자연과학연구소 주최 2014 학술행사

[1] 2014.03.11. 나노물리학과. Perpendicular magnetic tunnel junctions for spin-transfer-torque random access memory (연사 : 민병철 박사)
[2] 2014.03.19. 화학과. Electrochemical oral cancer diagnosis based on telomerase activity (연사 : Shigeori Takenaka 박사)
[3] 2014.03.25. 나노물리학과. 플라즈모닉 광학나노안테나 (연사 : 서민교 박사)
[4] 2014.03.26. 화학과. Ferroelectricity and magnetoelectric coupling in non-ferroic tri-color superlattices (연사 : 서지원 박사)
[5] 2014.03.28. 수학과. 초기치 문제에 대한 에러 내장형 에러보정 수치해법 (연사 : 김필수 박사)
[6] 2014.04.01. 나노물리학과. Let's calculate!! expanding our understanding through computing (연사 : 이주형 박사)
[7] 2014.04.04. 수학과. 로미오와 줄리엣의 심장 박동을 어떻게 동기화 시킬 것인가? (연사 : 하승열 박사)
[8] 2014.04.08. 나노물리학과. Single-molecule biophysics (연사 : 홍석철 박사)
[9] 2014.04.23. 의약과학과. Novel Recombinant technology for making proteins by Protein Science Corporation (연사 : Dan Adams 박사)
[10] 2014.05.09. 수학과. 대용량 유전체 데이터 분석을 위한 생명정보학 방법론 (연사 : 남덕우 박사)
[11] 2014.05.23. 수학과. 금융시장 이슈를 통해 본 수학과 금융의 관계 (연사 : 장봉규 박사)
[12] 2014.05.23. 멀티미디어과학과. 하이브리드 방송 서비스 현황 및 발전 방향 (연사 : 배병준 연구원)
[13] 2014.05.27. 의약과학과. USP Activities in Botanical Standard Development (연사 : 김남철 박사)
[14] 2014.05.30. 통계학과. A Power Set Based Statistical Selection Procedure to Locate Susceptible Rare Variants Associated with Complex Traits with Sequencing Data (연사 : 선호근 박사)
[15] 2014.06.11. 통계학과. Regression Analysis on Interval-Valued Data (연사 : 박철우 박사)
[16] 2014.06.23. 통계학과. Math in Music (연사 : 안홍식 박사)
[17] 2014.09.12. 컴퓨터과학과. OpenCV 버전 2.4.9 소개 및 Visual studio 2010에서의 설치 및 활용, 이를 이용한 이미지 연산 방법의 구현 등 사용법 습득 (연사 : 조선영 박사)
[18] 2014.09.15. 컴퓨터과학과. 아듀이노 마이크로프로세서 소개, 아듀이노 활용방법 소개(기본), 교육현장에서의 응용 방법 (연사 : 홍선학 박사)
[19] 2014.09.17. 화학과. van der Waals epitaxial integration of semiconductors on graphene (연사 : 홍영준 박사)
[20] 2014.09.23. 나노물리학과. 자기센서의 원리와 응용 (연사 : 손대락 박사)
[21] 2014.09.23. 통계학과. Sparse high dimensional varying coefficient models (연사 : 이은령 선생)
[22] 2014.09.24. 화학과. On the Basis of Molecular Rectification in Tunneling Junctions (연사 : 윤효재 박사)
[23] 2014.10.01. 화학과. 고강도 다기능 탄소나노복합소재 (연사 : 구본철 박사)
[24] 2014.10.02. 컴퓨터과학과. R을 이용한 통계 분석 (연사 : 김익태 연구원)

[25] 2014.10.10. 수학과. 금융과 수학 (연사 : 구형건 박사)
[26] 2014.10.14. 나노물리학과. 영구자석재료의 이해 (연사 : 권해웅 박사)
[27] 2014.10.31. 수학과. Undergraduate Research in Mathematics (연사 : 이남용 박사)
[28] 2014.11.06. 멀티미디어과학과. 창의적 여성리더십 (연사 : 김현주 선생)
[29] 2014.11.11. 통계학과. Local Deviations from the Uncovered Interest Rate Parity - Kernel Smoothing Functions and the Role of Macroeconomic Fundamentals (연사 : 김건호 박사)
[30] 2014.11.12. 화학과. Structures and Interactions of Amyloidogenic Proteins Related to Their Cytotoxic effects in Heterogeneous Systems (연사 : 김준곤 박사)
[31] 2014.11.14. 수학과. 착색문제에서 군(群)을 이용한 패턴 수 계산 (연사 : 유원석 박사)
[32] 2014.11.19. 화학과. Self-Assembled Nanostructures of Block Copolymers for Biomedical Applications (연사 : 양승윤 박사)
[33] 2014.11.26. 화학과. Visible Light-Induced Perfluoroalkylations (연사 : 조은진 선생)
[34] 2014.12.03. 멀티미디어과학과. 기계학습 Support Vector Machine(SVM)의 이해 (연사 : 유환조 박사)
[35] 2014.12.05. 수학과. 대수기하학 소개 (연사 : 박의성 박사)
[36] 2014.12.10. 화학과. Fabrication of Hollow Nanoparticles for Nanoreactor and Biomedical Applications (연사 : 이인수 박사)
[37] 2014.12.15. 멀티미디어과학과. CE 제조사 관점에서 본 웹 기술의 현재와 미래 (연사 : 이동영 연구원)

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 체재 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 연구활동을 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 ~ 2014년 8월 : 12대 소장 - 황 선 영 교수
2014년 9월 ~ 현재 : 13대 소장 - 이 진 호 교수

Contents

◈ 나노물리학과

◈ 화학과

◈ 수학과

◈ 통계학과

◈ 컴퓨터과학과

◈ 멀티미디어과학과

◈ 의약과학과

나노물리학과

Physica C 507 (2014) 55–58

Contents lists available at ScienceDirect

Physica C

journal homepage: www.elsevier.com/locate/physc

Transport properties along *c*-axis of $DyNi_2B_2C$

W.C. Lee *

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

ARTICLE INFO

Article history:
Received 5 June 2014
Accepted 1 October 2014
Available online 13 October 2014

Keywords:
Superconductivity
Anisotropy
Critical field
Transport property

ABSTRACT

We have measured the resistivity along *c*-axis $\rho_c(H, T)$ of $DyNi_2B_2C$ with the applied magnetic field *H* perpendicular to *c*-axis for 0 kG < *H* < 4 kG and temperature range 2 K < *T* < 300 K. From these, the superconducting upper critical field $H_{C2}(T)$ curve of $DyNi_2B_2C$ for the *c*-axis was constructed and our $H_{C2}(T)$ curve from $\rho_c(H, T)$ measurement has been compared with that from previous known $\rho_{ab}(H, T)$ result. With additional magnetization isotherms M(*H*, *T*) for $H \perp c$ and $H \parallel c$-axis, the anisotropy in $H_{C2}(T)$ curves of the magnetic structure $DyNi_2N_2C$, which has the superconducting transition temperature T_C is lower than the Néel temperatures T_N, might be originated from the additional anisotropic magnetic Dy^{3+} sublattice.

Since the discovery of the intermetallic quaternary compound superconductors, RNi_2B_2C (R = rare-earth elements), many studies have been done on the interplay between superconductivity and magnetism which is one of the most attractive topics in solid state physics and it is now well known that there is a considerable degree of interaction between the conduction electrons and the rare-earth ions in these compounds which brings about the reduction in T_C [1–10]. More detailed studies on single crystals of these compounds containing magnetic rare-earth ions show possess highly anisotropic normal state magnetizations and anisotropic and suppressed superconducting properties, which indicate the existence of a significant interaction between the local magnetic moments and superconducting electrons [5–10]. It has been reported that among RNi_2B_2C (R = rare earth elements) only antiferromagnetic state exists at low temperature region for R = Gd, Dy, Ho, Er, and Tm [11–13]. Also the neutron scattering measurement has contributed to find the magnetic structures and magnetic interactions among rare earth elements and it turned out that some of these compounds show magnetic and superconducting properties simultaneously at certain temperature [11–13]. Especially, RNi_2B_2C (R = Tm, Er, Ho, and Dy) have the superconducting transition temperatures T_C = 10.8 K, 10.6 K, 8.2 K, and 6.2 K, and the antiferromagnetic Néel temperatures T_N = 1.5 K, 6.0 K, 5.1 K, and 10.3 K, respectively [14,15] (see Table 1). As we can see at the table, the ratios of the superconducting transition temperatures (T_C) to the antiferromagnetic Néel temperatures (T_N), T_C/T_N, are 7.2, 1.8, 1.6, and 0.6 for R = Tm, Er, Ho, and Dy, respectively. This order of magnitude variation in T_C/T_N has a very useful information to study the relation of T_C, T_N and rare-earth elements R, and there is a crossover between T_C and T_N in the intermediate compound of $HoNi_2B_2C$ and $DyNi_2B_2C$ [3,14]. For example, the magnitude variation in T_C/T_N changes from 1.6 for $x = 0$ to 0.6 for $x = 1.0$ in $Ho_{1-x}Dy_xNi_2B_2C$ system.

Particularly, the $DyNi_2B_2C$ single crystal is the first such number of the RNi_2B_2C family with $T_C < T_N$, and was observed as the first crystallographic ordered compound except the heavy fermions family at $T_C < T_N$. Many works have been done on the transport and magnetic properties along ab plane of $DyNi_2B_2C$, which show the superconducting upper critical field $H_{C2}(T)$ and possible anisotropy [9,16].

However, up to now, the physical properties of the transport properties and anisotropy with different magnetic field configurations along *c*-axis of $DyNi_2B_2C$ have not been reported yet and the possible transport anisotropy due to the interplay between superconductivity and magnetism has been missed. In this paper, we present, for the first time, the *c*-axis resistivity with the applied magnetic field H perpendicular to the *c*-axis and have determined the $H_{C2}(T)$ curve of the $DyNi_2B_2C$ single crystal. We have compared our $H_{C2}(T)$ curve from $\rho_c(H, T)$ measurement for *H* perpendicular to *c*-axis with that from $\rho_{ab}(H, T)$ measurement for *H* parallel to *c*-axis and obtained the anisotropy in $H_{C2}(T)$ curves. Finally, the possible origin of anisotropy has been investigated with some kind of magnetic transitions far below the 3 dimensional long range antiferromagnetic Néel state.

Single crystals of $DyNi_2B_2C$ were grown by high-temperature flux method, using Ni_2B as a solvent and details are described in Ref. [5]. Single-crystal platelets extracted from the flux were examined by X-ray and were found to be single phase of high quality with *c*-axis perpendicular to their flat surface. Also, from the

* Tel.: +82 2 719 9402; fax: +82 2 718 2337.
E-mail address: wclee@sookmyung.ac.kr

http://dx.doi.org/10.1016/j.physc.2014.10.001
0921-4534/

resistivity measurement, the superconducting transition width $\Delta T = T_C^{onset} - T_C(\rho = 0)$ is 0.2 K for $H = 0$ G, which shows that the crystals are of high quality. The sample used in our work and one used in Ref. [9] were from the same batch.

A Liner Research Inc. LR 400 Four-Wire AC Resistance Bridge was used by using the SQUID magnetometer to measure the *c*-axis resistivity $\rho_c(H, T)$ as a function of temperature and applied magnetic fields with current I applied parallel to *c*-axis. The sample dimension was $2.43 \times 1.0 \times 0.31$ mm^3 with the crystallographic *c*-axis perpendicular to the plate surface and six gold electrical contacts by evaporation deposition were made on both surfaces of a crystal and finally this crystal was anneal at 300 °C for 24 h for the ohmic contacts.

