 민감한 정보가 노출되어 집단 소송에 연루된 사례가 있다.

이 논문에서는 개인정보를 포함하는 공개 데이터를 분석하여 데이터의 프라이버시 안전성을 측정한다. 이것은 데이터를 배포하기 이전에 프라이버시 문제가 발생할 여지를 미리 알아본다는 점과, 현재 배포되고 있는 데이터가 어느 정도의 프라이버시 안전성을 지니고 있는지 확인한다는 점에서 의미가 있다. 이를 위하여 우리는 프라이버시 보호 모델 k -익명성(k -anonymity)[3]과 l -다양성(l -diversity)[4]을 차용한다. k 와 l 값은 각각 ' k 개 이상의 구별 불가능한 레코드', ' l 이상의 다양성을 지닌 레코드'를 의미한다. 즉, 이 값들이 클수록 더 높은 수준의 프라이버시가 보장된다. 우리는 실제 공개 데이터의 k 값과 l 값을 구하고, 이를 바탕으로 해당 데이터가 어느 정도의 프라이버시 안전성을 갖는지 분석한다.

논문의 구성은 다음과 같다. 2장에서는 프라이버시 안전성을 판단하는 척도에 대해 알아본다. 3장에서는 프라이버시 안전성을 구하는 실험의 방법과 그 결과에 대하여 설명한다. 마지막으로, 4장의 결론을 통하여 논문을 마무리한다.

2. 프라이버시 안전성 판단 척도

이 장에서는 데이터의 프라이버시 안전성을 측정하는 방

* 이 논문은 2013년도 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임(No. 2013-022884)

<표 1> 실험에 활용한 속성

속성명	도메인 크기	속성 설명	배경지식 단계			
			L1	L2	L3	L4
연령	100	수진자의 연령	○	○	○	○
성별	3	수진자의 성별	○	○	○	○
입원도착경로구분	6	도착경로(1:타 기관 경유, 2:응급후송, 3:기타) + 입원경로(1:응급실, 2:외래)		○	○	○
진료결과구분	6	최종진료일의 환자상태 (계속, 이송, 회송, 사망, 기타, 퇴원)		○	○	○
진료과목코드	28	진료과목에 대한 코드		○	○	○
수술여부	2	수진자의 수술 여부		○	○	○
최초입원일	365	요양기관에 상병진료를 위하여 그달에 최초 내원한 년/월/일			○	○
요양개시일자	365	수진자가 진료를 받기 시작한 일자, 조제투여일자, 최초내방일자			○	○
요양만료일자	365	수진자가 진료 받기를 종료한 일자				○
요양일수	365	수진자를 진료한 총 일 수				○
입내원일수	365	수진자가 요양기관에 내원한 일수(초진+재진)				○
주상병	-	주된 상병분류기호	민감한 정보			
부상병	-	부상병코드				

법에 대하여 설명한다. 이를 위해 먼저 몇 가지 배경지식과 기존 프라이버시 보호 모델에 대해 알아본다.

개인정보를 포함하는 데이터 D 의 스키마는 다음과 같이 QI (준식별자)와 SI (민감한 정보)의 집합으로 표현된다.

$$D = \{QI_1, QI_2, \dots, QI_n, SI_1, SI_2, \dots, SI_m\}$$

여기서 준식별자는 데이터에서 개인의 특징을 표현하는 속성(attribute) 중 식별자를 제외한 것이다. 나이, 성별, 우편번호 등이 이에 속한다. 민감한 정보는 밝혀지게 되면 곤란한 개인의 사적인 정보로서, 질병, 개인적인 흥미, 수입 등이 이에 해당한다. 이 논문에서 다룰 프라이버시 문제는 공격자가 준식별자들의 조합($QI_1, QI_2, \dots, QI_n$)을 배경 지식으로 하여 특정인의 민감한 정보($SI_1, SI_2, \dots, SI_m$)를 알아내는 것으로 가정한다.

k -익명성과 l -다양성은 데이터의 개인정보를 보호하기 위해 널리 쓰이는 프라이버시 보호 모델이다. k -익명성은 데이터의 임의의 레코드가 최소한 $k-1$ 개의 다른 레코드와 구별되지 않음을 보장한다. 이 때, 구별되지 않음이란 ‘ QI 가 동일함’을 의미한다. 따라서 k -익명성은 $1/k$ 이하의 프라이버시를 유출 확률을 보장한다. l -다양성은 구별되지 않는 레코드 집합에서, 민감한 정보가 l 이상의 다양성을 가짐을 보장한다. 이는 QI 에 의해 레코드가 구별되지 않더라도 민감한 정보의 다양성 부족(lack of diversity)에 의해 프라이버시 문제가 발생할 수 있는 k -익명성 모델을 보완한 것으로, $1/l$ 이하의 프라이버시 유출 확률을 보장한다.

우리는 위의 두 모델을 프라이버시 안전성의 판단 기준으로 활용한다. k -익명성과 l -다양성의 목적인 프라이버시 보호를 역으로 활용하면, 해당 데이터가 어느 정도의 프라이버시 안전성을 지니고 있는지 알 수 있다. 데이터를 분석하여 QI 가 동일한 레코드를 찾는 방법으로 k 값을 구하고, 그 레코드 집합에서 민감한 정보의 다양성을 구하여 l 값을 도출한다. k 값과 l 값은 각각 $1/k$, $1/l$ 이하의 프라이버시 유출 확률을 의미하므로, 값의 크기는 프라이버시 보호 수준과 비례한다. 따라서 우리는 이 값들을 통해 데이터의 프라이버시 안전성을 파악할 수 있다.

3. 실험

실험을 위해 건강보험심사평가원의 전체 환자 표본 자료[5]가 사용되었다. 이는 2011년에 의료서비스를 이용한 모든 환자의 진료비 청구 명세서를 바탕으로 제작된 데이터이다. 데이터는 4600만 명의 모든 환자 중에서 3%(1,375,843명)의 환자를 층화 추출한 것으로 구성되어 있다. 우리는 전체 데이터 중 환자의 준식별자와 질병 정보(주상병, 부상병)가 포함되어 있는 명세서 일반테이블을 사용한다. 데이터의 크기는 약 3.82GB로, 약 2700만개의 레코드를 포함하고 있다. 데이터는 39개의 속성으로 구성되어 있으나, 우리는 이 중 현실적으로 QI 와 SI 로 활용 가능한 것들만 고려한다. 실험에 활용한 상세 속성 정보는 표 1과 같다.

전체 환자 표본 자료에는 개인정보 보호를 위해 단순한 정보보호 기법이 적용되어 있다. 먼저, 개인 식별자인 주민등록번호가 대체키로 암호화되어 있다. 따라서 공격자는 주어진 데이터에서 대체키를 통해 동일인을 구별할 수 있지만 개인을 식별할 수는 없다. 요양기관 고유번호 또한 이와 마찬가지로 암호화되어 있다.

실험에서는 평균 k , l 값과 k 값별 환자의 수를 측정하였다. (l 값별 환자의 수는 k 의 경우와 유사하여 제외하였다.) 각 실험은 임의로 가정한 QI 조합에 대하여 수행되었으며, 이 QI 의 조합은 공격자의 배경 지식을 고려하여 4단계($L1 \sim L4$)로 나누어 설정하였다. $L1$ 의 경우 가장 기본적인 배경지식인 연령, 성별만을 포함하며, $L2$ 의 경우 병원 도착 경로, 진료 과목 등의 보다 자세한 속성을 포함한다. $L3$ 는 최초 입원일 등의 속성을 추가적으로 포함하며, $L4$ 는 총 진료 일수 등을 포함한 가장 자세한 수준의 QI 의 조합으로 구성된다. $L4$ 단계에서 공격자가 가지는 배경 지식의 속성 개수는 총 11개이다. 한편, 속성 중 주상병과 부상병은 환자의 질병에 대한 것으로, 이 실험에서의 민감한 속성으로 가정한다.

3.1 평균 k , l 값에 의한 프라이버시 안전성 분석

맵리듀스에서 집계 질의 스트림의 효율적인 처리 기법

최현진, 이기용

숙명여자대학교 컴퓨터과학부

e-mail: gomsun09@naver.com, kiyonglee@sookmyung.ac.kr

Efficient Processing of an Aggregate Query Stream in MapReduce

Hyunjean Choi, Ki Yong Lee

Division of Computer Science, Sookmyung Women's University

요 약

최근 들어 맵리듀스는 빅데이터 처리의 표준 기술로 자리잡고 있다. 빅데이터 분석에 널리 쓰이는 질의 중 하나는 집계(aggregate) 질의이다. 본 논문에서는 서로 다른 집계 질의가 계속적으로 요청되는 환경에서, 맵리듀스를 사용하여 이들 질의를 효율적으로 처리하는 방법을 제안한다. 제안 방법은 여러 집계 질의를 하나의 효율적인 맵리듀스 잡(job)으로 묶어 일괄 처리함으로써, 단순 방법에 비해 시간당 처리되는 질의 수를 크게 증가시킨다. 성능 평가를 통해, 제안 방법은 단순 방법에 비해 처리 성능을 크게 향상시킴을 확인하였다.

1. 서론

최근 들어 Social Networking Service (SNS), WWW, 센서 네트워크, 네트워크 모니터링 등 다양한 분야의 데이터가 급증하면서 소위 빅데이터(Big Data)[1]에 대한 관심이 커지고 있다. 빅데이터란 용량이 매우 크거나, 증가 속도가 매우 빠르거나, 형태가 매우 다양해서 현존하는 기술로는 효율적으로 처리할 수 없는 데이터를 말한다[2].

빅데이터는 일반적으로 단일 컴퓨터로 처리하는 것이 불가능하다. 빅데이터 플랫폼이란 빅데이터를 다수의 컴퓨터에 분산하여 저장하고, 이를 다수의 컴퓨터로 병렬 처리할 수 있도록 하는 플랫폼을 말한다. 하둡(Hadoop)[3]은 현재 대표적인 산업계 빅데이터 표준 플랫폼으로 자리잡고 있다. 하둡을 비롯하여 대다수의 빅데이터 플랫폼은 빅데이터의 병렬 처리를 위해 맵리듀스(MapReduce)[4]라는 기법을 사용한다.

맵리듀스에서는 사용자가 수행하고자 하는 작업을 맵(Map)과 리듀스(Reduce)라는 두 개의 함수로 표현한다. 시스템은 이를 자동으로 다수의 컴퓨터에서 병렬로 수행한다. 더우기 수행 도중 어떤 컴퓨터에 이상이 발생하더라도, 해당 컴퓨터의 작업을 다른 컴퓨터에게 재할당함으로써 고장 감내(fault tolerance) 기능을 제공한다. 따라서 사용자는 복잡한 병렬 처리 메커니즘과 복잡한 고장 감내 메커니즘을 전혀 모르면서도, 대용량의 데이터를 다수의 컴퓨터를 활용하여 손쉽게 분석할 수 있다. 이러한 사용의 용이성과 편리성으로 인해, 맵리듀스는 빅데이터에 대한 표준 처리 기술로 자리잡고 있다.

한편 빅데이터의 분석을 위해 널리 쓰이는 질의 중 하나는 집계(aggregate) 질의이다. 집계 질의의 대표적인 예는 SQL의 group-by 질의로서, 데이터를 주어진 애트리뷰트(attribute) 값에 따라 그룹으로 나누고 각 그룹에 대해 합(sum), 개수(count), 평균(average) 등과 같은 집계치를 구하는 질의이다.

최근 빅데이터에 대한 접근성이 크게 향상되면서, 누구나 빅데이터에 접근하여 질의를 던지고 그 결과를 활용할 수 있도록 하는 연구가 활발히 진행되고 있다. 예를 들어, 여러 사용자의 스마트폰 사용 내역이 서버에 로그(log)로 저장되고 있을 때, 각 사용자는 로그 데이터에 대해 집계 질의를 던져 자신과 동일한 위치, 동일한 시간대, 혹은 동일한 연령대에 있는 사용자들의 사용 패턴에 대한 정보를 얻을 수 있다. 이 예는 로그라는 빅데이터에 대해 여러 집계 질의가 계속해서 요청되는 예로 볼 수 있다.

본 논문에서는 주어진 데이터에 대해 서로 다른 집계 질의가 계속해서 요청되는 경우, 맵리듀스로 이를 효율적으로 처리하는 방법을 제안한다. 앞서 언급한 바와 같이, 맵리듀스를 이용하면 병렬 처리와 고장 감내가 자동으로 지원되는 질의 처리기를 쉽고 효율적으로 구현할 수 있다. 제안 방법은 여러 집계 질의를 하나의 효율적인 맵리듀스 잡(job)으로 묶어 일괄 처리함으로써, 단순 방법에 비해 시간당 처리되는 질의 수를 크게 증가시킨다. 성능 평가를 통해, 제안 방법은 단순 방법에 비해 처리 성능을 크게 향상시킴을 확인하였다.

본 논문의 구성은 다음과 같다. 2 장에서는 본 논문

에서 다루는 문제를 정의하고, 3 장에서는 관련 연구를 간략히 살펴본다. 4 장에서는 제안 방법을 자세히 설명하고, 5 장에서는 성능 평가 결과를 보인다. 마지막으로 6 장에서는 결론을 맺는다.

2. 문제 정의

본 논문에서는 데이터가 다수의 튜플(tuple)로 이루어진 집합이며, 데이터의 스키마(schema)는 다음과 같다고 가정한다.

$$(D_1, D_2, \dots, D_b, M)$$

$D_1, D_2, \dots, D_b$ 는 시간, 위치 등과 같이 데이터의 그룹화에 사용되는 애트리뷰트이며, M 은 집계 대상이 되는 값을 저장하는 애트리뷰트이다.

시스템에는 여러 사용자가 각각 데이터에 대해 서로 다른 집계 질의를 계속해서 던질 수 있다고 가정한다. 각 질의 q 는 다음과 같이 표현된다.

$$q = (d_1, d_2, \dots, d_b, AGG(M))$$

위 질의 q 는 $D_1, D_2, \dots, D_b$ 애트리뷰트의 값이 각각 $d_1, d_2, \dots, d_b$ 인 튜플들에 대한 집계 함수 $AGG(M)$ 의 값을 요청하는 질의이다. 예를 들어, $q = ("2013-11", "서울", SUM(Requests))$ 는 2013년 11월에 서울에서 발생한 총 요청의 수를 구하는 데 사용되는 질의이다. 각 질의는 서로 다른 $d_1, d_2, \dots, d_b$ 값을 가질 수 있으며, 본 논문에서는 설명의 편의를 위해 AGG 는 SUM 이라고 가정한다. 하지만 제안 방법은 SUM 외에 $COUNT$, AVG , MAX , MIN 과 같은 일반적인 집계 함수에도 적용될 수 있다.

본 논문에서는 q 와 같은 형태의 질의가 계속해서 요청될 때(이를 질의 스트림이라고 하자), 이들을 맵리듀스를 사용하여 효율적으로 처리하는 방법에 대해 논의한다. 방법의 성능은 시간당 처리되는 질의 수로 측정한다.

3. 관련 연구

어떤 작업을 맵리듀스로 수행하기 위해서는 해당 작업을 맵과 리듀스 두 개의 함수로 표현해야 한다. 맵 함수는 입력으로 하나의 키-값 쌍(key-value pair) (k, v) 을 받아 하나 이상의 키-값 쌍 $(k_1, v_1), (k_2, v_2), \dots$ 을 출력으로 내보낸다. 맵 함수의 수행이 모두 끝나면, 시스템은 맵 함수가 출력한 키-값 쌍들을 정렬하여 같은 키 값을 가지는 키-값 쌍들을 모아 $(k', [v'_1, v'_2, \dots])$ 형태로 만든 후, 이를 리듀스 함수의 입력으로 전달한다. 리듀스 함수는 입력으로 하나의 키-값리스트 쌍 $(k', [v'_1, v'_2, \dots])$ 을 받아 하나 이상의 키-값 쌍 $(k'_1, v'_1), (k'_2, v'_2), \dots$ 을 출력으로 내보내며, 이것이 맵리듀스의 최종 수행 결과가 된다.

맵리듀스에서 데이터는 분산 파일 시스템의 파일 형태로 저장된다. 대용량의 파일은 고정된 크기(주로 64 MB)로 파티션되어 저장되며, 각 파티션을 split이라 부른다. 각 split은 고장 감내를 위해 하나 이상의

컴퓨터에 중복되어 저장된다.

맵리듀스에서 수행되는 작업을 맵리듀스 잡(job)이라 부른다. 맵리듀스 잡은 하나의 맵 함수와 하나의 리듀스 함수로 구성된다. 어떤 맵리듀스 잡의 수행이 요청되면, 시스템에는 정해진 수만큼의 맵 태스크(map task)와 리듀스 태스크(reduce task)가 생성된다. 각 맵 태스크는 맵 함수를 수행하는 주체이며, 각 리듀스 태스크는 리듀스 함수를 수행하는 주체이다. 각 맵 태스크는 입력 데이터를 저장하고 있는 파일의 한 split을 할당 받아, 해당 split에 저장된 데이터를 지정된 방식에 따라 키-값 쌍들로 변환하고, 각 키-값 쌍에 대해 사용자가 정의한 맵 함수를 호출한다. 맵 함수의 호출 결과로 생성된 키-값 쌍들은 다시 파일로 저장된다. 모든 맵 태스크의 수행이 완료되면, 각 맵 태스크는 맵 함수가 출력으로 내보낸 키-값 쌍들을 키 값으로 정렬하여 이들을 리듀스 태스크들에게 분배하여 전송한다. 이 때 어떤 키-값 쌍을 어떤 리듀스 태스크에게 전송하느냐는 사용자가 지정한 규칙에 따르며, 어떤 경우든 동일한 키 값의 키-값 쌍들은 동일한 리듀스 태스크로 전송된다. 맵 태스크의 수행 결과로 생성된 키-값 쌍들이 모두 리듀스 태스크들로 전달되고 나면, 리듀스 태스크의 수행이 시작된다. 각 리듀스 태스크는 전송받은 키-값 쌍들을 키 값으로 정렬한 뒤, 동일한 키 값을 가지는 키-값 쌍들을 모아 이로부터 키-값리스트 쌍을 생성하고, 각 키-값리스트 쌍에 대해 사용자가 정의한 리듀스 함수를 호출한다. 리듀스 함수의 호출 결과로 생성된 키-값 쌍들은 파일로 저장되며, 이것이 맵리듀스 잡의 최종 결과물이다.

지금까지 설명한 맵리듀스를 사용하여 다양한 형태의 질의를 보다 효율적으로 처리하는 방법에 대해서 현재 많은 연구가 진행되고 있다. 예를 들어, SQL에서 제공하는 Select, Project, Join, Group-By 등과 같은 질의를 맵리듀스로 어떻게 처리해야 하는지에 대해서 많은 연구가 이루어지고 있다[5][6]. 하지만 2장에서 정의한 형태의 집계 질의가 계속해서 요청되는 상황에 대해서는 아직까지 많은 연구가 이루어지지 않았다. 따라서 본 논문에서는 2장에서 정의한 집계 질의가 계속해서 요청되는 경우, 이들을 효율적으로 처리하는 방법을 제안한다.

4. 제안 방법

4.1 단순 방법

2장에서 정의한 집계 질의가 계속해서 요청되는 경우, 맵리듀스에서 내부적으로 정렬이 수행된다는 점을 이용하여 이들을 매우 간단히 처리할 수 있다. 질의가 계속해서 요청되는 경우, 각각을 개별적으로 하나의 맵리듀스 잡으로 만들어 처리하면 매우 큰 비용이 발생한다. 따라서 본 논문에서는 질의가 미리 정의된 개수만큼 도착하면(P 개라고 하자), 이들을 하나의 맵리듀스 잡으로 묶어 일괄처리한다.

P 개의 질의가 새로 도착했다고 하고, 이들 P 개 질의의 집합을 Q 라고 하자. 이들을 한꺼번에 처리하기 위해 하나의 맵리듀스 잡이 수행된다. 맵 함수는 데

새로운 형태의 사용자 인터페이스를 제공하는 안드로이드용 이미지 합성 애플리케이션 개발 *

한아름, 유은영, 이기용
숙명여자대학교 컴퓨터과학부
e-mail: gks0097@naver.com, yey5857@nate.com, kiyounglee@sookmyung.ac.kr

Development of an Image Synthesis Application for Android Providing a New Type of User Interface

Areum Han, Eunyoung Yu, Ki Yong Lee
Division of Computer Science, Sookmyung Women's University

요 약

안드로이드는 휴대용 장치를 위해 널리 사용되고 있는 공개 모바일 운영체제로서, 그를 기반으로 다양한 안드로이드용 애플리케이션들이 제작되고 있다. 그 중 이미지 합성 애플리케이션은 Social Networking Service (SNS), 블로그 등에 올라가는 다양한 이미지 제작을 위해 활발히 사용되는 애플리케이션이다. 본 논문에서는 기존의 이미지 합성 애플리케이션들과는 다른 방식으로 이미지를 합성하도록 해주는 새로운 사용자 인터페이스의 이미지 합성 애플리케이션을 개발한다. 본 논문에서 개발한 이미지 합성 애플리케이션은 보다 직관적이고 쉬운 방식으로 사용자가 이미지를 합성할 수 있도록 해준다.

1. 서론

최근 들어 안드로이드는 스마트폰, 태블릿 PC 등과 같은 휴대용 장치를 위한 모바일 운영체제로서 널리 사용되고 있다. 이와 함께 안드로이드에서 수행되는 애플리케이션(application)들의 개발도 활발하게 이루어지고 있다[1]. 휴대용 장치에서 널리 쓰이는 애플리케이션 중의 하나는 이미지 합성 애플리케이션이다. 이미지 합성 애플리케이션은 사용자가 Facebook[2]과 같은 Social Networking Service(SNS)나 블로그, 혹은 Wiki 등에 원하는 이미지를 제작하여 올리거나 할 때 흔히 사용된다.

본 논문에서는 다양한 기능을 제공하는 여러 이미지 합성 애플리케이션들 중에서, 두 이미지를 합성하여 새로운 이미지를 만들어내는 애플리케이션을 고려한다. 현재 이렇게 두 이미지를 합성하는 이미지 합성 애플리케이션들은 대부분 두 이미지 중 하나를 다른 이미지에 붙여 넣는 방식을 사용한다. 하지만 이러한 방식은 두 이미지를 합성하는데 불편함이 있다. 즉, 두 이미지를 합성하기 위해 각 이미지에서 원하는 부분을 직접 오려내야 하거나, 두 이미지의 크기가 맞지 않는 경우 각각의 크기를 별도로 조정해 주어야 한다는 불편함이 있다.

본 논문에서는 이러한 불편함을 해소하기 위해, 두 이미지를 쉽게 합성할 수 있도록 해주는 새로운 사용자 인터페이스의 안드로이드용 애플리케이션을 개발

한다. 본 논문의 이미지 합성 애플리케이션은 두 이미지를 합성하기 위해, 첫 번째 이미지에서 원하는 부분을 터치하여 잘라내고, 두 번째 이미지를 첫 번째 이미지의 아래에 위치시킨 후, 두 번째 이미지의 위치와 크기를 Drag 및 Pin-to-Zoom(이미지를 엄지와 검지로 벌리듯 확대하거나 조이듯 축소하는 기능) 방식으로 조정하여 원하는 부분이 첫 번째 이미지의 잘라낸 부분에 위치하도록 하는 방식을 사용한다. 이 방식은 사용자가 보다 직관적이고 쉬운 방식으로 이미지를 합성할 수 있도록 해준다.

본 논문의 구성은 다음과 같다. 2 장에서는 본 논문에서 개발한 이미지 합성 애플리케이션의 기능을 요약하고, 3 장에서는 개발한 애플리케이션의 실제 사용 예를 보여준다. 4 장에서는 개발에 사용된 기술들을 간략히 설명하고, 5 장에서 결론을 맺는다.

2. 개발 애플리케이션의 기능

본 논문에서 개발한 이미지 합성 애플리케이션은 두 개의 이미지를 불러와 첫 번째 사진의 원하는 위치에 두 번째 사진의 원하는 부분을 쉽게 위치시킬 수 있도록 해준다. 본 논문에서 개발한 이미지 합성 애플리케이션이 제공하는 기능은 크게 다음과 같다.

- **이미지 촬영 및 로드 기능:** 합성의 대상이 되는 이미지를 직접 촬영하거나 갤러리에서 불러올 수 있다.

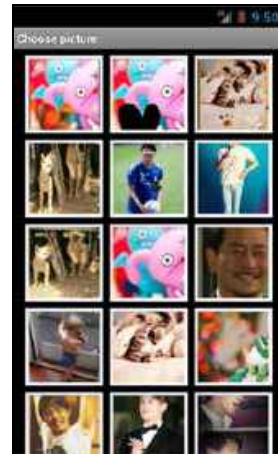
*이 논문은 2013년도 정부(미래창조과학부)의 지원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2012R1A1A1001269)

- **이미지 밝기 및 대비 조절 기능:** 합성의 대상이 되는 이미지를 불러온 후, 그의 밝기(brightness) 및 대비(contrast)를 슬라이더 바를 이용하여 손쉽게 조절할 수 있다.
- **이미지 잘라내기 기능:** 먼저 불러온 이미지에서 원하는 부분을 잘라내는 기능이다. 이 잘라낸 공간에 다른 이미지의 일부가 합성된다.
- **이미지 합성 기능:** 먼저 불러온 이미지 아래에 다른 이미지를 불러와 위치시킨 후, Drag 및 Pinch-to-Zoom 방식으로 아래에 위치한 이미지의 위치와 크기를 조정하여 원하는 부분이 먼저 불러온 이미지의 잘라낸 부분에 위치하도록 한 후, 합성을 완료한다.
- **이미지 저장 및 타 어플리케이션 연동 기능:** 합성이 완료된 이미지를 저장하거나 다른 어플리케이션과 연동하여 부가적인 편집을 할 수 있다.

본 논문에서 개발한 이미지 합성 애플리케이션이 기존 애플리케이션과 차별화되는 점은 다음과 같다. 현재 Google Play[3]에서 두 개의 이미지 합성에 사용되는 애플리케이션들은 대부분 두 이미지 중 하나를 다른 이미지의 원하는 부분에 붙여 넣는 방식을 사용한다. 이에 비해 본 논문에서 개발한 애플리케이션은 한 이미지의 일부를 잘라낸 뒤, 다른 이미지를 원래 이미지의 아래에 위치시킨 후 그의 위치와 크기를 조정하여 합성을 완료한다. 따라서 보다 손쉽고 직관적인 방식으로 합성을 수행할 수 있다.

3. 사용 예

본 장에서는 본 논문에서 개발한 이미지 합성 애플리케이션의 실제 사용 예를 보인다. (그림 1)은 합성될 두 이미지 중 첫 번째 이미지를 로드하는 장면이다. (그림 2)는 첫 번째 이미지를 로드한 후, 해당 이미지의 밝기와 대조를 슬라이드 바를 이용하여 조정하는 장면이다. 이렇게 첫 번째 이미지의 밝기와 대조를 조정한 후에는 해당 이미지에서 합성이 일어날 부분을 잘라내는 작업을 수행한다. (그림 3)은 간단한 터치를 통해 이미지에서 원하는 부분을 잘라내는 장면을 보여준다. 잘라낸 부분에 추후 다른 이미지가 합성된다. (그림 4)는 첫 번째 이미지와 합성할 다른 이미지를 불러와 첫 번째 이미지의 아래에 위치시킨 후, Drag 및 Pinch-to-Zoom 방식을 통해 해당 이미지의 위치와 크기를 조정하여 첫 번째 이미지와 합성하는 장면이다. 마지막으로 (그림 5)는 합성이 완료된 이미지를 저장하고 다음 작업을 선택하는 장면이다.



(그림 1) 이미지 로드



(그림 2) 이미지 밝기 및 대조 조절



(그림 3) 이미지 잘라내기

어학 공부와 전자책 읽기 기능을 제공하는 안드로이드용 한영 텍스트-음성 변환 애플리케이션 개발 *

유미연, 김수지, 김소민, 이기용

숙명여자대학교 컴퓨터과학부

e-mail: miyeon4385@hanmail.net, curywoman@naver.com, ardor1229@naver.com,

kiyonglee@sookmyung.ac.kr

Development of a Korean-English Text-to-Speech Application for Android Supporting Language Study and Book Reading

Miyeon Yu, Suji Kim, Somin Kim, Ki Yong Lee

Division of Computer Science, Sookmyung Women's University

요 약

최근 들어 스마트폰과 같은 휴대용 장치에 안드로이드 운영체제가 널리 탑재되면서, 안드로이드용 애플리케이션이 활발히 개발되고 있다. 본 논문에서는 텍스트를 음성으로 변환하는 안드로이드용 TTS(Text-to-Speech) 애플리케이션을 개발한다. 기존의 유사 TTS 애플리케이션들과 달리, 본 논문에서 개발한 애플리케이션은 사용자가 직접 입력한 임의의 텍스트를 음성으로 읽어주는 기능과 미리 저장되어 있는 전자책을 불러와 음성으로 읽어주는 기능을 모두 제공한다. 또한 한글과 영어를 모두 지원하기 때문에, 한글 텍스트와 영어 텍스트를 모두 사용할 수 있다. 따라서 본 애플리케이션은 여러 사용자에게 의해 다양한 용도로 사용될 수 있다.

1. 서론

최근 들어 스마트폰, 디지털 카메라 등과 같은 모바일 기기의 운영체제로 안드로이드의 탑재가 널리 확산되고 있다. 안드로이드의 사용 확대와 함께, 다양한 안드로이드용 애플리케이션(application)들도 활발히 개발되고 있다[1]. 그 중 TTS(Text-to-Speech)[2] 애플리케이션은 텍스트를 음성으로 변환하여 들려주는 기능을 제공하는 애플리케이션으로서, 아이들이나 글을 읽지 못하는 사람들에게 글을 읽어주거나, 운동 등 책을 읽기 어려운 상황에서 책의 내용을 듣거나, 혹은 자동으로 안내 방송을 내보내거나 등의 다양한 용도로 사용된다.

본 논문에서는 이러한 용도 외에 어학 공부의 용도로도 사용될 수 있는 안드로이드용 TTS 애플리케이션을 개발한다. 최근 들어 TOEIC Speaking 이나 OPIc 과 같은 영어 말하기 시험을 준비하거나 영어 회화를 공부할 때, TTS 애플리케이션을 사용하여 본인이 직접 입력한 문장의 발음을 들어보거나, 발음이 궁금한 문장을 입력하고 TTS 애플리케이션을 통해 그의 발음을 반복하여 듣는 방식으로 영어 듣기와 말하기를 공부하는 경우가 늘고 있다. 영어 외에도, 한류의 영향으로 한국에 대한 관심이 높아지면서 한국어를 배우고자 하는 외국인들이 많아지고 있는데, 이러한 외

국인들에게 한글이 지원되는 TTS 를 제공하면 그들의 한글의 읽기와 듣기 학습에 매우 큰 도움을 줄 수 있다.