Fig. 1 shows the temperature dependence of the *c*-axis resistivity $\rho_c(T)$ of a single crystal $DyNi_2B_2C$ for 2 K < T < 300 K with current $I = 10$ mA with the applied magnetic field $H = 0$ G. At room temperature, the *c*-axis resistivity ρ_c(300 K) was 113 μΩ cm, which is almost twice of the in-plane resistivity ρ_{ab}(300 K) [9] and the measured $\rho_c(T)$ shows the metallic behavior from room temperature down to T_C of 6.2 K as the reported $\rho_{ab}(T)$ does. It has been known that Néel temperature T_N of $DyNi_2B_2C$ is 10.3 K from $\rho_{ab}(T)$ measurement and is independent of applied magnetic fields [9]. Also, our result from $\rho_c(T)$ measurement in Fig. 2 shows the almost the same Néel temperature of 10.3 K and this temperature is independent of the applied magnetic fields, too. In Fig. 2, we can see same interesting features of coexistence of the superconducting state far below the antiferromagnetic state (see Table 1. Only $DyNi_2B_2C$ and $HoNi_2B_2C$ show $T_C < T_N$ among RNi_2B_2C systems.). The superconducting onset temperature T_C^{onset} for $H = 0$ G is about 6.4 K and zero resistivity temperature $T_C(\rho_c = 0)$ occurs at 6.2 K which is similar to 6.0 K obtained from $\rho_{ab}(T)$ measurement. As can be seen clearly in Fig. 2, $T_C(H, \rho_c = 0)$ is suppressed and its width of the superconducting transition is increased with increasing applied magnetic field. It has been also reported that the superconducting transition temperature is suppressed and the width of the superconducting transition is increased with increasing applied magnetic field H in $\rho_{ab}(T)$ measurement, too [9]. Even though we have measured $\rho_c(T, H)$ with applied magnetic field $H \parallel c$-axis (not shown in Fig. 2), but did not find the difference between $\rho_c(T, H)$'s with the applied magnetic field $H \perp c$-axis and $\rho_c(T, H)$ with the applied magnetic field $H \parallel c$-axis. Also from the Ref. [9], it has been reported that $\rho_{ab}(T, H)$'s with $H \parallel c$-axis and $H \perp c$-axis are almost same. In other words, there was no difference between $H \parallel c$-axis and $H \perp c$-axis cases for both $\rho_c(T, H)$ and $\rho_{ab}(T, H)$ measurements.

To figure out some possible origin for the different temperature- and applied magnetic field-dependence between $\rho_{ab}(T, H)$ in Ref. [9] and our work $\rho_c(T, H)$ far below Néel temperature, we measured the low-temperature static magnetization $M(T, H)$ of $DyNi_2B_2C$ for $H \perp c$-axis and for $H \parallel c$-axis shown in Figs. 3 and 4. Fig. 3 shows some interesting typical magnetization isotherms $M(T, H)$ with applied magnetic fields perpendicular to the *c*-axis.

Table 1
T_N (Néel Temperature) and T_C (superconducting transition temperature) of RNi_2B_2C (R = rare-earth elements), after Refs. [14] and [15].

RNi_2B_2C	T_N (K)	T_C (K)
YNi_2B_2C	–	15.5
$LuNi_2B_2C$	–	16.5
$TmNi_2B_2C$	1.5	10.8
$ErNi_2B_2C$	6.0	10.6
$HoNi_2B_2C$	5.1	8.2
$DyNi_2B_2C$	10.3	6.2
$TbNi_2B_2C$	15.0	–
$GdNi_2B_2C$	20.0	–

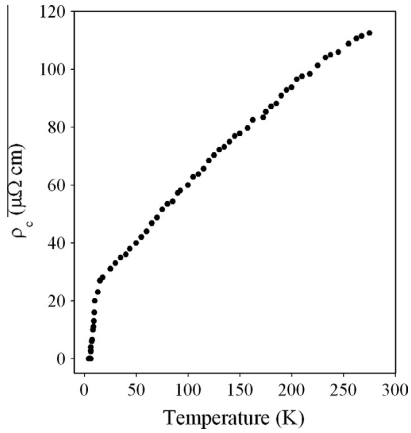


Fig. 1. The resistivity along *c*-axis of $DyNi_2B_2C$ with $H = 0$ G.

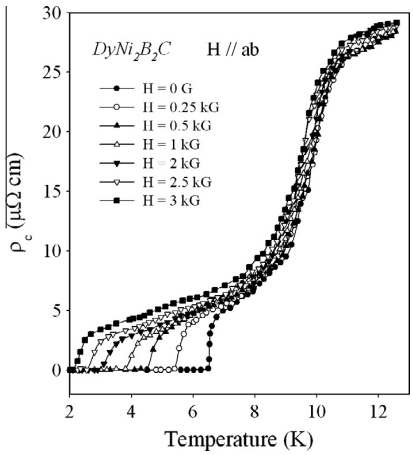


Fig. 2. The resistivity along *c*-axis of $DyNi_2B_2C$ with $H \parallel$ ab plane (There was no difference between $H \parallel c$ axis- and $H \parallel$ ab plane-measurements. See the text).

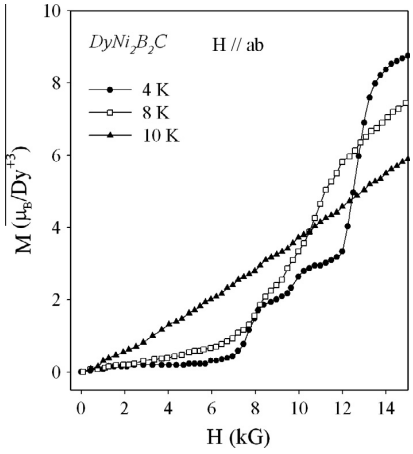


Fig. 3. The magnetization vs. applied magnetic fields of $DyNi_2B_2C$ with $H \parallel$ ab plane at various temperatures.

Effect on the Electronic, Magnetic and Thermoelectric Properties of Bi_2Te_3 by the Cerium Substitution

Tran Van Quang[1] and Miyoung Kim[2]

[1]Department of Physics, Ajou University, Suwon 443-749, Republic of Korea
[2]Department of Nano Physics, Sookmyung Women's University, Seoul 140-742, Republic of Korea

We investigated the effect of the Ce substitution in Bi_2Te_3 on its electronic, magnetic, and thermoelectric properties in first-principles using the precise full-potential linearized augmented plane-wave (FLAPW) method. Results revealed that $CeBiTe_3$ is a magnetic semiconductor with a very narrow energy band gap in the spin-polarized phase within GGA + U. The calculation of thermoelectric coefficients, which is determined by utilizing the Boltzmann's equation in a constant relaxation-time approach using the FLAPW wave-functions, shows that the Ce substitution causes a reduction of the thermoelectric power, as a result of the change in Seebeck coefficient and electrical conductivity due to the strongly localized $4f$ bands and the reduced band gap. The maximum figure of merit ZT is found to be about 0.29 at 450 K, which is in good agreement with the experiment.

***Index Terms*—Electronic structure calculation, first-principles, rare earth, thermoelectric (TE).**

I. Introduction

FOR many decades, thermoelectric (TE) technology has been studied intensively in theory and experiment, aiming for the practical application. Various TE application devices such as TE energy conversion, heat pump, space power generation, thermocouple, deep-space probe, solid state refrigerator using semiconductor lasers, and seat cooler have been feasibly investigated. Examining and optimizing materials to improve the TE efficiency is a critical and fundamental key to introducing the practical application. The TE efficiency of a material or a device is characterized by a dimensionless figure of merit, $\mathrm{ZT} = \sigma \mathrm{S}^2 \mathrm{T}/\kappa$, in which S is the Seebeck coefficient or thermopower, σ is the electrical conductivity, κ is the thermal conductivity determined from the lattice (κ_L) and the electronic (κ_e) thermal conductivities by $\kappa = \kappa_\mathrm{L} + \kappa_\mathrm{e}$, and TT is the temperature. With no obvious upper bound of ZT value claimed, the highest ZT reported are 2.4 at room temperature (RT) [1] and 3.2 at 300 °C [2], which are still far from the value for the TE efficiency factor to achieve ideal Carnot's efficiency. To improve ZT, the power factor $\mathrm{PF} = \mathrm{S}^2\sigma$ has to be high or S and σ have to be large while κ is kept simultaneously low. Bismuth telluride (Bi_2Te_3) and its alloys are well known as the best TE materials operating with $\mathrm{ZT} \sim 1$ at RT [3]. Various efforts to improve ZT have been made by optimizing the carrier concentration and also by making nanostructures or supperlattices. One of the methods to optimize the TE efficiency is introducing the rare earth elements into the material, resulting in the beneficial TE properties by lowering the thermal conductivities. It has been reported that the rare earth chalcogenides are potentially brilliant TE candidates for the high temperature application [4]. It has been also shown recently that lanthanum telluride in bulk n-type material has a high $\mathrm{ZT} \sim 1.2$ performing above 1000 K [5]. One of the alleged reasons for the improved TE efficiency in this class of materials is the contribution of the strong localized f- and d-states located near the Fermi level affecting on the TE transport properties at high temperature. Similar properties are also observed in the cerium compounds [6], [7]. In addition, exploiting the Mahan-Sofo theory with assumption of δ-shape of the density of states, it has been suggested lately that ZT can be huge at high temperature in cerium telluride [7], [8]. Therefore, taking advantage of the localized f-states near Fermi level by introducing a cerium atom into the conventional TE material to improve ZT could be of great interest. In this paper, employing the precise full-potential linearized augmented plane-wave (FLAPW) method [9], we examine the effect of the Ce substitution into the Bi_2Te_3 on its transport properties. We study the influence of the strongly localized f-states and its role on the TE transport properties by analyzing the electronic structures and computing the thermoelectric transport coefficients utilizing the Boltzmann's equation in the constant relaxation-time approach.

II. Computational Methods

The crystal structure of $CeBiTe_3$ is rhombohedral with the space group of $D^5_{3d}(R\bar{3}m)$ with broken inversion symmetry and broken time reversal symmetry. The experimental lattice constants in hexagonal form are $\mathrm{a} = 4.19$ Å and $\mathrm{c} = 31.6$ Å [10]. In the framework of the density functional theory (DFT), the electronic structure calculations are performed within the generalized gradient approximation (GGA) for the exchange potential [11]. The strong correlation is additionally treated by the Hubbard U correlation, via GGA + U [12].

The U and J values of 6.1 and 0.7 eV, respectively, are adapted in our calculation as widely used in other studies for Ce compounds [13]–[16]. Spin-orbit coupling (SOC) is included in the second variational way to describe the relativistic effect of valence electrons [17], [18]. The Brillouin zone (BZ) integration is performed using $17 \times 17 \times 17$ Monkhorst-Pack special k-points [19].

In the structural aspect, $CeBiTe_3$ is obtained by replacing one of the two Bi atoms in Bi_2Te_3 crystal substitutionally with a Ce

Manuscript received May 06, 2013; revised July 14, 2013; accepted August 18, 2013. Date of current version December 23, 2013. Corresponding author: M. Kim (e-mail: mykim.nu@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMAG.2013.2279854

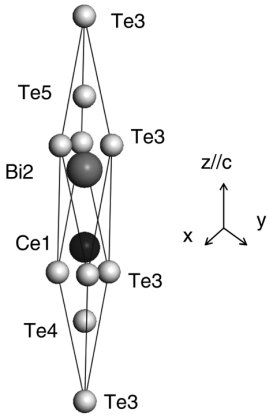


Fig. 1. Primitive rhombohedral cell of $CeBiTe_3$. Big grey ball stands for Bi, dark ball Ce, and small bright ball Te.

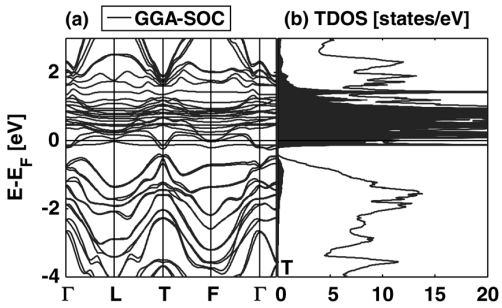


Fig. 2. Calculated results of (a) the electronic band structure and (b) the total density of states of $CeBiTe_3$ within GGA. Shaded DOS stands for the 4f-projection of DOS.

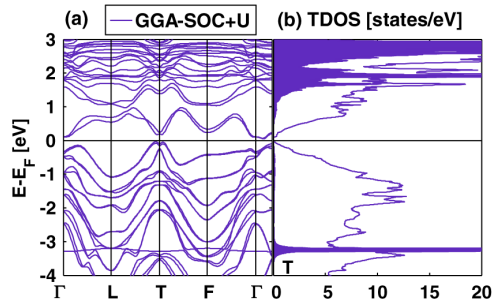


Fig. 3. Calculated results of (a) electronic band structure and (b) density of states of $CeBiTe_3$ within GGA + U. Shaded DOS stands for the 4f-projection of DOS.

atom as shown in Fig. 1. All atoms are fully relaxed by calculating the atomic forces [18] to obtain the equilibrium positions employing the experimental lattice constants. It is well known that the single crystal Ce can exist in two phases depending on the pressure. The γ-Ce phase has the Curie-Weiss susceptibility while the α-Ce phase has the Pauli paramagnitism with a collapse of volume [14], [20]. It has been shown that the α-phase could be well described using LDA while the γ-phase needs to adapt a beyond LDA approach to describe the strong exchange correlation effect. Therefore, we have shown the calculated results by both GGA and GGA + U approaches in this work.

III. Electronic Structure

The calculated total spin magnetic moments of $CeBiTe_3$ are found to be 0.9 and 1.0 μ_B per formula unit cell in the GGA and GGA + U approaches, respectively. However, the calculated electronic structures exhibit the big change obviously due to +U as can be seen from Figs. 2 and 3, where we showed the band structure and density of states (DOS) of $CeBiTe_3$ calculated in GGA and GGA + U. It was found that $CeBiTe_3$ is metallic in GGA while its DOS at Fermi level (E_F) reduces drastically, tending to be a semiconductor with a very narrow energy band gap of 48 meV in GGA+U. Within the GGA, the $4f$-electrons of Ce contribute dominantly to the total DOS around the Fermi energy. A single flat band right below E_F is found contributed mostly by the $4f$ electron of Ce atoms, which turns out to contribute meaningfully in the TE transport properties (will be shown later). Except for this single band, most of the spin-minority $4f$ electrons are found unoccupied and widely ranged up to 2 eV above E_F. As a result, the spin magnetic moment is found to be 0.9 μ_B. With the contribution of the sharp $4f$-peak, the total DOS at E_F is much greater than that of Bi_2Te_3, which could be a good sign promoting $CeBiTe_3$ to be a potential candidate for TE application according to Mahan-Sofo theory [8].

When we invoke the strong correlation by GGA + U, the $4f$ states split into two distinct peaks located far away from each other in energy as can be seen from Fig. 3. The single f-band which was observed just below E_F in GGA is now shifted deep below Fermi energy and localized at -3.0 $\sim$ -3.5 eV. The unoccupied f-states are shifted up to the energy region of 1 eV from E_F and above. As a result, the spin magnetic moment is slightly enhanced to 1.0 μ_B. Now with the $4f$-bands far from E_F, we expect that the contribution to the transport properties is mainly by s and p states of both Te and Bi, and d states of Ce instead of the localized Ce $4f$-states. Compared with the Bi_2Te_3, the magnitude as well as the slope of the DOS for $CeBiTe_3$ near E_F is found to be smaller than that of Bi_2Te_3, which may cause a reduction of the TE coefficients. Another distinct effect of +U correction is an opening of the energy band gap of 48 meV. Therefore, the band gap is reduced due to the Ce substitution from the value of Bi_2Te_3 (0.162 eV in experiment [20] and 0.154 eV in first-principle calculation [21]). Opening of the band gap along with the strong localization is known to be beneficial for the improved TE properties. We also note that the shift of $4f$-states due to the strong correlation effect is similar to what was observed in the electronic structure change of the γ-phase Ce due to the external pressure [14], [22]. The pressure was found to govern the behavior of the spin-minority $4f$-states for the γ-phase Ce resulting in a band split and the enhanced band localization which were not observed in the α-Ce. Therefore, while the distinct band localization at E_F is not explicitly found in our GGA+U calculation, our observation of a metal-to-semiconductor transition through a band gap opening upon the strong correlation-effect correction implies that the strong correlation effect via Ce substitution, probably together with the applying an external strain effect, can be used to tune the TE properties.

Computational Materials Science 81 (2014) 510–516

Contents lists available at ScienceDirect

Computational Materials Science

journal homepage: www.elsevier.com/locate/commatsci

The surface electronic properties of newly designed half-metallic ferromagnets: GeKCa and SnKCa

Beata Białek [a], Jae Il Lee [a,*], Miyoung Kim [b]

[a] Department of Physics, Inha University, Incheon 402-751, Republic of Korea
[b] Department of Physics, Sookmyung Women's University, Seoul 140-742, Republic of Korea

ARTICLE INFO

Article history:
Received 11 June 2013
Received in revised form 28 August 2013
Accepted 30 August 2013
Available online 5 October 2013

Keywords:
Half-metal
First-principles calculations
Half-Heusler alloy
Surface electronic structure

ABSTRACT

The electronic and the magnetic properties of the (001) surfaces of GeKCa and SnKCa with half-Heusler structure are studied with the use of a full-potential linearized augmented plane wave method. It is shown that although the two compounds are half-metals in their bulk structures, only the surfaces terminated with a carbon group atom retain the half-metallic properties. The magnetic properties of the surfaces terminated with layers containing group 14 atoms are enhanced compared with the properties of the bulk. The calculated magnetic moments on the Ge atom in GeKCa are $0.38\mu_B$ in the bulk and $0.97\mu_B$ at the (001) surface. In the SnKCa surface, the value of the magnetic moment on the Sn atom increases from $0.28\mu_B$ in the bulk to $0.75\mu_B$ at the surface. In the Ge (Sn) terminated Ge(Sn)KCa surfaces, the metal atoms are also polarized. In addition to destroyed half-metallicity at the surfaces terminated with metal atoms we also find a strong demagnetization of the systems.

1. Introduction

Half-metallic ferromagnets (HMFs) are characterized by a complete spin polarization of electrons at the Fermi level (E_F) [1] even in the ground state. Such materials are rare in nature, but nowadays there are many known compounds possessing these properties. Many of them were computer designed first and synthesized later. For example, Akinaga et al. synthesized CrAs in zinc-blende (ZB) structure by molecular beam epitaxy [2]. ZB CrSb thin films were also grown by that method [3], and epitaxial layers of ZB (Ga,Mn)As were deposited on the Si (001) surface [4]. Those experiments prove that it is possible to obtain materials of desired magnetic properties, even though their crystal structure is not the most stable one.

The latest trend in search of new materials in the group of HMFs is studying substances that possess magnetic properties, in spite of the lack of a transition metal in their structure. For example, Kasukabe et al. reported p-electron magnetism in a ZB CaAs [5]. The magnetic moment in the material appeared due to 4p electrons of As; 4p As orbitals transformed using a proper symmetry group so they could hybridize with the t_{2g} d orbitals of Ca [5]. That finding opened a possibility for materials design where the topology of p (or π) orbitals was a guiding principle.

Since then, many similar materials were proposed as candidates for application to spintronics; a vast group of them are compounds of group 2 and group 14 or 15 elements. Gao et al. found p-electron half-metallic ferromagnetism in ZB CaSi and CaGe [6], and in rock-salt SrC and BaC [7], as well as in several other compounds (see for example [8,9]). Recently, attention has also been drawn to potential sp-electron HMFs of Heusler or half-Heusler structure. This is mainly because the Heusler-based materials have T_C much higher than the room temperature [10]. High T_C is very important for application to spintronic devices since it stabilizes the half-metallicity of the material through a small reduction of the spin polarization at room temperature. Unfortunately, the same compound prepared in the form of a thin film no longer possessed the half-metallic properties of the bulk, which was attributed to a disorder in the crystal structure [11]. We surmise, however, that the alteration can also be attributed to surface effects, which destabilize the properties of the bulk.

In an attempt to combine the desired properties of half-Heusler alloys and sp-electron HMFs, Chen et al. investigated a new group of materials, i.e. GeKCa and the SnKCa in the half-Heusler structure [12]. Using the first-principles pseudopotential plane wave method, they found that both the GeKCa and SnKCa are HMFs and the property is preserved even if their lattice constants are contracted by 10% (GeKCa) or 13% (SnKCa) of their equilibrium values, 7.58 and 7.99 Å, respectively. These values are much larger than the ones determined for the compounds containing 3d metals, such as CoMnSb or NiMnSn [13,14]. Hence, it is likely that while applied in the form of thin films, these new materials would preserve their half-metallic properties at a wide range of temperatures.

* Corresponding author. Fax: +82 32 872 7562.
E-mail address: jilee@inha.ac.kr (J.I. Lee).

0927-0256/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.commatsci.2013.08.061

Even though sp-electron ferromagnetic solids have not yet been synthesized, the challenge will most likely be addressed due to the promising properties. Geshi et al. have already proposed a synthesis process of ferromagnetic nitrides, i.e. CaN and SrN, in RS structure [15]. Alkaline earth carbides, such as CaC_2, are synthesized in a high yield and with a high purity in monoclinic structure [16]. In one of the possible methods, synthesis of the carbides involves soaking highly oriented pyrolitic graphite in a Li–Ca melt [17]. Perhaps similar methods may be used in the future to synthesize the compounds of K, Ca, and Ge or Sn as bulk materials. Thin films of KCa compounds with Ge or Sn could form interfaces with other sp-electron magnetic materials, since they happen to have quite a large lattice constant; for example, the lattice constants of a few alkaline earth silicides in ZB structure range from 6.55 to 7.55 Å [18].

In this paper, we present and discuss the results of calculations of the electronic and magnetic properties of (001) surfaces of GeKCa and SnKCa in the half-Heusler structure carried out with the use of a first-principles method.

2. Computational method

In half-Heusler alloys the three elements constituting the compound form three interpenetrating face centered cubic sublattices. GeKCa and SnKCa form the most energetically stable configurations when group 14 atom (Ge or Sn) occupies the (0,0,0) position while K and Ca occupy the $(\frac{1}{4},\frac{1}{4},\frac{1}{4})$ and $(\frac{3}{4},\frac{3}{4},\frac{3}{4})$ positions, respectively [12]. The calculated equilibrium lattice constant of the half-Heusler GeKCa in the most stable arrangement is 7.58 Å and that of the SnKCa is 7.99 Å [12]. We adapted the lattice constants to the study of (001) surfaces of GeKCa and SnKCa.

The surfaces were simulated by repeated slabs consisting of 13 atomic layers. A building block of these slabs is shown in Fig. 1. The separation between the slabs was twice the slab thickness, which made the interactions between the slabs negligible. We considered two possible terminations of the surfaces. Fig. 1a illustrates the termination with a carbon group atom, i.e. Ge in GeKCa and Sn in SnKCa. We refer to these terminations as Ge-term and Sn-term, respectively. In the other termination, illustrated in Fig. 1b, the topmost layer consists of K and Ca atoms. We call this termination KCa-term. Since the predicted half-metallicity of the materials in their bulk structures is not sensitive to the lattice constant contraction within 10% and 13% of their equilibrium values [12], the geometrical parameters were not optimized. All the calculations were carried out with the use of a first-principle full potential linearized augmented plane wave (FLAPW) method [19,20] with the generalized gradient approximation [21] to the exchange–correlation potential. Lattice harmonics with $l \leqslant 8$ were employed to expand the charge density, potential, and wave functions inside the muffin-tin radii (MT) of 2.50, 2.60, 3.70, and 3.00 a.u. for Ge, Sn, K, and Ca atoms, respectively. The number of basis functions was about 200 per atom. A plane wave cut-off of 16 Ry was employed. Integrations were performed over a $7 \times 7 \times 1$ mesh of k-points inside the irreducible three-dimensional Brillouin zone. All core electrons were treated fully relativistically, while valence states were treated scalar relativistically. The self-consistent calculations were considered to be converged only when the integrated charge difference per formula unit between the input and the output charge densities was less than 1×10^{-4}.

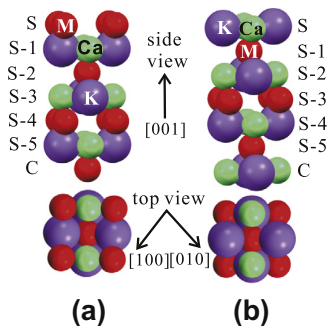


Fig. 1. Arrangement of atoms in a building block of the MKCa (M = Ge, Sn) (001) surfaces (a) M-terminated and (b) KCa-terminated. Only half slabs are shown. C dentoes the center layer, S – the topmost layer; S – n ($n = 1, \ldots, 5$) numerate the other layers from the subsurface ($n = 1$) toward the bulk.

3. Results and discussion

3.1. The total densities of states

As bulk materials, both GeKCa and SnKCa are HMFs with energy gaps in the spin-up electron channels [12]. Cleaving the bulk structures and exposing (001) planes results in slabs whose electronic properties depend on surface termination. As seen in Fig. 2, in which the calculated total density of states (DOS) of the GeKCa surfaces are presented, the half-metallic properties of the bulk are sustained only in the Ge-term surface. The calculated width of the energy gap in the spin-up electron channel in this system is 0.80 eV, which is smaller than the reported 1.21 eV for the bulk structure.