하지만 지금까지 개발된 대부분의 안드로이드용 TTS 애플리케이션들은 전자책과 같이 미리 저장되어 있는 텍스트만을 불러와 읽어주는 기능만을 가지고 있거나, 또는 사용자가 직접 입력한 텍스트를 읽어주는 기능을 제공하더라도 지원되는 텍스트의 길이가 매우 짧거나 (1~2 문장), 미리 저장되어 있는 텍스트를 불러와 읽어주는 기능은 제공하지 않는다[3]. 또한 한국어와 영어를 모두 지원하면서, 사용자가 직접 입력한 텍스트와 전자책과 같이 미리 저장되어 있는 텍스트를 모두 읽어주는 애플리케이션은 찾아보기 어렵다.

따라서 본 논문에서는 여러 사용자들이 다양한 용도로 사용할 수 있는 새로운 안드로이드용 TTS 애플리케이션을 개발한다. 본 논문에서 개발한 TTS 애플리케이션은 어학 공부 등을 위해 사용자가 직접 입력한 텍스트를 읽어주는 기능을 제공하는 한편, 전자책과 같이 미리 저장되어 있는 텍스트를 불러와 읽어주는 기능도 동시에 제공한다. 또한 한글과 영어를 모두 지원하기 때문에 한글 텍스트와 영어 텍스트를 모두 사용할 수 있다.

*이 논문은 2013 년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2012R1A1A1001269)

본 논문의 구성은 다음과 같다. 2 장에서는 본 논문에서 개발한 애플리케이션이 제공하는 기능들을 소개한다. 3 장에서는 개발된 애플리케이션의 일부 화면을 보여주고, 4 장에서는 본 애플리케이션 개발에 사용된 기술들을 간략히 설명한다. 5 장에서는 결론을 맺는다.

2. TTS 애플리케이션 기능

본 논문에서 개발한 안드로이드용 TTS 애플리케이션은 어학 공부 및 전자책 읽기의 용도로 모두 사용될 수 있다. 또한 한글과 영어를 모두 지원하며, 음성의 빠르기 및 속도도 조절이 가능하다. 본 논문에서 개발한 애플리케이션의 세부 기능들은 다음과 같다.

- **텍스트 직접 입력:** TTS 로 읽어주기를 원하는 임의의 텍스트를 사용자가 직접 입력할 수 있도록 해주는 기능이다.
- **파일에서 불러오기:** 전자책과 같이 미리 저장되어 있는 텍스트를 불러오는 기능이다. SD 카드에 저장된 텍스트도 불러올 수 있다.
- **저장 DB 에서 불러오기:** 본 애플리케이션은 사용자가 직접 입력한 텍스트를 DB 에 저장하는 기능을 제공한다. (아래에서 설명) 본 기능은 이렇게 DB 에 저장된 텍스트를 다시 불러오는 기능이다.
- **텍스트 읽기:** 본 애플리케이션의 핵심 기능으로서, 위의 세 가지 방법으로 읽기가 준비된 텍스트를 실제로 읽어주는 기능이다.
- **텍스트 저장:** 화면에 나타나 있는 모든 텍스트를 내부 DB 에 저장하는 기능이다. 이렇게 저장된 텍스트는 언제든지 나중에 다시 불러올 수 있다.
- **음성 인식:** 텍스트 창에 사용자가 텍스트를 직접 입력할 때, 타이핑을 통해서가 아니라 음성을 통해서도 음성 인식을 통해 텍스트를 입력할 수 있게 해준다.
- **음성 빠르기 조절:** 텍스트를 음성으로 들려줄 때, 음성의 빠르기를 조절한다.
- **음성 높낮이 조절:** 텍스트를 음성으로 들려줄 때, 음성의 높낮이(pitch)를 조절한다.

1 장에서 언급한 바와 같이, 기존에 출시되어 있는 애플리케이션 중에서도 TTS 기능을 제공하는 애플리케이션들이 많이 있다. 하지만 대부분은 전자책 읽기 용도로서, 미리 저장되어 있는 텍스트만을 읽어주거나, 또는 사용자가 직접 입력한 임의의 텍스트를 읽어준다 하더라도 지원하는 텍스트의 길이가 매우 짧다 (단어 또는 1~2 문장). 이에 비해 본 논문에서 개발한 애플리케이션은 두 가지 기능 모두를 제공하며, 사용자가 1,000 자 이상의 텍스트를 직접 입력하는 것도 가능하다. 따라서 뉴스 기사와 같은 긴 문장을 음성으로 듣고자 할 때 유용하게 사용될 수 있다. 또한 본 애플리케이션은 영어와 한글을 모두 지원하기 때문에, 영어를 공부하거나 한글을 공부할 때 모두 사용될 수 있다.

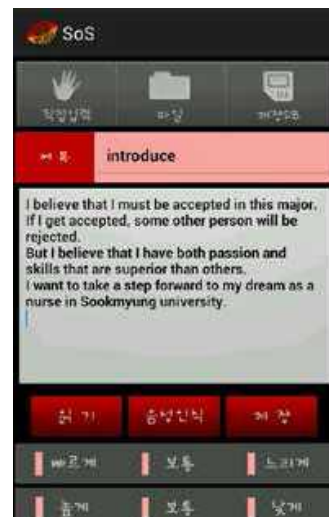
3. TTS 애플리케이션 화면

본 장에서는 본 논문에서 개발한 TTS 애플리케이션의 일부 화면을 보여준다. (그림 1)은 본 TTS 애플리케이션의 시작 화면이다. 한글과 영어를 모두 지원하며, 사용자는 버튼을 눌러 원하는 언어를 선택할 수 있다.



(그림 1) 시작 화면

(그림 2)는 본 애플리케이션의 메인 화면이다. 사용자는 '직접입력', '파일', '저장 DB' 버튼을 통해 TTS 로 읽을 텍스트를 준비할 수 있으며, 각 기능은 2 장에서 설명한 바와 같다. 읽을 텍스트가 준비되면 사용자는 '읽기' 버튼을 눌러 텍스트 읽기 기능을 구동할 수 있다. 사용자는 '음성인식' 버튼을 눌러 음성을 통해 텍스트를 입력할 수도 있다. '저장' 버튼을 누르면 화면에 나타나 있는 텍스트가 내부 DB 에 저장된다. 화면 하단에는 텍스트를 읽는 음성의 빠르기와 높낮이를 조정하는 버튼이 있다.



(그림 2) 메인 화면

POWER-EFFICIENT CONFIGURATION CACHE STRUCTURE FOR COARSE-GRAINED RECONFIGURABLE ARCHITECTURE*

YOONJIN KIM

*Department of Computer Science,
Sookmyung Women's University,
Seoul, 140-742, South Korea
ykim@sookmyung.ac.kr*

Received 17 December 2010

Accepted 17 April 2012

Published 19 February 2013

Coarse-grained reconfigurable architectures (CGRA) require many processing elements (PEs) and a configuration memory unit (configuration cache) for reconfiguration of its PE array. Although this structure is meant for high performance and flexibility, it consumes significant power. Specially, power consumption by configuration cache is explicit overhead compared to other types of IP cores. Reducing power in configuration cache is very crucial for CGRA to be more competitive and reliable processing core in embedded systems. In this paper, I propose a power-efficient configuration cache structure based on two design schemes — one is a reusable context pipelining (RCP) architecture to reduce power-overhead caused by reconfiguration and another is a dynamic context management strategy for power saving in configuration cache. This power-efficient approach works without degrading the performance and flexibility of CGRA. Experimental results show that the proposed approach saves 56.50%/86.84% of the average power in write/read-operation of configuration cache compared to the previous design.

Keywords: System-on-chip (SoC); embedded systems; digital signal processing; coarse-grained reconfigurable architecture (CGRA); configuration cache; context word.

1. Introduction

As the market pressure of embedded systems compels the designer to meet tighter constraints on cost, performance, and power, the application specific optimization of a system becomes inevitable. On the other hand, the flexibility of a system is also important to accommodate rapidly changing consumer needs. In addition, with the growing demand for high quality multimedia, especially over portable media, there has been continuous development on more sophisticated algorithms for audio, video, and graphics processing. These algorithms have the characteristics of complex

*This paper was recommended by Regional Editor Majid Ahmadi.

data-intensive computation. To accommodate these incompatible demands while efficiently supporting the complex embedded-applications, domain-specific design has emerged as a suitable solution for embedded systems. Coarse-grained reconfigurable architecture (CGRA)-based embedded system is the domain-specific design in that it can boost the performance by adopting specific hardware engines while it can be reconfigured to adapt to ever-changing characteristics of the applications.

In spite of the above advantages, the deployment of CGRA is prohibitive due to its significant power consumption. This is due to the fact that CGRA is composed of many computational resources such as ALU, multiplier, divider and configuration cache to perform frequent memory-read operations for dynamic reconfiguration in every cycle. The configuration cache is the main component in CGRA that provides distinct feature for dynamic configuration. Even though configuration cache plays an important role for high performance and flexibility, it suffers from large power consumption. Therefore, reducing power consumption in the configuration cache has been a serious concern for reliability of CGRA. This paper addresses the power-saving approach to configuration cache and provides design schemes to achieve this.

In this paper, I propose a power-efficient configuration cache structure based on two design schemes — one is a reusable context pipelining (RCP)¹ to support power-conscious loop pipelining and another is a dynamic context management (DCM) strategy for power saving in configuration cache. RCP is a universal approach in reducing power and enhancing performance for CGRA because it can be achieved by closing the power-performance gap between low power-oriented spatial mapping and high performance-oriented temporal mapping. In addition, DCM is a novel control mechanism of configuration cache to manage context words in both cases at transfer time and run time with removing redundant context words for power-saving. The validation of the proposed approaches is demonstrated through the use of real application benchmarks and gate level simulations. The CGRA including the proposed configuration cache has shown to save 56.50%/86.84% of the power in write/read-operation of configuration cache compared to conventional CGRA.

This paper is organized as follows. After describing the related work in Sec. 2, I present CGRA, its context architecture and loop pipelining as preliminary in Sec. 3. In Secs. 4 and 5, I present the motivation of the proposed approach and necessity of power-efficient configuration cache structures for CGRA. Then I describe power-conscious loop pipelining and DCM for low power CGRA in Secs. 6 and 7. I show the experimental results in Sec. 8 and conclude the paper in Sec. 9.

2. Related Work

Many CGRAs have been suggested as summarized in Ref. 2. Most of them comprise a set of specialized processing elements (PEs) and interconnection fabrics between them and the run-time control of the operation of each PE and the interconnection provides the reconfigurability. These CGRAs can be classified into two cases:

멀티미디어과학과

Mobile Advertising Service Based on 4R Mode

Arto Kaasinen and Yong-Ik Yoon

Department of Multimedia Science, Sookmyung Women's University, Seoul 140-742, South Korea

Received: April 16, 2013/ Accepted: May 06, 2013 / Published: May 31, 2013.

Abstract: As people own more and more various smart devices, providing advertising to those devices has become a lucrative platform for different advertisement providers. However, the way advertisements are delivered to customers' smart devices needs to be improved. Many variables determine the most effective way to provide the mobile advertisement to the end-user. So a mobile advertising service based on the 4R model is needed to make the mobile advertising interesting and also useful for the customers. The correct time, location, contents and target device need to be determined in order for the advertisement to be contextual and the most effective. The 4R model is beneficial for the advertisers and the end-users alike.

Key words: Smart mobile device, mobile advertising, context-awareness, situation-awareness, user behavior.

1. Introduction

In these days people use many different devices to consume various kinds of contents. Majority of these contents are consumed through a device which has a display. The devices that people typically have are a television, a personal computer/laptop, a tablet computer and a smartphone. People can also own gaming consoles with a touchscreen controller, mp3 players and so on. People actively use those devices either concurrently or in sequence. The multiple display environment of the user can also be taken advantage of in advertising. However, the user's context needs to be taken into consideration. The user might not be able to use all of the different types of devices. And the devices that the user is using at any given time also depend on the context of the user. So the user context needs to be analyzed in order to provide the most effective advertisement to the user. The focus of this paper is in finding the correct way of delivering advertisements to the end-user's smart device. And also on how to take advantage of the user's various smart devices in order to provide

effective advertising. There are more possibilities than just showing the same advertisement on multiple displays.

Context basically means any information which can be used to define the situation of various things. Those things can be people, objects, places, weather, etc.. In computer science those can be users, other applications, networks etc. [1]. The context needs to be relevant, because useless information only makes things unnecessarily complicated. For example, when a user is using a mobile device indoors, then the weather is not really relevant information that needs to be processed, because the weather has no effect on the usage of the mobile device. Though, it will become relevant information when the user wants to go outside. When there are other users nearby, then they are more relevant information, because they can be maybe interacted with in some way.

Context awareness in technology and computer science means that a computer is able to sense the surrounding environments and react to it or its changes. A context aware computer uses context to manage services and applications and change their states accordingly and it also uses the context to provide useful information to the user [1].

Corresponding author: Yong-Ik Yoon, Ph.D., professor, research fields: mobile advertising, smart device context awareness. E-mail: yiyoon@sm.ac.kr.

In this paper the authors propose a Mobile Advertising Service Based on 4R Model. The big idea behind the 4R model is that the delivery of the correct advertising needs to be carefully planned in order for the advertisement to be the most effective. The 4R's are the right time of delivery, right location of delivery, right advertising contents and finally the right target device. When an advertisement gets all of these conditions correct, it will also make sense to the customer in the current context. So the advertisement will probably also be useful to the end-user and provide some value. Also, multiple devices can be used creatively as an advantage in the delivery of the advertisement. Multiple devices add the value in advertising, because they give the advertisers many ways to interact with the customers.

The paper is organized as follows: Section 2 details the related work; Section 3 offers a general overview of the model; Section 4 gives a detailed description of the different aspects of the model; Section 5 provides analysis; Section 6 gives conclusions.

2. Related Works

2.1 Mobile Advertising

Currently the advertising for smartphones is not working so well for the advertisement companies. The smaller size of the advertisements is a big issue, because it is hard to present a product or a service on a smaller screen. Tablets and computers on the other hand are much more profitable. There are other problems too. Currently it is more difficult to gather information about the smartphone users in order to provide more accurate and contextual advertisements [2].

However, there are many advantages in smartphone advertising. Being a mobile platform means that the users move around with them and their location can also be tracked. Location-based mobile advertising already exists, but for the advertisements to be truly contextual, the smartphone needs to sense its surroundings in some way or contact another device in

its surroundings. Advertising based only on the location data might not be that effective, because users' interests and past behavior are not taken into account. And when the users are for example going on to a journey then the advertisements in their current location might be completely useless.

2.2 Contextual Advertising

Contextual advertising is the advertisements that are displayed in various media. It is a form of targeted advertising. The advertisements are selected automatically based on the content that the user is currently seeing. Google is a famous contextual advertising service provider with its AdSense program, which automatically displays advertising in a web browser based on the site that the user is currently viewing or the browser history. In smartphones, the contextual advertisements are usually displayed in the various apps such as web browsers, video apps, games and more [3].

The problem with the current contextual mobile advertising is that it usually displays advertisements based on what the user is currently viewing on their smartphone. The users' previous behavior might not be taken into account. The even bigger problem is that the current physical, actual environments of the user are also not taken into account. It would make more sense for example to advertise sports equipment special offers when the user is nearby the sports equipment store and the user has a history of buying sports equipment.

2.3 Context-Aware Smartphones and Applications

In late 2010, Intel revealed that it was working on a context-aware API for smartphones and tablets. Their API will use the smartphone's sensors and also the user's social networks and personal tastes in order to analyze user behavior and to provide content. There was a prototype application, which guided users in the city they were by providing info about local attractions and by recommending restaurants based on

스마트 학습 공간 구성을 위한 Timed Button 기반의 다중스크린 동기화 기법

윤 용 익*, 조 윤 아*

A Study on Multi-Screen synchronization techniques based on Timed Button configuration for smart learning space

YongIk Yoon*, YoonAh Cho*

요 약

스마트 디바이스의 개발과 발달, 그리고 스마트 디바이스의 보편화에 따라 사용자들은 한 개 이상의 스마트 디바이스를 소지하게 되었다. 기존의 온라인 학습 서비스는 하나의 스크린을 분할시켜 다양한 정보를 제공해 왔다. 그러나 기존의 PC환경과 다르게 스마트 디바이스가 발달되고 보편화됨에 따라, 사용자의 학습공간은 하나의 스크린 환경에서 다중스크린 환경으로 변화할 것이다. 본 논문에서는 다중스크린을 기반으로 한 스마트 학습공간을 구성하고, 스마트 학습공간에서 미디어 콘텐츠를 융합하여 서비스하는 시스템을 제고하고자 한다. 이에 다중스크린 관련 연구 분야와 다중스크린 환경에 합당한 온라인 학습 서비스를 분석·설계하여 Timed Button이라는 스마트 디바이스 간의 동기화 기법을 제시한다.

▶ Keywords : 다중스크린, 스마트 학습 서비스, 스마트폰 서비스

Abstract

In these days, the smart devices are being developed and used repeatedly. But almost E-learning System did not provide these smart devices environment and user's E-learning space is changing from Single-screen environment to multi-Screen environment. Accordingly we configure smart E-learning space based on multi-screen and integrate the multi-media contents for the Smart E-learning service. We present the synchronization techniques to make the multi-screen

•제1저자 : 윤용익 •교신저자 : 윤용익

•투고일 : 2013. 6. 23, 심사일 : 2013. 7. 23, 게재확정일 : 2013. 8. 21.

* 숙명여자대학교 멀티미디어과학과(Dept. of Multimedia Science, Sookmyung Women's University)

※ 본 연구는 숙명여자대학교 2012학년도 교내연구비 지원에 의해 수행되었음.

environment for smart space with Timed Link and design the Timed Button to relate each Screen which is using multi-media e-learning contents.

▶ Keywords : multi-screen, smart e-learning system, smart phone service

I. 서 론

최근 스마트 디바이스의 개발과 발달, 그리고 스마트 디바이스의 보편화에 따라 사용자들은 한 개 이상의 스마트 디바이스를 소지하게 되며 사용자들은 인터넷을 통해 다양한 미디어를 접할 수 있는 기회를 가질 수 있게 되었다[1]. 그 뿐만 아니라, 안정적인 유무선 네트워크 통신망의 제공과 개인이 소지한 다양한 스마트 디바이스를 연결함으로써 개인적인 스마트 네트워크 공간을 구성하고 각각의 스마트 디바이스에 적합한 이미지 콘텐츠를 동시에 사용할 수 있는 환경이 조성되었다[2]. 이에 스마트 네트워크 공간에서 사용할 수 있는 맞춤형 서비스들의 필요성이 대두된다 할 수 있다[3][4].

기존의 온라인 학습 서비스는 대부분이 단순한 미디어 콘텐츠만을 제공한다. 그렇기 때문에 사용자가 더 많은 정보를 얻고자 하면, 하나의 스크린에서 다중 프레임을 사용한 일차적인 학습 서비스를 제공하게 된다. 그러나 정보가 제공되는 프레임의 크기가 분할되면, 정보의 중요도 또는 프레임의 크기가 감소하여 정보의 전달성이 떨어지게 된다. 스마트 디바이스의 개발과 발달, 보편화됨에 따라 기존의 단순 영상 미디어 콘텐츠나, 하나의 스크린에서 제공되어 각 콘텐츠의 중요도가 떨어지는 온라인 학습이 아닌, 다양한 미디어 콘텐츠가 융합되어 개인의 다양한 스크린을 적합하게 사용할 수 있는 다중스크린을 기반으로 한 스마트 학습 공간이 요구된다. 또한 미디어 콘텐츠를 융합하여 스마트 학습 시스템을 구성하기 위해서는 사용자의 스마트 디바이스(스크린)의 동기화가 필수적이다[5].

본 시스템은 이용 중인 학습시스템의 메인 미디어 콘텐츠에 심층적인 학습 효과를 주기 위하여, 적합한 시점에 보조적으로 동기화된 미디어 콘텐츠가 제공된다. 기존의 일차적인 정보를 담은 콘텐츠 제공에서 더 나아가, 해당 강의에 대한 능동적인 학습을 진행하도록 유도하기 위하여, 스마트 학습 환경에 적합한 형태의 강의와 그에 적합한 융합 미디어 콘텐츠를 제공한다. 이를 위하여 Timed Link라는 디바이스 연결을 통해 N개의 스크린을 연결하고, 사용자를 위한 스마트 학

습 환경을 구성한다. 또한 Timed Button 기반의 다중스크린 동기화 기법을 제시하여 N개의 스크린을 동기화한다. 즉, 본 시스템은 스마트하게 정보를 분석하고, 사용자의 최소한의 제어로 다양한 콘텐츠를 이용할 수 있는 스마트 학습 공간 서비스를 제공한다. 이 스마트 학습 시스템은 다중스크린 체제를 기반으로 하여 학습자에게 맞춰진 융합 미디어 콘텐츠를 제공하기 때문에, 보다 효율적인 스마트 학습 공간을 구성할 것이다.

본 논문의 2장에서는 관련된 기술동향을 언급하고, 3장에서 Timed Link 기반의 다중 스크린 모델을 제시한다. 4장에서는 Timed Button 기반의 동기화 기법의 구현 전략 및 결과를 제시하고, 5장에서는 본 논문의 결과를 기술한다.

II. 관련 연구

2.1 다중스크린 기법

다중스크린은 N개의 스크린을 사용하여 사용자들에게 서비스를 제공하는 기술로써 '여러 개의 정보기기로 관련된 콘텐츠를 이용할 수 있는 차세대 네트워크 서비스'이다[6][7]. 즉, 다중스크린은 하나의 콘텐츠에 대한 많은 데이터(빅 데이터)를 스마트 폰, PC, 스마트 TV, 태블릿 PC 등 다양한 디지털 디바이스의 특성에 맞춰 동시에 제공해주는 차세대 컴퓨팅 네트워크 서비스로 한 콘텐츠에 대한 다양한 정보를 동시에 적합하게 파악할 수 있게 해주는 서비스이다[8].

국내외 이동통신사들은 다중스크린을 통해 사용자들이 자사 서비스를 이용하는 시간이 늘어날수록 영화, 드라마 등 콘텐츠를 구입할 가능성이 높아진다. 궁극적으로 콘텐츠 매출을 늘릴 수 있기 때문에 미래 핵심서비스로 인지해 시장 선점을 위해 경쟁하고 있다. 국내외 이동통신사들은 다중스크린을 미래 핵심서비스로 인식해 시장 선점을 위해 경쟁하고 있다[9].

기존의 온라인 학습 시스템에 다중스크린 기술을 응용하게 된다면, 사용자들이 소지한 다양한 스마트 디바이스들을 융합하여 보다 효과적인 교육 콘텐츠 제공이 가능해질 것이다.

A model of Self-reliant Mobile Cloud for Collaboration Services

¹Sumi Song, ²YongIk Yoon

^{1,2}Multimedia Science, Sookmyung Women's University, Seoul, Korea,
songsm0328@naver.com, yiyoon@sm.ac.kr

Abstract

Collaboration between mobile devices is again attracting attention because mobile devices are getting more powerful and more widespread. In addition, Cloud Computing technology is being applied to mobile devices. A variety of tasks is made possible despite the technical limitations. Also, the Cloud can be treated as a virtual space between devices. So, each of the mobile devices can share their own superfluous resources in the Cloud. It is expected to be a good solution for preventing the waste of IT resources and increasing working efficiency. So, in this journal, models of collaboration with mobile cloud are proposed to equip environment with cooperative work and group service. We expect that this proposition will be a help in understanding the mobile cloud collaboration.

Keywords: Mobile Cloud; Collaboration; Mobile Cloud Application, smart service model.

1. Introduction

Users can similarly work for their desired actions with mobile devices like using personal computer (PC) [2][7]. But it is a fact that mobile devices still do not have enough computing power and application abilities because they still have many limitations such as small size, heavy weight, insufficient energy supply and etc. For solving this dissatisfaction, through collaboration, mobile devices are allowed to share IT resources. For the cooperation with mobile devices, the cloud computing can be more a systematic and reliable tool. The mobile devices connect to service cloud(SCs) individually for getting service data or saving own device information. This means that different mobile devices can be managed in cloud. Then, devices can be is stuffed needful abilities by collaborative action with near devices. Through collaboration, we expect to increase of performance and capability of a single device. Also if users have group work, this work can be performed easier and more efficiently because users are able to share their progress in mobile cloud.

For this reason, in this paper, we propose a collaborative service model for sharing the resources between mobile devices. It is not simply about making connections with mobile devices. New models of collaboration with mobile cloud aim to reduce the burden within a productive working relationship and rationality.

2. Main Issues for Collaborative mobile Service

2.1. Mobile Cloud service Interworking

The mobile cloud service interworking can be defined as services that can allow sharing various contents across multiple mobile devices in real time [3]. In other words, users can take advantage of the resources and the variety of content with their owned devices through a connection with cloud server. The same content can be enjoyed on multiple devices. Thus, many services are being provided on the interworking such as (1) synchronization service (Apple Mobile Me, Google Google Sync, Microsoft My Phone, LG U+, etc.), (2) streaming service (SK Hoppin in Korea) and (3) web storage service. A wider variety of service interworking studies have been performed. Also, the service interworking between mobile cloud and devices, between content and content as well as service and service in the cloud have attracted interest. And the approach from interworking to collaboration is to be expected.

2.2. Data Conversion

Users can be provided a variety of data and content, because they can overcome the limitations of mobile devices through mobile access to cloud. In addition to this, several actions that were previously

only possible only on a stationary PC through a network connection became available in mobile devices. The searching with Google, Bing and Yahoo, NAVER, etc. and of course Twitter, Facebook, Flickr, and a variety of social networking services are also available in the mobile devices. Also, office infrastructure access is available through cloud with mobile devices and working within cloud may be performed. Moreover, application services such as navigation, LBS-based services are possible without large crowd.

The Cloud has nearly infinite amount of services and nearly an infinite amount of space to store a lot of resources in. But the definition of a way to connect to the cloud, receive data and share the data is not clear. Users cannot ignore the problem with the heavy traffic which is due to a lot of people accessing the cloud simultaneously. In addition, security problems which might occur in this process cannot be overlooked.

Thus, various kinds of services as well as stable and efficient access methods and way of sharing data in cloud should be presented. In all of this process, the security technologies should be considered as essential.

2.3. Mobile Cloud Device resource Searching

Mobile cloud has a big advantage in the collaboration between the mobile devices unlike existing clouds. Users have a large number of mobile devices instead of just one. This trend is expected to be increasing. In addition, communication mobile devices as well as the consumer serviced devices also have function of network. So, collaboration between several devices becomes much more expected using cloud. This is a technique for using mobile cloud and generating a small mobile cloud. It will be necessary that the clear definition and research for system about generating and maintaining of personal cloud via various devices. Also, studies for domain navigation of mobile cloud are worth exploring to represent N-screen and the fusion media. Likewise, resource discovery for cloud/device and device/device are issued.

3. Mobile Cloud technology research

Table 1. The essential technology for mobile cloud service[9]

<i>Technology(Tx)</i>	<i>contents</i>
T1	Storage space & management
T2	Synchronization
T3	Mobile Office Infrastructure
T4	Hardware Solution
T5	Web-based Application
T6	Security
T7	Connection-holding
T8	Devices Collaboration

3.1. Storage / management

When users want to save their data in a serviced cloud server, they should be able to access easily to the cloud with application or web serviced- platform. Naturally, the cloud should keep, sort and update with maintaining the integrity.

3.2 Synchronization

In the cloud, users can use their data or contents on multiple smart devices or with other people who can access the same cloud with devices. In addition, since the cloud is owned by a number of

Abnormal Behavior Tracking Model Based on Sensors of Mobile User

Yong-Ik Yoon¹ and Syed Usman Asghar²

¹Department of Multimedia Science, Sookmyung Women's University, Seoul, Korea

E-mail: yiyeon@sm.ac.kr

²Department of Information and Communication Technology, Helsinki Metropolia
University of Applied Sciences, Helsinki, Finland

E-mail: syed.asghar@metropolia.fi

Abstract

The recent advancement in mobile fabrication technology allows adding large scale of sensor networks which can be the source of pervasive monitoring and tracking. One of the most important areas where the advantages of sensor networks can be exploited is for tracking mobile targets. The smartphone world is expanding at a rapid pace since the technology was first introduced. This paper suggests a tracking system for mobile users based on sensors in smart device, where tracking information can be sent from mobile to server so that security approaches can work more specifically and accurately by tracking and recognizing abnormal situations of objects, like mobile users. A safety insurance approach is proposed, in which a user's current situation can be estimated by information from mobile sensors if he/she is in a potential threat. In case of emergency, location, nature of problem and possible difficulties can be determined in comparatively less time by authorities as they have already monitoring the situation.

Keywords: Mobile sensors, sensor network, smartphones, mobile tracking, observation, recognizing objects.

1. Introduction

Human activity recognizing based on sensors information is one of the most promising research topics for a variety of areas, including pervasive and mobile computing, surveillance-based security, and context-aware computing. In particular, the use of wearable sensors allows an uninterrupted flow of information from the users [1]. The use of mobile phone has become an essential part of our lives. Statistics show that 80% of the world's population now has mobile phone. There are more than 5 billion mobile phones in the world, out of which more than 1.08 billion smartphone are users [2]. The increasing use of affordable smartphones enabled with sensors, such as accelerometers and some others including gyroscopes, magnetometer and biosensor, have pushed the research focuses of the context processing and activity recognition and interference.

There are many mobile applications which are being used for tracking purposes. As technology has been touching to new horizons by every passing day so many ideas are also popping up to the surface for betterment of mankind. Object tracking, in particular, is essential for retrieving the trajectory of a moving target in a sensor network. So here, an idea is being proposed which may help the user to secure his security when being in dangerous situation. By using this application, situation and location of a user can be get. The best use of this application might be in that case when a user feels danger around or not sure about the security of surroundings.

In proposed application, tracking can be characterized into three basic steps. These steps include (1) to collect, store and process the information which is received from the appropriate sensors according to environments and object's situation, (2) to create a computational model to perform reasoning and manipulating on processed information, and (3) to develop or select an algorithm which can estimate about the object behavior and predict next move.

2. Related Work

With the widespread market penetration of smartphones one of the major applications is their use for personal navigation. A smartphone holds numerous extra features, like inertial sensors, compared to that of a mobile cell phone. Inertial sensors are used measures and reports on a device's velocity, orientation, and gravitational forces, using a combination of accelerometers and gyroscopes, sometimes also magnetometers. Zero-velocity updates (ZUPT) and zero angular rate updates (ZARU) are further improvements in inertial sensor technology. [3] [4]

The inertial sensors employed in today's smartphones are typically low cost sensors mainly intended for games and enhanced user experience. Promising results in the field of smartphone navigation are obtained by step counting and activity classification algorithms [5]. As a complementary system, GPS and map matching is often used to further improve the localization accuracy [5].

3. System Overview

The key component of the proposed positioning system is an Android smartphone, which hosts the applications for navigation and sensor fusion. It has multiple hardware components of which we mainly use the GPS receiver, the gyroscope and the accelerometer.

3.1. Global Positioning System

The built-in GPS receiver delivers position information in the NMEA 0183 string format. GPS information is the main source of positioning information. In case of GPS signal loss, the last GPS position information available before the dropout is used to initialize the IMU-based positioning.

3.2. Gyroscope

The gyroscope delivers information about the angular velocities of the smartphone in three orthogonal directions. However, the measurements of the inexpensive smartphone sensor suffer from bias errors and sensor data drifts such that a post filtering is required.

3.3. Accelerometer

The accelerometer is a built-in electronic component that measures tilt and motion. It is also capable of detecting rotation and motion gestures such as swinging or shaking. The main function of this is to adapt the orientation change when the position is changed from vertical to horizontal and vice-versa.

4. Model

We propose abnormal behavior tracking by means of computation of acquired data from smartphone's sensors. The main idea of the application is described in figure 1. As it is shown in figure 1 that application working principle can be divided into four parts; *Data Capturing*, *Computational Model*, *Estimation Model* and *Main Algorithm*.

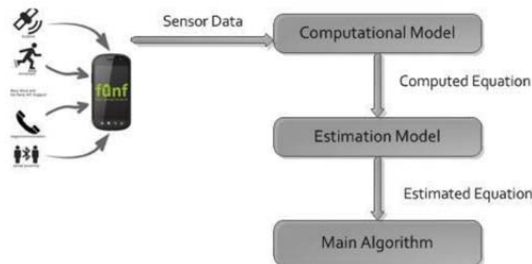


Figure.1 Block diagram of application model

모바일 환경에서 감성을 기반으로 한 영상 합성 기법 연구 및 개발

심 승 민 *, 이 지 연* 윤 용 익*

Research and Development of Image Synthesis Model Based on Emotion for the Mobile Environment

SeungMin Sim*, JiYeon Lee*, YongIk Yoon*

요 약

최근 스마트폰 카메라의 성능은 디지털 카메라 못지않게 발전하고 있다. 이에 따라 점점 더 많은 사람들이 사진 촬영을 하게 되고, 사진 관련 어플리케이션에 대한 관심은 꾸준히 증가하고 있다. 하지만 현재 나와 있는 사진 합성 프로그램은 여러 장의 사진을 배치하는 방법, 기존 이미지를 플러그인 이미지에 포개는 방법 등으로 단순한 합성 프로그램에 머물러 있는 실정이다. 본 논문에서 제안하는 모델은 얼굴 표정에서 추출한 감정을 기반으로 이에 맞는 배경을 합성하고 효과 필터를 적용하여 기존 사진 합성 프로그램보다 다양한 분야에서 활용할 수 있는 확장성을 가진 기법을 제시하였다.

▶ Keywords : 사진 합성, 감정 추출, 큐레이션, 퓨전 미디어 콘텐츠

Abstract

Camera performance of smartphone recently has been developed as much as the digital camera. Interest in applications As a result, many people take pictures and the number of people who are interested in application according to photos has been steadily increasing. However, there are only synthesis programs which are arraying some photos, overlapping multiple images. The model proposed in this paper, base on the emotion that is extracted from the facial expressions by combining the background and applying effects filters. And it can be also utilized in various fields more than any other synthesis programs.

▶ Keywords : Image Synthesis, Extract Emotion, Curation, Fusion Media Contents

•제1저자 : 심승민 •교신저자 : 윤용익

•투고일 : 2013. 6. 3, 심사일 : 2013. 8. 9, 게재확정일 : 2013. 9. 27.

* 숙명여자대학교 멀티미디어과학과(Dept. of Multimedia Science, Sookmyung Women's University)

※ 본 연구는 숙명여자대학교 2013학년도 교내연구비 지원에 의해 수행되었음.

I. 서 론

인간은 이성적인 동물이자 감성적인 동물이지만, 그동안 인류는 역사를 통해 이성적인 측면만을 지나치게 강조해 왔다. 그러나, 감성은 인간의 본성으로 이성 못지않게 중요한 속성이다.

각종 정보를 디지털로 기호화하고 분류하면서 사람들은 빠르고 간편한 속성에 매료되었고, 이는 높은 경제적 가치를 창출하였다. 하지만, 디지털 기술의 발전만으로는 미래를 지배할 수 없음을 인식하게 되었고, 디지털 기반과 아날로그 정서가 융합하는 첨단기술을 의미하는 디지로그라는 개념이 나오게 되었다. 그 결과 디지털 기술의 제품이나 서비스를 아날로그로 보완하였다. 새로운 '틈새'의 영역을 장악하였고 사회, 문화, 산업 전반의 흐름을 형성하는데 있어 큰 영향을 주고 있다.

이제 시장에서도 디지털이 제 역할을 다하기 위해서는 아날로그가 풍부하게 포함되어야 한다는 인식이 확산되고 있다. 가장 좋은 디지털이란 감성적이고 따뜻하며 인간적인 것이어야 한다는 인식이 중요시되고 있는 것이다. 이러한 시대적 흐름을 바탕으로 최근 감성을 담은 스마트 기술이 신 성장 동력으로 떠오르고 있음을 그림 1[1]에서 확인할 수 있다.

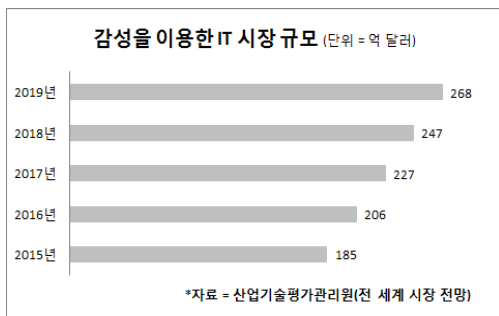


그림 1. 감성을 이용한 IT 시장 규모
Fig. 1. IT Market Size Using Emotion

이에 본 연구에서는 영상 속 인물의 감정을 파악하여 자동으로 배경과 합성하고 효과를 적용하는 모델을 제안한다. 단순한 영상 합성이 아닌 사용자 개개인의 상황에 맞는 맞춤형 서비스를 제공하는 큐레이션(Curation)과 퓨전 미디어(Fusion Media) 개념이 적용된다.

II. 관련 연구

1. 표본 기반 영상 분리

영상을 이루는 픽셀들은 객체와 배경의 영역으로 나눌 수 있다. 다양한 알고리즘을 사용하여 영상에서 객체를 추출할 수 있는데, Bayesian Matting, Knockout, Grab Cut은 픽셀의 값을 이용한 표본 기반 영상 분리 방법이다.

Bayesian Matting 알고리즘은 픽셀의 표본 데이터 분포를 Gaussian Mixture Model(GMM)으로 모델링한 정보를 이용한다. Bayesian Matting 알고리즘은 사용자가 입력한 초기점이 많으면 많을수록 정교한 결과를 얻을 수 있다. 하지만, 매번 픽셀의 투영도를 계산하기 위해 표본을 선정하고 GMM을 생성해야 하므로 연산 시간이 오래 걸린다. 또한 초기점의 정보가 부족하거나 객체와 배경이 복잡한 영상일 경우 잡음이 많이 발생하여 정밀도가 급격히 떨어지는 단점이 있다.

Knockout 알고리즘은 찾아낸 객체 픽셀과 배경 픽셀을 Red, Green, Blue의 각 축으로 사용해 혼합 비율을 활용한다. 먼저 확실한 객체 부분, 배경 영역, 그리고 정교한 분리가 필요한 불확실한 부분으로 나눈다. 이후 불확실한 픽셀 주변에서 가장 가까운 객체와 배경 색상들의 가중치로 계산된 결과로 객체인지 배경인지 결정된다. 지정되지 않은 픽셀 주변에서 거리에 따른 가중 합으로 표본을 선택하기 때문에 표본의 신뢰도에 따라 정밀도가 떨어질 수 있다.

Grab Cut 알고리즘은 영상 데이터에서 객체의 영역을 Grab(테두리로 표시)한다. 그 후 객체가 아닌 배경 영역의 데이터 표본 분포를 분석하여 배경에 대한 Gaussian Mixture Matrix(GMM)을 생성한다. 생성된 GMM을 이용하여 객체 영역에서 배경 영역이라 추정되는 픽셀을 제거해 나가는 방식으로 객체를 추출한다. Grab Cut 알고리즘은 사용자가 객체와 배경의 두 번의 초기점을 정할 필요 없이 객체에 대한 단순한 영역을 지정함으로써 영상 분리가 가능한 장점이 있다. 그러나 입력 영상에서 객체가 차지하는 영역이 많아 배경의 영역이 부족할 경우 배경의 GMM이 제대로 생성되지 않아 전체 성능이 떨어지는 문제점이 있다. 또한 반복에 의해 Grab한 영역의 테두리에서부터 한 픽셀씩 연산한다. 따라서 최대 영역의 테두리에서 객체 영역까지 픽셀 개수의 차이만큼 연산을 반복해야 하는 단점이 있다[2].

Elastic Service Model for Smart Learning based on Cloud Environment

Svetlana Kim

Department of Multimedia Science
Sookmyung Women's University
Seoul, Korea
xatyna@nate.com

YongIk Yoon

Department of Multimedia Science
Sookmyung Women's University
Seoul, Korea
yiyoona@sm.ac.kr

Abstract— One of the interesting and realistic application areas for context-aware and cloud computing technologies are the learning systems. In order to add those technologies to the e-learning services that already exist, there is a need for a service architecture model that transforms the already existing situation-aware e-learning environment to an environment that understands context also. We propose using elastic computing in cloud computing for a new notion in smart learning. Our suggestion is the elastic conductor which including the four smarts concept to the cloud services in order to make smart learning services possible. The benefits of smart computing for e-learning solution, which use the handle intelligently to meet the users' needs, are the focus of this paper. This is achieved by going through the collecting user's behavior, prospecting, building, delivering and rendering steps. The smart-learning model that is proposed in this paper will show that it is possible to make customized and personalized learning services in different fields.

Keywords—elastic computing, e-learning, context-awareness, synchronization, fusion content, cloud environment

I. INTRODUCTION

The smart learning (s-learning) has become an important way of learning during the last few years. It has been made possible by the recent advancements in the Mobile Internet and Information technologies. The S-learning has a major role in creating a good and personalized learning environment, and also being well adapted to the current education model wherever possible [1][2][3]. Usually, the teaching and learning that e-learning offers is only inside of a lecture-style classroom with desktop computers. Though the students could download information and browse through the existing e-learning platform regardless of time and place, they were still confined to the limits of the wired classroom-setups.

The concept of s-learning plays an important role in the creation of an efficient learning environment that offers personalized contents and easy adaptation to current education model. It also provides learners with a convenient communication environment and rich resources. However, the existing-learning infrastructure is still not complete. For example, it does not allocate necessary computing resources for s-learning system dynamically [4][5]. Currently, the majority of s-learning systems have difficulty in interfacing and sharing data with other systems, i.e., it falls short of systematic arrangement, digestion and absorption of the

learning contents in other systems. This may lead to duplication in creating teaching resources and low utilization of existing resources. To resolve this problem, it is recommended to use cloud computing to support resource management.

The cloud computing environment provides the necessary foundation for the integration of platform and technology. It integrates teaching and research resources distributed over various locations by utilizing existing conditions as much as possible to meet the demands of the teaching and research activities. The cloud computing environment with respect to s-learning offers new ideas and solutions in achieving interoperability among heterogeneous resources and systems. The cloud services mean that the Internet can be used as huge workspace, repository, platform, and infrastructure.

Learners can access to the Internet from anywhere at any time, using widely spread mobile devices but the existing cloud computing technologies are only passively responsive to users' needs. This situation necessitates proactive cloud services rather than passive services. Since learners typically carry mobile devices of some kind at their hands, the volume of information and services processed through the devices continues to increase.

In this paper, we propose a elastic learning conductor for smart learning contents through the four smarts concept based on the user behavior acquired by the sensors in the user' mobile device. The elastic conductor is making a collection all information of user. The information of user is included the information about the user action, device and total user's environment received by sensing data. Sensing data coming from various sensors mounted on a smart phone, because the users always is have smart phone in their general life time. This application is running data within the smart phone and it is good information for the appropriate situation awareness. In addition, the web browsing is access information identify the main tasks of the users, preference and MAC-address connections are useful in identifying the location. At this point, the correct recognition of the obtained user's situation it is very important.

II. RELATED WORK

Advances in mobile and wireless technologies have changed the way learners' access and share resources, acquire knowledge, and collaborate with each other. Such technologies may include various mobile devices such as

hand-held computers and smart phones, embedded sensors in those devices, high-speed wireless networking technologies such as 4G networks that allow those heterogeneous devices to interconnect together. The platform advancements enable context awareness in a smart cloud computing environment and smart services for innovative learning processes. The rest of this section discusses a cloud computing project that uses cloud-based applications in s-learning and an overview of context-awareness in a learning environment.

The cloud computing is defined as a technology that provides its users with IT resources by using the Internet as a medium. The users can use IT resources such as application software or storage space from the cloud without needing to own them. The users only need to pay per usage charges for the resources they used. The concept of cloud computing is not new. It is the combination of distributed computing, grid computing, utility computing, *etc.* [6][7][8].

There are a number of cloud-based applications available in the e-learning sector as well. Casquero *et al.* [9] presented a framework based on iGoogle and using the Google Apps infrastructure for the development of a network of cooperative personal learning environments. They discussed the integration of institutional and external services in order to provide customized support to faculty members in their daily activities. They also take advantage of the framework as a test-bed for the research, implementation and testing of their educational purpose services. Marcnzi *et al.* [10] investigated how educational software can be used in an academic or corporate learning environment. They integrated models and tools that they developed into an open source environment for the creation, storage and exchange of learning objects as well as learning experiences. They presented the “LearnWeb 2.0” infrastructure to support lifelong learning and to enhance the learning experience. This infrastructure brings together information stored on institutional servers, centralized repositories, learners’ desktops, and online community—sharing systems like Flickr and YouTube. Sedayao [11] proposed an online virtual computing lab that offers virtual computers equipped with numerous applications such as Matlab, Maple, SAS, and many others that can be remotely accessed from the Internet.

III. ELASTIC LEARNING SYSTEM

The cloud computing enable cloud servers to provide smart learning service to users through additional intelligent processes on existing cloud system. The most important responsibility of the servers is to perform elastic processes such as Elastic Computing for Infrastructure as a Service (IaaS), Elastic Management for Platform as a Service (PaaS) and Elastic Deployment for Software as a Service SaaS) constantly. The elastic processing can be described as collecting user information that is pulled by the sensors in the users’ mobile devices and process the pulled information in real-time so that it can accommodate users’ changing situation dynamically[12].

To satisfy the requirements of IaaS, such as storage limitation and processing power, the server must be equipped with virtualization capabilities for mass storage and

infrastructure in the cloud by incorporating elastic computing, which means that the elastic computing supports dynamic adaptation for the needs of the computing environment. This can be accomplished by a decision policy that can handle users’ changing situations.

In this paper the Elastic Conductor performs provisioning and scheduling for the decision of smart activities. The provisioning and scheduling are performed through an inference engine that utilizes the rules based on three attributes—an object id for user context, a predicate relationship for user behavior, and a value for thinking. The elastic conductor is utilized in PaaS as a smart activity.

The Elastic Conductor it can also generate user interface configurations, such as personalized views and content rendering for each individual user. These processing tasks include four smart concepts: Behavior sensing, Behavior matching, Synchronization and Push – for displays multi-contents on the multi-device or all content on the one device to render learning content. Using sensing context-awareness through the location and IP address of each device, it can orchestrate all devices. Therefore, the multi learning contents can be played in different devices separately to form a “virtual class”. The architecture of the model is show in the “Fig.1”.

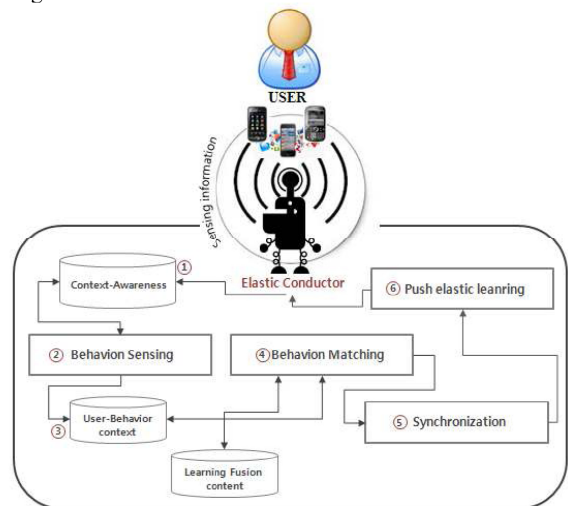


Figure 1. Elastic conductor architecture

The “Fig. 1” shows how the conductor provides smart learning to the user. The information of the user included the information about the user and the device, received by context-awareness sensor. Context-awareness monitoring monitors user request(s) and the kind of device(s) that the user is currently using. By using the information collected by the sensors, conductor (1) pulls the sensing information and analyses the extractable contents. The behavior sensing (2) concept is functioning as an information filter that extracts only the intended information from sensing data and (3) storage in the user-behavior context DB. There can be multiple contexts in sensing data depends on the services available in the learning management system. The behavior

A Study on CCTV-Based Dangerous Behavior Monitoring System

Young-Bin Shim¹, Hwa-Jin Park¹, Yong-Ik Yoon¹

*¹ Dept. of Multimedia, Sookmyung Women's University,
Seoul, Korea
{ rindasim, phj2000,yiyoon}@sookmyung.ac.kr

Abstract. Proposed in this paper is a detection system that, by determining whether a dangerous act is being carried out in the images captured using CCTV, provides pre-warnings and establishes emergency measures. To determine the presence of a dangerous act, after setting zones of interest and danger zones within those zones of interest, the danger is detected when an object trespasses the danger zone. If a situation is deemed as warning-worthy and emergency, the integrated control center is immediately alerted to facilitate prevention in advance.

Keywords: CCTV, blob segmentation, object detection, determination of danger level

1 Introduction

Recently, due to a sharp increase in sexual offenses that threaten social safety or due to suicide incidents caused by stresses such as anxiety and depression, there has been a rapid rise in interest in social safety systems. In particular, since most of the cases occur in deserted areas or during times of darkness, relatively low-cost CCTVs have been installed and used, instead of manning via personnel, for monitoring and preparation. However, even though these precautions have played an important part in post-incident investigations, unless someone continues to monitor in real-time CCTV systems do not help to prevent incidents at all. As such, this study is purposed in studying a danger detection system that is not intended for providing post-incident reference but for providing pre-incidence prevention; to achieve such, the system would examine the danger level by detecting a person's dangerous act and, in dangerous cases, it would alert the emergency situation with beeping or via the system.

Situations and criteria are very diverse and wide for detecting all dangerous acts and the technical difficulty of image analysis varies depending on the number of

* This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2013S1A5B6043996)

target persons involved; therefore, in this study, specific dangerous acts were pre-classified according to the number of target persons.

Table 1. Dangerous acts in accordance with the number of target persons.

Number of Main Target Persons	Examples of Dangerous Acts
1 Person	Suicide, Self-infliction, Smoking in non-smoking area, Unlawful entry
2 Persons	Fighting, Making threats, Sexual attack, Assault, Stalking, Assault on children
3 Persons or more	Group assault, etc.

2 Dangerous Action Detection System

2.1 Overall Process

The system this study is attempting to implement is a CCTV-based system for detecting bridge suicide acts and it has the following characteristics. It is always in a fixed state as long as the position and angle are not changed. Utilizing this characteristic, the overall process of this system can be summarized as follows. First, on the bridge CCTV footage the area of interest for detection is set and cut, and within the area of interest, danger zones are selected via the coordinates. The overall process is composed of image segmentation, object detection, and danger detection. **Fig 1.** shows a flow diagram of object and danger detection system.

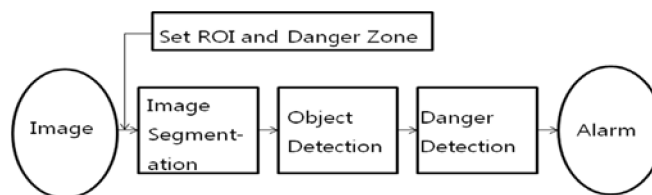


Fig. 1. Object detection and danger determination system flow

2.2 Image Segmentation Unit

The background image in a CCTV-based bridge suicide detection system is always constant once the CCTV has been installed. The background image is extracted from the rows of sequential images. We set ROI (region of interest) in order to exclude

다중 객체의 위험 행동 감시 시스템 연구

심영빈*, 박화진**

요 약

CCTV를 이용하여 획득한 영상 내에서 다중 객체의 위험한 행위를 판단하여 사전에 미리 경고 및 긴급대책을 세워주는 감시 시스템을 제안한다. 위험한 행위의 판단여부를 위해 관심지역 및 관심지역 내에 위험지역을 설정한 후, 위험 행동 객체를 검출하여 객체의 위험지역 침범 범위에 따라 안전, 경고, 긴급 등의 위험도를 판단한다. 특히 본 연구는 위험 행동 중 교량에서 투신하는 행위를 감지하는 것을 목표로 하며 기존의 연구에서 단일객체의 행동검출에만 제한했던 연구를 여러 보행자 속에서 투신 행동 하는 객체를 감지하는 것까지 확대하여 구현한다. 한 객체의 위험지역 침범의 정도에 따라 안전, 경고 및 긴급 상태로 분류하고 상황에 따라 긴급 상태로 판단되면 통합관계 센터에 즉시 알려 위험행위를 사전에 예방 할 수 있도록 한다.

키워드 : 객체검출, 객체 구성 분석, 블롭 분할, 위험도 판단

A Study on the Surveillance System of Multiple Object's Dangerous Behaviors

Young-Bin Shim*, Hwa-Jin Park **

Abstract

This paper proposes a detection system that, by determining whether a dangerous act is being carried out among other pedestrians in the images captured using CCTV, provides pre-warnings and establishes emergency measures. To determine the presence of a dangerous act, after setting zones of interest and danger zones within those zones of interest, the danger level is determined in accordance with the range of encroachment upon detecting an object. Especially, this research aims at detecting a suicide jump from the bridge and extends to detecting a dangerous act among pedestrians from detecting a dangerous act of only one person with no one in the previous research. This system classifies the status into 3 levels as safe, alert, and danger according to the amount of part being over the bridge railing. If a situation is deemed as warning-worthy and emergency, the integrated control center is immediately alerted to facilitate prevention in advance.

Keywords : analysis of object's composition, blob segmentation, determination of danger level, object detection

1. 서 론

※ 교신저자(Corresponding Author): Hwa-Jin Park
접수일: 2013년 11월 19일, 수정일: 2013년 12월 07일
완료일: 2013년 12월 18일

* 숙명여자대학교 멀티미디어학과
email: rindasim@sm.ac.kr

** 숙명여자대학교 멀티미디어학과
Tel: +82-10-6309-0029, Fax: +82-2-710-9704
phj2000@sm.ac.kr

■ 본 연구는 숙명여자대학교 교내연구비지원에 의해 수행되었음 (과제번호 1 - 1203 - 0141)

최근 들어 사회 안전을 위협하는 성범죄 사고나 부녀자 납치, 강도, 살인부터 불안 우울과 같은 정신적 스트레스로 인해 자살사고 등이 급증하고 있다. 중대한 사회범죄뿐만 아니라 사소한 경범죄 예를 들면 바깥에 주차한 차를 긁는다던가, 쓰레기 투기, 급연 지역에서 흡연 및 불량학생들의 집단 폭력 등보이지 않는 틈을 이용하여 일어날 수 있는 사건 사고들이 급증하고 있어 사회 안전 시스템에 대한 관심이 증폭되고 있다. 특히 대부분의 사건이 인적이 드문 곳이나 어두

운 시간을 이용하여 발생하므로 CCTV를 설치하여 사고 예방 및 진상조사에 효율적으로 대비하여 왔다. 고비용의 인력배치보다 저렴하고 정황 파악 및 증거물로 유용한 부분이 많지만 누군가 계속 실시간으로 모니터링을 하지 않는다면 사건 예방에는 전혀 도움이 될 수 없는 단점이 존재한다. 따라서 본 연구는 사후 참조용뿐만 아니라 사전에 사건 발생 억제를 위한 목적으로 영상안의 행인의 위험한 행동을 감지하여 위험도를 검사하고 위험한 경우 경고음이나 시스템을 통하여 긴급사항을 알려주는 위험 감지 시스템을 연구하고자 한다.

위험한 행동을 전부 감지하기에는 상황 및 범위가 매우 넓고 영상분석 시 대상인원수에 따라 기술의 난이도가 다르기 때문에 본 연구에서는 특정한 위험 행동을 주요 대상 인원수 별로 미리 분류하였다.

<표 1> 주요대상인원수 별 위험행동.

Number of Main Target Persons	Examples of dangerous acts
1 Person	Suicide, Self-infliction, Smoking in non-smoking area, Unlawful entry
2 Persons	Fighting, Making threats, Sexual attack, Assault, Stalking, Assault on children
3 Persons or more	Group assault, etc.

<Table 1> Dangerous acts in accordance with the number of target persons.

본 연구는 다중 객체의 위험행동 감지 연구로서, 주요대상은 사람이 행할 수 있는 위험한 행동 중 교량위에서의 투신행위를 감지하는 지능형 감시 시스템을 연구한다. 실제적으로 서울시 소방 재난 본부로 부터의 자료에 따르면 최근 5년간 한강에서 하루 평균 약 0.5명이 투신했고, 0.23명이 사망 하였다고 한다. 따라서 본 위험행동 감지 시스템의 필요성과 맞물려 시기적으로 적절한 연구라고 할 수 있다. 특히 기존 연구

인 단일객체의 투신행위 감지연구는 매우 제한적 상황 즉 단순히 한 객체만 있었던 상황에서 감지하는 것을 확장하여 보행자가 많은 경우에도 위험한 행동을 하는 객체를 인지할 수 있도록 하였다. 정리하면, 본 연구는 다수의 행인들 속에서의 한 객체가 (혹은 두 객체 이상) 교량에서 투신하는 것을 감지하여, 상태는 3단계로 나누어 안전, 경고, 긴급으로 구분하여 위험행동을 감지하여 예방해 주는 시스템을 구현하였다. 특히 만약 보행자가 자살 의도 없이 무의식적으로 교량에서 손이나 몸을 약간 내미는 등의 경우에는 무조건 경고음을 내는 것이 아니고 위험한 지 여부에 따라 경고음 및 긴급성을 차별하는 등의 조치로 보다 정확한 대처를 할 수 있도록 하였다.

본 연구의 구성은 다음과 같다. 2장은 객체 검출 및 추적 관련 기술과 행동 감지 모니터링 시스템 사례연구, 3장은 보행자들 중 단일 객체의 위험행동 감지 시스템, 4장은 실험결과 및 분석, 5장은 결론 및 향후 과제를 서술 한다.

2. 관련연구

영상인식을 위해 2.1절에서는 객체 검출 및 추적에 관한 기존의 기술 연구들을 조사하였고, 2.2절에서는 영상 처리 분야의 적용사례로 지능적 영상처리를 하여 정보를 제공하는 지능형 감시시스템의 현황에 대해 조사하였다.

2.1 객체검출 및 추적에 관한 기술

객체의 검출로 많이 사용하는 방법은 연산이 간단한 배경차분법과 장면차분법이 있다. 차분법이란 현재 프레임과 미리 생성된 배경 모델과의 차이를 기반으로 이동 객체를 검출하는 방법이다. 검출된 객체를 추적하기 위한 방법에는 검출에서와 마찬가지로 차영상을 이용하는 방법과 배경영상을 이용한 방법이 있고 그 외 블록정합, 특징점을 이용한 방법들이 있다. 배경영상을 이용한 방법 중 배경프레임을 추출하는 방법에 따라 시간적평활법과 시간적 중간치법으로 나누어진다. 시간적평활법은 이전 프레임들의 화소 값을 평균하는 방법을 사용하는 것이고 시간적 중간치법은 화소값을 크기순으로 정렬하여 빈도가

객체 추적을 통한 이상 행동 감시 시스템 연구

박화진*

요 약

사회의 범죄율 증가와 더불어 지능형 보안 시스템강화에 대한 관심이 높아지고 있다. 이에 본 연구에서는 CCTV에 획득되는 영상으로부터 객체의 이상 행동을 감지하는 시스템을 제안한다. 배경영상과의 차연산 및 모폴로지를 통해 객체를 검출하고 객체의 특징 정보를 이용해 각각의 객체를 인식하여 추적하여 이를 통해 이상행동을 탐지한다. 객체가 영상 내에서 일정시간 이상을 배회했을 때 이를 이상행동으로 판단하여 사전에 관제센터에 알려 미연에 방지할 수 있도록 한다. 특히 본 연구는 이상 행동 중 객체의 배회행위를 감지하는 것을 목표로 하며 영상 내에서 사라진 객체가 다시 영상 내로 들어 왔을 때의 이전 객체와의 동일여부를 판단할 수 있도록 하였다.

키워드 : 객체검출, 객체추적, 배회, 이상행동

A Study on Monitoring System for an Abnormal Behaviors by Object's Tracking

Hwa-Jin Park*

Abstract

With the increase of social crime rate, the interest on the intelligent security system is also growing. This paper proposes a detection system of monitoring whether abnormal behavior is being carried in the images captured using CCTV. After detection of an object via subtraction from background image and morphology, this system extracts an abnormal behavior by each object's feature information and its trajectory. When an object is loitering for a while in CCTV images, this system considers the loitering as an abnormal behavior and sends the alarm signal to the control center to facilitate prevention in advance. Especially, this research aims at detecting a loitering act among various abnormal behaviors and also extends to the detection whether an incoming object is identical to one of inactive objects out of image.

Keywords : abnormal behavior, loitering, object detection, trace of object,

1. 서론

※ 교신저자(Corresponding Author): Hwa-Jin Park
접수일:2013년 12월 08일, 수정일:2013년 12월 21일
완료일:2013년 12월 27일

* 숙명여자대학교 멀티미디어학과

Tel: +82-2-710-9204, Fax: +82-2-710-9704

email: phj2000@sm.ac.kr

■ 본 연구는 숙명여자대학교 교내연구비지원에 의해 수행되었음 (과제번호 1 - 1203 - 0142)

최근 범죄예방과 사회 안전을 위해 폐쇄회로 TV (혹은 CCTV-closed circuit television) 설치 및 사용이 급증하고 있다. 개인의 자산보호용 설치 뿐 아니라 각 지자체에서는 CCTV 통합관제 센터를 설립하여 기존의 방법, 교통 및 주차단속, 쓰레기 무단투기 등 다양한 목적으로 운영되던 CCTV를 하나로 통합, 연동해 관리함으로써 효율성을 높이고 있다. 이로써 각종 범죄예방과 치안유지, 생활안전 등의 대처를 통합적으로 관리함과 동시에 인력 및 비용의 절감효과를 거두고 있다. <표 1>은 2012년도까지의 CCTV 통합관

제센터를 구축한 전국의 시군구 수와 사용된 예산을 보여주고 있다.

<표 1> CCTV 통합관제센터를 구축한 전국의 시군구 수와 사용된 예산.

Classification	total	2010	2011	2012
city, gun-district, gu-borough (quantity)	88	27	34	27
government expense (one hundred million won)	426	79	204	143

<Table 1> The number of municipalities across the country where CCTV integrated control center has built and their budget

<표 2>는 각 지자체 별 CCTV 수 및 관제인원을 보여주고 있다. 하지만 <표 2>에 나타나 있듯이 평균 1,000여대에 달하는 CCTV를 관제하는 요원의 수가 20여명에 불과해 이를 제대로 관제하기는 어려운 실정이다. 사실상, CCTV가 사건 후에 참고자료로 중요한 역할을 하는 것은 사실이지만 사전에 사건을 예방할 수 있는 역할을 기대하기는 역부족이다.