In the KCa-term surface, there are electronic states present at E_F in both spin channels. As discussed later, the new states at E_F are contributed by the electrons of surface and subsurface atoms. Polarization of electrons at E_F is 70% and the exchange splitting of the spin-up and spin-down electronic states is small. The ferromagnetic order of the bulk is supressed at the surface. In the half-metallic Ge-term GeKCa surface, the surface states are as near E_F as 0.04 eV. This distance in energy is called a half-metallic gap and is a measure of the excitation energy for an electron in the top of the valence band. The narrow half-metallic band gap may be disadvantageous from an application point of view, since the spin-up elec-

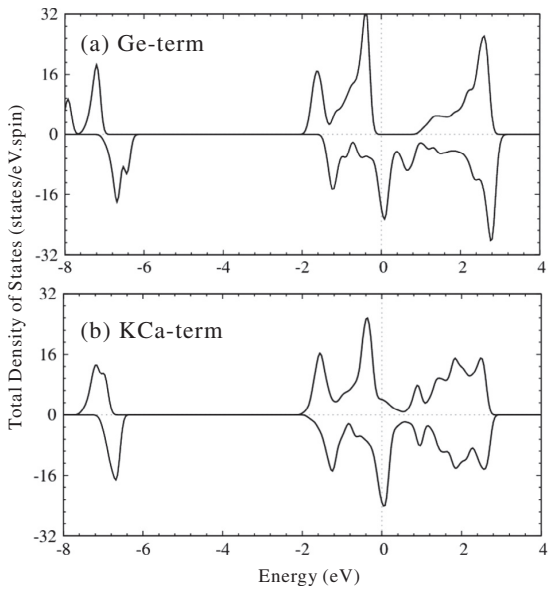


Fig. 2. The total spin-polarized density of states (DOS) calculated for GeKCa (001) surfaces of (a) Ge-term, and (b) KCa-term. The spin-down DOS values are multiplied by -1, and the Fermi levels are set to zero.

Effects of polycrystallinity in nano patterning by ion-beam sputtering

Sun Mi Yoon,[1] J.-S. Kim,[1,a)] D. Yoon,[2] H. Cheong,[2] Y. Kim,[3] and H. H. Lee[3]
[1]*Department of Physics, Sook-Myung Women's University, Seoul 140-742, South Korea*
[2]*Department of Physics, Sogang University, Seoul 121-742, South Korea*
[3]*Beamline Research Division, Pohang Accelerator Laboratory (PAL), Pohang 790-784, South Korea*

(Received 4 May 2014; accepted 30 June 2014; published online 10 July 2014)

Employing graphites with distinctly different mean grain sizes, we study the effects of polycrystallinity on the pattern formation by ion-beam sputtering. The grains influence the growth of the ripples in a highly anisotropic fashion; both the mean uninterrupted ripple length along the ridges and the surface width depend on the mean size of the grains, which is attributed to the large sputter yield at the grain boundary compared with that on the terrace. In contrast, the ripple wavelength does not depend on the mean size of the grains, indicating that the mass transport across the grain boundaries should efficiently proceed by both thermal diffusion and ion-induced processes. © *2014 AIP Publishing LLC*. [http://dx.doi.org/10.1063/1.4889976]

I. INTRODUCTION

The incidence of energetic ion beams causes the displacement and removal of atoms near the surface of a substrate. This process is conventionally referred to as ion-beam sputtering (IBS). Simultaneously, the healing kinetics proceeds via mass transport to minimize the surface free energy of the modified surface. Those two competing processes often produce patterns with nanodots/holes or ripples, depending on the incidence angle of the ion beam. IBS is one of the most versatile tools available to fabricate nano patterns in various sizes and shapes by controlling physical variables, and it is applicable to a wide range of materials from metals and insulators to organic materials. Thus, IBS has drawn attention as a representative method for triggering physical self-assembly.[1,2]

Continuum models have mostly elucidated the pattern formation by IBS. In their seminal work, Bradley and Harper (BH) (Ref. 3) proposed a model of the pattern evolution: they took the erosion into account according to Sigmund's model[4] in the linear approximation, and considered the diffusion of adatoms according to Mullin's model.[5] According to Sigmund, the sputter yield at a position on the surface is proportional to the energy deposited at that position by an incident ion. The energy depends on the Gaussian ellipsoidal function of the distance from a terminal position of an incident ion to the position at the surface.[4] According to Mullin's model, the adatom current is proportional to the gradient of the surface free energy, which is in proportion to the curvature at the position of interest.[5] The BH model and its nonlinear extensions[6–9] that incorporate surface-confined mass flow, redeposition, and surface damping qualitatively well reproduce many features of the pattern formation by IBS. Those models, however, tacitly assume an amorphous surface. Refined models have also been proposed taking the crystallinity[10] of the substrate and the anisotropy[11] of the surface and sputter geometry combined into account.

Polycrystals consist of the grains, whose mean size is comparable with the characteristic size of the structures formed by IBS.[12,13] Thus, polycrystals offer a unique environment for the pattern formation by IBS contrasted with the homogeneous substrates such as the amorphous and single crystals. A challenging question is, then, whether and how the grained structure of the polycrystalline substrate affects the pattern formation and its temporal evolution.

Recently, Toma *et al.*[13] have studied the ripple evolution of polycrystalline Au films by IBS. With continued sputtering, the initially rough surface leveled off, and the pattern evolved as it did on a single-crystalline Au surface in regards to the surface width and ripple wavelength. This observation led them to the conclusion that the polycrystallinity did not influence the pattern evolution. However, Škereň *et al.*[14] reported that the variation of the sputter yield, depending on the orientation of the grains, induced the surface instability during IBS, and reproduced the morphological evolution of polycrystalline Ni films well without invoking the instability owing to the curvature-dependent erosion.[3] Thus, the two conclusions on the effects of the polycrystallinity contradict each other.

Highly oriented pyrolytic graphite (HOPG) is polycrystalline and composed of large grains, compared with metallic films such as Au (Ref. 13) and Ni.[14] HOPG would, thus, offer the opportunity to study the effects of the grained structure on the pattern formation in the limit opposite to that of the metallic films. Several groups have previously worked on patterning HOPG by IBS.[15–18] Their studies are, however, not focused on an interest in the effects of polycrystallinity on pattern formation.

In this work, we study the effects of the grained structure on the pattern formation, employing two different kinds of graphites, HOPG and natural graphite (NG). The mean grain size of NG is distinctly larger than that of HOPG. This controlled experiment clearly shows that the grain boundaries play a critical role in determining the mean uninterrupted length of the ripples along their ridges and the surface width, while they little influence the ripple wavelength. These highly anisotropic effects on the ripple evolution are

[a)]jskim@sm.ac.kr

attributed to the intricate roles of the grain boundaries in the temporal evolution of the primordial islands to the ripples during the pattern formation.

II. EXPERIMENTS

The IBS of both HOPG (ZYA grade, SPI) and NG (donated from Union Carbide) samples was performed in a high vacuum chamber, whose base pressure was 5×10^{-10} Torr. The ion-irradiated surface is characterized *ex situ* by atomic force microscopy (AFM) in both the contact (PSI, Autoprobe CP) and the noncontact modes (Park Systems, XE-100). Sputtering is performed by irradiation of the Ar^+ ion beam with a beam diameter ~10 mm and an incident ion energy E_{ion}, 2 keV at a polar angle of incidence θ, 78° from the global surface normal. The partial pressure of Ar^+, P_{Ar}, and the ion flux f are 1.2×10^{-4} Torr and 0.3 ions nm^{-2} s^{-1}, respectively. The sample temperature is kept around room temperature by limiting each sputter period to 1 min in intervals of 10 min.

The Raman spectra of the samples were obtained by using a micro-Raman spectroscopy system.[19,20] The 514.5 nm (2.41 eV) line of an Ar ion laser was used as the excitation source, and the laser power was kept below 1 mW to avoid unintentional heating. The laser beam was focused with a spot size (diameter) < 1 μm onto the graphite sample by a 50× microscope objective lens (0.8 N.A.), and the scattered light was collected and collimated by the same objective. The collected Raman scattered light was dispersed by a Jobin-Yvon Triax 550 spectrometer and detected by a liquid-nitrogen-cooled charge-coupled-device detector. The spectral resolution was about 1 cm^{-1}.

The grazing incidence x-ray diffraction (GID) of the samples was investigated with 20 keV photons ($\lambda_{ph} = 0.62$ Å) in the 5A beam line of a Pohang light source in Korea. The incident angle of the x-ray was kept around 0.1° from the global sample plane to reduce both the beam penetration depth and bulk diffuse scattering.

III. RESULTS

In Fig. 1(a), the Raman spectrum of HOPG before IBS shows three bands, labeled G, G*, and 2D, which is typical of *pristine* HOPG.[21] After an extended sputtering with an ion fluence ψ, flux times total sputter time, $\psi = 5625$ ions nm^{-2}, its surface is patterned by ripples as shown in the inset of Fig. 1. Two new bands, D and D′, develop as shown in Fig. 1(b), while both G and 2D bands notably weaken as previously reported.[17,18] Both D and D′ bands originate from defective carbon atoms.[21,22] The observed spectral changes show that sputtering produces defective carbons at the cost of the *pristine* carbons.

The sample, however, largely remains crystalline; the spectral shape of the *amorphous* carbon shows broad adjoined peaks in the D and G bands,[23] while the present spectrum in Fig. 1 shows a well-defined D band with a still intense and sharp G band. Since the Raman spectrum could also sample the *pristine* layers beneath the damaged surface layers, we exfoliated the surface layers of the HOPG after transferring it to a silicon oxide substrate. Fig. 1(c) shows a

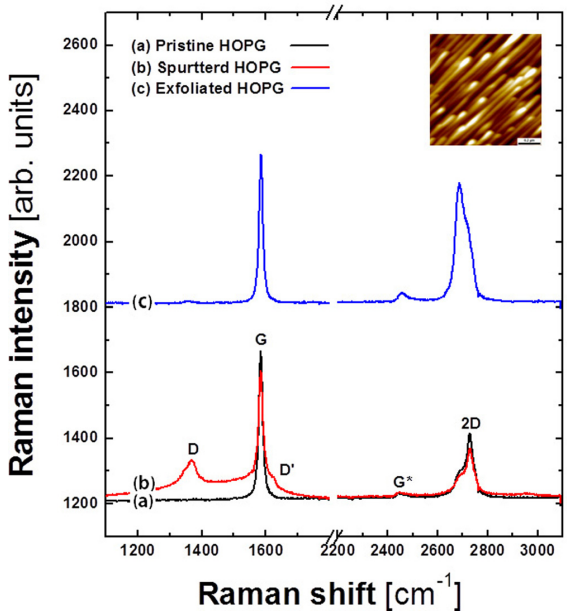


FIG. 1. Raman spectra (a) before and (b) after sputtering HOPG with $\psi = 5625$ ions nm^{-2}. (c) Raman spectrum of the surface layers peeled off from the HOPG giving spectrum (b). The graph (c) is vertically shifted for clear presentation. Inset: A nano ripple pattern corresponding to spectrum (b).

Raman spectrum of the exfoliated surface layers. They should be thinner than 10 layers, because the intensity of the 2D band is now comparable with that of the G band.[19] In Fig. 1(c), we still see a sharp G band with a negligible D band, confirming the crystallinity of the sputtered surface.

Takahiro *et al.*[17,18] also observed the development of both D and D′ bands with IBS by Ar^+. Their intensities monotonically increase with the increase of E_{ion} (≥10 keV) for the same ψ, indicating that the ion beam creates defects in the deeper region and/or more effectively for the larger E_{ion}. Then, such defects could form in the terrace region with the increased frequency, where the carbon atoms are more tightly bonded due to their larger coordination number than those around the grain boundaries. The proliferation of the defect sites exposing the graphene layers would escalate the surface modification, and even lead to the amorphization of the graphite as observed in the previous reports.[17,18] For the present sputtering condition, E_{ion}, 2 keV, seems too small to modify the surface region into the amorphous state, and leaves the sample in a largely (poly) crystalline state.

Under the sputtering condition similar to the present one, the morphological evolution of Au(001) is governed by erosion.[24] However, the erosion rate on the graphite surface should be lower than on Au(001) due to the stronger bonding within a layer of the graphite than that between Au atoms. Moreover, the adspecies such as carbon and carbon platelet should diffuse very swiftly on the graphite surface due to a strong intraplanar sp^2 bond of the graphite surface, and heal the defects quite efficiently. In short, the sputter condition belongs to the diffusive regime, limiting the amorphization of the graphite by IBS.

Applied Surface Science 324 (2015) 240–244

Contents lists available at ScienceDirect

Applied Surface Science

journal homepage: www.elsevier.com/locate/apsusc

Temporal evolution of the chemical structure during the pattern transfer by ion-beam sputtering

N.-B. Ha, S. Jeong, S. Yu, H.-I. Ihm, J.-S. Kim*

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

ARTICLE INFO

Article history:
Received 1 May 2014
Received in revised form
19 September 2014
Accepted 13 October 2014
Available online 22 October 2014

PACS:
81.16.Rf
79.20.Rf

Keywords:
Ru
Si
Metallic glass
Ion beam sputtering
Nano patterning

ABSTRACT

Ru films patterned by ion-beam sputtering (IBS) serve as sacrificial masks for the transfer of the patterns to Si(1 0 0) and metallic glass substrates by continued IBS. Under the same sputter condition, however, both bare substrates remain featureless. Chemical analyses of the individual nano structures simultaneously with the investigation of their morphological evolution reveal that the pattern transfer, despite its apparent success, suffers from premature degradation before the mask is fully removed by IBS. Moreover, the residue of the mask or Ru atoms stubbornly remains near the surface, resulting in unintended doping or alloying of both patterned substrates.

1. Introduction

The incidence of energetic ions on substrates results in the ejection, redeposition, and displacement of substrate materials. Those processes as a whole are conventionally referred to as ion-beam sputtering (IBS). If a surface is modified by IBS, a nanoscale pattern can spontaneously form on the surface [1,2] in the course of curing the damaged surface via various types of mass displacement [3–5]. The shapes and sizes of the patterns can be easily controlled by adjusting the physical variables, such as the angle of incidence of the ion beam, the beam energy, the beam flux, and the temperature of the substrate [1,2]. Moreover, the patterns can be fabricated on a macroscopic scale by IBS with relatively low cost, and the target materials allowing IBS-induced nanopatterning range from metal [6–9], semiconductor [10–14], oxides [15], to organic materials [16]. Therefore, IBS is a versatile tool for nanopatterning and can complement lithographic methods.

Well-ordered Si nanostructures have drawn a great deal of attention because of their great potential in applications in electronic and electro-optic devices [17,18]. Numerous theoretical and experimental works on nanopatterning Si by IBS have also been reported [2]. One notable result is that the parameter space where the ordered patterns form is quite restrictive for Si [19] compared with other cases e.g., for metals. In addition, the chemical defects can exert unpredictable, but very significant effects on the pattern formation on Si [20,21]. Thus, nanopatterning Si by IBS still presents a challenge.

Similarly, metallic glasses have been reported to remain flat and featureless under IBS, possibly due to low efficiency in mass displacement [22,23]. Nanopatterning the metallic glass has thus been performed mainly by lithographic methods employing focused ion beams and chemical etching [24].

Thus, an important issue is to find remedies to fabricate nanopatterns on substrates such as Si and metallic glass by IBS under sputter conditions that are adverse to pattern formation. In order to address the aforementioned issues, we give attention to the recent work reporting the transfer of a pattern from an overlayer to its substrate by IBS [25]. Here, the pattern in the Au film formed by IBS works as a sacrificial mask, which is removed by continued sputtering, leaving its pattern transferred to its glass substrate as depicted in Fig. 1.

In the present work, we transfer the patterns formed on 100-nm-thick Ru films by IBS to both Si(1 0 0) and metallic glass substrates by continued IBS. The patterns formed on the Ru films apparently transfer very well to both substrates by IBS. However,

* Corresponding author. Tel.: +82 27109406.

http://dx.doi.org/10.1016/j.apsusc.2014.10.053
0169-4332/

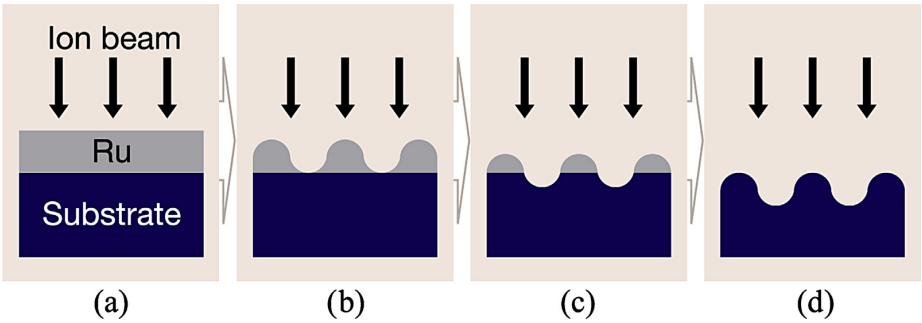


Fig. 1. Schematic of the pattern transfer.

the transferred pattern starts to deteriorate before the overlayer is fully removed. Moreover, the residue of the overlayer stubbornly remains near the patterned surface, resulting in unintended doping or alloying of the patterned substrates. Despite its promise, the pattern transfer by IBS via the sacrificial mask should be performed with caution.

2. Experimental

Both Si(1 0 0) (prime grade, B-doped, ρ = 1–5 Ω) and $Zr_{57.5}Cu_{15.6}Ni_{12.8}Al_{10.3}Nb_{2.8}$ bulk metallic glasses were used as the substrates for this pattern transfer. Before depositing the Ru film, Si(1 0 0) was etched by HF and then rinsed with de-ionized water. The bulk metallic glass grown by arc melting of its mother material was cut into slices by a diamond saw and extensively sputtered by an ion beam to expose a clean surface.

Ru films were deposited on the clean substrates by the magnetron sputtering in a high vacuum chamber whose base pressure was 2.4×10^{-7} Torr. Thicknesses of the Ru films were around 100 nm as judged by a cross-sectional image taken by a scanning electron microscope. The surface width W after the deposition of the Ru film is less than 0.5 nm for both substrates as estimated by atomic force microscopy (AFM) (PSI CP) in the contact mode.

IBS was performed in the two different chambers at room temperature. One is a high vacuum chamber whose base pressure is 5×10^{9} Torr and used to perform nano-patterning by using a Kauffman-type ion source with a beam diameter of ~10 mm. The ion energy and the beam flux at the sample position were respectively, 2 keV and 0.56 ions/nm^{2}s.

Another chamber housed a scanning electron microscope (SEM) with scanning Auger microscope (SAM) capability (PHI Nanoprobe 700) with its nominal spatial resolution 8 nm, allowing *in situ* investigation of both the surface topography and the chemical structure. The base pressure of the chamber was maintained around 2.0×10^{-9} Torr. IBS in the chamber was conducted employing a rastering Ar^+ beam; the beam energy was 2 keV, and the area of the raster was 2 mm × 2 mm. The ion flux was equivalent to 0.56 ions/nm^{2}s that is determined by the electric current at the sample position. Since the secondary electrons contribute to the current, the abovementioned ion flux is nominal, giving the upper limit of the real ion flux.

3. Results and discussion

Fig. 2 shows the typical AFM images after sputtering bare (a) Si(1 0 0) and (b) the metallic glass by using the wide ion beam incident at 75° from the surface normal (θ). Ion fluence F, ion flux times the total sputter time were 1076.6 and 3594.8 ions/nm^2 for Si(1 0 0) and the metallic glass, respectively. A pattern is observed neither on Si(1 0 0) nor on the metallic glass.

Fig. 3(a) shows Auger electron spectra of the 100-nm-thick Ru covered Si(1 0 0) with increasing F. The inset under each spectrum is a corresponding SEM image showing a ripple pattern and a dot on the pattern indicating where the Auger spectrum is taken. The SEM image with F = 1056.0 ions/nm^2, the top inset, shows a ripple pattern with the ridges of the ripples running along the ion beam direction. The corresponding Auger spectrum I shows only Ru peaks, indicating that the ripples are still covered by Ru.

Further sputtering with F = 2112 ions/nm^2 produces the ripple pattern as shown in the middle inset of Fig. 3(a). The corresponding Auger spectrum II indicates that now the Si peak has become the major one. The remaining Ru film may be very thin in the scale of

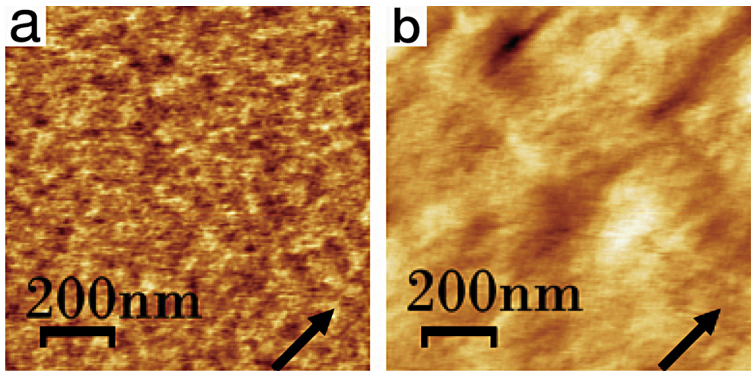


Fig. 2. AFM images of the sputtered (a) Si(1 0 0) and (b) a metallic glass show no pattern. The image size is 1 μm × 1 μm. The arrow shows the direction of the incident ion beam projected to the surface. The incident angle and energy of Ar ion are, respectively, 75° from the surface normal and 2 keV.

Journal of the Korean Physical Society, Vol. 64, No. 4, February 2014, pp. 543∼546

Atomic and Electronic Structures of ZnO Divacancy in Hexagonal ZnO

Eun-Ha Shin and Hanchul Kim*
Department of Physics, Sookmyung Women's University, Seoul 140-742, Korea

(Received 7 January 2014)

We report density functional theory calculations for the bulk ZnO divacancy in hexagonal ZnO and compare the results with the monatomic bulk vacancies. We find that the ZnO divacancy could be viewed as a tightly-bound state of a doubly-negative Zn vacancy and the doubly-positive O vacancy from the structural and the electronic points of view. The singly-positive ZnO divacancy is found to carry a magnetic moment of $1\mu_B$, but it can only be stabilized in very high p-type samples. Otherwise, the non-magnetic neutral divacancy is stable. These results suggest that the ZnO divacancies in the bulk region cannot play a significant role in stabilizing the recently-observed ferromagnetism in ZnO nanoparticles.

PACS numbers: 61.72.Ji, 71.55.Gs, 71.55.-i, 75.50.Pp
Keywords: ZnO, Divacancy, Magnetism, Electronic structure
DOI: 10.3938/jkps.64.543

I. INTRODUCTION

Recently, metal oxides that are intrinsically nonmagnetic, such as HfO_2, TiO_2, and ZnO, were reported to show ferromagnetism (FM) without doping of magnetic- or transition-metal atoms at room temperature (RT) [1, 2]. Because these compounds do not have unpaired p- and/or d-electrons, the observation of FM in these metal-oxides has triggered extensive studies to determine the origin of the FM. Among different oxides, ZnO is an important semiconductor in optoelectronics for its wide band gap, ~ 3.4 eV, and its large exciton binding energy, ~ 60 meV [3]. In addition to these desirable optoelectronic properties, the recently observed RT-FM may lead to potential applications of ZnO in future spintronic devices.

Sundaresan and Rao conjectured that the surface O vacancies (V_O) and their exchange interaction would be the origin of the observed RT-FM in nonmagnetic metal-oxide nanoparticles [2]. Theoretically, there are suggestions that the bulk Zn vacancy (V_{Zn}) is an important candidate for the origin of the FM at RT. For instance, a V_{Zn} in a small supercell can induce ferromagnetism due to the p-orbitals of O atoms which are the nearest-neighbor atoms to the Zn vacancy [4, 5]. Wang *et al.* proposed that V_{Zn} energetically preferred to reside on the surfaces of thin films and nanowires having a finite magnetic moment [6]. On the other hand, V_O has been reported to be thermodynamically more stable than V_{Zn} over a wide range of atomic chemical potentials of ZnO [7,8]. If the stability of V_O and the capability of V_{Zn} to induce a magnetic moment are considered, the diatomic vacancy, made up of adjacent V_{Zn} and V_O, also might be a potential candidate for the origin of the observed RT-FM. In addition, the negatively-charged V_{Zn} is known to be more stable than the neutral vacancy and carry a reduced magnetic moment [7, 9]. These charged defects have been observed in experiments [10] and may play a role in stabilizing the RT-FM. As a whole, the Zn vacancy with different charge states and/or the ZnO divacancy (DV) may be important structural ingredients in understanding the experimentally-observed RT-FM of ZnO nanoparticles.

In this paper, we report first-principles calculations of ZnO DV, as well as the monatomic vacancies (V_{Zn} and V_O), in order to achieve a comprehensive understanding of the atomic and the electronic properties of the intrinsic vacancies in bulk hexagonal ZnO. The DV can be generated in two configurations: one along the a-axis and the other along the c-axis. The two different DVs are slightly different in energetic stability, but have similar electronic structures. The electronic density of states (DOS) of DVs can be regarded as a mixture of the DOSs of isolated V_O and V_{Zn}. Positively-charged states of a DV, which resemble the V_O, are available. Although DVs possess a finite magnetic moment in their singly-positive charged state, they are stable only in a very narrow region above the valence band maximum (VBM), indicating that the bulk DVs are not responsible for the RT-FM.