<표 2> 각 지자체별 CCTV수 및 관제인원

Classification	Gwangmyoung-city	Gunpo-city	Suweon-city	Ansung-city
CCTV (quantity)	1527	705	1,131	574
control personnel (persons)	33	13	36	18

<Table 2> The number of CCTV and traffic control personnel per each municipality

따라서 CCTV 모니터를 육안으로 관제하는 한계를 극복하기 위해 지능형 CCTV 서비스를 연구 개발하는것이 요즘의 추세이다. 시범적으로 일부기능을 개발하여 지능형 CCTV 서비스라는 이름으로 운용되고 있으나 개발되어 적용되는

기술은 아직 초보적인 단계이다. 이상행동으로 인식되는 상황과 인지해야 할 사람들의 형태, 행동 양식 등에 따라 다양한 이상행동이 인식되어 질 수 있어야 한다. 특히 CCTV영상을 실시간으로 분석해 특정인의 얼굴 및 표정 등을 인식해야 하는데 현재의 기술은 지나가는 객체가 사람인지 자동차인지를 구별하는 정도, 자동차 번호 인식등의 초보적인 지능형 기술들로 구성 되어 있다. 특히 현재의 기술은 고정되어있는 한 개의 CCTV의 출력영상에만 의지한다는 한계가 있다.

예를 들면 한 집 앞이나 골목에서 한 사람이 일정시간 이상 서성거리면 이상한 사람이 있다는 것을 알려주는 감시 시스템의 경우이다. 서성이다 보면 한 구역만을 비춰주는 CCTV 영상에서 벗어날 확률이 매우 많은데도 불구하고 기존 시스템의 경우 한 특정인이 CCTV영상에서 사라졌다가 다시 재등장 할 때도 다시 새로운 인물로 인식하게 되는 문제점을 가지고 있었다.

그리하여 본 연구는 이러한 지능형 관제 기술 중 재등장하는 인물을 포함하여 특정인의 행동 인식에서 이상행동인 배회에 대하여 감시하는 시스템을 구현하고자 한다. 특히 배회 시 CCTV 촬영 가능 영역 외에서의 배회까지 인식한다는 점에서 기존의 방법과는 다르다고 할 수 있다.

즉 본 연구는 객체추적을 통해 객체가 영상 내와 영상 외에서 다시 재등장하여 일정시간 이상 배회했을 때 재등장인물로 인지하고, 이를 이상 행동으로 판단하여 알려 사전에 인지할 수 있도록 한다. 또한 이 연구는 향후 CCTV간의 긴밀한 협조시스템의 초석이 될 수 있다.

본 연구의 구성은 다음과 같다. 2장은 객체추적 관련 기술과 지능형 감시 시스템 현황, 3장은 객체의 추적을 통한 이상행동 감지 시스템, 4장은 실험결과 및 분석, 5장은 결론 및 향후 과제를 서술 한다.

2. 관련연구

영상인식을 위해 2.1절에서는 객체 추적에 관한 기존의 기술 연구들을 조사하였고, 2.2절에서는 영상 처리 분야의 적용사례로 지능적 영상처리를 하여 정보를 제공하는 지능형 감시 시스템의 현황에 대해 조사하였다.

객체 애니메이션 제작을 위한 스케치 기반의 사용자 인터페이스 기법

(Sketch-based User Interfaces for Authoring 3D Object Animation)

전 재 웅[†] 장 현 호[†] 손 의 성[†] 임 순 범^{††} 최 윤 철^{†††}
(Jaewoong Jeon) (Hyunho Jang) (Eisung Sohn) (Soon-bum Lim) (Yoon-chul Choy)

요 약 오늘날 컴퓨터 애니메이션은 많은 매체에서 쉽게 볼 수 있는 대중화 된 콘텐츠이다. 하지만 컴퓨터 애니메이션을 제작하는데 사용되는 저작 도구들은 대부분 초보자를 비롯한 일반인이 사용법을 익히는데 오랜 시간이 걸리고 사용하기 어렵다는 단점을 가지고 있다. 이런 문제점을 해결하기 위하여 본 논문에서는 다양한 사용자가 손쉽게 3D 객체 애니메이션을 제작할 수 있는 스케치 기반의 인터페이스를 제안한다. 3차원 객체 애니메이션을 효과적으로 제작하기 위하여 본 논문에서는 객체의 움직임을 수평 운동과 수직 운동이 결합된 주요 움직임과 3D 공간 속의 복잡한 이동을 포함하는 세부 움직임으로 구분하였다. 시스템은 수평 이동과 수직 운동을 제어하는 수평, 수직 운동 관리 시스템과 사용자의 스케치와 객체의 동작을 결합시켜주는 스케칭 모듈을 제공한다. 또한 객체의 주요 움직임뿐만 아니라 객체의 세부 움직임을 제어하기 위하여 교차 동작 윈도우 시스템을 제안한다. 초보자를 포함한 다양한 사용자들이 참여한 사용자 평가에서 제안 연구는 쉽고 빠르게 3D 객체 애니메이션을 제작할 수 있다는 평가를 받았다.

키워드: 스케치기반 사용자 인터페이스, 객체 애니메이션, 사용자 인터페이스

Abstract Computer-generated three-dimensional (3D) animations are now commonplace. Despite their popularity, however, such animations are largely created by experts because most computer animation tools are complex and difficult to learn and use. To produce 3D object animations, the proposed system divides the spatial movements of an object into horizontal, vertical, and detailed movements. The horizontal movement system defines the movement of an object in the ground plane. The vertical movement system controls the movement in a plane perpendicular to the ground plane. In order to control the detailed movement of an object, the proposed system includes a cross-motion window with a camera controller that provides the user with an interactive first-person camera view. Experimental results show that various user groups including novice participants can easily and quickly produce 3D object animations.

Keywords: sketch-based user interface, object animation, 3D user interface

· 이 논문은 서울시 산학연협력사업(PA110072) 지원과 2011년도 및 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(355-2011-1-D00059, 2012R1A6A3A01040446)

† 학생회원 : 연세대학교 컴퓨터과학과
demiblu@gmail.com
kencot@gmail.com
(Corresponding author임)
essohn@gmail.com

†† 종신회원 : 숙명여자대학교 멀티미디어학과 교수
sblim@sookmyung.ac.kr

††† 종신회원 : 연세대학교 컴퓨터과학과 교수
ycchoy@rainbow.yonsei.ac.kr
논문접수 : 2012년 7월 24일
심사완료 : 2013년 2월 4일

Copyright©2013 한국정보과학회 : 개인 목적이거나 교육 목적인 경우, 이 저작물의 전체 또는 일부에 대한 복사본 혹은 디지털 사본의 제작을 허가합니다. 이 때, 사본은 상업적 수단으로 사용할 수 없으며 첫 페이지에 본 문구와 출처를 반드시 명시해야 합니다. 이 외의 목적으로 복제, 배포, 출판, 전송 등 모든 유형의 사용행위를 하는 경우에 대하여는 사전에 허가를 얻고 비용을 지불해야 합니다.

정보과학회논문지 : 소프트웨어 및 응용 제40권 제4호(2013.4)

1. 서 론

애니메이션은 영화와는 달리 가상의 주제를 살아있는 것처럼 동작시킴으로써 주제를 표현하는 매체이다. 오늘날 컴퓨터 그래픽스와 하드웨어의 발달로 애니메이션은 영화와 TV, 컴퓨터 게임 및 멀티미디어 콘텐츠 등 다양한 분야에서 중요한 역할을 담당하고 있다. 그러나 현재 3D 애니메이션은 일반적으로 전문적인 도구를 사용해야만 제작할 수 있기 때문에 아무리 간단한 애니메이션의 저작이라 할지라도 초보자를 비롯한 일반 사용자가 쉽게 참여할 수 없는 상황이다.

일반적으로 3D 객체 애니메이션 역시 3ds Max[1]나 Maya[2], Blender[3]와 같은 전문 제작 도구를 사용하여 제작된다. 3D 객체 애니메이션 저작의 경우 3D 캐릭터 애니메이션의 제작보다 상대적으로 난이도가 낮은 편이나, 간단한 경로의 객체 이동뿐만 아니라 복잡한 경로를 포함하는 객체 애니메이션을 제작할 경우 시간이 오래 걸리고 기본적으로 전문적인 도구의 사용법을 익혀야 한다는 근본적인 문제점을 안고 있다. 또한 그동안 3D 객체 애니메이션 제작에 관한 많은 연구가 이뤄져 왔지만 캐릭터 애니메이션 분야에 비해서 상대적으로 부족한 편이며, 3D 입력을 위해 일반적 사용자가 갖추기에 쉽지 않은 전문적인 기구를 필요로 하거나 복잡한 입력 방식 등으로 인해 초보자가 3차원에서 자유로운 객체의 움직임을 입력하는 방식을 제시하기에는 부족한 것이 현실이다.

따라서 본 연구에서는 이를 보완하기 위하여 사용자가 일반적으로 접하는 마우스와 태블릿 PC의 스타일러스 또는 터치와 같은 입력 도구를 사용하여 쉽고 편리하게 3D 객체의 애니메이션을 제작할 수 있는 스케치 기반의 인터페이스를 제시한다. 제시하는 스케치 기반의 3D 객체 애니메이션 시스템은 객체 움직임을 멀티 패스(multi-pass) 방식으로 입력 받아 처리하며, 빠른 시간 안에 복잡한 3차원 공간 상의 움직임을 스케치 방식으로 입력할 수 있다. 또한 사용자의 원활한 스케치 입력을 돕기 위하여 각 스케치 인터페이스에 맞는 카메라 모듈을 제공한다.

2. 관련연구

본 연구와 관련된 3D 객체 애니메이션 및 인터페이스에 대한 연구를 살펴보면 현재까지 3D 객체 애니메이션을 제작하기 위한 다양한 방식의 연구들이 진행되어 왔다. 초기의 연구 중 Balaguer의 연구[4]에서는 3D 스케칭 포인트로부터 3D 애니메이션 패스를 생성하기 위해 데이터 리덕션 알고리즘을 사용하는 시스템을 선보였다. Popovic[5]은 그의 연구에서 강체(rigid body) 시뮬레이

션의 조절을 위한 인터랙티브한 컨트롤 기법을 제시하였다. 이 시스템에서 사용자는 객체의 시퀀스 속에서 특정 물체를 선택하여 객체의 상태를 변경할 수도 있고, 물리적 속성을 변경하여 원하는 객체 애니메이션을 제작할 수도 있다. 또한 Popovic[6]은 강체 시뮬레이션을 컨트롤하기 위해 스케치 기반의 인터페이스를 포함한 시스템도 제시하였다. Terra[7]와 Igarashi[8]는 초보 사용자가 3차원 공간에서 동작을 생성하기 위한 키프레밍 방식의 시스템을 제안하였다. 특히 Terra[7]의 연구에서는 사용자가 제스처를 사용하여 키프레임을 설정하는 방식을 사용하였다. 하지만 본 논문에서는 기존 연구들과 비교하여 객체의 3차원 움직임을 분리하여 멀티 패스로 처리하고, 각 단계에서 스케칭 인터페이스와 카메라 모듈을 제공하여 사용자의 편의를 돕는다는 점과 객체 중심의 시점에서 경로를 스케치 할 수 있는 인터페이스를 제시한다는 점에서 이전의 연구와 크게 다르다고 볼 수 있다.

스케치 기반 인터페이스에 대한 연구는 그 응용분야가 제스처 인터페이스, 인터페이스 디자인, 3D 모델링, 3D 애니메이션의 캐릭터 동작 조절에 이르기까지 매우 다양하다[9-11]. 특히 본 연구와 관련하여 3D 환경에서 선을 입력하는 방법에 대한 연구도 계속적으로 이뤄지고 있다. Xin[12]은 태블릿 PC와 핸드헬드 혼합 현실 기술을 이용한 3D 스케칭 인터페이스를 제시하였다. 이 시스템은 3D 평면에 스케치를 하기 위하여 두 단계로 이뤄진 스케칭 인터페이스를 사용하였다. Fei[13]는 Space canvases 라 불리는 수직형 평면을 사용하여 3D 스케칭을 구현하고 있다. 하지만 [12,13] 연구의 시스템에서 3D 스케칭 방식은 모두 예술적인 선의 입력을 위한 것으로 3D 공간 상의 객체 동작 정보를 입력하기 위한 본 논문의 스케칭 방식과는 목적을 달리하고 있다.

Bowman[14]은 그의 연구에서 몰입형 가상 현실에서 1인치 시점의 동작 컨트롤과 이동을 다루고 있으며 McCrae[15]은 3D 환경에서의 패스와 패스 네트워크를 구성하는 방식을 연구하였다. 본 논문의 시스템에서는 1인치 시점에서 사용자의 스케치를 입력받기 위한 3D 평면 인터페이스를 제시하고 있으며 [15]의 시스템보다 직관적인 방식으로 객체의 수평 움직임을 입력받는 모듈을 사용하고 있다.

3. 3D 객체 애니메이션을 위한 스케치 기반 인터페이스

본 논문에서 제시하는 3D 객체 애니메이션 제작을 위한 스케치 기반의 인터페이스 기술과 시스템은 객체 크기 조절 모듈, 수평 움직임 조절 모듈, 수직 움직임 조절 모듈, 그리고 세부 움직임 제어 모듈로 구성된다. 먼

A sketch-based interface to modify and reproduce motion sequences

Hyunho Jang · Jaewoong Jeon · Eisung Sohn ·
Soon-Bum Lim · Yoon-Chul Choy

© Springer Science+Business Media New York 2013

Abstract We present a sketch-based user interface, which was designed to help novices to create 3D character animations by multi-pass sketching, avoiding the ambiguities usually present in sketch input. Our system also contains sketch-based editing and reproducing tools, which allow paths and motions to be partially updated rather than wholly redrawn; and graphical block interface permits motion sequences to be organized and reconfigured easily. A user evaluation with participants of different skill levels suggest that novices using this sketch interface can produce animations almost as quickly as users who are experienced in 3D animation.

Keywords Sketch-based interface · Multiple-pass sketching · Computer animation · Human computer interaction

1 Introduction

The current demand for 3D animations is difficult to meet because it takes time to learn how to use current animation authoring tools. Among the varied user interfaces that have been developed for 3D animation, sketch-based techniques come nearest to addressing the problem of accessibility by novices. This type of interface allows a user to draw the path to be taken by a character [2, 10, 12] or the trajectory of a rigid body [17, 19]. These approaches offer relatively accessible interfaces, but do not permit detailed 3D motions to be created without the use of a 3D input device such as a tracker or 3D mouse. Exceptions are Motion Doodles [20] and the system developed by Jeon et al. [8], which do allow a user to create the motion of a character using cursive gestures. Motion Doodles has a one-pass sketching interface, which is simple, but the subset of possible motions it can offer is limited

H. Jang · J. Jeon (✉) · E. Sohn · Y.-C. Choy
Multimedia Graphics Lab—Department of Computer Science, Yonsei University, 134 Shinchon-Dong
Sudaemoon-ku, Seoul 120-749, Republic of Korea
e-mail: demibblue@gmail.com

S.-B. Lim
Department of Multimedia Science, Sookmyung Women's University, Seoul, Republic of Korea

Published online: 19 February 2013

 Springer

by 3D ambiguities. Jeon et al. [8] avoid these ambiguities through the use of a multiple-pass sketching interface to specify the motion of a character or object. Both systems only offer controls for a single subject and lack the tools to edit motion sequences after they have been created.

Although users can input motion information easily with a sketch-based interface, this naturally commits them to perform the sketching necessary to specify the motions. This can be tedious if the interface required the repeated creation of identical motions, which is likely to occur in creating similar motions sequences for multiple characters.

In previous research, an entire motion sequence has been constructed from a single sketched stroke. But this means that a whole stroke has to be redrawn to make a change, even if the user only wants to change part of it. This suggests the need for a partial editing function, not only for changing the type and properties of the motion, but also for editing its location and length.

We address this need with a new sketch system for authoring motion sequences. Our system use multiple-pass sketching to avoid spatial ambiguities problems; but it offers new editing and reproducing techniques, and a motion block interface, for modifying motion sequences. The system contains an authoring tool for motion sequences, an interactive path editor, an editing and reproducing module for motion sequences, and a motion block interface, as shown in Fig. 1.

The user of this new system firstly selects a character to be animated, and then draws the path for that character to follow. Next, the system presents purpose-specific windows and camera modules which allow the user to specify the character's detailed motions. A user can

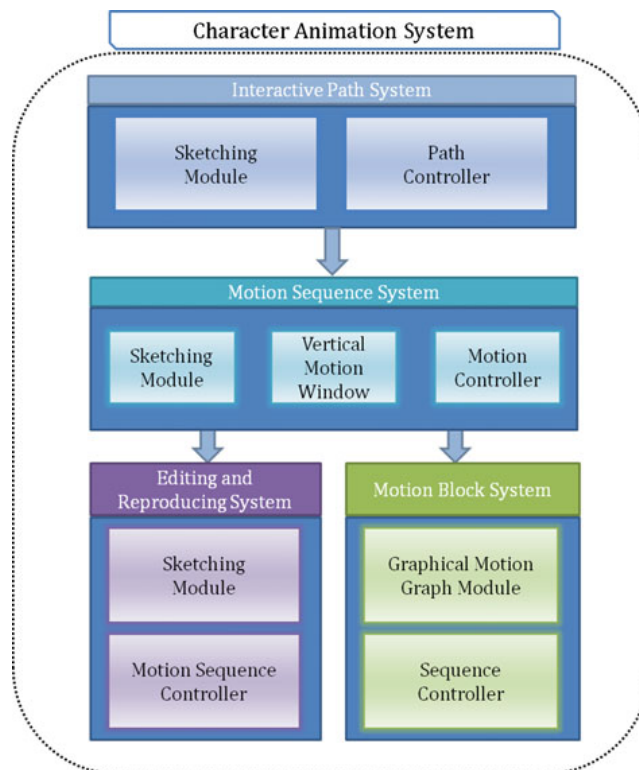


Fig. 1 Overview of our new sketch-based 3D character animation system

독서장애인을 위한 음성 도서 어노테이션 검색 기법

박주현^{*}, 임순범^{**}, 이종우^{***}

요 약

본 연구에서는 독서장애인을 위한 모바일 전자책에 저장된 어노테이션을 탐색하고 그 결과를 재생하는 기법을 제안하고, 이를 음성 어노테이션 브라우징 시스템이라 칭하였다. 제안된 음성 어노테이션 브라우징 시스템은 명령 입력, 중요도 분석 및 추천, 검색, 출력 단계로 구성된다. 특히 본 논문에서는 대상 사용자가 청각 의존도가 높은 독서장애인들이기 때문에 완전히 청각에 의존해서 사용할 수 있도록 모든 단계에서 음성인식 인터페이스를 제공한다. 제안된 음성 어노테이션 브라우징 시스템의 효율성을 검증하기 위해 모바일 안드로이드 환경에서 실행되는 전자책 소프트웨어와 음성 어노테이션 브라우징 시스템을 설계하고 구현하였다. 구현된 시스템은 전맹인으로 구성된 다수의 사용자들로 하여금 검증하도록 하였으며, 그 결과 이들이 원하는 어노테이션을 찾는 과정이 성공적으로 이루어 질 수 있음을 확인 하였다.

A Voice Annotation Browsing Technique in Digital Talking Book for Reading-disabled People

Joo Hyun Park^{*}, Soon-Bum Lim^{**}, Jongwoo Lee^{***}

ABSTRACT

In this paper, we propose a voice-annotation browsing system that make the reading-disabled people to be able to find and play the existing voice-annotations. The proposed system consists of 4 steps: input, ranking & recommendation, search, and output. For the reading-disabled people depending only on the auditory sense, all steps can accept voice commands. To evaluate the effectiveness of our system, we design and implement an android-based mobile e-book application supporting the voice-annotation browsing ability. The implemented system is tested by a number of blind-folded users. As a result, we can see almost all the reading-disabled people can successfully and easily reach the existing voice-annotations they want to find.

Key words: Annotation(어노테이션), Voice Annotation Browsing(음성 어노테이션 브라우징), Reading-disabled People(독서장애인)

1. 서 론

‘눈으로 보는 책’에서 벗어나 ‘읽어주는 책’이 대중

성을 얻고 있는데, 이 음성지원 책의 출현은 장애로 인하여 각종 도서나 다양한 정보에 소외되었던 시각 장애인 같은 독서장애인들도 불편함 없이 전자책을

※ 교신저자(Corresponding Author): 임순범, 주소: 서울 시 용산구 청파동 2가 숙명여자대학교 명신관 511호(140-741), 전화: (02) 710-9379, FAX: (02) 710-9704, E-mail: sblim@sm.ac.kr

접수일: 2012년 7월 27일, 수정일: 2012년 9월 28일
완료일: 2013년 2월 14일

^{*} 준회원, 숙명여자대학교 멀티미디어과학
(E-mail: park.joohyun5@gmail.com)

^{**} 종신회원, 숙명여자대학교 멀티미디어과학

^{***} 종신회원, 숙명여자대학교 멀티미디어과학
(E-mail: bigrain@sm.ac.kr)

※ 본 연구는 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No.20120007897)

※ 본 연구는 숙명여자대학교 2012학년도 교내연구비 지원에 의해 수행되었음.

이용할 수 있게 해주었다. “책에 사용자가 노트나 주석 등을 기록하는 것”을 의미하는 어노테이션(annotation) 기능 역시 독서장애인이 사용하려면 음성 어노테이션이 필수적이다. 독서장애인 역시 일반 사용자와 같은 경험을 하고 싶어 하기 때문이다. 단순히 녹음된 음성으로 책의 내용을 듣는 것에 만족하지 않고 자신이 책을 읽으며 직접 생각이나 의견을 메모하고 싶어 하는 욕구가 있는 것이다. 또한 자신이 기록한 생각이나 의견을 재열람 하고 공유하는 등 일반 사용자와 차별 없는 경험 역시 요구한다. 따라서 독서장애인이 어노테이션을 탐색하고 재생할 수 있게 해주는 전자책 어노테이션 브라우저의 필요성 또한 요구되고 있다. 음성 어노테이션의 기록 방법에 대한 연구는 현재 진행되고 있지만, 기록된 음성 어노테이션을 독서장애인이 탐색하는 방법에 대한 연구는 거의 없는 형편이다.

이에 본 연구에서는 독서장애인의 효율적인 독서 환경 개선을 위한 음성 어노테이션 브라우저 시스템을 설계하고 구현한다. 제안하는 시스템은 음성지령을 통해 사용자가 원하는 어노테이션을 단어 검색 등 다양한 방법을 통해 쉽고 간편하면서도 정확하게 찾아가는데 중점을 두고 있다. 기록된 어노테이션을 찾아가는 방법으로 키워드 검색 방법과 추천어 검색 방법을 지원하며, 이를 달성하기 위해 어노테이션 키워드 입력, 검색, 출력, 중요도 분석 및 추천 등의 모듈들을 설계하고 구현한다.

2. 관련연구

본 장에서는 전자문서에서의 기존 어노테이션 입력 방법을 살펴보고 아울러 어노테이션 브라우징 사례도 살펴본다.

2.1 전자문서에서의 음성 어노테이션 입력

기존의 어노테이션에 관한 연구들은 어노테이션 표현 방식에 따라 텍스트, 스타일, 펜, 컨텍스트 순서로 발전되어 왔으나 이 모든 발전은 사운드 환경을 고려하지 않은 비장애이용 전자책 기반 어노테이션 기법이었다. 반면 음성 기반 어노테이션 시스템은 간단한 음성 노트를 녹음하여 오디오 파일로 저장하면 본문의 원하는 위치에 음성 녹음 아이콘이 생기는 방식이다[1]. 이 방법은 녹음하고 저장하는 간단한 원리로 작동함으로 사용자가 사용하기 간편하지만 녹음 당시 주변 소음으로 인한 음질 훼손의 위험이 있다. DENIM은 스케치 기반 웹사이트 디자인 툴로써 마우스 펜 드로잉과 음성 어노테이션을 지원하는 데, 그 중 음성 어노테이션은 위치 설정 후 새 다이얼로그 창에서 음성을 녹음하면 해당 위치에 스피커 아이콘이 생성되고, 추후 아이콘을 클릭하여 녹음된 음성파일을 재생하는 방식으로 동작한다[2]. 그림 1은 DENIM에서의 음성 어노테이션 입력, 녹음, 아이콘 생성과정을 차례대로 보이고 있다. DENIM에서는 음성 어노테이션 입력이 가능하지만, 독서장애인으로 사용 시 역시 시력이 없는 독서장애인은 조작을 전혀 할 수 없다는 불편함이 있다. 또한 일반 사용자에게 맞추어진 화면 인터페이스 역시 독서장애인이 사용하기에 어려움이 있다.

이를 개선하기 위하여 독서장애이용 음성 기반 어노테이션 입력 지원 연구가 진행되었는데 [3,4]가 대표적이다. 이 연구들은 음성 인식 기술을 이용하여 사용자가 음성 어노테이션을 입력하고 이를 텍스트로 변환하여 저장한다. 사용자는 본문의 내용을 오디오로 듣는 도중 ‘노트’라고 음성지령하면 음성인식기가 실행되면서 사용자가 말한 내용을 텍스트로 변환하는 방식으로 동작한다. 본 연구에서는 연구 [3,4]의

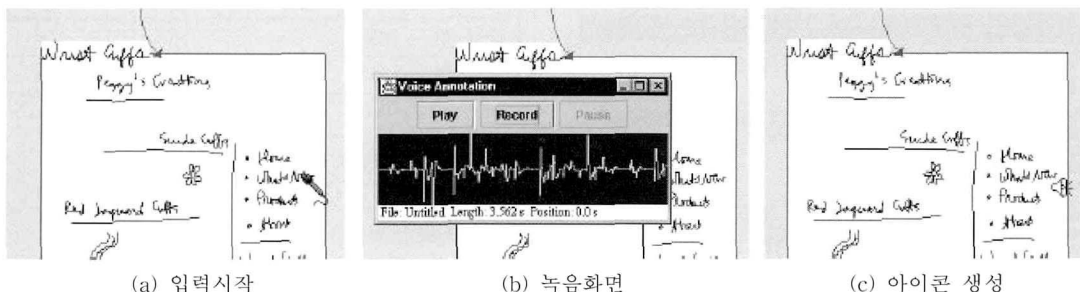


그림 1. DENIM에서의 음성 어노테이션 입력 과정

3DTV에서 스테레오 웹 콘텐츠를 위한 CSS3 사양 확장 및 에뮬레이터 연구

이희진, 임현정, 임순범

숙명여자대학교 멀티미디어학과, ETRI, 숙명여자대학교 멀티미디어학과
wmfvv@hanmail.net, hjyim@etri.re.kr, sblim@sm.ac.kr

A Study on CSS3 Stylesheet Extension and Emulator for Representation of Stereo Web Content in 3DTV

Hee-Jin Lee⁰, Hyun-Jeong Yim, Soon-Bum Lim

Dept. of Multimedia Science, Sookmyung Women's University,
ETRI,

Dept. of Multimedia Science, Sookmyung Women's University

요 약

본 논문에서는 브라우저 기반의 선언적 콘텐츠 CSS3 스타일시트를 이용하여 HTML5로 작성된 웹 페이지를 3차원 공간에서 3D 스테레오로 표현하기 위해 기존의 CSS 사양을 확장하여 새로운 CSS Stereo 3DTV 프로파일을 제안한다. 먼저, 웹브라우저상의 웹 콘텐츠를 3차원 공간에 효과적으로 재구성할 수 있도록 다양한 웹 페이지 분리 방법을 제시하였다. 다음, 확장한 CSS3 모듈을 이용하여 3D 뷰 볼륨 설정과 분리한 콘텐츠를 뷰 볼륨 내에 배치하는 방법을 제안하였다. 이후, 전처리기에서는 확장 CSS 사양으로 작성된 샘플 콘텐츠를 웹킷 기반 브라우저에서 출력 가능하도록 기존 CSS 코드로 변환한다. 마지막으로 스테레오 이미지 생성을 위한 시스템 렌더링 엔진 에뮬레이터는 웹브라우저상에서 간단히 실행 가능하도록 자바스크립트로 구현되었으며, 자동으로 스테레오 뷰 볼륨을 생성하여 가상의 좌안, 우안 카메라로부터 각각의 영상을 획득한다. 이를 통해 각 샘플 콘텐츠들이 3D 스테레오로 표현된 결과를 확인하였다.

Abstract

In this paper, we represented the HTML5 webpage in 3D space for the 3D stereoscopic display by using CSS3 stylesheet; browser-based declarative contents, and proposed CSS Stereo 3DTV Profile. First, we suggested various webpage separation methods for reconstructing webpages in the 3D space effectively. Next, we suggested 3D view volume setting method by using extended CSS3 modules. And then, pre-processor converts sample contents which is written extended CSS stylesheet into the present CSS stylesheet for displaying in the webkit based browser. For the resulting stereoscopic images, we developed a rendering engine emulator which is implemented in JavaScript for simple display in the web browser, which produced dual images from virtual left and right-eyed cameras. And we have checked the sample contents displayed on the 3DTV.

키워드: 3DTV, 웹 콘텐츠, 스테레오스코픽, HTML5, CSS3

Keywords: 3DTV, Web Contents, Stereoscopic, HTML5, CSS3

* 본 논문은 숙명여자대학교의 2012학년도 교내 연구비 지원에 의해 수행되었음.
* 본 연구는 방송통신위원회의 지상파 양안식 3DTV 방송시스템 기술 개발 및 표준화 사업의 연구결과로 수행되었음 (KCA-2012-10912-02001).

1. 서론

최근 스마트폰, 스마트 TV 등 3D 스테레오를 지원하는 스마트 디바이스가 급증하면서 스마트 TV 및 데이터 방송에서의 웹 콘텐츠 활용 요구가 증가하고 있다[1]. 또한 국내에서는 웹 표준기술 HTML5를 기반으로 하는 TV 플랫폼 표준화 작업이 진행 중이다. 이에 따라 W3C(World Wide Web Consortium)에서는 3D 스테레오 디스플레이 관련 연구[2] 및 표준화 작업이 활발히 진행 중이며, 하이브리드 방송에서는 브라우저 기반의 선언적 콘텐츠를 이용하여 데이터를 표현하는 연구를 진행하고 있다. 브라우저 기반의 선언적 콘텐츠로는 XML(Extensible Markup Language), HTML(HyperText Markup Language), CSS(Cascading Style Sheet) 등이 활용되고 있으며, 이러한 웹 표준을 이용함으로써 단말기의 종류에 관계없이 플랫폼 독립적으로 콘텐츠를 기술할 수 있게 되었다.

최근 가장 주목을 받고 있는 브라우저 기반의 선언적 콘텐츠 기술로는 HTML5와 CSS가 있다. 먼저 HTML5는 웹 표준을 제정하는 국제 표준화 기구인 W3C에서 제정중인 차세대 웹 표준으로 다양한 태그들이 확장 정의되었다. CSS는 웹 페이지의 논리적 구조와 화면상의 프리젠테이션을 분리하고 기술하기 위한 목적으로 제정된 W3C 표준으로 미디어별로 다양한 스타일을 지원하기 위한 여러 레벨과 프로파일을 정의하고 있다. 특히 CSS에서는 다양한 디바이스 환경을 고려하기 위한 다양한 프로파일을 지원하고 있는데, 이중 CSS TV 프로파일은 TV 장치의 요구사항 및 제약 조건들을 고려하여 정의한 프로파일로, TV에서 브라우저 기반 콘텐츠를 표현할 때 이용된다. 그러나 현재 CSS TV 프로파일 규격은 CSS3 등장 이전에 정의되어 3차원 공간 개념을 포함하고 있지 않다. 또한 현재 개발 되었거나 개발 중인 HTML5 및 CSS 관련 규격에서는 콘텐츠를 3차원 공간상에 배치하거나 이를 스테레오 입체 영상으로 표현하기 위한 내용은 포함하고 있지 않다.