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

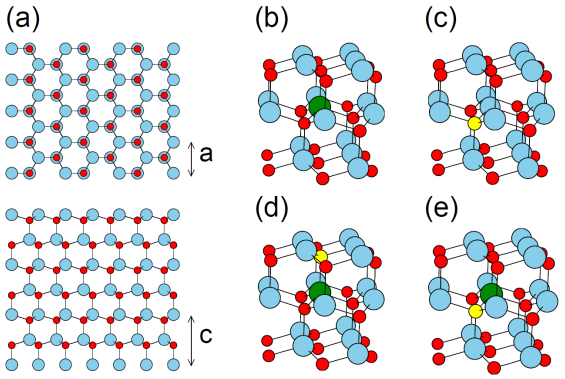


Fig. 1. (Color online) Structure of wurtzite ZnO and atomic vacancies: (a) top (upper) and side (lower) view of the $4a \times 3\sqrt{3}a \times 3c$ supercell, (b) $V_{\rm Zn}$, (c) $V_{\rm O}$, (d) DV_A, and (e) DV_Z. The Zn atom, Zn vacancy, O atom, and O vacancy are represented by light blue, green, red, and yellow balls, respectively.

II. COMPUTATIONAL METHODS

We performed density functional theory (DFT) calculations by using the Vienna *ab anitio* simulation package (VASP) [11]. We employed the plane-wave basis set with a kinetic energy cutoff of 450 eV, the projector-augmented-wave (PAW) potentials [12], and the Perdew-Burke-Ehrenkof (PBE) generalized-gradient approximation (GGA) exchange-correlation functional [13]. As the GGA functionals seriously underestimate the band gap of ZnO in comparison with the experiments, we used the DFT+U scheme to include the on-site Coulomb repulsion with the effective Coulomb repulsion parameter of $U = 6$ eV [14]. With this, the band gap was increased to 1.6 eV from the normal GGA value of 0.7 eV. To model the monatomic and the diatomic vacancies in bulk ZnO, we used a $4a \times 3\sqrt{3}a \times 3c$ supercell containing 288 atoms, as shown in Fig. 1(a). The theoretical lattice constants were $a = 3.18$ Å and $c = 5.12$ Å (with an a/c ratio of 0.388) and are in good agreement with the experimental values [15]. The **k**-space integration was performed using a Γ-centered $2 \times 2 \times 2$ **k**-point mesh [16]. The structures were relaxed until the quantum-mechanical Hellman-Feynman forces were smaller than 0.01 eV/Å.

The formation energy of an intrinsic defect, $E_{\rm form}$, in its charge state (q) can be calculated by using

$$E_{\rm form} = E_{\rm defect} - E_{\rm sc} + \sum_i n_i \mu_i + q(E_F + E_{\rm VBM}),$$

where $E_{\rm defect}$ is the total energy of a supercell with a defect, $E_{\rm sc}$ is the total energy of a pristine supercell, E_F is the Fermi level referred to the valence band maximum (VBM), $E_{\rm VBM}$ is the electronic energy at the VBM, and μ_i is the chemical potential of the element i (i = Zn, O). In thermodynamic equilibrium, the Zn and the O chemical potentials should satisfy

$$\mu_{\rm Zn} + \mu_{\rm O} = \mu^{\rm bulk}_{\rm ZnO},$$
$$\mu_{\rm Zn} \le \mu^{\rm bulk}_{\rm Zn},$$
$$\mu_{\rm O} \le \frac{1}{2}\mu^{\rm gas}_{\rm O_2}.$$

With the formation enthalpy defined by $\Delta H_f \equiv \mu^{\rm bulk}_{\rm ZnO} - (\mu^{\rm bulk}_{\rm Zn} + \frac{1}{2}\mu^{\rm gas}_{\rm O_2})$, which is calculated to be -3.42 eV in good agreement with the experimental value of -3.6 eV [15], the available chemical potential range of Zn is

$$\mu^{\rm bulk}_{\rm Zn} + \Delta H_f \le \mu_{\rm Zn} \le \mu^{\rm bulk}_{\rm Zn}.$$

Under a Zn-rich condition, $\mu_{\rm Zn} = \mu^{\rm bulk}_{\rm Zn}$ and $\mu_{\rm O} = \frac{1}{2}\mu^{\rm gas}_{\rm O_2} + \Delta H_f$, whereas under an O-rich condition, $\mu_{\rm Zn} = \mu^{\rm bulk}_{\rm Zn} + \Delta H_f$ and $\mu_{\rm O} = \frac{1}{2}\mu^{\rm gas}_{\rm O_2}$.

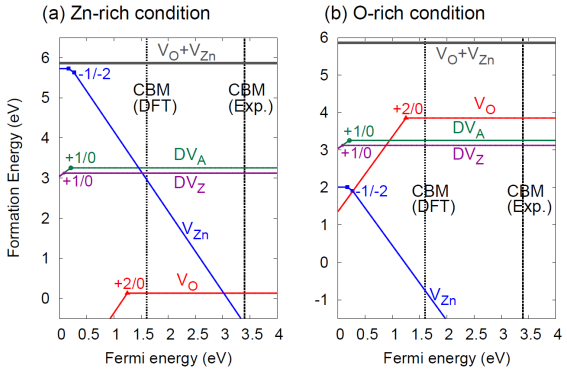


Fig. 2. (Color online) Formation energies of $V_{\rm O}$, $V_{\rm Zn}$ and DVs.

III. RESULTS AND DISCUSSIONS

The directions of the bonds in the hexagonal ZnO can be divided into two kinds: one is along the armchair chain in the c direction with a bond length of 2.01 Å, and the other is along the zigzag chain in the a direction with a bond length of 1.91 Å. Thus, the ZnO DV can be created along two different crystallographic directions, as shown Figs. 1(d) and 1(e). We will call the two DV structures as DV_A [Fig. 1(d)] and DV_Z [Fig. 1(e)].

The calculated energetics of the monatomic vacancies and the ZnO DVs in different charge states is shown in Fig. 2. The O vacancy is found to have stable charge states $q = 0$ and $q = +2$, and the charge transition level $\varepsilon(+2/0)$ is located at 1.3 eV above the VBM. As for the Zn vacancy, the stable charge states are $q = 0$, -1, and -2, and the $q = -2$ state is stable for most electronic chemical potentials (or the Fermi level). The charge transition levels are $\varepsilon(0/-1) = 0.17$ eV and $\varepsilon(-1/-2) = 0.27$ eV. As expected, the O vacancy is

Journal of the Korean Physical Society, Vol. 64, No. 4, February 2014, pp. 547∼553

Effect of Van der Waals Interaction on the Structural and Cohesive Properties of Black Phosphorus

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

(Received 14 January 2014)

We investigated the structural and the binding properties of black phosphorus (black-P) by employing density functional theory (DFT) calculations in combination with various implementations of the van der Waals (vdW) interaction. Both the binding energy curve of the two isolated puckered layers and the equation of states of the bulk black-P suggest that the conventional generalized gradient approximation (GGA) functional is incapable of describing the interlayer vdW interaction. From the comparison of the seven different vdW implementations, the appropriate vdW schemes in describing layer-structured black-P are found to be either the Grimme's dispersion correction (DFT-D2) or the optB86b vdW denisty-functional approach.

PACS numbers: 61.50.Lt, 61.66.Bi, 64.30.+t, 71.15.Nc
Keywords: Density functional theory, Van der Waals interaction, Black phosphorus, Layered structure
DOI: 10.3938/jkps.64.547

I. INTRODUCTION

Black phosphorus (black-P) has many benefits as an anode material for lithium ion batteries because of its atomic and electronic properties. A composite of black-P and carbon has been demonstrated to result in an enhanced electrochemical discharge/charge efficiency with a good cyclability [1]. Due to the layered structure, black-P can guarantee large specific capacity for lithium ion batteries. In addition, the semiconducting nature of black-P with a small band gap (0.19 eV) will allow many potential applications in electronics technology [2].

In nature, phosphorus is found in three allotropic forms with different dimensionalities, reminiscent of carbon allotropes such as the zero-dimensional (0D) fullerene, the one-dimensional (1D) carbon nanotube, and the two-dimensional (2D) graphene. The three phosphorus allotropes are classified by their characteristic colors. White phosphorus (white-P) is a reactive molecular crystal whose building unit is the 0D P_4 molecule. White-P is further classified into three structural modifications (α-, β-, and γ-P_4) due to different arrangements of P_4 molecules [3–5]. Red phosphorus (red-P) is made up of 1D polymeric chains. Finally, the most stable allotrope, black-P, is characterized by a staggered stacking of puckered layers (PLs) [6]. Judging from the structural building units, the van der Waals (vdW) interaction seems to play an important role in the structural and the thermodynamic properties of P allotropes.

Density functional theory (DFT) [7, 8] calculations have been successfully applied in describing the covalent and the ionic bonds in various molecules and crystals and has become the preferred tool in condensed matter physics. However, it has experienced problems in predicting the structural and the thermodynamic properties of weakly-bound molecular aggregates because the DFT is not capable of describing the long-range dynamical correlation between induced dipoles. Many efforts have been made to remedy this shortcoming. One class of approaches is to devise non-local density functionals describing the vdW interaction and is called the vdW density-functional (vdW-DF) scheme [9–15]. Kliměs *et al.* showed that the accuracy of the original vdW-DF could be improved by using alternative generalized-gradient-approximation (GGA) exchange components and proposed three variations of vdW-DF, optPBE, optB88, and optB86b [16, 17]. The other class is the semi-empirical approach to describe the vdW interaction by adding a pairwise interatomic dispersion correction (C_6R^{-6}) [18–22]. Within this class, depending on the way in which the C_6 coefficients are determined, the approaches can be further divided into the parametrized dispersion corrections, where the most widely used is the so-called DFT-D method by Grimme *et al.* [18–21], and the parameter-free methods represented by a recent work done by Tkatchenko and Scheffler (TS) [22].

The vdW forces are very weak compared to chemical interactions, but they play a crucial role in weakly-bound systems such as noble-gas elements and non-polar molecules [23], π-electron systems like carbon allotropes

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

Table 1. Binding properties of the two PLs from LDA, PBE, and different DFT+vdW calculations: equilibrium distance d_0 ($\AA$) and binding energy ε_b (eV/P_4).

	LDA		PBE							
	w/o vdW	D2	w/o vdW	D2	TS	DF	DF2	optPBE	optB88	optB86b
d_0	2.93	2.89	3.42	3.09	3.32	3.53	3.51	3.30	3.19	3.09
ε_b	0.17	0.34	0.03	0.17	0.19	0.15	0.17	0.20	0.22	0.24

(C_{60}, carbon nanotubes, graphite, *etc.*), and organic molecules in biological systems [11,12]. As a few examples, the interaction of the in-plane benzene molecules [24], the interactions between benzene and naphthalene molecules on graphite [25], and the interlayer binding in graphite [26] were studied to verify that the vdW interaction was crucial in stabilizing such materials. Very recently, the vdW-DF method was applied to bulk crystals, for which the vdW forces had previously been thought not to be that important [17].

To our knowledge, there has not been a comprehensive investigation of the role of the vdW interaction in P allotropes except for our recent study on the intermolecular interaction between P_4 molecules [27]. We found that the GGA predicted too weak an intermolecular binding whereas the local density approximation (LDA) results in too strong an intermolecular binding. Only with the use of the dispersion correction (DFT-D2, a modified version of DFT-D) in combination with the GGA, was a reasonable binding property obtained. As for black-P, early theoretical studies were focussed either on the electronic structure [28, 29] or on the structural phase transitions [30]. In a few recent studies, the researchers tried to consider the effect of the vdW interaction [31–33]. Du *et al.* [31] claimed that the Perdew-Wang GGA functional (PW91) [34] described the vdW interaction between two PLs with the interlayer distance of 5.1 Å and the binding energy of 0.06 eV/P_4, which is an incorrect proposition because the interlayer vdW distance in black-P is about 3.1 Å [35, 36]. Prytz and Flage-Larsen [32] examined the Heyd-Scuseria-Ernzerhof (HSE03) [37] and Perdew-Burke-Ernzerhof (PBE0) [38] hybrid functionals, and found that neither hybrid functionals could properly describe the vdW interaction in black-P. The HSE03 predicted an interlayer vdW distance (3.48 Å) shorter than the GGA value (3.52 Å), which is much larger than the experimental value (3.1 Å), and the PBE0 failed to relax the structure. Very recently, Appalakondaiah and coworkers [33] investigated the effect of the vdW interaction on the structural and the elastic properties of black-P. Among the studied semi-empirical vdW corrections, the DFT-D2 scheme [19] was found to yield structural parameters which are in closest agreement with experimental data.

In this work, we investigated the effect of the vdW interaction on the structural and the thermodynamical properties of black-P. First, we calculated the binding energy curve of the two PLs by employing various implementations of the vdW interaction. Then, we calculated the equation of states of black-P, where one can obtain the fundamental structural, elastic, and cohesive properties. Based on those calculated results, we were able to address the efficacy of different vdW implementations in describing the inter-PL interaction of black-P.

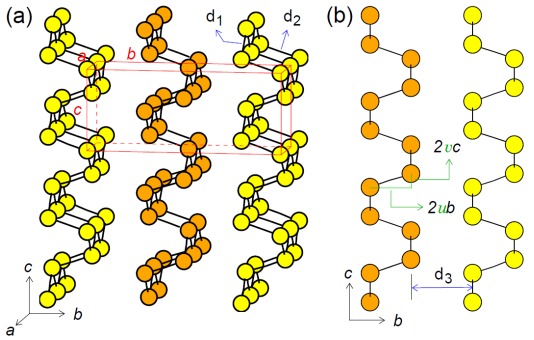


Fig. 1. (color online) (a) Structure of black-P showing the three lattice constants, a, b, and c, and the orthorhombic cell. d_1 and d_2 are the in-plane and the out-of-plane bonds, respectively. (b) Side view of black-P where the interlayer vdW distance d_3 and the internal coordinates u and v are defined. The alternating staggered PLs are represented by using different colors for distinction.

II. COMPUTATIONAL METHODS

We used the Vienna *Ab initio* Simulation Package (VASP) [39–42] to perform the DFT calculations with the vdW interaction, as well as the conventional DFT calculations. We employed two semi-empirical dispersion corrections, DFT-D2 [19] and TS [22], the original vdW-DF scheme of Dion *et al.* [10] and its modified version (vdW-DF2) [15], and three variations of vdW-DF (optPBE, optB88, and optB86b) [16, 17]. The phosphorus atoms were represented by using the projector-augmented-wave potential [43, 44]. The exchange-correlation was described within the Ceperley-Alder scheme [45] for the LDA and within the Perdew-Burke-Ernzerhof (PBE) scheme [46] for the GGA. An orthorhombic cell, which is shown in Fig. 1(a) as a red rectangular parallelepiped and is twice as large as the primitive unit cell, was used for the calculations of bulk

Stability and electronic structures of native defects in single-layer MoS_2

Ji-Young Noh,[1,2] Hanchul Kim,[2] and Yong-Sung Kim[1,3,*]

[1]*Korea Research Institute of Standards and Science, Yuseong, Daejeon 305-340, Korea*

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

[3]*Department of Nano Science, Korea University of Science and Technology, Daejeon 305-350, Korea*

(Received 15 July 2013; revised manuscript received 17 March 2014; published 19 May 2014)

The atomic and electronic structures and stability of native defects in a single-layer MoS_2 are investigated, based on density-functional theory calculations. Native defects such as a S vacancy (V_S), a S interstitial (S_i), a Mo vacancy (V_{Mo}), and a Mo interstitial (Mo_i) are considered. The S_i is found to have S-adatom configuration on top of a host S atom, and the Mo_i has Mo-Mo_i split interstitial configuration along the c direction. The formation energies of the native defects in neutral and charged states are calculated. For the charged states, the artificial electrostatic interactions between image charges in supercells are eliminated by a supercell size scaling scheme and a correction scheme that uses a Gaussian model charge. It is found that the V_S has a low formation energy of 1.3–1.5 eV in the Mo-rich limit condition, and the S_i has 1.0 eV in the S-rich limit condition. The V_S is found to be *a deep single acceptor* with the $(0/-)$ transition level at 1.7 eV above the valence-band maximum (VBM). The S_i is found to be *an electrically neutral defect*. The Mo-related native defects of V_{Mo} and Mo_i are found to be high in formation energy above 4 eV. The V_{Mo} is *a deep single acceptor* and the Mo_i is *a deep single donor*, of which the $(0/-)$ acceptor and $(+/0)$ donor transition levels are found at 1.1 and 0.3 eV above the VBM, respectively.

DOI: 10.1103/PhysRevB.89.205417 PACS number(s): 73.61.Le, 71.55.Ht, 73.20.Hb, 91.60.Ed

I. INTRODUCTION

Molybdenite (MoS_2) is an abundant mineral in the earth crust and has been widely used as a mechanical lubricant, a desulfurization catalysis in petrochemisty, and a principal ore for Mo extraction. The molybdenite crystal (2H-MoS_2) has a layered atomic structure [Fig. 1(a)], from which a single-layer MoS_2 can be fabricated by, for example, mechanical exfoliation [1–3]. There are a variety of interesting properties [4]. The MoS_2 is a semiconductor electrically. The 2H-MoS_2 bulk and thicker MoS_2 than a single-layer have an indirect band gap, while the single-layer MoS_2 has a direct band gap [5–7]. The two-dimensional nature of the single-layer MoS_2 implicates a strong Coulomb interaction between charged species therein, which has remarkable consequences on the correlation of electrons and charged defects. The fundamental (GW) band gap of a single-layer MoS_2 is 2.8 eV [8,9], but the large excitonic binding energy of 0.9 eV renders the optical band gap measured at 1.9 eV in absorption and photoluminescence spectroscopy [5,6]. In n-type single-layer MoS_2 thin-film transistors, the optically excited trions comprising two electrons and one hole have been also identified in presence of carrier electrons [10–13]. Besides the large electronic correlation effects, the intrinsic broken inversion symmetry in a single-layer MoS_2 induces spin-orbit split bands at the K and $-K$ points in the hexagonal Brillouin zone [14] and is now opening a new era for the spin- and valley-controlled electronics [15–19].

The carrier mobility of electrons in a single-layer MoS_2 is as high as $\sim$200 $cm^2\ V^{-1}\ s^{-1}$ in $HfO_2/MoS_2/SiO_2$ stacked (top-gate) thin-film transistors [1], while it lowers to about $\sim$1 $cm^2\ V^{-1}\ s^{-1}$ in an air/MoS_2/SiO_2 stacked bottom-gate structure [3]. The measured high electron mobility makes the single-layer MoS_2 promising for next generation high-speed thin-film electronics. The role of the dielectric encapsulation layers is to screen the Coulomb scattering interaction between the carrier electrons and charged impurities or defects [20]. A single-layer MoS_2 exfoliated from a naturally grown MoS_2 bulk crystal shows typically n-type electrical conductivity [1,3], which is doped unintentionally. A certain amount of defects and impurities are always present in single-crystalline materials in nature, and even with the small imperfections, the electrical and optical properties, such as the electrical conductivity, are largely affected. However, understanding of defects and impurities in single-layer MoS_2 is still premature, even though an efficient and controllable n- and p-type doping technology as well as the mobility engineering is an essential constituent for the electronics application of a single-layer MoS_2 [21–25].

In this study, we investigate the native defects such as a S vacancy (V_S), a S interstitial (S_i), a Mo vacancy (V_{Mo}), and a Mo interstitial (Mo_i) in a single-layer MoS_2, through density-functional theory (DFT) calculations. We find that V_S and S_i have a low formation energy, about 1 eV, in Mo- and S-rich conditions, respectively, and the Mo-related defects of V_{Mo} and Mo_i have a high formation energy above 4 eV. The atomic structure of S_i is a S-adatom on top of a host S atom, and that of Mo_i is a Mo-Mo_i split interstitial along the c direction. The V_S is found to be a deep single acceptor having the $(0/-)$ transition level at 1.7 eV above the valence band maximum (VBM), and the S_i is an electrically neutral defect. The V_{Mo} is a deep single acceptor with the $(0/-)$ transition level at 1.1 eV above the VBM, and the Mo_i is a deep single donor, with the $(+/0)$ transition level at 0.3 eV above the VBM.

II. METHODS

The density-functional theory (DFT) calculations are performed as implemented in the Vienna *ab initio* simulation package (VASP) [26,27]. The projector augmented wave (PAW) pseudopotentials [28,29] and the local density approximation

*yongsung.kim@kriss.re.kr

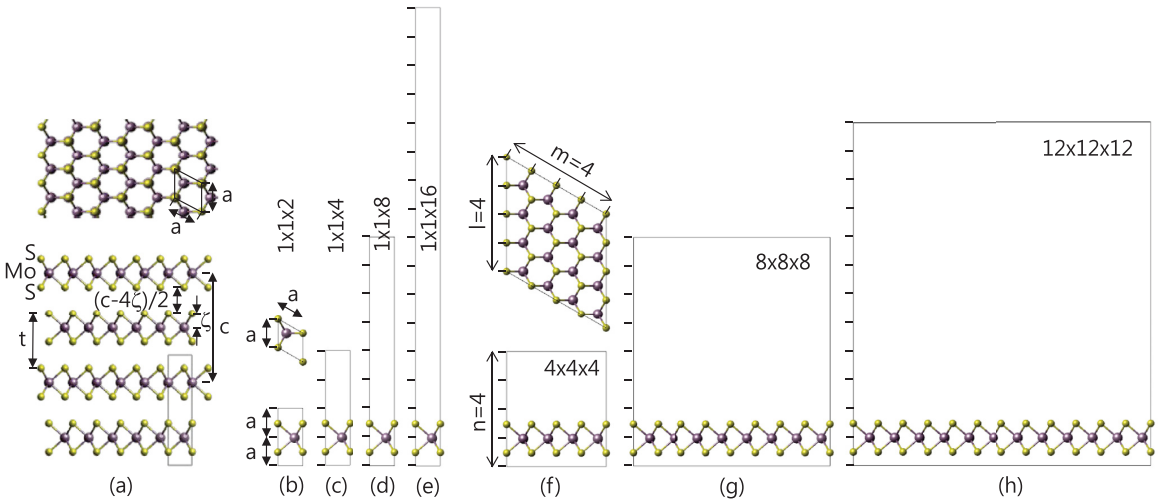


FIG. 1. (Color online) (a) Atomic structure of the bulk crystalline 2H-MoS_2. The (b) $1 \times 1 \times 2$, (c) $1 \times 1 \times 4$, (d) $1 \times 1 \times 8$, (e) $1 \times 1 \times 16$, (f) $4 \times 4 \times 4$, (g) $8 \times 8 \times 8$, and (h) $12 \times 12 \times 12$ supercell structures, in which a single-layer MoS_2 is placed. In (a), (b), and (f), the top views (the c-axis projection) are shown at the top. The large (purple) and small (yellow) balls indicate Mo and S atoms, respectively.

(LDA) [30] for the exchange-correlation energy of electrons are employed. A kinetic energy cutoff of 350 eV for the plane-wave basis set expansion is used. The atomic structures are relaxed until the Hellmann-Feynman forces are less than 0.02 eV/Å. For the single-layer MoS_2, the calculated band gap is direct at the K point in the hexagonal Brillouin zone with the value of 1.851 eV within the LDA.

The atomic structure of the bulk crystalline 2H-MoS_2 is shown in Fig. 1(a). The calculated lattice constants and internal structural parameters of the bulk and single-layer MoS_2 are compared in Table I. The in-plane lattice constant (a) is calculated to be 3.124 Å both in the bulk and single-layer MoS_2. The distance between the top and bottom S layers (2ζ) in a S-Mo-S layer [see Fig. 1(a)] is 3.115 Å in bulk and 3.112 Å in an isolated single-layer MoS_2. In the bulk 2H-MoS_2, the lattice constant (c) perpendicular to the S-Mo-S layers is 12.066 Å, and the interlayer spacing $[(c - 4\zeta)/2]$ between adjacent two S-Mo-S layers [see Fig. 1(a)] is 2.918 Å. The thickness (t) of a S-Mo-S layer including the interlayer spacing $[t = c/2 = 2\zeta + (c - 4\zeta)/2]$ is 6.033 Å, and the thickness t of a single-layer MoS_2 is arbitrary chosen to be the same as the t of bulk 2H-MoS_2.

TABLE I. Calculated lattice and internal structural parameters (in Å) of bulk 2H-MoS_2 and single-layer MoS_2 in LDA. †The thickness (t) of a single-layer MoS_2 is chosen for convention to be the same as the t in bulk 2H-MoS_2, which includes the distance between the top and bottom S layers (2ζ) in a S-Mo-S layer and the interlayer spacing $[(c - 4\zeta)/2]$ between adjacent two S-Mo-S layers.