이에 따라 본 논문에서는 웹 환경에서 차세대 마크업 언어인 HTML5로 표현되어 있는 웹 페이지를 3D 스테레오 표현하기 위해 필요한 CSS 사양을 확장하여 새로운 CSS Stereo 3DTV 프로파일을 제안하고, 이를 이용하여 작성한 샘플콘텐츠를 3DTV에서 스테레오로 표현해본다.

2. 관련 연구

2.1 스테레오스코픽 TV에서의 콘텐츠 표현 연구

현재 국내외에서는 3D 콘텐츠를 스테레오 TV에 표현하는 다양한 연구를 진행하고 있다. W3C에서는 <표 1>과 같이 3D 스테레오 디스플레이 출력에 필요한 사양 확장 및 표준화 작업이 진행 중이며[3], W3C Web and TV Interesting Group에

서는 스테레오 디바이스에서 3D 스테레오 비디오를 지원하는 방법에 대한 논의가 진행 중이다[4]. 또한, 웹브라우저 환경에서 3D를 선언하는 방법[5]과 모바일 및 TV 사용자에게 스테레오 이미지를 제공하는 3DTV Broadcasting and Distribution System 연구가 진행된 바 있다[6].

<표 1> 스테레오 3D 지원에 대한 확장

구성	CSS에서의 표현 사항
CSS 3D 확장 속성	perspective-baseline
스테레오 3D 콘텐츠 지정을 위한 스테레오 속성	stereo-content, stereo-render-option, stereo, stereo-size-type, stereo-order-type, stereo-format
Media type '3d-display' for Media Query	non-normative
Display Interface	non-normative

2.2 TV 용 웹 콘텐츠 표현 관련 표준 기술

TV 용 웹 콘텐츠 표현 관련 표준으로는 CE-HTML[7]과 CSS TV 프로파일 1.0[8]이 있다. CE-HTML은 일반적인 PC와 달리 소비자 가전기기에서의 HTML 표현 문제를 해결하기 위해 개발된 언어로 <표 2>와 같이 구성되어 있으며, TV와 같은 가전기기의 조작을 위한 사용자 인터페이스를 정의하기 때문에 소비자는 알맞은 인터페이스를 제공 받을 수 있다는 장점이 있다.

<표 2> CE-HTML의 구성

표준 대상	구성
CE-HTML	Ecmascript 262, 3rd edition
	XHTML 1.0 transitional/strict
	CSS TV Profile 1.0
	XMLHttpRequest object
	DOM Level 2.0(Core, Style, Events, HTML.)
	여러 소비자 가전기기를 위한 특정 확장 언어

CSS TV 프로파일 1.0은 셋톱박스, 양방향 텔레비전 세트 등과 같은 TV 디바이스에 적합한 색상사양과 사용자 에이전트(UA)가 최소한의 규격으로 정의된 기능들의 하위 집합을 정의한다. 속성, 속성 값, 선택자, 캐스캐이딩 등의 방법으로 TV 장치에 대한 적합한 프로파일을 지정하고 있으며, 주요 모듈은 CSS1으로 구성되고 부분적으로 CSS2와 CSS3를 포함한다. 사양은 <표 3>과 같으며, 표 이외에도 Speech Module, Template Layout Module, Fragmentation Module 등의 마우스 이벤트와 조건검색에 관련된 속성들이 TV 프로파일에서 제외되었다.

이 외에도 방송망 또는 IP 망을 이용하여 3D 영상 데이터를 미리 사용자 단말기에 다운로드하는 방식으로 실시간으로 전송되는 기존 방송 콘텐츠와 동기식 및 비동기식 방법으로 임의의 시간에 3DTV 서비스를 제공하는 하이브리드망 기반 3DTV 서비스 기술이 연구되고 있다.

An Annotation System for Reading-disabled People on Mobile e-book

Joo Hyun Park^{*}, KyungHee Lee^{**}, Soon-Bum Lim^{*}, Jongwoo Lee^{*}

^{*} Dept. of Multimedia Science, Sookmyung Women's University, Korea

^{**} UX Planning, LG Electronics

E-mail: park.joohyun5@gmail.com, cibong@naver.com, sblim@sm.ac.kr, bigrain@sm.ac.kr

Abstract

For the reading-disabled people who have any visual impairment or dyslexia, visual-based e-books are of no use. That is why DTBs(digital talking books) are developed for and warmly welcomed by them. DTBs have been developed to enlarge the reading experiences for reading-disabled people. They also provide convenience through navigation functionality. The existing digital talking books however, support bookmarks or notes only through a screen interfaces. This screen annotation interface is useless for reading-disabled people. In this paper, we suggest a voice annotation system that can insert highlights or notes during playing a talking book by using only hearing sense and voice. Our annotation system implemented on the android platform consists of 5 steps of input, analysis, rendering, storing, and output. We found by performing blindfolded users group's usability tests that our voice annotation system can effectively create annotations only through hearing sense and voice.

1. Introduction

An annotation of digital documents is very important information that can be reused because it allows you to share and exchange additional information between users[1]. However, existing electronic document annotation services provide screen-based annotation. So it is useless function to users who are visually impaired. Even in the DTBs[2] for the visually impaired users, it can be inserted through the screen interface.

In this paper, we suggest a voice annotation system to improve the efficiency of the DTBs reading experience. We implement the system for efficient annotation content's reuse, interactions between e-book and users, and user-to-user information sharing by supporting the voice annotation input to the reading-disabled people who have high dependence on hearing. The annotation system

consists of 5 steps of input, analysis, rendering, storing, and out. Especially, this system suggested time-positioning technique to determine the precise location of input voice annotation.

2. Related Research

Digital Talking book is a form that converts analog talking book to digital file. DAISY(Digital Accessible Information System)[3] is a typical e-book standard format. DAISY defines bookmarks, highlights, and notes that are required for the implementation of digital talking book playback system's annotations, but it does not mention specific annotations input and output. The standard does not set a limit on the annotation file's storage, manipulation and etc., but player that supports annotation should follow bookmark DTD. Bookmark DTD defines <bookmark>, <note>, < highlightw > as annotation type.

In general electronic document, voice annotation methods can be divided into two, depending on the storage method. One is a saving the media files without any manipulation after voice recording. Another is a text form conversion based on speech recognition.

The first method is relatively simple, but it has a risk of the compromised sound quality due to the ambient noise recorded at the time. Also it is inefficient when you want to reuse the annotation content information later. This method cannot expect the diversity of annotations such as highlights because it services only simple audio note function. In addition, the performance degradation can be incurred by the storage waste according to the increasing of recording files. Patent of the United States supporting the voice annotation in e-book reader, for example, E-BOOK READER WITH VOICE ANNOTATION/US2010-0324709[4] provides voice annotation input. It allows users to select a specific area and highlights them. Also provides voice note input on it. Namely, voice note file is linked to the highlighted text.

The second method is a way to save the converted text file with Voice-Note which is based on speech

Approaching to read the Book's content has a big difference between general electronic document and DTBs in annotation inserting environment. Plain document requires the sense of sight, so it can accept a variety of information at once. But talking Book used only auditory sense has a linear playback environment and it accepts only one piece of information at a time. For such differences, input method of the various annotations which is supported to general electronic document has limitation to apply equally in DTBs environment.

3. A Voice-Annotation System for Digital Talking Book

In addition, existing annotation systems are using a display-positioning focused on the visual, however, in this paper, our voice annotation proposal is applied the time positioning technique based on sound environmental. The time positioning technique can make the annotation with auditory sense only. This is using time information about each sentence of the media files in the book.

INPUT, ANALYSIS, RENDERING, STORE and OUTPUT sequentially. The speech inputted is converted into text format and passed to the analysis module. Analysis module analyzes the type and location of the annotation inputted. Based on the results of the analysis, the Rendering module applies sound and display effect depending on the type of each annotation. After storing all of the annotation's information and managing the annotation's database, Output module prints the result of the annotation to the user.

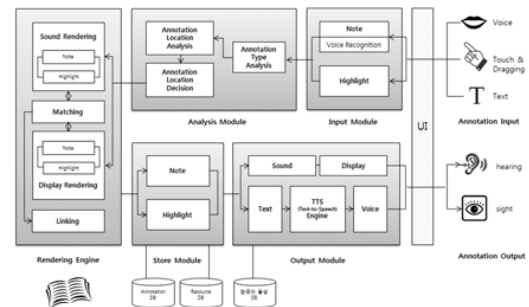


Figure 1: Configuration of the system

```
<?xml version="1.0"?>
<annotation>
  <note xlink:type="extended" id="01">
    <identity>n01</identity>
    <name>note01</name>
    <location xlink:type="x" xlink-label="annotation"/>
    <date>20100517</date>
    <comment>comment.</comment>
  </note>
</annotation>
```

(a) XML Schema

```

<element name="annotation">
  <complexType>
    <all>
      <element name="note">
        <complexType>
          <sequence>
            <element name="identity">
              <complexType>
                <sequence>
                  <element name="name" minOccurs="1"
                    maxOccurs="1">
                    <complexType>
                      <sequence>
                        <element name="location" minOccurs="1"
                          maxOccurs="1">
                          <complexType>
                            <sequence>
                              <element name="comment" minOccurs="1"
                                maxOccurs="1">
                                <complexType>
                                  <sequence>
                                    <element name="highlight">
                                      <complexType>
                                        <sequence>

```

(b) Annotation Object File

Design a Motion Library with Kinetic Typography for Messengers

Yoon-Ah Cho*, Joo-Eun Kwak*, Soon-Bum Lim*

* Dept. of Multimedia Science, Sookmyung Women's University, Korea

E-mail: fanwith@naver.com, lovelygirlje@naver.com, sblim@sm.ac.kr

Abstract

An existing messenger has limitations of expression due to a static text. In this paper, we will use Kinetic Typography to help expression of communication. We systematically designed a motion library for efficient usage of the Kinetic Typography on messenger system. Select the motion templates, design by defining the components of the kinetic typography motion, and configure the library. After selecting the motion, we implement the interface with HTML5 to set the motions' properties.

1. Introduction

Recent Mobile Messenger is composed of static text-only, therefore we wanted to provide the emoticon separately and convey the emotion. In this paper, we gave the motion which is called 'kinetic typography' to existing static text. It effects evaluating the contents of the message and increases the expressiveness. Defining the motion of kinetic typography provides motion library that is suitable for mobile environment; it enables users to set the properties of the motion.

2. Design a Motion Library for Messenger

In this study, we designed the motion library that can be used effectively in the messenger. Users input the contents of the conversation in the chat messenger, select the motion in the motion library, and set a permanent of the motion. And the library apply the kinetic typography to the contents that user input.

2.1 Design a Motion Library

Designing motion library for motion typography production proceed with two steps. First, conceptually design the kinetic typography motion to define the properties and parameters of movement in the library. Secondly, design the motion library based on the first step.

2.2 Conceptual design of the movement for the kinetic typography

Component of kinetic typography is divided into two parts, motion and motion attribute. Refer the <Table 1> explaining this concept. Motion is divided into basic part and detailed part depending on the direction and movement of kinetic typography. Properties of motion are decided by properties which affect movement of the kinetic typography.

Table 1: Components of the kinetic typography

Basic motion : by method	Detailed motion : by direction	Properties of the motion
Movement within fixed position	Movement within fixed position	speed / position / size / Transparency / character separation /Basic elements of typography
Linear movement	Up / down / left / right / front / back	
Curved movement	Parabolic / wave / bound / Concentric	

2.3 Set the Kinetic typography motion and attribute

Set the attributes of speed, and determine the detailed motion by the motion factor like moving direction, size, and transparency of the basic motion, the linear movement, then complete the motion.

3. Conclusions

We defined the concept of motion that will be utilized as a kinetic typography for messenger, designed the motion library through the evaluation of user based on this concept, and finally implemented the interface. It is expected that easier production of the kinetic typography in the Messenger environment would be possible by using the motion library.

Acknowledgements This work was supported by the IT R&D program of MKE/KEIT. [10041794, Development of the Kinetic Typography Technology for Emotional Representations in Social Media Environment]

4. References

- [1] Zhizuan Yeo, and Scott. E. Hudson, "KTE2: An Engine for Kinetic Typography", ACM CHI 09, 2009
- [2] S.Ford, J.Forlizzi, S.Ishizaki, "Kinetic Typogeaphy: Issues in time-based presentation of text", ACM CHI 97, 1997

POI 데이터 검색을 위한 문자열 유사도 측정 기법

(Measuring Method of String Similarity for POI Data Retrieval)

진 아 연 * 이 재 원 ** 이 종 우 ***
(AhYeon Jin) (JaeWon Lee) (Jongwoo Lee)

요 약 사람들은 목적지까지 빠르고 정확하게 도달하기 위해 네비게이션을 사용한다. 그러나 목적지 검색에서 네비게이션이 사용자가 원하는 결과를 출력하지 못해 문제가 되는 경우가 많다. 목적지 검색 결과의 재현율과 정확률을 떨어뜨리는 가장 큰 이유는 사용자가 데이터베이스에 저장된 목적지의 명칭을 정확하게 알지 못하기 때문이다. 본 논문에서는 POI 데이터 검색의 재현율과 정확률을 높이기 위하여 추정과 블록이라는 개념에 기반한 문자열 유사도 측정 기법을 제안한다. 실험 결과, 본 논문에서 제시하는 기법이 모든 POI 데이터에 대한 검색 재현율과 정확률을 상당수준 향상시켰음을 알 수 있었다.

키워드: POI 검색, 네비게이션, 문자열 유사도 측정, 한글 검색, 알고리즘

Abstract People take advantage of navigations in order to arrive at destination quickly and accurately. It happens frequently, however, that the user cannot acquire the correct retrieval result in the destination retrieval stage. Most of the retrieval failures result from the false queries which are mainly due to the situation that the users can not exactly memorize the correct names of the destinations stored in the database. In this paper, a novel method for measuring POI string similarity based on the concept of presumption and matching block is proposed in order to enhance the recall and precision of POI retrieval. From the experimental data, by using the proposed method, the recall and precision of the POI retrieval are remarkably enhanced than those of the existing retrieval methods.

Keywords: POI retrieval, navigation, string similarity measure, the Korean alphabet retrieval, algorithm

1. 서 론

교통이 발달함에 따라 현대인들은 점점 더 자주, 그리

고 더 멀리 이동한다. 목적지까지 빠르고 정확하게 도달하기 위한 많은 노력의 일환 중 가장 성공적이고 보편화된 것이 네비게이션이다. 네비게이션이 보급된 이래로, 보다 정밀하고 빠르게 도움을 줄 수 있는 다양한 기술들이 개발되고 있으나, 정작 가장 기본적인 기능인 목적지 검색에서 사용자가 원하는 결과를 출력하지 못해 문제가 되는 경우가 많다.

목적지 검색은 사용자가 입력한 검색어를 POI 데이터와 비교하는 것을 통해 이루어진다. POI(Point Of Interest, 관심 지점)[1]란 주요 시설물 등을 좌표 형태로 전자 수치지도에 표시하는 데이터를 의미하며 이는 검색 데이터와 바탕 데이터로 나눌 수 있는데, 상호명이나 도로명 등의 명칭으로 저장된 검색 데이터를 이용하여 목적지 검색을 수행한다.

목적지 검색의 재현율과 정확률을 떨어뜨리는 가장

* 학생회원 : 숙명여자대학교 멀티미디어학과
ayjin@sookmyung.ac.kr

** 정 회 원 : 성신여자대학교 IT학부 교수
jwlee@sungshin.ac.kr
(Corresponding author임)

*** 종신회원 : 숙명여자대학교 멀티미디어학과 교수
bigrain@sookmyung.ac.kr

논문접수 : 2012년 10월 18일

심사완료 : 2013년 1월 26일

Copyright©2013 한국정보과학회 : 개인 목적이나 교육 목적인 경우, 이 저작물의 전체 또는 일부에 대한 복사본 혹은 디지털 사본의 제작을 허가합니다. 이 때, 사본은 상업적 수단으로 사용할 수 없으며 첫 페이지에 본 문구와 출처를 반드시 명시해야 합니다. 이 외의 목적으로 복제, 배포, 출판, 전송 등 모든 유형의 사용행위를 하는 경우에 대하여는 사전에 허가를 얻고 비용을 지불해야 합니다.

정보과학회논문지: 컴퓨팅의 실제 및 레터 제19권 제4호(2013.4)

큰 이유는 사용자가 목적지의 명칭을 정확하게 알지 못하기 때문이다. 주로 사용자가 명칭을 기억하지 못하거나, 외래어 표기법 및 틀린 맞춤법으로 인해 명칭을 혼동할 때, 사용자는 원하는 목적지의 명칭과 일치하지 않는 검색어를 입력하게 된다. 예를 들면, ‘금강사시유리’가 원하는 목적지의 명칭이지만 사용자가 검색어를 ‘금강유리샷시’로 입력할 경우 기존 시스템에서는 검색 결과를 찾을 수 없다. 이러한 문제점을 해결하기 위해 본 논문에서는 POI 데이터 검색을 위한 문자열 유사도 측정 기법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서 상용 네비게이션의 문제점과 문자열 간 유사도를 측정하는 일반적인 방법에 대해 설명하고, 3장에서는 본 논문에서 제안하는 알고리즘을 적용할 시스템인 POI검색 시스템에 대해 기술한다. 4장에서는 제안하는 문자열 유사도 측정 방법을 설명한다. 5장에서는 실험을 통해 제안하는 기법의 성능을 검증하고, 마지막으로 6장에서 결론을 맺는다.

2. 관련 연구

현재 상용되는 네비게이션에서 POI 검색은 서버 사용 여부에 따라 두 가지로 분류할 수 있다. 일반적으로, 서버를 이용하지 않는 POI 검색법은 입력조건과 정확하게 일치하는 결과만 제공하는 하드매칭 검색 방법을 사용하며, 서버를 이용하는 POI 검색법은 키워드와 별칭 데이터베이스(alias database)에 기반한 매칭 알고리즘을 사용하는 것으로 알려져 있다.

서버 미사용 검색법은 통신이 불가능한 상황에서도 검색을 가능하게 해준다는 장점이 있으나, 사용자가 목적지의 정확한 명칭을 모두 기억해야 하므로 오질의어에 매우 취약하다는 단점이 있다.

서버를 사용하여 POI 데이터를 검색하는 것으로는 네이버 지도, T map, 올레네비 등이 있다. 서버를 사용하는 알고리즘은 공개가 되어있지 않지만 일반적으로 키워드와 별칭 데이터베이스를 이용하여 검색을 수행하는 것으로 알려져 있다. 별칭 데이터베이스를 이용하면, ‘숙명여자대학교’를 찾기 위해 ‘숙명여대’를 입력하는 것처럼 자주 발생하는 오질의어가 입력되는 경우에도 기존 하드매칭보다 나은 검색 결과를 제공한다는 장점이 있다.

네이버의 검색엔진을 예로 들면, 사용자가 질의어를 입력했을 때 색인기라는 장치를 이용하여 키워드만을 추출해낸다. 검색 엔진은 이렇게 추출한 키워드를 통해 별칭 데이터베이스를 이용하여 검색을 수행하는 것으로 판단된다. ‘종로방통위’라는 검색어를 입력하는 경우 ‘서울특별시 종로구 종로’ 근처에 있는 ‘방통위’를 검색한다는 안내말이 뜨며 ‘방송통신위원회’가 검색 결과로 출력되는 것을 볼 수 있다. 다음의 그림 1은 네이버 지도에

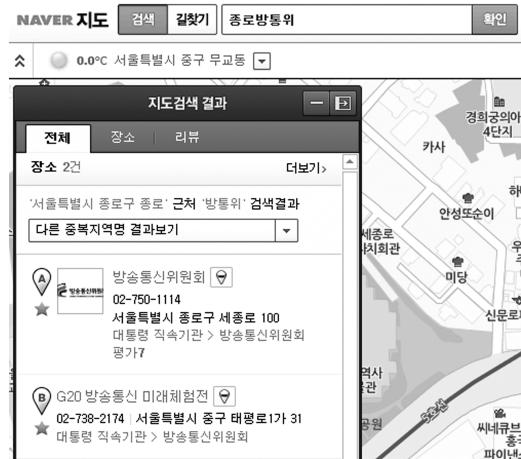


그림 1 네이버 지도에서의 ‘종로방통위’ 검색 결과



그림 2 네이버 지도에서의 ‘방통위원회’ 검색 결과

서 ‘종로방통위’를 검색한 결과이다.

하지만, ‘방통위원회’라는 질의어를 입력하면 검색 결과가 출력되지 않고, 관련 없는 키워드를 질의어로 제안하는 것을 볼 수 있다. 이는 별칭 데이터베이스에서 모든 별칭을 관리하지는 못하기 때문에 발생하는 문제로 판단된다. 그림 2는 네이버 지도에서 ‘방통위원회’를 검색한 결과이다.

서버 사용 검색법은 하드매칭 기법보다 나은 검색 결과를 제공할 수 있고, POI 데이터베이스 및 알고리즘 등을 서버에서 처리하기 때문에 클라이언트 측 프로그램이 가벼워진다는 장점이 있다. 그러나 서버에 접속해야 하므로 상시 통신이 가능한 상태여야 하며, POI 데이터의 모든 별칭을 관리할 수는 없기 때문에 여전히 오질의어에 취약하고 비효율적이라는 단점이 있다.

기존 검색 방법들의 문제점들을 해결하면서 각 장점을 유지할 수 있는 검색 방안은 유사도를 이용한 검색 방법이다. 대부분의 기존 유사도 검색 기법은 문자부터 문서까지의 매우 긴 문자열 간의 비교를 수행한다. 하지

효율적인 집합-기반 POI 데이터 검색 알고리즘 (An Efficient Set-based POI Data Search Algorithm)

고 은 비 ^{*} 이 종 우 ^{**} 이 재 원 ^{***}
(EunBi Go) (Jong woo Lee) (Jae Won Lee)

요약 위치 기반 서비스는 특정 위치의 지리 정보를 파악하기 위해 POI 데이터베이스를 사용한다. 모바일 환경에서의 위치 기반 서비스의 경우에는 검색 성능 향상을 위해 사용자의 유동성에서 기인하는 부정확한 POI 쿼리를 개선해야 한다. 이를 위해 기존의 시스템에서는 쿼리 자동 확장 기술과 하드매칭 기법을 사용하지만, 이는 외부 자원을 필요로 하고 성능이 떨어진다는 단점을 지닌다. 그러므로, 본 논문에서는 시스템 자체 내에서 부정확한 POI 쿼리로 인한 성능 저하 문제를 해결할 수 있는 새로운 POI 데이터 검색 알고리즘을 제시한다. 본 알고리즘은 n 개의 문자로 이루어진 POI 쿼리를 m 개의 블록으로 균등 분할한 뒤, 각 블록에 대해 집합 기반 연산을 적용하고, '차수'라는 개념을 사용하여 블록 간 집합 연산을 수행한다. 실험을 통해 본 논문에서 제시한 집합 기반의 POI 검색알고리즘의 성능이 기존 기법에 비해 83%~96% 우수함을 확인하였다.

키워드: 관심지역정보, POI 데이터베이스, POI 검색 알고리즘, 집합-기반 알고리즘

Abstract Location-based services (LBS) deploy the Point of Interest (POI) databases to locate the positions users want. LBS should support a query correctness enhancing technique because the POI queries are likely to be incorrect due to the user's mobility. The existing systems commonly use an automatic query expansion technique and a hard matching technique, but these techniques have weaknesses such as necessity of outer resources and low performance, respectively. Thus, in this paper, we propose a new POI data search algorithm working well even when the POI queries are incorrect. The algorithm regards a POI query as a sequence of n characters, and divides it into m chunks. The algorithm applies the set-based operation for each chunk and between chunks. We can find by the performance evaluation using real POI database that our algorithm improve the POI search performance for 83%~96%.

Keywords: Point of Interest, POI database, POI search algorithm, Set-based algorithm

1. 서론

위치 기반 서비스(Location-based services)는 특정

지리적 위치를 파악한 뒤, 이를 바탕으로 사용자에게 관련 정보를 제공하는 플랫폼이다. 일반적인 위치 기반 서비스는 파악해야 할 특정 위치에 대한 위도, 경도 등의 지리 정보를 얻기 위해 POI (Point of Interest) 데이터베이스 또는 GPS (Geographical Positioning System)을 사용한다.

위치 기반 서비스를 제공하는 모바일 기기 중 스마트폰과 차량 내비게이션 시스템이 일반적이다. 스마트폰은 위치 기반 서비스를 지도/내비게이션 어플리케이션의 형태로 제공하는데, 그 중 구글 지도 모바일 서비스¹⁾가 대중적이다. 차량 내비게이션 시스템은 GPS로 현 위치를 파악하고, 파악한 위치와 사용자의 POI 간의 최단 거리를 계산한다[1].

POI는 점으로 표현될 수 있는 모든 지리적 객체[2]이

^{*} 학생회원 : 한국과학기술원 전산학과
eunbi_go@kaist.ac.kr

^{**} 종신회원 : 숙명여자대학교 멀티미디어학과 교수
bigrain@sookmyung.ac.kr

^{***} 정 회 원 : 성신여자대학교 컴퓨터정보학부 교수
jwlee@sungshin.ac.kr
(Corresponding author임)

논문접수 : 2012년 12월 17일

심사완료 : 2013년 2월 12일

Copyright©2013 한국정보과학회 : 개인 목적이거나 교육 목적인 경우, 이 저작물의 전체 또는 일부에 대한 복사본 혹은 디지털 사본의 제작을 허가합니다. 이 때, 사본은 상업적 수단으로 사용할 수 없으며 첫 페이지에 본 문구와 출처를 반드시 명시해야 합니다. 이 외의 목적으로 복제, 배포, 출판, 전송 등 모든 유형의 사용행위를 하는 경우에 대하여는 사전에 허가를 얻고 비용을 지불해야 합니다.

정보과학회논문지 : 컴퓨팅의 실제 및 레터 제19권 제5호(2013.5)

1) Google Maps Mobile, <http://www.google.com/mobile/maps/>

며, 사용자가 정보를 얻고자 하는 관심 지점을 의미한다. POI는 POI 데이터베이스에 인구 통계 자료와 함께 위도, 경도, 국가 등의 지리적 정보의 형태로 저장된다 [3]. 이는 사용자에게 직접 입력 받은 쿼리로 POI 데이터베이스를 검색하여 지리적 위치 정보를 추출하는 지리 정보 시스템(Geographical Information System)에 기반한다.

GPS는 인공위성으로 데이터 신호를 송수신하여 기기의 현 위치를 파악하는 방법으로, 애플의 3G 아이폰이 출시된 2008년[4]부터 급격히 활성화되었다. 애플 사(社)에서 본사의 모바일 기기에 GPS 센서를 내장한 이후 통합 포지셔닝 센서를 내장한 다양한 범용 모바일 기기가 출시되었기 때문이다.

모바일 환경에서 위치 기반 서비스의 성능을 향상시키기 위해서는 사용자의 유동성을 고려하여 POI를 빠르고 정확하게 검색해야 한다. GPS를 이용한 POI 검색 서비스는 내장된 센서를 사용하므로 사용자의 유동성보다 기기의 성능이 서비스의 성능에 많은 영향을 미친다. 그러나 POI 데이터베이스를 이용한 POI 검색 서비스는 사용자의 이동성에 직접적인 영향을 받는다. 유동적 환경은 사용자가 POI 쿼리를 부정확하게 작성할 확률을 증가시켜 데이터 접근성을 떨어뜨리기 때문이다. 그러므로 사용자의 유동적 환경은 GPS보다 POI 데이터베이스를 이용한 지리 정보 검색에 고려되어야 하며, POI 데이터베이스를 이용한 위치 기반 서비스의 경우 성능 향상을 위해 부정확한 POI 쿼리를 개선해야 한다.

부정확한 POI 쿼리 입력에서 기인한 서비스 성능 저하를 개선하기 위해 쿼리 자동 확장 기술이 대부분의 시스템에 적용된다. 쿼리 자동 확장 기술에는 전역적 쿼리 확장(Global Query Expansion)과 지역적 쿼리 확장(Local Query Expansion)[5], 그리고 적합성 피드백(Relevance Feedback)[6]과 같은 방법이 존재한다. 대부분의 쿼리 확장 기술은 계산이 복잡하고 방대한 양의 데이터를 필요로 하기 때문에 차량 내비게이션과 같은 독립형 시스템에는 부적합하다. 그러므로, 대부분의 독립형 시스템에서는 하드매칭(hard matching) 기법이 적용되는데, 이는 POI 쿼리와 POI 데이터베이스 간의 유사도를 고려하지 않은 단순 검색 방법이므로 성능이 매우 떨어진다.

본 논문에서는 부정확한 POI 쿼리 입력으로 인한 POI 검색 서비스 성능 저하 문제를 시스템 자체 내에서 해결하기 위한 새로운 알고리즘을 제시한다. 이는 외부 자원을 사용하지 않고 사용자의 부정확한 POI 쿼리를 알고리즘 차원에서 재구성 하는 방법을 사용하여 POI 검색 서비스 성능을 향상시킨다.

본 알고리즘의 개략적인 동작 과정은 다음과 같다.

POI 데이터베이스를 메인 메모리에 로딩한 뒤, 레드-블랙 트리를 이용하여 글자 별 아이디를 생성한다. 사용자의 POI 쿼리가 n 개의 문자로 이루어져 있다면, 이를 m 개의 블록으로 균등분할 하여 집합의 연산을 적용한다. POI 쿼리의 각 블록에 대해 연산을 수행한 뒤, 블록 간 연산을 수행한다. 각 블록의 글자가 모두 포함된 POI 데이터베이스의 레코드 개수와 번호를 파악하여 블록 내 연산을 수행하고, '차수'라는 개념을 사용하여 블록 내 연산을 통해 얻은 검색 후보에 대해 다른 블록의 정보를 활용하여 추가적인 제약을 가함으로써 검색 후보의 수를 축소시키는 블록 간 연산을 한다.

본 논문은 POI 검색 시스템의 쿼리 처리에서 다음의 공헌을 제시한다:

- 사용자의 POI 쿼리를 m 개의 블록으로 나누어 집합의 연산을 적용한다. 집합의 교환법칙에 따라 쿼리 블록의 모든 조합을 통해 검색 후보의 수를 축소한다.
- n 개의 문자로 이루어져 있는 POI 쿼리를 m 개의 블록($n \geq m$)으로 나눔으로써, 기존 시스템보다 향상된 성능으로 검색 할 수 있도록 한다.

본 논문의 구성은 다음과 같다. 2장에서는 서비스 성능 저하를 개선하기 위해 기존의 POI 검색 시스템에서 사용하는 쿼리 자동 확장 기술을 설명하고, 3장에서는 본 논문에서 제안하는 알고리즘과 이에 대한 실제 예를 기술한다. 4장에서는 제안한 알고리즘의 실효성을 실험으로 입증한 뒤, 5장에서는 결론을 기술한다.

2. 관련 연구

본 장에서는 부정확한 POI 쿼리 입력으로 인한 성능 저하를 개선하기 위해 기존의 시스템에서 사용하는 쿼리 자동 확장 기술을 설명한다. 쿼리 자동 확장은 정보 검색 분야에서 쿼리와 검색 결과의 불일치를 해결하기 위해 연구된 기술이다. 이 기존 기술은 사용자의 쿼리 내 단어와 유사한 단어를 자동으로 찾아내어 이를 쿼리에 대치, 추가 시키는 방법을 사용하며, 사용하는 데이터의 양에 따라 전역적 쿼리 확장 기술과 지역적 쿼리 확장 기술로 구분된다[5].

2.1 전역적 쿼리 확장 기술

전역적 쿼리 확장 기술은 단어 간 관계를 찾기 위해 통계적 연산을 실시한다. 통계적 연산은 가능한 두 단어의 조합에 대한 정보를 계산하는 것으로, 모든 데이터를 대상으로 연산을 실시한다. 전역적 쿼리 확장 기술은 데이터 전체를 이용하여 연산하므로 연산 결과에 대한 신뢰도가 높지만, 데이터를 축적하고 연산하는 데에 시간이 오래 걸려 효율성이 떨어진다. 전역적 쿼리 확장 기술에는 용어 클러스터링 기법(term clustering)[7], 차원 축소 기법(dimensionality reduction)[8] 등이 있다.

사진 콘텐츠 분류를 위한 태그 클러스터링 기법 및 태그 추천[☆]

A Tag Clustering and Recommendation Method for Photo Categorization

원 지 현¹ 이 중 우¹ 박 희 민^{2*}
Ji-Hyeon Won Jongwoo Lee Heemin Park

요 약

개인이 사용할 수 있는 스마트 기기가 다양해지면서 여러 기기로 생산된 사진 콘텐츠가 어떤 기준이나 규칙 없이 분산되어 있어 콘텐츠를 관리하고 원하는 콘텐츠를 검색하는 것이 어려워졌다. 따라서 본 논문에서는 개인 사진 콘텐츠를 효과적으로 분류하기 위하여 의미적 유사도를 기반으로 한 태그 클러스터링 기법과 개인이 사진에 태그를 넣을 때 초기 클러스터를 기반으로 태그를 추천하는 방법을 제안한다. 태그들 사이의 유사도를 계산하여 서로 관련이 있다고 판단되는 태그들을 클러스터링 하는데, 태그가 같은 클러스터에 포함되어 있으면 그 태그를 가진 사진들도 유사성을 가진다고 볼 수 있으므로 개인 사진들을 의미에 따라 분류하는데 이용할 수 있다. 또한 분류된 초기 클러스터로 태그를 추천하여 개인 사용자가 태그를 분류에 맞게 추가할 수 있어 사진 분류 관리가 용이해진다.

주제어 : 태그 분류, 태그 추천, 클러스터링, 사진 콘텐츠, 사진 관리

ABSTRACT

Recent advance and popularization of smart devices and web application services based on cloud computing have made end-users to directly produce and, at the same time, consume the image contents. This leads to demands of unified contents management services. Thus, this paper proposest tag clustering method based on semantic similarity for effective image categorization. We calculate the cost of semantic similarity between tags and cluster tags that are closely related. If tags are in a cluster, we suppose that images with them are also in a same cluster. Furthermore, we could recommend tags for new images on the basis of initial clusters.

☞ keyword : Tag Categorization, Tag Recommendation, Clustering, Photo Contents, Photo Management

1. 서 론

스마트 미디어 환경에 따라 스마트 폰, 태블릿, 스마트 TV 등 스마트 기기가 대중적으로 보급됨에 따라 개인 사진이 다수의 기기와 저장장치에 분산되어 있으므로 사진 콘텐츠 모음의 파편화, 중복성이 발생하여 효율적인 검색, 열람이 어려워지게 되었다. 사람들은 매일의 일상을 기록해 두고 싶어 하며 쉽게 찾아서 꺼내보거나 인화 혹은 인터넷 상 업로드를 원하기 때문에 산재해있는 사진 콘텐츠를 통합하여 관리할 수 있도록 하는 통합 관리 서

비스에 대한 요구사항이 증가하였다.

이러한 문제를 해결하고 효율적으로 사진을 관리하기 위해서는 사용자가 일일이 수작업으로 분류하고 저장하는 방법보다는 자동으로 사진을 분류하고 관리하는 방법을 필요로 한다. 현재 클라우드 컴퓨팅 기반의 서비스가 증가하여 여러 종류의 서비스를 제공하고 있지만 사진을 통합하여 저장하고 열람하는 기능만을 제공하고 있다. 하지만 스마트 기기 사용자는 점점 더 까다로워지고 있으며 사용자에게 맞춤형 서비스나 조금 더 편리하고 사용성이 높은 서비스를 원한다. 따라서 사용자가 수작업을 하지 않아도 주제별, 이벤트별로 사진을 분류하고 그에 따라 검색할 수 있게 하는 것이 필요하다.

이에 따라 이 논문에서는 사진을 주제별, 이벤트별로 효과적으로 분류할 수 있기 위한 태그 클러스터링 기법을 제안하고 시스템을 구축하였다. 개인 사용자가 웹 서버에 사진을 업로드 할 때 사진을 자동으로 분류하여 저장하고, 업로드 시 사진에 맞는 태그를 추천해주어 사진 콘텐츠의 관리를 용이하게 하였다.

¹ Dept. of Multimedia Science, Sookmyung Women's University, Seoul, 140-742, Korea

² Dept. of Computer Software Engineering, Sangmyung University, Cheonan, 330-720, Korea

* Corresponding Author (heemin@smu.ac.kr)

[Received 3 December 2012, Reviewed 5 December 2012, Accepted 20 March 2013]

☆ 본 논문은 상명대학교 2013학년도 교내연구비에 의하여 수행되었음

2. 관련 연구

2.1 메타데이터 및 이미지 시각정보를 이용한 사진 분류

먼저 메타데이터 중 위치 데이터의 연관성을 이용하여 콘텐츠를 추천하는 시스템의 예로 GeoLife [1]가 있다. 이 연구는 GPS (Global Positioning System) 정보를 이용하여 여행자의 위치 패턴을 저장하고 여러 사람의 위치 패턴을 분석하여 연관성을 찾아내 다른 여행자에게 여행 코스를 추천해준다. 사용자의 GPS 로그로 사용자의 궤적을 저장하고 위치와 사용자 그래프, 각 사용자간의 관계 그래프, 위치와 위치간의 그래프를 통해 추천 여행코스를 추천하게 된다.

또한 Direct Share [2]에서는 사진을 분류할 때 사용자 설문을 통해 사용 습관과 인터페이스 조사 후 적용하는 방법을 제안한다. 사진을 공유할 때 사용자들의 선호도에 따라 공유할 사진에 점수를 주어 업로드 할 사진을 결정하고 공유하는 사진관리 시스템 제안한다. 라운드-로빈 방식 (round-robin concept)을 이용하여 여러 사용자에게 사진을 보여주고 점수를 받아서 공유할 만한 사진인지 결정한다.

사진 콘텐츠를 웹에 업로드 하여 관리하는 대표적인 서비스로는 구글의 Picasa [3]와 야후의 Flickr [4] 서비스가 있다. 두 서비스 모두 사진을 수정할 수 있는 사진 편집 기능을 제공하고 업로드 한 사진을 메타데이터 기반으로 분류하고 검색할 수 있게 해준다. 그림 과 같이 업로드 한 개인 사진 데이터를 열람할 때 날짜 별로 열람이 가능하다. 사진을 생성한 날짜와 업로드 한 날짜를 구분하여 열람하여 사진을 찾을 수 있다.

하지만 이는 단순히 날짜만을 이용한 분류 방법으로, 사용자가 예전 사진을 검색할 때 날짜를 기억하지 못하는 경우가 많아 의미있는 분류 방법이라고 할 수 없다. 이에 따라 사용성이 떨어지는데 다른 분류 기준을 제시하지 않고 있어서 방대한 양의 사진 중 원하는 사진을 쉽게 찾을 수 있는 사용자 인터페이스나 분류 방법이 없다고 할 수 있다.

시각 정보 (visual information)를 이용한 사진 분류에 관한 기존 연구는 지도를 타일로 나누어 위치 태그 (geo-tag)에 따라 비슷한 위치에서 찍힌 사진을 분석하고 시각적 유사도 (visual similarity)에 따라 같은 물체나 사물이 있는 사진을 모아주는 연구 [5]가 있다. 이 연구에서는 영상처리 기법을 통해 이미지의 유사성을 판단하여

분류기준을 정한다. 사진 간의 시각적 유사도를 계산하여 물체 (object)를 추론하고 추론된 이미지 정보와 위치 정보를 종합하여 이벤트를 추론하여 분류하게 된다.

또한 SheepDog [6]에서는 연구에서 미리 정해놓은 컨셉 (태그) 중에 하나를 키워드로 검색하여 검색 결과로 나온 사진과 그룹을 클러스터링하여 그 시각적 특징을 학습시킨다. 그리고 새로 들어오는 사진에서 특징을 추출하고 학습된 데이터를 이용하여 비교한 후 각 사진에 맞는 상위 n 개의 그룹 (top- n)과 태그를 추천한다. 추천된 그룹에 따라 사진을 분류할 수 있고 원래 정해진 컨셉 별로 분류 기준이 정해지게 된다. 이 연구는 미리 정해져 저장해놓은 컨셉에 관련된 사진인 경우에만 적용이 가능한 단점이 있어서 일반화하기에는 적합하지 않다.

2.2 태그 클러스터링 및 태그 추천 관련 연구

태그를 이용한 사진 분류에 초점을 맞추어 살펴보면 태그 클러스터링 관련 연구들을 볼 수 있다. 먼저 의미적 태그 클러스터링 방법을 제안하고 태그를 클러스터링하여 의미에 따라 계층적으로 트리를 만든 연구가 있다 [7]. 이 연구에서는 주어진 조건에 맞춰서 모든 태그에 대해 각 클러스터에 포함될 확률을 계산하고 유효한 클러스터 (effective cluster)가 될 때까지 클러스터를 분할한다. 실제 클러스터가 되면 그 클러스터를 대표하는 태그를 리딩 태그 (leading tag)로 지정하고 의미적 클러스터링을 하게 된다.

또한 태그들의 동시 발생 빈도수 (co-occurrence)에 따라 유사성을 측정하는 연구도 있는데 [8] 어떤 웹 페이지에서 태그들 사이의 동시 발생 빈도수를 세어서 높게 나타나는 태그들을 강한 연관성이 있는 태그 (strongly related tag)라고 정하여 태그의 의미 사이의 연관성을 알아보는 연구이다. 이를 통해 태그들간의 유사성이나 관련성을 알아내어 의미 있는 계층적 태그 구조를 만드는 데 활용할 수 있다. 예를 들면 RSS라는 태그가 주어졌을 때 한 페이지에 동시에 다른 단어가 나타나는 횟수를 저장하게 된다. 횟수에 따라 가중치를 부여하여 태그 간의 유사성을 파악하게 된다. 태그간의 유사성을 파악한 후에 그 유사도에 따라 각 태그와 연관된 태그를 추천할 수 있다.

시각 정보와 메타데이터를 함께 이용하여 태그를 추천할 수도 있다. Tag Suggestr [9]에서 제안한 방법으로 사용자가 이미지를 업로드 할 때 초기 태그를 붙이게 되고 이것과 기존 태그 정보를 이용하여 유사성을 판단하

부호 분포 변동성에 기반한 한국어 음소 ‘ㅅ’, ‘ㅈ’, ‘ㅊ’ 음성인식 (Speech Recognition of Korean Phonemes ‘ㅅ’, ‘ㅈ’, ‘ㅊ’ based on Sign Distribution Volatility)

이 지 현 ^{*} 이 재 원 ^{**} 이 종 우 ^{***}
(Jihyeon Lee) (JaeWon Lee) (Jongwoo Lee)

요 약 본 논문은 최근 스마트폰의 보급이 확산됨에 따라, 요구가 급속하게 증가하고 있는 음성 인식의 일부로써 한국어 음소 중 ‘ㅅ’, ‘ㅈ’, ‘ㅊ’ 인식에 대한 연구이다. 음소는 음성을 구성하고 있는 최소단위로써 음성을 인식하는데 매우 중요한 역할을 한다. 음소를 기본단위로 하는 음성 인식 시스템은 음절이나 단어단위의 인식보다 매우 효과적인 방법이지만 각각의 음소들을 정확하게 분리하고 인식하려면 발음의 다양성 등으로 인해 많은 어려움이 존재한다. 본 논문에서는 음소를 인식의 기본단위로 하는 음성 인식 시스템 구현의 일부로써 한국어 음소 중 인식하기 어려운 음소로 알려진 ‘ㅅ’, ‘ㅈ’, ‘ㅊ’을 인식하기 위한 새로운 지표로 부호 분포 변동성을 제안한다. 2음절 고립단어를 대상으로 실험을 수행한 결과, 기존 방법인 영교차율(ZCR)에 비해 부호 분포 변동성을 이용한 인식 기법이 에러율을 최소 52.8% 이상 감소시키는 성과를 보였다.

키워드: 음성인식, 파찰음, 파열음, 마찰음, 부호 분포

Abstract This paper describes a study on the recognition of ‘ㅅ’, ‘ㅈ’, ‘ㅊ’ among Korean phonemes as a part of speech recognition which is arousing much demand of people while the smart phones are gaining wide popularity. Phoneme is the minimal unit consisting of speech and plays a very important role in recognizing speech. Although speech recognition systems using phonemes as recognition unit work in a very efficient way than those using syllables or words, there remain many difficulties including various variations of speech unresolved in exactly separating and recognizing each constituting phoneme. In this paper, sign distribution volatility is proposed as a novel measure for recognizing ‘ㅅ’, ‘ㅈ’, ‘ㅊ’ which are known to be very difficult to recognize among Korean phonemes, as a part of the implementation of speech recognition system using phonemes as the basic unit of the recognition. According to the experimental results using bi-syllables isolated words, the proposed measure reduced by about 52.8% in the error rate over the existing zero crossing rate.

Keywords: speech recognition, affricate, plosive, fricative, sign distribution

· 이 논문은 2013년도 정부(교육부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임(2013R1A1A2013155)

^{*} 학생회원 : 숙명여자대학교 멀티미디어학과

cong891@naver.com

^{**} 정 회 원 : 성신여자대학교 IT학부 교수

jwlee@sungshin.ac.kr

(Corresponding author임)

^{***} 종신회원 : 숙명여자대학교 멀티미디어학과 교수

bigrain@sookmyung.ac.kr

논문접수 : 2012년 10월 18일

심사완료 : 2013년 4월 18일

Copyright©2013 한국정보과학회 : 개인 목적이나 교육 목적인 경우, 이 저작물의 전체 또는 일부에 대한 복사본 혹은 디지털 사본의 제작을 허가합니다. 이 때, 사본은 상업적 수단으로 사용할 수 없으며 첫 페이지에 본 문구와 출처를 반드시 명시해야 합니다. 이 외의 목적으로 복제, 배포, 출판, 전송 등 모든 유형의 사용행위를 하는 경우에 대하여는 사전에 허가를 얻고 비용을 지불해야 합니다.
정보과학회논문지: 컴퓨팅의 실제 및 레터 제19권 제7호(2013.7)

1. 서 론

음성 인식 기술은 컴퓨터 등 기계가 사람의 말을 알아들을 수 있도록 하는 기술이다. 1950년대에 나타난 최초의 모음인식 시스템[1]에서 시작해서 지난 60여년간 꾸준한 개발로 지속적인 발전을 이루어왔다[2]. 최근 스마트폰이 일반인에게 급속히 확산되기 시작하면서 음성 인식 기능에 대한 요구가 빠르게 증가하고 있다. 스마트폰의 경우 PC와는 달리 문자 입력 수단이 불편하며 대부분 이동 중 활용하기 때문에 음성 인식과 같은, 손의 사용을 최소화 시켜주는 입력 기능이 절실히 필요한 환경이기 때문이다[3].

음성 인식 시스템에서 인식의 기본 단위로는 단어, 어절, 음소 등이 있으나, 인식단위를 음소로 할 경우, 한국어에 있어서는 겨우 30여 개의 음소만이 존재하기 때문에 구별되는 음소 수가 한정되어 있고, 어휘 수의 증가에 따라 기억용량이나 계산량의 증가가 거의 없게 된다[4-6]. 음향적인 특성을 인식기에 고르게 반영할 수 있기 때문에 현재에는 대용량 어휘인식시스템, 화자독립 인식 시스템 등 대부분의 음성인식 시스템이 인식의 기본단위를 음소로 하고 있다[7]. 그러나 발성자의 습관이나 발음의 변화 등 여러 가지 요인으로 인해 각각의 음소들을 정확하게 인식하는 것은 무척 어려우며, 일관성이 보장되지 않는 등의 문제점을 안고 있다[8].

따라서 본 논문에서는 음소를 인식의 기본단위로 하는 소규모 음성인식 시스템을 구현하기 위해 그 일부로써 한국어 음소 중 ‘ㅅ’, ‘ㅈ’, ‘ㅊ’을 인식하고자 한다. 국어음운학에서 파찰음에는 ‘ㅈ’, ‘ㅊ’, ‘ㅊ’ 3개만 있으나 이는 조음방법에 의거한 분류이고, 본 논문에서는 아날로그 파형에 기초하여 ‘ㅅ’의 음성 파형을 분석한 결과, 파찰음과 비슷하다는 점에 착안하여 ‘ㅅ’, ‘ㅈ’, ‘ㅊ’만을 묶어서 인식하고자 하는 것이다. 그리고 파형에서 일정한 간격으로 부호 변화가 발생한다는 특징을 발견하여 부호 분포 변동성이라는 새로운 지표를 제안한다.

본 논문은 총 5장으로 구성되어 있다. 2장에서 기존 연구에 대해 살펴본다. 3장에서는, 파찰음의 음성학적 특징에 대하여 알아보고, 본 논문에서 제안하는 ‘ㅅ’, ‘ㅈ’, ‘ㅊ’ 인식 방법을 기술한다. 4장에서는 실험을 통해 제안하는 방법의 성능을 평가하고, 마지막으로 5장에서 결론을 맺는다.

2. 관련 연구

음성의 경우 동일한 단어라도 발성자의 습관이나 성별, 지속 시간의 차이, 그리고 조음 환경에 따른 조음 결합 현상 등의 영향으로 동일한 단어라도 그 특성이 다르다. 이러한 특성 때문에 컴퓨터가 음성을 정확히 인식하는 것은 매우 어렵다[3,4,9]. 현재까지 개발된 음성 인식 시스템은 부분적으로 음성을 어느 정도 인식하는 수준에 머물러 있으며, 인식된 음성을 분석하는 영역만이라도 높은 성능을 나타내도록 하는 연구가 활발히 이루어지고 있다[4,5].

그러나 본 논문에서는 음성을 정확히 인식하는 것에 초점을 맞추어 음소 단위 기반의 고립 단어 음성 인식을 대상으로 한정시켜 인식의 정확도를 높이기 위한 단계별 연구의 첫 단계로 고립단어를 대상으로 ‘ㅅ’, ‘ㅈ’, ‘ㅊ’ 인식을 하게 되었다.

기존의 논문에서는 파찰음을 인식하는 방법에는 ZCR (Zero Crossing Rate)과 RMSE(Root Mean Square

Error)를 이용한 방법이 있으며[8,10], LPC(Linear Prediction Coding) 계수를 이용하여 파열시점을 검출하는 방법이 있다[11].

ZCR은 음성 신호 파형의 위상이 중심 축을 통과하는 횟수이고[12], RMSE는 음성신호의 에너지 검출 방식 중 하나로 음성 신호의 크기를 제공하여 더하고 이를 평균한 다음 제곱근을 구한다. 무성음의 경우, 진폭이 크지 않고 불규칙적인 진동이 계속 있어 ZCR이 크며 RMSE는 모음보다 작고 다른 자음보다는 크다. 하지만 이 두 방법은 유성음이나 다른 자음에서는 효율적이지 몰라도 무성음에 속하는 파찰음과 파열음, 마찰음을 구분하는 데에는 변별력이 없어 비효율적이다[8,10].

3. ‘ㅅ’, ‘ㅈ’, ‘ㅊ’ 인식

한국어 마찰음으로는 ‘ㅅ’, ‘ㅆ’, ‘ㅎ’, 파찰음으로는 ‘ㅈ’, ‘ㅊ’, ‘ㅊ’이 있다. 마찰음은 폐강에서 나온 공기가 구강의 좁아진 어느 한 부분을 통과하면서 마찰시키는 언어음이고, 파찰음은 두 조음기관을 폐쇄시켜 공기를 압축하였다가 공기를 일시에 파열시키지 않고 천천히 내보내면서 마찰시키는 언어음이다[13]. 즉, 파찰음은 조음점부분을 혀로 막은 다음, 숨을 그쳐 이를 터트리면서 발성을 시작하게 되며, 치아에서 강한 마찰작용이 생겨나기에 마찰음의 성질을 가지고 있다[11]. 따라서 마찰음 및 파찰음의 파형을 분석하면 전 구간에 걸쳐서 일정한 간격으로 중심을 교차한다는 동일한 특징이 있기에 본 논문에서는 ‘ㅅ’, ‘ㅈ’, ‘ㅊ’을 묶어서 이들을 인식하고자 하였다. 이후에 파찰음이라고 작성된 부분은 편의상 사용된 것이며, 인식하고자 하는 음소인 ‘ㅅ’, ‘ㅈ’, ‘ㅊ’을 의미한다.

파찰음은 공기를 일단 막았다가 터뜨리면서 내는 파열음 ‘ㄱ’, ‘ㄷ’, ‘ㅂ’의 성질을 가지고 있기에 비슷한 파형을 가진다. 예를 들면, 그림 1은 발음 “금촌”에 대한 원본 wav 파형인데, (a)는 ‘ㄱ’에 해당하고 (b)는 ‘ㅊ’에 해당한다. 두 음소의 파형이 비슷하게 조밀한 간격으로 진동이 계속 있어 이 두 파형을 구별해내기 어렵다.

본 논문에서 제안하는 파찰음 인식 기법을 요약하면 다음과 같다. 우선 음성 신호에서 추출된 파형을 정수값으로 변환하고, 잡음구간에서는 정수값 변동이 작다는 것을 이용하여 변동성 지표를 통해 음성이 입력된 구간을 검출한다. 그런 후, 파찰음을 인식하기 위하여 일정한 간격으로 부호 변화가 발생한다는 것에 착안하여 부호 분포와 부호 분포 변동성이라는 새로운 지표를 제안하였으며, 이를 이용해 파찰음과 파열음을 구분하고자 하였으며 새로운 지표의 성능을 평가하기 위해 기존의 파찰음 인식방법인 영교차율(ZCR)을 사용하였다. 영교차율의 경우 단순히 0을 교차하는 횟수만을 계산하여

안드로이드 NDK(Native Development Kit)를 이용한 3D 큐브 게임 이식 및 구현

고은별*, 김녹희*, 황성미*, 이종우**

요 약

휴대폰 사용자들 대부분이 스마트폰으로 옮겨가면서 사용자들은 이제 스마트폰으로 다양한 게임을 즐기고 있다. 그런데 현재 앱스토어에 기기의 성능을 고려한 2D 게임은 다양하게 올라오고 있지만 3D 게임의 수는 아직도 많이 부족한 것이 실정이다. 이는 3D 게임이 많은 컴퓨팅 성능을 필요로 하기 때문인데, 본 논문에서는 안드로이드에서 NDK를 이용하여 3D큐브 게임을 구현하였다. 기존 비주얼C++ 일리히트 3D 큐브 소스를 NDK를 이용해 안드로이드 시스템 라이브러리로 이식하고 편의 기능을 추가하는 방식으로 구현하였다. NDK를 이용했으므로 기존 C++ 코드의 대부분이 자바 코드로 변환되지 않고 그대로 실행될 수 있어서 실행 시험 결과 비교적 구형 스마트폰 상에서도 3D 그래픽 동작들이 무난하게 이루어짐을 확인할 수 있었다.

키워드 : 안드로이드, NDK, 3차원 큐브 게임, 이식

Porting and Implementation of a 3D Cube Game using Android NDK(Native Development Kit)

Eunbyul Koh*, Nokhee Kim*, Sungmi Hwang*, Jongwoo Lee**

Abstract

Almost all the mobile phone users already moved or are now moving away to smartphones for their various applications like games. If we are to speak about game applications, due to the performance limits of smartphones, 2D games are predominant over 3D games in every app. store. In this paper, we implement a 3D cube game application by porting an existing visual c++ irrlicht cube application to android platform library using the android Native Development Kit. After the porting is done, we add a few new features for more fun. Because the android NDK makes the existing C/C++ codes run directly on the android operating systems, we found by real execution tests that our 3D cube app. is well executed on a low-end android smartphone without any performance problem.

Keywords : Android, NDK, 3D Cube, irrlicht, Porting

1. 서론

스마트폰이 출시되면서 사람들의 생활중심은

하루가 다르게 스마트폰으로 옮겨 가고 있으며 심지어 어린이 장난감까지도 모바일 기기로 나오는 추세이다. 또한 하루가 멀다 하고 출시되는 핸드폰기와 업데이트되는 운영체제 버전으로 인해 사용자들의 그래픽 응용에 대한 요구도 날로 높아지고 있다.

하지만 스마트폰에서 지원할 수 있는 그래픽 기능은 아직까지 제한되어 있고, 또한 고품질의 그래픽 화면을 구현하기에는 여전히 많은 한계가 있다[1]. 안드로이드 앱은 보통 객체지향 언어인 자바(JAVA)로 개발되는데 자바는 웹 활용

※ 교신저자(Corresponding Author): Jongwoo Lee
접수일:2013년 09월 24일, 수정일:2013년 09월 30일
완료일:2013년 09월 30일
* 숙명여자대학교 멀티미디어과학과
** 숙명여자대학교 멀티미디어과학과 교수
Tel: +82-2-710-9952, Fax: +82-2-710-9704
email: bigrain@sm.ac.kr

능력은 뛰어나지만 게임을 구현하기에는 처리 성능 측면에서 많은 한계를 가지고 있다.

스마트폰용 OpenGL/ES는 PC용 3D 그래픽스 엔진인 OpenGL의 서브셋으로 스마트폰이나 PDA 등과 같은 임베디드 기기를 위한 API 이다. 따라서 OpenGL에 비해 지원하는 API 타입의 수가 적고 즉시 모드를 지원하지 않으며 텍스처 맵핑은 2D만 지원하는 등의 단점이 있다.

그래서 본 논문에서는 안드로이드 NDK(Native Development Kit)를 이용하여 목표하는 게임을 구현하였다. NDK를 이용하면 JNI(Java Native Interface)를 통해 안드로이드 커널 위의 C 라이브러리를 이용할 수 있어 성능이점이 있기 때문이다. 또한 OpenGL/ES의 단점을 보완하면서도 상대적으로 가벼운 안드로이드용 그래픽 라이브러리인 irrlicht[2]를 사용하였다. Irrlicht는 C++ 다중 플랫폼 3D 그래픽 라이브러리로써 게임이나 과학적 시각, 2D/3D 앱을 위한 API를 지원한다.

본 논문에서는 안드로이드 NDK와 irrlicht 라이브러리를 이용해 ‘똑똑이 큐브’라고 이름 지은 3D 큐브 게임을 구현하고 그 구현 방법을 제안한다. 본 논문의 구성은 다음과 같다. 2장에서는 현재 관련기술 동향을, 3장에서는 구현 환경과 구현 과정을, 4장에서는 위에서 설명한 기술들을 이용하여 구현한 똑똑이 큐브 게임의 실행화면을 기술한다. 마지막으로 5장에서는 결론을 맺는다.

2. 관련 기술 및 관련 연구

2.1 안드로이드 NDK 동향

안드로이드 NDK는 안드로이드 애플리케이션(이하 앱)에서 C 와 C++ 같은 네이티브 코드를 사용할 수 있게 하는 도구이다. 즉 JNI를 통해서, SDK가 기본적으로 지원하지 않은 기능들을 추가 개발해서 상위 자바 응용 단에서 쓸 수 있게 해줌으로써 게임과 같이 속도에 민감한 앱을 개발하는 데에 도움을 주는 도구인 셈이다.

하지만 현재 NDK를 활용한 앱 개발은 그리 활발한 편은 아닌데, 그 이유는 비교적 배우기 쉬운 자바 외에 C/C++ 언어를 알아야 한다는 점과 가상머신 상에서 수행되는 앱의 성능 저하를 극복하고자 하는 필요성이 있어야 하기 때문인 것으로 판단된다. 또한, NDK를 사용하여 고품질 그래픽 요소를 지원하는 모바일 게임 앱을 제작하려 해도 모바일 기기의 성능 자체가 따라주지 못한다면 상용화하는데 큰 걸림돌로 작용하기 때문일 것이다.

하지만 하루가 다르게 발전하는 모바일 기기의 성능은 최근 들어 노트북 못지않은 성능과 메모리 크기를 가지게 되었다. 또한 하루에도 800만개의 앱이 쏟아지는 요즘, 사용자들의 고품질 그래픽 앱에 대한 요구도 날로 증가 하고 있다 하겠다.

본 연구를 진행했던 2012년 3월 ~ 7월 말 동

<표 4> NDK 버전별 차이점 요약

ndk r3	ndk r4	ndk r7
ndk폴더/apps/폴더/프로젝트 배치	프로젝트 배치 상관없음	
폴더에 Application.mk작성 필수	Cygwin/home/(사용자계정)/.bashrc파일 마지막에 Export ANDROID_NDK_ROOT=/home/(사용자 계정)/(팩키지버전) 추가	
make APP= 명령 사용	ndk-build 명령 사용	
r3와 r4,r7의 차이는 ndk-build 명령문의 차이 r4,r7에서도 make APP 명령을 쓰기 위해서는 ndk폴더/apps/폴더/프로젝트 배치하고 폴더에 Application.mk작성 필수 ndk-build 명령문의 등장으로 ndk-build명령을 프로젝트 배치에 상관없이 사용가능		
		Application.mk에 APP_PLATFORM := android-7 추가해야 make APP명령 사용가능

<Table 1> Summary of differences between NDK versions

중복글자 구분을 지원하는 집합 기반 POI 검색 알고리즘 구현

고은별*, 이종우**

요 약

집합 기반 POI 검색 기법은 부정확한 질의어 입력 시 검색결과와 재현율과 정확도가 현저히 떨어지는 기존 하드매칭 기법에 비해 우수한 성능을 보인다. 하지만 집합 개념을 바탕으로 했기 때문에 한 POI 레코드 내에 중복으로 포함된 동일 글자들을 구분하지 못하는 문제점이 있다. 본 논문에서는 이러한 문제를 해결하여 한 POI 내에 동일 글자가 여러 개 등장해도 동작하는 개선된 집합 기반 POI 검색 기법을 제시하고 구현하였다. 또한 개선된 집합 기반 POI 검색 기법의 검색결과와 기존 집합 기반 POI 검색 기법의 검색결과를 비교하는 실험을 통해 레코드 내에 중복으로 포함된 동일 글자가 있는 레코드에 대한 검색 성능이 향상되었음을 확인하였다.

키워드 : POI 데이터베이스, POI 검색 알고리즘, 집합-기반 알고리즘, 중복 글자

Implementation of A Set-based POI Search Algorithm Supporting Classifying Duplicate Characters

Eunbyul Ko*, Jongwoo Lee**

Abstract

The set-based POI search algorithm showed better performance than the existing hard matching search when inaccurate queries are entered. In the set-based POI search algorithm, however, there is a problem that can't classify duplicate characters within a record. This is due to it's 'set-based' search property. To solve this problem, we improve the existing set-based POI search algorithm. In this paper, we propose and implement an improved set-based POI search algorithm that is able to deal duplicate characters properly. From the experimental results, we can find that our technique for duplicate characters improves the performance of the existing set based POI search algorithm

Keywords : POI database, POI search algorithm, Set-based algorithm

1. 서론

스마트폰이 대중화되면서 위치 기반 서비스를

이용한 지도/네비게이션 어플리케이션이 활성화되고 있다.[1] 과거 지도/네비게이션에 대한 활발한 연구의 결과로[2] 현재는 언제 어디서나 가고자 하는 목적지를 입력하면 원하는 정보를 얻을 수 있다. 또한 추출된 정보를 사용자가 이해하기 쉽게 전달하기 위한 인터페이스 연구도 활발하다.[3,4,5] 하지만 사용자가 목적지의 명칭을 정확하게 입력하지 못하는 경우가 많아 목적지 검색의 재현율과 정확률이 떨어진다. 전통적인 하드매칭 기법에서는 이런 부정확한 질의어를 제대로 처리하지 못해 검색 서비스의 성능 저하 문제를 야기한다. 이 문제를 해결하기 위해 집합 기반 POI 검색

※ 교신저자(Corresponding Author): Jongwoo Lee
접수일:2013년 11월 18일, 수정일:2013년 12월 20일
완료일:2013년 12월 25일

* 숙명여자대학교 멀티미디어학과 학과

** 숙명여자대학교 멀티미디어학과 교수

Tel: +82-2-710-9952, Fax: +82-2-710-9704

email: bigrain@sm.ac.kr

■ 이 논문은 2013년도 정부(교육부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임 (2013R1A1A2013155)

기법[6]에서는 의미있는 실험결과를 보인 바 있다.

하지만 집합 기반의 POI 검색 기법은 집합 개념을 바탕으로 하기 때문에 한 레코드 내에 동일 글자가 2개 이상 있는 경우를 처리하지 못한다. 같은 글자가 레코드에 여러 개 포함되어 있어도 집합 개념 검색이므로 단 한 개만 포함되어 있는 것으로 처리하는 것이다. 그래서 이런 중복 글자가 있는 명칭을 질의어로 입력하면 검색 정확도가 현저하게 떨어진다. '동성자동차'처럼 같은 글자를 포함하고 있는 질의어가 입력될 경우 중복된 '동'을 구별하지 못하고 하나로 보기 때문에 '한성자동차', '성원자동차'같은 결과가 검색 일치로 나오게 된다. 문자열 유사도 측정을 통해 이 같은 문제를 일부 보완했지만 근본적인 해결책은 되지 못한다.[7,8]

이런 문제를 해결하기 위해 본 논문에서는 기존 집합 기반 POI 검색 기법을 보완한 중복 글자 검색을 지원하는 집합 기반 POI 검색 기법을 제안한다. 동일 글자가 한 레코드 내에 여러 번 등장할 경우 최대 등장 횟수만큼 글자 아이디를 더 생성한다. 동일 글자이지만 글자 별로 부여하는 아이디를 달리 하는 것이다. 레코드 내 중복을 반영한 글자 아이디는 역 인덱스 생성에 반영된다. 그래서 기존 집합 기반 POI 검색 기법에서는 인식하지 못했던 중복 글자들을 정확하게 인식하여 검색 성능을 향상시킨다.

본 논문의 구성은 다음과 같다. 2장에서는 기존 집합 기반 POI 검색 기법에 대해 설명하고, 3장에서는 기존 집합 기반 POI 검색 기법의 한계에 대해서 설명한다. 4장에서는 본 논문에서 제안하는 중복 글자 지원 집합 기반 POI 검색 기법을 제시하고, 5장에서는 본 논문에서 제시된 알고리즘의 실효성을 보인 뒤, 6장에서 결론을 맺는다.

2. 기존 집합 기반 POI 검색 기법 소개

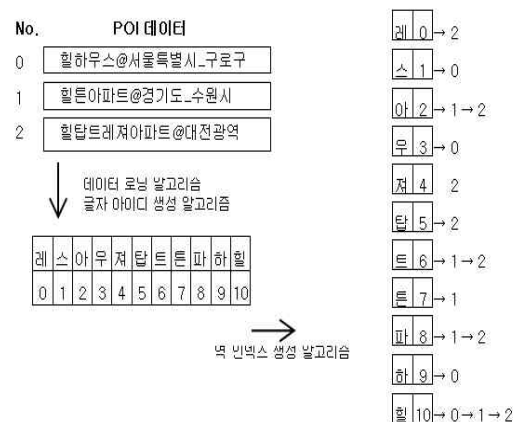
본 장에서는 기존 집합 기반 POI 검색 기법에 대해서 간략히 설명한다. 집합 기반 POI 검색

기법[6]이란 집합 개념을 적용하여 POI 데이터를 검색하는 기법으로, 전역적 쿼리 확장[9,10], 지역적 쿼리 확장[9,10], 적합성 피드백[11]과 같은 계산이 복잡하고 방대한 양의 데이터를 필요로 하는 기법을 사용하기 힘든 차량 내비게이션과 같은 독립형 시스템을 위한 기법이다. 부정확한 POI 질의어 입력으로 인한 POI 검색 서비스 성능 저하 문제를 시스템 자체 내에서 해결하기 위해 제시한 새로운 알고리즘이다. 검색 과정을 간략히 소개하면 다음과 같다.

기존 집합 기반 POI 검색 기법은 POI 디버로딩 알고리즘, 글자 아이디 생성 알고리즘, 역 인덱스 생성 알고리즘, 텍스트 검색 알고리즘으로 구성되며, 텍스트 검색 알고리즘은 전처리 과정, 블록 내 연산, 블록 간 연산으로 구성된다.

먼저, POI 디버로딩 알고리즘에서는 POI 데이터베이스를 로딩하며, 글자 아이디 생성 알고리즘에서는 레드-블랙 트리를 이용하여 로딩된 데이터베이스의 모든 글자를 중복없이 파악하고 이를 통해 각 글자의 아이디를 생성한다. 이 아이디는 역 인덱스 생성 알고리즘에서 사용된다. 역 인덱스에는 아이디가 부여된 각 글자가 포함되어 있는 POI 데이터베이스 내 레코드 번호를 저장한다. (그림 1)은 세 개의 레코드에 대해 글자 아이디와 역 인덱스를 생성하는 과정 예를 표현한 것이다.

(그림 1) 역 인덱스 생성 과정



(Figure 1) Steps for generating inverted indexes

중복글자 구분을 지원하는 집합 기반 POI 검색 알고리즘 설계*

고은별, 이종우**

숙명여자대학교 멀티미디어학과

e-mail : iks1614@naver.com, bigrain@sm.ac.kr

Design of A Set-based POI Search Algorithm
Supporting Classifying Duplicate Characters

Eunbyul Ko, Jongwoo Lee

Dept of Multimedia science, Sookmyung Women's University

요약

집합 기반 POI 검색 기법은 부정확한 질의어 입력 시 검색결과와 재현율과 정확도가 현저히 떨어지는 기존 하드매칭 기법에 비해 우수한 성능을 보인다. 하지만 집합 개념을 바탕으로 했기 때문에 한 POI 레코드 내에 중복으로 포함된 동일 글자들을 구분하지 못하는 문제점이 있다. 본 논문에서는 이러한 문제를 해결하여 한 POI 내에 동일 글자가 여러 개 등장해도 동작하는 개선된 집합 기반 POI 검색 기법을 제시한다.

1. 서론

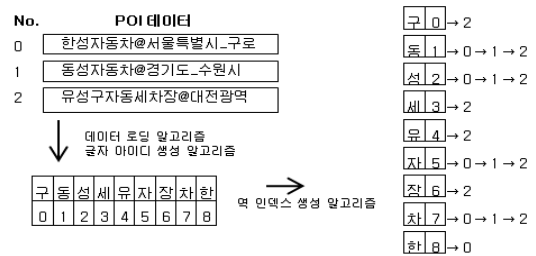
스마트폰이 대중화되면서 위치 기반 서비스를 이용한다. 지도/네비게이션 어플리케이션이 활성화 되고 있다[1]. 하지만 사용자가 목적지의 명칭을 정확하게 알지 못하고 부정확한 질의어를 입력할 경우 목적지 검색의 재현율과 정확도가 떨어진다. 이런 문제를 해결하기 위해 제안된 집합 기반 POI 검색 기법[2]은 오질의어에 대해 향상된 성능을 보이지만 POI 레코드나 쿼리에 동일 글자가 포함되어 있는 경우를 처리하지 못한다. 본 논문에서는 이를 수정 보완한 기법을 제시한다.

2. 기존 집합 기반 POI 검색 기법의 한계

집합 기반 POI 검색 기법[2]은 전역적 쿼리 확장[3], 지역적 쿼리 확장[3], 적합성 피드백[4]과 같은 계산이 복잡하고 방대한 양의 데이터를 필요로 하는 기법을 사용하기 힘든 차량 내비게이션과 같은 독립형 시스템을 위한 기법이다. n개의 문자로 이루어진 POI 질의어를 m개의 블록으로 균등 분할한 뒤, 각 블록에 대해 집합 기반 연산을 적용하고, '차수'라는 개념을 사용하여 블록 간 집합 연산을 수행하는 기법이다. 실험을 통해 집합 기반 POI 검색 알고리즘이 기존 하드매칭 기법에 비해 83%~96%우수함을 확인하였다[2].

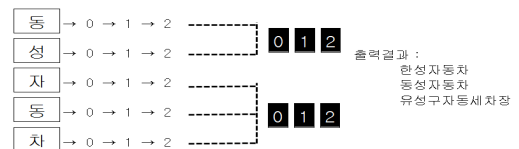
하지만 집합 개념을 바탕으로 했기 때문에 한 POI 레코드 내에 중복으로 포함된 동일 글자들을 구분하지 못하는 문제점이 있다. 그림 1은 한 POI 내 동일 글자를 구분하지 못하는 기존 집합 기반 POI 검색 기법에서 POI 디비

에 대해 역 인덱스를 생성하는 과정을 보이고 있다.

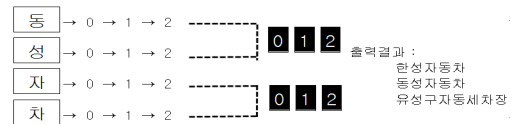


<그림 1> 기존 집합 기반 POI 검색 역 인덱스 생성 과정

그림 2는 이렇게 생성된 역 인덱스를 바탕으로 질의어 '동성자동차'와 질의어 '동성자동차'의 검색결과를 비교한 것이다.



a) 질의어가 '동성자동차'일 경우



b) 질의어가 '동성자동차'일 경우

<그림 2> 질의어 '동성자동차'와 '동성자동차'의 검색결과 비교

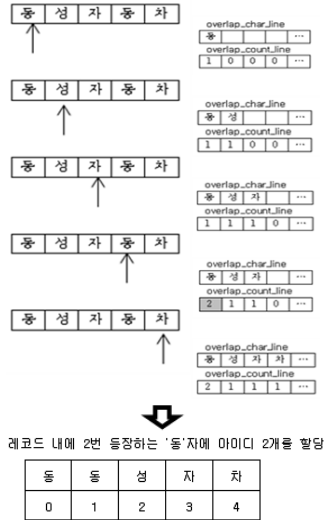
기존 기법에서는 질의어 '동성자동차'의 첫 번째 '동'과 두 번째 '동'을 구분하지 못하기 때문에 검색결과와 차이가 없다. 기존 기법에서는 질의어 내에 같은 글자가 2번 등장할 경우 검색 정확도가 현저히 떨어지는 것이다.

* 이 논문은 2013년도 정부(교육부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임(2013R1A1A2013155)

** 교신저자

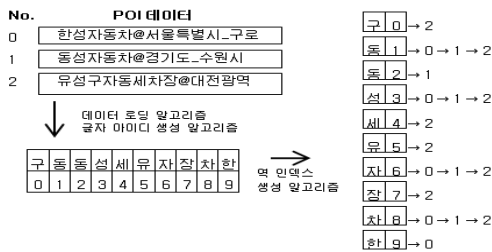
3. 개선된 집합 기반 POI 검색 기법

기존 집합 기반 POI 검색 기법의 문제를 해결하기 위해 글자 아이디 생성과 역 인덱스 생성 알고리즘을 중점적으로 개선하였다. 한 POI 레코드 내에 2번 이상 등장하는 글자에 아이디를 중복된 횟수만큼 추가로 부여한다. 기존 기법에서는 여러 개의 동일 글자에 같은 글자 아이디를 부여하였다. 이렇게 생성된 글자 아이디로 역 인덱스를 생성하면 레코드 내에서 2번 이상 중복으로 등장하는 글자를 고려한 검색을 수행할 수 있다. 그림 3은 개선된 알고리즘을 통해 글자 아이디를 생성하는 과정이다.



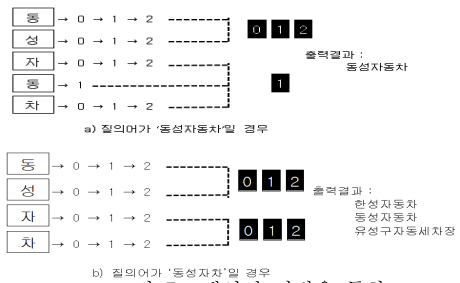
<그림 3> 개선된 글자 아이디 생성 과정

1번 레코드 '동성자동차'에서 중복되어 등장한 '동'자에 대해 글자 아이디를 2개 부여한다. 그림 4는 이렇게 생성된 글자 아이디로 역 인덱스를 생성하는 과정이다.



<그림 4> 개선된 집합 기반 POI 검색
역 인덱스 생성 과정

그림 5는 개선된 집합 기반 POI 검색 기법을 통해 질의어 '동성자동차'와 '동성자차'의 검색결과를 비교한 것이다. 기존 집합 기반 POI 검색 기법에서는 검색결과가 차이가 없었지만 개선된 기법에서는 검색결과가 차이가 분명함을 알 수 있다.



<그림 5> 개선된 기법을 통한

질의어 '동성자동차'와 '동성자차'의 검색결과 비교

5. 실험결과

본 논문에서 제시한 개선된 집합 기반 POI 검색 알고리즘의 성능을 평가하기 위해 실험을 수행하였다. 질의어는 2번 이상 등장하는 글자를 적어도 1개 이상 포함한 단어로 무작위로 추출했다. 검색 성공 여부는 정답 데이터의 랭킹을 기준으로 판단했다. 상위 5위안에 들 경우 성공, 상위 20위 안에 들 경우 준성공, 그렇지 않은 경우에는 실패라 간주한다. 표 1은 실험결과를 정리한 것이다.

<표 1> 실험결과

구분	총	성공		준성공		실패	
기존 검색 기법	50	11	22%	13	26%	26	52%
개선된 검색 기법	50	40	80%	9	18%	1	2%

수행한 실험을 통해 개선된 집합 기반 POI 검색 기법이 기존 집합 기반 POI 검색 기법보다 중복글자가 포함된 경우 검색 성공률과 효율이 향상되었음을 알 수 있다.

6. 결론

기존 집합 기반 POI 검색 기법은 오질의어에 취약한 하드매칭 기법에 비해 정확한 검색결과를 출력하여 성능을 향상시켰다. 하지만 집합 개념을 기반으로 했기 때문에 동일 글자가 2번 이상 등장하는 질의어나 레코드에 대한 처리가 미흡했다. 이 문제를 해결하기 위해 본 논문에서는 개선된 집합 기반 POI 검색 기법을 제안하였다.

개선된 집합 기반 POI 검색 기법에서는 한 레코드 내에 2번 등장하는 중복 글자를 구분해서 글자 아이디를 부여하고, 이를 통해 중복 글자를 인식할 수 있는 역 인덱스를 생성한다. 이 역 인덱스를 통해 중복 글자가 있는 질의어가 들어와도 정확한 정보를 출력할 수 있었고, 이를 실험으로 평가하였다.

집합 기반 POI 검색 알고리즘을 활용한 In-시스템 검색 엔진 설계*

안혜영, 이종우**

숙명여자대학교 멀티미디어학과

e-mail : hyeyeong@sookmyung.ac.kr, bigrain@sookmyung.ac.kr

Design of An In-System Search Engine Using Set-based POI Search Algorithm

Hyeyeong Ahn, Jongwoo Lee

Dept of Multimedia Science, Sookmyung Women's University

요 약

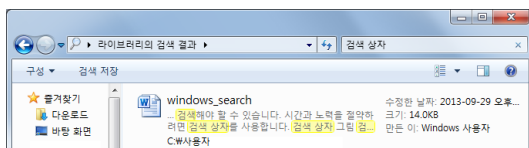
기존 OS의 파일 검색 기능에서는 사용자가 입력한 키워드와 완벽하게 일치하는 파일만을 검색하여 결과로 보여준다. 따라서 오타자나 외래어와 같이 표기가 모호한 검색어 등 부정확한 쿼리에 대한 처리에 취약하다. 이에 본 논문에서는 집합 기반 POI 검색 알고리즘을 활용하여 검색 엔진을 설계하여 파일 검색에 대한 검색 서비스의 성능을 개선하고자 한다.

1. 서론

윈도를 비롯한 다양한 운영체제는 방대한 양의 파일에 대한 검색 서비스를 제공한다. 하지만 사용자가 입력한 문자열과 완벽히 일치하는 파일 또는 내용만 검색이 가능하므로 오타자나 외래어 등 부정확한 쿼리에 대한 처리가 부족하다. 이에 본 논문에서는 기존 집합 기반 검색을 활용하여 다양한 오인식 쿼리에 대하여도 사용자가 원하는 검색 결과를 보여줄 수 있도록 하고자 한다.

2. 관련 기술

윈도와 리눅스는 폴더 또는 파일의 이름뿐만 아니라 파일 내용에 대한 검색을 지원한다. 윈도의 경우 폴더 내의 검색 상자에 찾고자 하는 파일의 이름 또는 내용을 입력하여 검색한다(그림 1). 이 때 검색 필터를 사용하여 여러 개의 검색어를 모두 포함하는 경우 또는 여러 검색어 중 하나만 포함하는 경우 등 다양한 검색 조건을 적용할 수 있다. 하지만 파일 내용 검색의 경우, 검색 대상 파일 형식이 MS Office에서 사용하는 파일 등으로 제한된다.[1]



(그림 1) 윈도 환경에서 파일 검색 화면

* 이 논문은 2013년도 정부(교육부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임(2013R1A1A2013155)

** 교신저자

리눅스의 경우, 주로 명령어를 통해 검색을 실행한다. find 사용 시 파일명 또는 폴더 명으로 검색을 수행한다. 파일 내용으로 검색하기 위해선 'grep' 명령어를 사용한다. grep은 텍스트 검색 명령어로 사용자가 입력한 정규 표현식을 기준으로 검색을 수행한다.[2]

두 운영체제 모두 다양한 조건으로 파일 검색을 할 수 있어 검색의 편의성을 제공하고 있지만, 사용자가 입력한 키워드와 정확하게 일치하는 파일만을 결과로 보여주므로 사용자가 검색하고자 하는 키워드를 정확하게 알고 있어야만 파일 검색이 가능하다. 예를 들어, 외래어의 경우에는 표기를 여러 형태로 할 수 있는데, 'contents'란 영단어를 '콘텐츠' 또는 '컨텐츠' 등으로 표기할 수 있다. '콘텐츠'란 단어를 포함하고 있는 문서를 찾기 위해 '컨텐츠'로 검색했을 경우, 존재하지 않는 파일이라는 검색 결과를 보여줄 것이다.

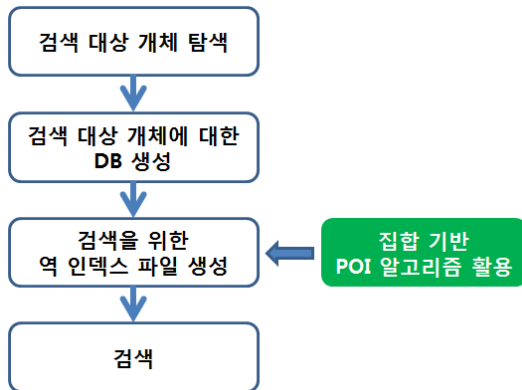
3. 집합 기반 POI 검색 알고리즘

집합 기반 POI 검색 알고리즘[3]은 부정확한 POI 쿼리로 인한 검색의 성능 저하 문제를 해결할 수 있는 알고리즘이다. n개의 문자로 이루어진 POI 쿼리를 m개의 블록으로 균등 분할한다. 각 블록에 대해 집합 기반 연산을 적용 후 '차수'라는 개념을 사용하여 블록 간 집합 연산을 수행한다.

이 알고리즘은 사용자의 부정확한 쿼리를 알고리즘 차원에서 재구성하므로 외부 자원을 사용하지 않고도 잘 못된 입력에 대한 처리가 가능하다. 이는 검색 서비스 성능 저하를 개선하고 검색 성공율을 향상시킬 수 있다.

4. 내부 검색 시스템 설계

기존 OS의 검색 시스템의 문제점을 개선하기 위하여 집합 기반 알고리즘을 활용한 새로운 검색 시스템을 설계하고자 한다. 그림 2와 같이 내부 검색 시스템을 위한 데이터베이스를 만드는 과정, 생성된 DB의 역 인덱스를 만드는 과정, 입력된 쿼리의 전처리 과정, 검색 과정까지 각각에 대하여 설명한다. 시스템의 원활한 설계를 위해 사용 OS는 리눅스 기반의 Ubuntu로 가정한다. 검색 대상 파일은 사용 빈도가 높고, 쿼리를 통해 쉽게 검색할 수 있도록 텍스트 문서 파일로 제한한다.



(그림 2) In-시스템 검색 엔진의 검색 과정

4.1. 검색 대상 개체 탐색 및 DB 생성

데이터베이스를 생성하기 위해 파일 시스템을 탐색하는 과정과 이를 데이터베이스에 저장하기 위한 시스템 프로그래밍이 필요하다. 먼저 파일 시스템을 탐색하며 대상 파일을 찾기 위해 리눅스의 find 명령어를 사용한다. 파일 시스템은 그 내용이 자주 변경되기 때문에 주기적으로 내용을 갱신하기 위하여 find 작업을 crontab에 등록하여 일정 시간에 실행하도록 한다. find 명령어를 통해 파일 시스템 내의 텍스트 파일을 찾고 파일 명과 경로를 수집한다. 이렇게 주기적으로 수집한 파일 시스템의 데이터로 데이터베이스를 만든다.

본 시스템에서 활용하는 POI 데이터베이스는 각각의 레코드가 장소의 이름인 명칭 부분과 주소인 주소 부분으로 이루어져 있으며, 이 둘을 구분자인 특수문자 '@'로 구분한다. 이를 OS 내부 검색 시스템에 적용하여 명칭 부분은 파일 내용으로, 주소 부분은 파일의 경로로 하여 레코드를 저장한다. 즉, '파일 내용'@'파일 위치'의 형태로 레코드를 생성한다.

4.2. 역 인덱스 생성 및 검색 과정

검색 시 사용되는 역 인덱스 파일을 만들기 위한 선행 작업으로 레코드의 파일 내용 부분에 해당하는 모든 글자에 아이디를 부여한다. 다음으로 모든 글자에 대하여 역

인덱스를 생성한다. 역 인덱스에는 각 글자가 등장하는 데이터베이스의 레코드 번호(줄 번호)가 글자 별로 저장된다. 역 인덱스 생성이 끝난 후 사용자로부터 쿼리가 입력되면 각 글자에 대한 아이디를 알아낸 후 역 인덱스를 이용하여 각 글자를 포함하고 있는 레코드의 개수와 번호를 파악한다. 이 과정이 끝나면 블록 내 연산으로 검색 결과를 추린 다음, 블록 간 연산을 통해 추가적인 제약을 가하여 검색 후보의 수를 축소한다.

기존 알고리즘과 비교하였을 때, 레코드의 명칭 부분이 파일 내용으로 확장됨에 따라 아래와 같은 추가적인 처리가 필요하다.

- 기존의 POI 데이터보다 크기가 증가하므로 더 많은 메모리의 사용이 요구된다.
- 파일 내용 안에서 중복되는 글자 또한 증가하므로 중복 글자 처리가 추가적으로 필요하다.
- 명칭 부분에는 없는 '~가(이), ~를'과 같은 어조사와 공백에 대한 처리가 필요하다.

5. 결론

윈도와 같은 OS는 다양하고 방대한 양의 파일을 포함하고 있다. 이러한 파일 시스템에서 원하는 파일을 찾기 위해서는 검색 서비스가 필수적이다. 하지만 사용자가 입력한 키워드와 완벽히 일치하는 파일만 검색 결과로 나오기 때문에 부정확한 쿼리에 대한 추가적인 처리가 필요하다.

본 In-시스템 검색 엔진은 짧은 단어 위주의 POI 데이터베이스를 확장하여 크기가 큰 파일 내용을 검색 대상으로 사용하기 때문에 메모리 문제 및 POI 데이터에는 없는 공백이나 특수문자와 중복되는 글자에 대한 처리가 기존 알고리즘에 추가적으로 필요하다.

따라서 구현 전 검색 엔진 설계를 통해 위와 같이 예상되는 문제점에 대하여 파악하였다. 본 설계를 In-시스템 검색 엔진의 토대로 삼아 문제점 해결에 주력하여 검색 엔진을 구현할 예정이다.

참고문헌

- [1] Microsoft Windows, "Windows 검색", <http://windows.microsoft.com/ko-kr/windows7/products/windows-search>
- [2] ubuntu documentation, "find", <http://help.ubuntu.com/community/find>
- [3] Eunbi Go, JaeWon Lee, Jongwoo Lee, An Efficient Set-based POI Search Algorithm [Online]. Available: <http://mm.sookmyung.ac.kr/~zinc/SetPOISearch.pdf>

개인 맞춤형 장학금 검색을 위한 알고리즘 연구

강다혜, 송지영, 최지수, 고은별, 이종우
숙명여자대학교 멀티미디어학과

e-mail : dahea_a@naver.com j0506i@naver.com aoipure@naver.com iks1614@naver.com bigrain@sm.ac.kr

Algorithm project for personalized scholarship search

Dahea Kang, Jiyeong Song, Jisu Choi, Eunbyul Koe, Jongwoo Lee
Dept of Multimedia science, Sookmyung Women's University

요 약

회원가입을 통해 자신의 학적 정보를 데이터베이스에 저장한다. 그 후 로그인을 통해 서버에 자신의 정보가 보내지면 장학금 알고리즘을 통해 자신의 장학 수혜 정보 및 추천 장학금을 알 수 있다. 웹과 모바일에서 모두 동작이 가능하여 언제 어디서나 사용자가 접속할 수 있도록 웹 접근성을 살렸다. 후에 자신의 학적 정보를 수정할 수 있다.

1. 서론

국가장학금과 같은 다양한 장학금 제도가 있지만 학생들이 미처 알지 못해 수혜 받지 못하는 경우가 적지 않다. 우리 프로그램은 학생들에게 자신이 받을 수 있는 장학금을 추천해주어 학생들의 장학금 수혜율을 높이고 가계부담을 덜어주고자 한다. 본 프로그램은 숙명여자대학교 내에서 제공하는 교내 장학금에 한정하여 제작되었다.

2. 관련연구



그림 1. 한국장학재단 웹 페이지 중 스마트 맞춤형

한국장학재단은 웹 페이지에서 국가장학금 신청, 학자금 대출 및 이자 안내, 장학금 안내 등을 지원하고 있다. 그 중 스마트 맞춤형 장학금 시뮬레이션은 개개인의 학생들에게 맞춤형 장학금 정보를 제공하고 있다. 사용자 본인이 입력한 학적, 성적, 개인 정보 등을 바탕으로 신청 가능한 학자금 대출이나 장학금 정보를 알려준다. 맞춤형 시뮬레이션에서는 수혜 가능한 장학금의 정보를 알 수 있지만 몇 가지 조건만 충족하면 받을 수 있는 후보 장학금은 알 수 없다.

한국장학재단이 운영하는 모바일 어플리케이션은 안드로이드와 애플 iOS에서 구동이 가능하다. 해당 모바일 어플리케이션은 장학금 신청 현황 조회나 수혜 현황 조회 등이 가능하다. 또 기업 및 민간재단의 장학금 정보 검색이 가능하다. 하지만 가장 핵심적인 기능인 맞춤형 시뮬레이션의 기능이 부재되어있다.

3. 설계

DB sookolarship

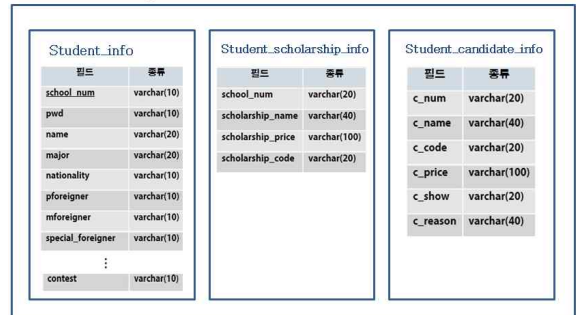


그림 2. 데이터베이스 및 테이블

데이터베이스 Sookolarship은 Student_info, Student_scholarship_info, Student_candidate_info, 이렇게 세 가지 테이블을 가진다. student_info는 사용자의 정보를 저장한다. Student_scholarship_info는 수혜 장학금의 정보를, Student_candidate_info는 후보 장학금 정보를 저장한다.

검색 알고리즘은 장학금의 종류별로 제작하였으며 크게 수혜 장학금 알고리즘과 후보 장학금 알고리즘으로 나뉜다.

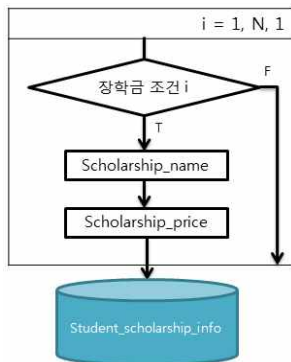


그림 3. 수혜 장학금 알고리즘

수혜 장학금 알고리즘은 장학금의 모든 조건을 만족할 때 검색되는 알고리즘이다. 사용자가 회원 가입 시 입력한 정보를 토대로 모든 수혜 장학금 알고리즘을 검사하고, 조건을 만족하는 장학금을 보여주게 된다.

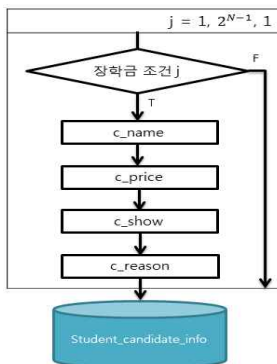


그림 4. 후보 장학금 알고리즘

후보 장학금 알고리즘은 장학금 조건 중 몇 가지 조건만 만족하는 경우 검색되는 알고리즘이다. 사용자가 입력한 정보를 후보 장학금 알고리즘으로 검사하여 장학금 명, 장학금 액수, 현재 조건 만족 상태, 그리고 부족한 조건을 알려준다.

금액에 따라 99부터 85까지의 코드 값을 부여하고 출력시에 데이터베이스 SQL문을 통하여 sorting하였다.

4. 구현

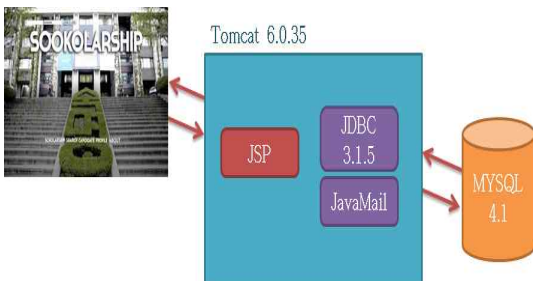


그림 5. 개발 환경 도식화

본 프로그램은 JSP(Java Server Page) 웹 프로그래밍 언어를 사용하여 구현하였다. 또한 HTTP와 같은 프로토콜에 따라 클라이언트의 요청을 처리하고 응답하기 때문에 로그인과 로그아웃 처리가 용이하다. JSP는 웹브라우저의 요청에 따라 알맞은 HTML 결과 화면을 생성한다.

웹페이지를 구동하기 위한 서버로 Apache Tomcat 6.0.35를 사용하였다. Database로는 MYSQL 4.1을 활용하였고 JSP페이지와 Database의 연결을 위해 커넥터로써 JDBC 3.1.5를 사용하였다. 또, 관리자에게 문의하기 메일을 보내는 기능 구현을 위하여 SMTP프로토콜과 함께 JavaMail 라이브러리를 프로젝트에 사용하였다.

5. 결론

본 논문에서는 숙명여대 학생들에게 맞춤형 장학금 정보를 제공하기 위해 알고리즘을 설계하고 구현하였다. 프로젝트의 주 기능인 검색 알고리즘은 수혜 장학금 검색 알고리즘과 후보 장학금 검색 알고리즘 나뉘어 설계 구현했다. 데이터베이스에는 미리 장학금 이름, 액수, 정보 등을 저장해놓고 알고리즘을 사용해 검색했을 때 기존의 데이터베이스에서 불러와 출력했다. 부가적인 기능으로 관리자에게 메일 보내기 기능이나 사용자의 이해를 돕기 위해 시각화하는 작업도 함께했다. 또 서버와 JSP페이지를 연결해 어디서나 장학금을 쉽게 검색할 수 있으며 웹과 모바일에서 모두 같은 기능으로 동작한다.

본 연구에서는 회원가입을 통해 사용자의 정보를 일일이 입력해야한다. 앞으로는 학교 학생 정보 시스템과 연동을 통해 간단한 인증을 거치면 학적 정보와 성적 정보를 받아올 수 있게 해 편리하게 검색할 수 있는 방법을 생각해보고 개선할 것이다. 또한 제공되는 장학금의 범위를 확대하여 교내뿐만 아니라 교외의 다양한 장학금 추천도 가능하도록 할 예정이다.

참고문헌

- [1] 김윤명, “뇌를 자극하는 JSP SERVLET”, 한빛미디어, 2010.
- [2] 한연희, 손진곤, “SERVLET 포함한 JSP 2.1 웹 프로그래밍”, 이한출판사, 2011.
- [3] 이지선, “JSP PROGRAMMING WITH A WORKBOOK”, 이한출판사, 2007.
- [4] 김은옥, “은노기의 JSP 웹 프로그래밍 입문”, 삼양미디어, 2008.
- [5] “교내장학금”, 숙명여자대학교 장학팀, 2013,
<http://scms.sookmyung.ac.kr/wiz/wizard/frames/server_index.html?home_id=service&menu_seq=23&menu_seq_open=>>
- [6] “학생 중심 교육 시스템으로 ‘여풍당당’ 인재 배출”, 중앙일보, 2012년 12월 12일.

HTML5를 이용한 멀티미디어 엽서 하이브리드 어플리케이션 구현

김지현, 허서윤, 고은별⁰, 이종우*

숙명여자대학교 멀티미디어학과

e-mail : kimjihhi@gmail.com, seoyoon.heo@gmail.com, iks1614@naver.com, bigrain@sm.ac.kr

Implementation of A Multimedia Post-mail Hybrid Application Using HTML5

Jihyun Kim, Seoyoon Heo, Eunbyul Koe, Jongwoo Lee

Dept of Multimedia Science, Sookmyung Women's University

요 약

기존의 카드 어플리케이션은 대부분 이미지만 전송할 수 있었다. 본 논문에서는 이미지뿐만 아니라 영상, 이미지, 음성, 텍스트 등의 멀티미디어 콘텐츠를 전송할 수 있는 카드 어플리케이션을 제안한다. 개발한 앱을 사용하면 사용자가 멀티미디어 콘텐츠를 포함한 카드를 직접 제작하고 저장한 후 다른 사용자에게 전송할 수 있다. 또한, 모바일 OS 종류에 관계없이 사용할 수 있도록 하기 위해 HTML5와 폰갭 라이브러리를 이용한 하이브리드 어플리케이션 형태로 개발하였기 때문에 적은 개발비용으로 다양한 모바일 OS에서 이용할 수 있다.

1. 서론

스마트 기기가 보편화됨에 따라 모바일 어플리케이션을 이용하여 카드를 보내게 되었다. 하지만 현존하는 타 카드 어플리케이션은 단순히 이미지만을 전송한다는 한계가 있다. 이미지, 동영상, 소리 등 다양한 멀티미디어 콘텐츠를 포함한 카드를 사용자가 직접 제작하여 전송할 수 있는 앱을 만들고자 한다. 응용프로그램 플랫폼으로 발전한 HTML5와 모바일 크로스 플랫폼인 PhoneGap을 이용하여 한 번의 개발로 모바일 OS와 상관없이 동일한 결과물을 볼 수 있는 하이브리드 카드 어플리케이션 구현 방법을 제시한다.

2. 관련연구

2.1. 하이브리드 앱

하이브리드 앱(Hybrid App)은 스마트폰으로 도메인을 입력해서 접속하는 '모바일 웹사이트'와 어플리케이션을 다운받아 접속하는 '앱'이 합쳐진 형태이다. 하이브리드 웹의 원리는 웹 표준을 준수한 모바일 웹을 만든 이후에 앱에 연동하는 형태이다. 따라서 어플리케이션을 다운 받거나 또는 북마크, 도메인을 입력해서 접속해도 동일한 형태의 모바일 페이지를 볼 수 있다. 보통 하이브리드 웹이라는 이름으로 통용되지만 앱을 설치하는 형태의 경우는 하이브리드 앱이라는 이름으로 사용한다.

하이브리드 앱의 장점은 OS별로 제작형식이 다르지

않고 HTML 국제 표준화 기구인 W3C에서 규격화한 'HTML5'를 사용한다는 점이다. 이 형식은 어느 OS, 플랫폼에서도 동일한 형태로 볼 수 있으며 별도의 플러그인 없이 플래시나 동영상, 멀티미디어 파일을 구동시킬 수 있어 매우 편리하다[1].

2.2. PhoneGap

PhoneGap은 웹 콘텐츠를 보여주는 웹뷰(WebView)에서 스마트폰 디바이스에 접근하여, 카메라나 마이크 제어 같은 기능을 실행하기 위한 인터페이스를 제공하는 대표적인 하이브리드 앱 개발 솔루션이다. 캐나다에 있는 Nitobi라는 회사에서 처음 제작했으며 2011년 10월에 Adobe사가 PhoneGap을 인수했다. Adobe사의 인수 후에 Apache 라이선스 적용으로 오픈소스 정책을 더욱 강화했으며, 각종 추가 기능이 플러그인 형태로 개발되어 커뮤니티에 공유되고 있는 것이 장점이다.

PhoneGap은 현재 하이브리드 앱 플랫폼 중 가장 큰 경쟁력을 가지고 있다. 가장 많은 스마트폰 플랫폼을 지원하고, 제품 완성도도 높기 때문이다[2].

PhoneGap에서 제공하는 기능은 표 1과 같다.

표 1 PhoneGap이 제공하는 기능

	iPhone 3GS	Android	BlackBerry OS 6.0*	WebOS	Windows Phone	Symbian	Bada
Accelerometer	✓	✓	✓	✓	✓	✓	✓
Camera	✓	✓	✓	✓	✓	✓	✓
Compass	✓	✓	×	✓	✓	×	✓
Contact	✓	✓	✓	×	✓	✓	✓
File	✓	✓	✓	×	✓	✓	✓
Geolocation	✓	✓	✓	✓	✓	✓	✓
Media	✓	✓	×	×	✓	×	×
Network	✓	✓	✓	✓	✓	✓	✓
Notification(Alarm)	✓	✓	✓	✓	✓	✓	✓
Notification(Notification)	✓	✓	✓	✓	✓	✓	✓
Storage	✓	✓	✓	✓	✓	✓	×

* 교신저자

※ 이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No.20120007897)

3. 설계

본 논문에서 구현한 어플리케이션은 엽서제작 인터페이스, 멀티미디어 콘텐츠 엽서 저장, 엽서 전송으로 구성되어 있다.

엽서제작 인터페이스는 사용자가 직접 엽서를 제작할 수 있는 도구를 제공한다. 영상 삽입, 이미지 삽입, 텍스트 삽입, 음성 녹음 등의 기능과 다양한 레이아웃을 제공한다.

멀티미디어 콘텐츠 엽서 저장은 사용자가 만든 엽서를 디바이스 내에 HTML형태의 파일로 저장한 뒤 이를 서버로 전송한다.

엽서 전송은 서버에 저장된 엽서를 URL을 사용하여 MMS, 카카오톡, 이메일 등으로 원하는 상대방에게 전송한다.

4. 구현

본 논문에서 제안한 인터페이스를 기본으로 하여 어플리케이션 프로토타입을 구현하였다. 하이브리드 어플리케이션이므로 모든 기본 언어는 HTML5를 사용한다.

4.1. 이미지 삽입

디바이스 내의 이미지를 불러오기 위해 PhoneGap의 Camera API를 이용하였다. PhoneGap의 Camera 객체를 통해 디바이스의 기본 카메라 어플리케이션에 접근할 수 있다. 이미지를 삽입한 화면은 그림 1과 같다.



그림 1 이미지를 삽입한 화면

4.2. 음성 삽입

디바이스 내에 새로운 음성 파일을 생성하기 위해 PhoneGap의 Media API와 File API를 이용하였다. PhoneGap의 Media 객체를 통해 오디오 파일을 녹음하거나 재생할 수 있다. 또한 이 때, File 객체를 사용하여 디바이스 내에 오디오 파일을 저장한다.

4.3. 동영상 삽입

jQuery와 PhoneGap의 File 객체를 이용하여 동영상 파일 탐색기 기능을 구현하였다. File 객체를 통해 기기의 파일 시스템에 접근하여 동영상 파일 확장자인 파일 명들을 찾아와 이들을 목록으로 보여준다. 사용자가 동영상 파일 명을 누르면 이를 재생하거나 카드에 포함시

킬 수 있도록 하였다.

HTML5의 Video 태그는 PhoneGap에서 잘 작동되지 않는다. 오디오 파일은 PhoneGap에서 제공하는 Media API를 사용할 수 있지만 동영상 파일은 따로 지원하는 API가 없다. 하지만 플러그인을 사용하여 동영상 문제를 해결할 수 있다.

동영상을 재생하는 화면은 그림 2와 같다.



그림 2 동영상 재생 화면

4.4. 카드 저장

사용자가 제작한 엽서는 PhoneGap의 File API를 이용하여 디바이스 내에 HTML파일로 저장한다. 저장이 완료된 파일은 포함된 멀티미디어 파일과 함께 jsp를 이용하여 서버에 전송된다.

4.5. 카드 전송

서버에 저장한 엽서를 상대방에게 보내기 위해 URL을 지정한다. 저장한 URL을 상대방에게 보내는 형식이다. PhoneGap의 전화번호부 API를 사용하여 전화번호를 불러온 후 SMS 플러그인을 사용하여 SMS를 전송한다. 카카오톡과 이메일 역시 플러그인을 사용하여 연동한다.

5. 결론

본 논문에서는 HTML5를 이용하여 멀티미디어 엽서 하이브리드 어플리케이션을 구현하였다. PhoneGap을 사용하였기 때문에 안드로이드, 아이폰, 블랙베리 등 다양한 OS에서 사용가능하다. 또한 음성, 사진, 동영상, 텍스트 등 다양한 멀티미디어 파일을 이용하여 엽서 제작이 가능하기 때문에 보다 화려하고 전달력 있는 엽서를 제작할 수 있다.

참고문헌

- [1] 정우진, 오장훈, 윤동원, “하이브리드 모바일 앱 프레임워크 설계 및 구현”, 한국정보통신학회논문지, 16(9), 2012
- [2] 장래영, 정성재, 배우미, 성경, 소우영, “모바일 하이브리드앱 개발도구에 관한 비교연구”, 한국정보통신학회, 2012

Parameter Optimization for Local Structure-Based Fingerprint Alignment

Heemin Park*, Jongwoo Lee** and Jae Won Lee***

**Department of Computer Science, Yonsei University,
Seoul 120-749, South Korea*

***Department of Multimedia Science, Sookmyung Women's University,
Seoul 140-742, South Korea
E-mail: bigrain@sookmyung.ac.kr*

****School of Computer Science, Sungshin Women's University,
Seoul 136-742, South Korea*

Abstract

We propose a method to obtain optimized parameters for fast fingerprint alignment based on corresponding local structures. In previously published research, we demonstrated that the major trend of the transformations could be efficiently extracted from the corresponding local structures using graph representations, even for noisy fingerprints. However, these methods can be improved using calibrated model parameters that directly affect the method's performance. In this paper, we develop a systematic approach to determine the parameters based on a training subset of fingerprint impressions. Experimental results show that the average alignment error obtained with the proposed method is less than half of the error obtained with existing algorithms.

Key Words: Fingerprint Matching, Alignment, Minutiae, Parameter Optimization

1. Introduction

As embedded computing systems advance, privacy and security are becoming more and more important. For example, the identity of a user must be checked for admission to some facilities and/or access to mobile systems [1, 2, 3]. Because of its reliability and ease of use, fingerprint verification is one of the most prevalent methods for user identification.

In fingerprint matching, minutiae-based local structures are often used for feature extraction because they can consider both local and global features and are robust and flexible enough to handle warps and noises [4, 5]. Many local structure-based fingerprint matching approaches use the triangle or triplet structure because triangles are more rigid and robust than couplet or single features. The global alignment plays a very important role because it directly affects the performance of local structure-based fingerprint matching.

Previous work on global alignment algorithms can be grouped into two categories: 1) fast and simple, but low-performance methods; and 2) computationally intensive approaches that consider noise. The transformation of the best corresponding (or matched) local structure pair

[6] is an example of a fast and simple approach. This transformation is computationally simple, but the alignment results may fall into local minima or outliers. In [7 and 8], the authors created a histogram of scores based on each of the possible transformations for the local structures. They selected the best transformations at the peak of the histogram for the global alignment. These methods [6, 7, 8] worked well for fingerprints with low noise. However, when transformations of corresponding local structures are distorted by noise, such approaches may not find the correct alignment. Additionally, the methods are computationally intensive due to an exhaustive search of the transformation space. To address noise, Tong *et al.* [9] determined the best transformation by the simultaneous local matching of two minutiae in the template and input fingerprints. This method can lead to a more robust alignment; but it also demands more computations. To solve the aforementioned problems, we previously proposed a fast and efficient method to find global fingerprint alignments based on corresponding local structures using a graph representation [10]. The approach in [10] showed that the major trend of transformations could be efficiently and effectively extracted from the *corresponding* triangles using neighborhood relations in a graph representation. However, the performance of this method depends sensitively on parameter selection. In this paper, we present an improved global alignment method and devise a systematic approach to set the parameters for optimized performance based on a training subset of fingerprint impressions.

To make the paper self-contained, we briefly explain the minutiae triangle matching and the alignment method using graph representations in Section 2. Section 3 presents the new alignment method and Section 4 proposes a systematic way to calibrate the parameters. Experimental results are presented in Section 5. We conclude in Section 6.

2. Global Fingerprint Alignment using Corresponding Triangles

2.1 Local Structure Matching

The fingerprint alignment algorithm [10] in our approach was based on matching the local structures of minutiae triangles. Like most verified methods, we used minutiae triangles (triangle hereafter) as local structures. When three minutiae a , b , and c , form a triangle, a nine-component feature vector can be defined by three side lengths, three interior angles, and three relative angles with respect to the direction of ridges. Fig. 1 illustrates the nine components of a feature vector, $F_{\Delta abc}$.

The signature of each triangle of the input fingerprint $\mathbf{I}$ were calculated from the feature vector and compared to the pre-calculated signatures of triangles in the template fingerprint $\mathbf{T}$. The pairs of matched local structures are termed *corresponding* triangles. The objective of a

파형의 시퀀스 탐지에 기반한 모음 'ㅏ' 인식¹

최수경^(*), 이재원^(**), 이종우^(**)

(*) 숙명여자대학교 멀티미디어학과, sukyungchoi@sookmyung.ac.kr

(**) 성신여자대학교 IT학부, jwlee@sungshin.ac.kr

(***) 숙명여자대학교 멀티미디어학과, bigrain@sookmyung.ac.kr

Korean Vowel 'ㅏ' Recognition based on Wave Sequence Detection

Choi Su-kyung(*), Lee Jae-won(**), Lee Jong-woo(**)

(*) Dept. of Multimedia Science, Sookmyung Women's University

(**) School of Information Technology, Sungshin Women's University

요약

본 논문은 최근 스마트폰의 보급이 확산됨에 따라, 요구가 급속하게 증가하고 있는 음성 인식의 일부로써 한국어 음소 중 모음 'ㅏ' 인식에 대한 연구이다. 음소는 음성을 구성하고 있는 최소 단위로써 음성을 인식하는데 매우 중요한 역할을 한다. 그러나 각각의 음소들을 정확하게 인식 하려면 발음의 다양성 등으로 인해 많은 어려움이 존재한다. 본 연구에서는 음소를 인식의 기본 단위로 하는 음성 인식 시스템 구현의 일부로써 한국어 음소 중 모음 'ㅏ'를 인식하기 위한 새로운 알고리즘을 제안한다.

1. 서론

음성 인식 기술은 컴퓨터 등 기계가 사람의 말을 알아들을 수 있도록 하는 기술이다. 1950년대부터 시작된 음성 인식 연구는 지난 60년 이상의 연구개발 역사를 가지고 있으며, 꾸준한 개발로 지속적인 발전을 이루어왔다[1, 2].

음성 인식 시스템에서 분석의 기본 단위로는 단어, 어절, 음소 등이 있으나, 음소 단위는 단어 및 어절 단위보다 그 종류가 작고 음향적인 특성을 인식기에 고르게 반영할 수 있기 때문에 많이 사용된다[3]. 그러나 발음의 변화 등 여러 가지 요인으로 인해 각각의 음소들을 정확하게 인식하는 것은 무척 어렵다[4].

따라서 본 논문에서는 음소를 기본 단위로 하는 음성 인식 시스템의 일부로써 한국어 음소 중 모음 'ㅏ'를 인식하고자 한다. 모음 'ㅏ'의 파형들을 관찰한 결과 일정한 패턴을 가지

고 있음을 발견하여 이를 바탕으로 모음 'ㅏ'의 파형을 탐지 하는 연구를 수행하였다.

본 논문의 구성은 다음과 같다. 2장에서 음성 인식과 관련된 기존 연구에 대해 살펴본다. 3장에서는, 모음 'ㅏ'의 파형 특징들을 자세히 분석한다. 4장에서는 본 논문에서 제안하는 시퀀스 탐지에 기반한 모음 'ㅏ' 인식 알고리즘을 기술하고, 마지막으로 5장에서 결론을 맺는다.

2. 관련 연구

기존의 음성 인식 연구에는 대표적으로 HMM(hidden markov model)이 있다. HMM은 음운, 단어와 같은 음성의 단위를 통계적으로 모델화한 것으로, 1980년대 후반부터 성행한 음성 인식 기술이다. 이는 정확도가 높지만 미리 방대한 양의 학습을 필요로 하고, 확률 모델을 사용하여 복잡하다.

음소를 기본 단위로 하는 음성 인식에는 음소 'ㅏ', 'ㅑ',

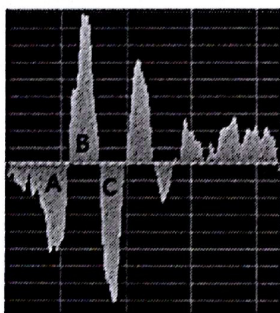
1 이 논문은 2013년도 정부(교육부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임(2013R1A1A2013155)

‘ㅈ’ 인식에 대한 연구가 있다[5]. 이는 부호 분포 변동성이라는 지표를 활용하여, 기존의 영교차율(ZCR) 방법에 비해 에러율을 크게 감소시켰다.

3. 모음 ‘ㅏ’의 특성

본 장에서는 모음 ‘ㅏ’의 파형 특성을 분석한다.

[그림 1]은 전형적인 모음 ‘ㅏ’의 파형을 나타내고 있다. 모음 ‘ㅏ’들을 관찰해 본 결과 위의 파형이 계속 반복됨을 알 수 있었다. 그리고 A, B, C 파형이 모음 ‘ㅏ’를 다른 음소들과 구분해 주는 주요한 특성임을 알 수 있었는데, 먼저 두꺼운 모양의 A가 아랫부분에 나타나고, 다음에 상대적으로 길고 뾰족한 B가 윗부분에 나타난다. 그리고 좀 더 가는 C가 아래에 나타난다.



[그림 1] 전형적인 모음 ‘ㅏ’ 파형

이런 모음 ‘ㅏ’ 특성 관찰 결과에 기반하여 모음 ‘ㅏ’의 파형을 탐지하는 알고리즘을 도출하였다. 알고리즘은 4장에서 다룬다.

4. 시퀀스 탐지에 기반한 모음 ‘ㅏ’ 인식 알고리즘

본 장에서는 모음 ‘ㅏ’ 인식을 위한 알고리즘을 설명한다. 이는 3장에서 관찰한 파형의 특성을 바탕으로 파형의 시퀀스 패턴을 분석하여 경험적으로 도출하였다.

A의 길이 > B의 길이 + 2 &&

A의 길이 > C의 길이 + 2 &&

B의 길이 > 11 && B의 길이 < 22 &&

C의 길이 > 9 && C의 길이 < 21 &&

A의 위치 = 아래 &&

B의 위치 = 위 &&

C의 위치 = 아래

[그림 2] 모음 ‘ㅏ’ 인식 휴리스틱 알고리즘

[그림 2]는 모음 ‘ㅏ’ 인식을 위한 알고리즘을 나타낸 것이다. [그림 2]에 기술된 조건을 만족시키면 모음 ‘ㅏ’로 인식된다.

이 알고리즘은 아주 간단하고 빠른 장점을 지닌다. HMM의 경우 미리 방대한 양의 학습을 필요로 하고, 확률 모델을 사용하여 복잡하다. 그러나 본 알고리즘은 HMM보다 정확도는 떨어질 수 있으나, 간단한 패턴만으로도 80% 이상의 모음 ‘ㅏ’를 인식할 수 있다. 또한 속도가 빠른 장점을 활용하여 실시간 음성인식 어플리케이션의 보조적 탐지 기법으로도 사용이 가능하다.

5. 결론

본 논문에서는 한국어 음소 중 모음 ‘ㅏ’의 특성을 관찰하고, 이를 바탕으로 모음 ‘ㅏ’의 파형을 탐지하는 연구를 수행하였다. 먼저 파형이 일정한 시퀀스 패턴을 가지고 있음을 관찰을 통해 파악하고, 직관에 기반하여 모음 ‘ㅏ’ 인식을 위한 알고리즘을 도출하였다. 이는 기존의 음성 인식 방법과 달리 매우 간단하면서도 빠른 장점을 지녀서, 최근 스마트폰 환경의 음성 인식에서 중요한 역할을 할 것으로 기대한다.

향후 연구과제로는 구체적인 실험을 통해 본 알고리즘의 성능을 평가하고, 본 논문에서 다루지 않은 파형의 높이나 비율, 파형 간의 간격 등 추가 사항을 고려하여 모음 ‘ㅏ’의 인식률을 높이는 것이다.

자연과학연구소 규정 (준칙)

제 1 장 총 칙

- 제1조 (명칭 및 위치) 이 연구소는 숙명여자대학교 자연과학연구소 (이하 ‘연구소’라 한다)라 하며 숙명여자대학교 내에 둔다.
- 제2조 (목적) 이 규정은 연구소의 조직과 운영에 관한 사항을 규정함을 목적으로 한다.
- 제3조 (설립목적) 연구소는 자연과학의 기초 및 응용분야에 관한 연구를 수행함으로써 자연과학의 발전에 기여함을 목적으로 한다.
- 제4조 (사업) 연구소는 제3조의 목적을 달성하기 위하여 다음의 사업을 한다.
1. 자연과학의 기초 및 응용분야의 연구, 개발 및 조사
 2. 학술 강연회와 연구발표회 및 공개강좌 개최
 3. 연구자료의 확보 및 연구 논문집 발간
 4. 위탁된 연구조사 및 감정 분석 실험
 5. 기타 연구소 설립목적에 부합되는 사업
- 제5조 (연구센터) 연구소는 제4조의 사업을 원활히 수행하기 위하여 연구센터를 둘 수 있다.

제 2 장 조 직

- 제6조 (소장)
- ① 연구소에 소장을 두며, 소장은 연구소를 대표하고 연구소 운영에 관한 제반업무를 통할한다.
 - ② 소장의 임명은 본 대학교 직제규정이 정하는 바에 의한다.
- 제7조 (연구센터장)
- ① 연구센터에 연구센터장을 두며, 연구센터장은 연구센터의 운영에 관한 제반 업무를 통할한다.
 - ② 연구센터장은 제9조에 규정된 연구위원 중에서 총장이 임명한다.
 - ③ 연구센터장의 임기는 2년으로 하되 연임할 수 있다.
- 제8조 (간사)
- ① 소장의 업무를 보좌하기 위하여 연구소에 간사를 둘 수 있다.
 - ② 간사의 제9조에 규정된 연구위원 중에서 소장의 제청으로 총장이 임명한다.
 - ③ 간사의 임기는 2년으로 하되 연임할 수 있다.

제9조 (연구위원)

- ① 연구소의 제반 연구활동을 수행하기 위하여 연구위원을 둔다.
- ② 연구소의 설립목적과 사업에 동의하는 본 대학교 소속 교원은 본인의 동의가 있는 경우 소장의 동의에 의해 연구위원으로 위촉할 수 있다.

제10조 (연구원)

- ① 연구소에 전임 및 비전임연구원을 둘 수 있으며 책임연구원, 선임연구원, 연구원으로 구성할 수 있다.
- ② 책임연구원은 박사학위를 소지한 자로서 소장의 제청으로 총장이 임명한다.
- ③ 선임연구원은 박사과정 또는 수료한 자로서 소장의 제청으로 총장이 임명한다.
- ④ 연구원은 석사학위과정 또는 소지자로서 소장의 제청으로 총장이 임명한다.
- ⑤ 연구원의 임기는 2년으로 하되 연임할 수 있다.

제11조 (연구조교)

- ① 연구소에 연구조교를 둘 수 있으며 제반 업무에 종사한다.
- ② 연구조교는 학사학위 이상학력의 소지자로서 소장의 제청으로 총장이 임명한다.

제12조 (객원연구원)

- ① 전문분야에 연구경력이 탁월한 해외 및 국내 학자들로서 본교 연구소의 연구활동에 참여하고자 하는 자는 제15조에 규정된 연구소 운영위원회의 심의를 거친 후 총장의 결재를 받아 객원연구원으로 소장이 위촉할 수 있다.
- ② 객원연구원의 임기는 1년으로 하되 연임할 수 있다.

제13조 (박사학위 소지자 연구과정 : Post-Doc. course) 연구소는 국내외에서 박사학위를 취득한 소장연구자들에게 연구의 기회를 제공하기 위하여 박사 후 연구과정을 운영할 수 있으며, 제15조에 규정된 연구소 운영위원회의 심의를 거친 후 총장의 결재를 받아 박사학위소지자 연구과정요원으로 소장이 위촉할 수 있다.

제14조 (감사)

- ① 연구소에 감사를 두며, 감사는 연구소의 회계결산에 대한 내부감사를 수행한다.
- ② 감사는 간사가 아닌 연구위원 중에서 총장이 위촉하며, 임기는 1년으로 하되 연임할 수 있다.
- ③ 감사는 감사종료 후 10일 이내에 감사보고서를 작성하여 총장과 제15조에 규정된 연구소 운영위원회에 제출하여야 한다.

제 3 장 운영위원회

제15조 (운영위원회)

- ① 연구소의 운영에 관한 중요사항을 심의하기 위하여 연구소 운영위원회(이하 “운영위원회”라 한다)를 둔다.
- ② 운영위원회는 소장과 10명 이내의 연구위원으로 구성하고 소장이 위원장이 된다.

제16조 (기능)

1. 사업계획에 관한 사항
2. 예산 및 결산에 관한 사항
3. 연구소 관련규정의 제정 및 개폐에 관한 사항
4. 객원연구위원의 위촉에 관한 사항
5. 기타 연구소 운영에 관한 중요사항

제17조 (회의)

- ① 운영위원회 회의는 위원장이 소집하고 그 의장이 된다.
- ② 운영위원회는 재적위원 2/3 이상의 출석과 출석위원 과반수의 찬성으로 의결한다. 단, 가부동수인 때에는 의장이 결정권을 갖는다.

제 4 장 재 정

제18조 (재정)

1. 대학, 정부기관, 관련단체의 보조금
2. 기업 및 개인의 찬조금
3. 연구위원의 기여금
4. 기타 수입

제19조 (회계연도) 연구소의 회계연도는 본 대학교의 회계연도와 같다.

제20조 (장부 및 물품관리)

- ① 소장은 연구소의 재정에 관한 회계장부를 기록, 유지하여야 한다.
- ② 소장은 연구소내의 제반 시설과 물품을 성실히 관리하여야 한다.

제 5 장 보 칙

제21조 (재산의 귀속) 연구소가 해산할 경우 그 재산은 본 대학교에 귀속한다.

제22조 (세칙) 이 규정 시행에 필요한 세부사항은 소장이 운영위원회의 심의와 총장의 승인을 얻어 세칙으로 정한다.

부 칙

- (1) (시행일) 이 준칙은 1996년 4월 29일부터 시행한다.
- (2) (경과조치) 연구기관의 장은 이 준칙에서 시행되는 날로부터 1개월 이내에 연구기관의 규정을 이 준칙에 저촉되지 않도록 개정해야 한다.

부 칙

- (1) (시행일) 이 준칙은 1996년 10월 14일부터 시행한다.

부 칙

- (1) (시행일) 이 준칙은 2004년 4월 9일부터 시행한다.