Parameters	2H-MoS_2	Single-layer MoS_2
a	3.124	3.124
c	12.066	–
2ζ	3.115	3.112
t ($c/2$)	6.033	6.033†
$(c - 4\zeta)/2$	2.918	–

Some of the supercell structures used in our calculations are shown in Figs. 1(b)–1(h) for a single-layer MoS_2. We represent the supercell structures as $l \times m \times n$, as indicated in Fig. 1(f), where the $l \times m$ is the in-plane periodicity and n is the out-of-plane supercell periodicity along the c direction perpendicular to the S-Mo-S layers. The in-plane unit length is the in-plane lattice constant ($a = 3.124$ Å) of a single-layer MoS_2, and the out-of-plane unit length ($a_\perp$) along the c direction is chosen to be the same: $a_\perp = 3.124$ Å. The supercell length along the c direction (L_z) is then $L_z = n \times a_\perp$, and the supercell volume (Ω) is $\Omega = la \times ma \times na_\perp$. The choice of $a_\perp = a$ is for the purpose of *nearly* uniform scaling of the supercell sizes in all the directions. We are interested in the properties of an *isolated* defect, which is placed in an (in-plane) infinite-size single-layer MoS_2 in between the (out-of-plane) infinite-size vacuum. They can be effectively extrapolated from the finite-size supercell calculations by the (nearly) uniform scaling scheme, which is especially useful for charged defect calculations [31].

The static dielectric constant tensor (ε) of a single-layer MoS_2 is calculated by density-functional perturbation theory (DFPT) with the $1 \times 1 \times n$ ($n = 2$, 4, 8, 16) supercells [Figs. 1(b)–1(e)] [32,33]. The DFT calculations for the native defects in single-layer MoS_2 are performed with the $\alpha(1 \times 1 \times 1)$ ($\alpha = 4$, 6, 8, and 10) supercells. Electrostatics calculations that solve the Poisson equations are done with the $\alpha = 4$–80 supercells. Figures 1(f)–1(h) show the supercell structures of $\alpha = 4$, 8, and 12, representatively. In the DFPT calculations, the $8 \times 8 \times 1$ k-point mesh is used for the $1 \times 1 \times n$ supercells. In the DFT calculations, the $2 \times 2 \times 1$ k-point mesh is used for the $\alpha = 4$ and 6 supercells, and the Γ point is used for the $\alpha = 8$ and 10 supercells. In the $\alpha = 6$ supercell, the single-layer MoS_2 contains 108 host atoms and the vacuum thickness ($L_z - t$) is 12.711 Å.

Initial stages of oxygen adsorption on In/Si(111)-4×1

Hyungjoon Shim, Heechul Lim, Younghoon Kim, Sanghan Kim, and Geunseop Lee*
Department of Physics, Inha University, Incheon 402-751, Korea

Hye-Kyoung Kim, Chiho Kim, and Hanchul Kim†
Department of Physics, Sookmyung Women's University, Seoul 140-742, Korea

(Received 16 February 2014; revised manuscript received 19 June 2014; published 16 July 2014)

The In/Si(111)4×1 surface with In wires is a prototypical one-dimensional electron system showing a temperature-induced structural phase transition. While most impurities are known to decrease the transition temperature (T_c), oxygen was reported to increase the T_c. In order to understand the exceptional influence of oxygen, we have examined the initial stages of oxygen adsorption on the surface by scanning tunneling microscopy (STM) and density-functional theory (DFT) calculations. Oxygen molecules were found to dissociate and adsorb as atoms on the surface. STM revealed three distinct O adsorption features as well as occasional transformations from one to another. The high-resolution images showed that all three adsorbed O features were located at the center of the In trimer at both edges of the In wire. This registry is in contradiction with the previous theoretical predictions [S. Wippermann *et al.*, Phys. Rev. Lett. **100**, 106802 (2008)]. The present DFT calculations identified two of the features as O atoms incorporated in the *interstitial* region between the In wire and subsurface Si layer.

DOI: 10.1103/PhysRevB.90.035420 PACS number(s): 68.37.Ef, 68.43.Fg, 68.43.Bc

I. INTRODUCTION

The phase transitions in solids have been a subject of fundamental research in both fields of statistical physics and condensed matter physics. Recently, the phase transitions in reduced-dimensional systems have attracted renewed interest since both structural and electronic transitions were reported in a number of metal-induced superstructures formed on the surfaces [1–9]. The nature of the surface phase transition and the influence of defects and impurities on the phase transition have been the focus of intensive research in these reduced-dimensional systems.

The In/Si(111)-4×1 surface at room temperature (RT) is a low-dimensional system, where its phase transition has been studied extensively [3,10–17]. This surface undergoes a structural phase transition to an 8×2 structure when the temperature is reduced to below 130 K [3,12,13]. A metal-insulator transition was also reported to occur at the same temperature, accompanying the structural phase transition [3,13]. Owing to the (quasi-)one-dimensional nature of this surface, the phase transition was originally interpreted in terms of the charge-density-wave transition [3]. In contrast, other studies proposed a different explanation [18,19].

Global characteristics such as the transition temperature (T_c) and transport property of this In/Si(111)-4×1 surface have been reported to be affected significantly by small amounts of defects [13,20–28]. In a recent series of diffraction experiments, the defects created by depositing metal atoms (In and Na) or by exposing the surface to gases (hydrogen and oxygen) were found to change the transition temperature [23–25,27]. Of particular interest was the adsorption of oxygen, which increased the T_c of the phase transition of the In/Si(111)-4×1 surface, whereas the other defects reduced it. The effect of oxygen defects increasing T_c is unique and extraordinary [24]. Two possibilities have been suggested to explain the O-induced increase in T_c; hole doping or a correlated arrangement of the O-induced defects. The former possibility was based on the property of oxygen taking electrons from the surface upon adsorption. A theoretical calculation, however, indicated that both electron and hole doping reduced T_c [29]. A recent angle-resolved photoelectron spectroscopy study found that the oxygen adsorption increases the band filling of the one-dimensional surface metallic bands, which is in contrast to the expected hole doping [30]. Therefore the hole doping effect of oxygen adsorption is unlikely to account for the increase in T_c. On the other hand, the latter possibility, correlated arrangement of the O-induced defects, requires further an atomic-level investigations.

Knowledge of the local properties, such as the identities of the adsorbed species and adsorption sites are essential to understanding both the local structural changes and the role of defects in the phase transition. While the adsorption site and local perturbation of the atomic H have been determined by scanning tunneling microscopy (STM) and first-principles theoretical calculations [22], little is known regarding the adsorption of oxygen. In a previous STM study, dark features were observed after exposing the surface to the oxygen gas and interpreted as vacancies due to surface etching [22]. Recently, a theoretical study considered the adsorption of atomic oxygen and predicted its adsorption site as well as its influence on surface conductance [31].

In the present study, the adsorption of oxygen on the In/Si(111)-4×1 surface in the initial stages was examined by STM and first-principles calculations. Experimentally, three distinct O adsorption features were observed in the STM images after exposing the surface at RT. All these adsorbed features were attributed to the atomic O adsorbed dissociatively from the molecules. Density-functional theory calculations identified two of the features: O atoms incorporated in the interstitial region between the In wire and Si substrate. The third adsorption feature remains to be identified.

*glee@inha.ac.kr
†hanchul@sookmyung.ac.kr

II. EXPERIMENT

The STM experiments were performed in an ultrahigh vacuum chamber with a base pressure below 1.0×10^{-10} Torr. The In/Si(111)-4×1 surface was prepared by the deposition of 1 ML of In from a Ta-wrapped In source onto a clean Si(111) surface at approximately 700–750 K [32]. The formation of the In/Si(111)-4×1 surface was confirmed by STM scanned at RT. Oxygen gas was introduced into the chamber through a variable leak valve during scanning. This ensured that the initial adsorption of oxygen as various adsorption features could be identified with certainty. The partial pressure of the O_2 gas, which back-filled the chamber was maintained at 2×10^{-9}–1×10^{-8} Torr. The amount of oxygen gas introduced in the chamber was controlled by the continued time during scanning at a fixed pressure and is given in units of L (1 L = 1×10^{-6} Torr sec) for a relative comparison.

III. RESULTS

Figure 1(a) presents STM images of the In/Si(111)-4×1 surface exposed to 4.7 L of O_2 gas in total (1×10^{-9} Torr, 4700 s). These images were taken at a negative sample bias ($V_s = -1.0$ V), representing the spacial distribution of the occupied surface electronic states. Two types of adsorption features, one bright (B) and the other dark (D), were clearly distinguishable. Both were attributed unambiguously to the oxygen adsorption features, because they appeared newly in the images [see Figs. 1(b)–1(d)] taken successively over time during the O_2 dosing. These adsorption features occasionally transform to each other, as shown in Figs. 1(b)–1(d).

Both the B and D adsorption features increased in number as the exposure to O_2 proceeded, indicating that they were certainly related to oxygen adsorption. They occupied either side of the In chain and showed a preference for one side over the other. A comparison with a Si(111)-7×7 surface showed that the preferential occupation side corresponds to the side of the unfaulted half unit of the 7×7 cell, i.e., $\langle 11\bar{2}\rangle$ side. This preference in the adsorption side is the same as that of the adsorption of H atoms [22]. By counting the adsorption features altogether, the ratio between the major and minor adsorption sides was approximately 2:1. This preference indicates the influence of the subsurface layer on the adsorption, which is not symmetrical under the reflection with respect to the centerline of the In chain. The apparent glide plane symmetry of the In wires is absent in the Si(111) substrate.

Although the dark features in Fig. 1 appear similar, they are divided into two different types, D_1 and D_2. This is clearly shown in Fig. 2, where the filled- and the empty-state images of the same area are compared. The D_1 and D_2 oxygen features, both appearing dark in the filled state, were clearly distinguished in the empty-state image [see Fig. 2(b)]. D_1 appeared as a bright protrusion with two bright satellites (marked with arrowheads) at the opposite side of the chain. In contrast, D_2 appeared dark (depressed than the background). The opposite side of the D_2 feature within the chain also appeared bright but it was not as prominent as the satellites of the D_1 feature.

The oxygen can be adsorbed on the surface both in the atomic (O) and molecular (O_2) forms. Determining unambiguously whether the adsorbed oxygen is an atom or a molecule is quite difficult. However, it has been well known that oxygen molecules are dissociated and adsorbed as atoms on metal surfaces at RT [33,34]. Adsorption of molecular oxygen in physisorbed or chemisorbed form has been observed only when the temperature was significantly reduced [35,36]. In our case, a good agreement between the experimental adsorption features with the calculated atomic oxygen species (as it will be shown later) validates the adsorption of oxygen as atoms on the In/Si(111)-4×1 surface at RT. Therefore we rule out the possibility of the observed STM features as O_2 molecules.

Information on the adsorption sites of the atomic O can be obtained from the high-resolution images compared with the atomic structure model, as shown in Fig. 3. In the STM images of the clean surface with low-bias voltages ($|V_s| < 0.5$ V) [Fig. 3(a)], bright protrusions form zigzag chains. Each bright protrusion coincides with the center of a triangle formed by one inner-row In atom and two outer-row (edge-row) In atoms [solid circle in Figs. 3(a) and 3(b)]. In Figs. 3(c)–3(e), which show STM images ($V_s = +0.36$ V) of the O-adsorbed surface, all three oxygen features are located on the tops of these bright protrusions. The B feature appears much brighter [Fig. 3(c)], D_1 is similar [Fig. 3(d)], and D_2 is slightly darker [Fig. 3(e)] than the surrounding protrusions at this low bias. D_1 and D_2 appear similar in that their protrusions are surrounded by dark depressed area. On the other hand, they are distinguishable by the relative brightness of the central protrusions, i.e., D_2 is darker than D_1. In the vicinity of these defects, ×2 modulations are developed and decay with the distance along the row from the defects. These O-defect-induced ×2 modulations are

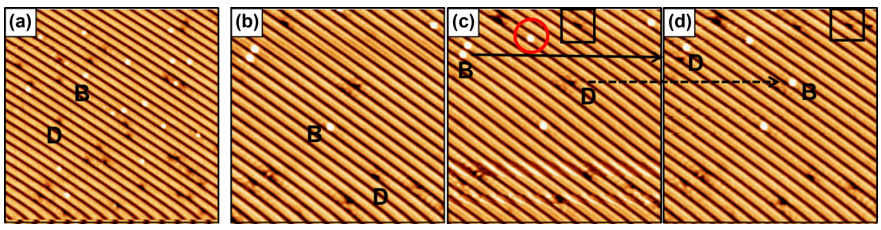


FIG. 1. (Color online) (a) Representative STM image of the In/Si(111)-4×1 surface exposed to O_2 gas (4.7 L) at RT. Two types of adsorption features, bright (B) and dark (D), are clearly visible. (b)–(d) STM images taken successively over time at approximately two-minute intervals during the O_2 dosing with 1×10^{-9} Torr. During dosing, new adsorption features appeared [B (circle) and D (rectangle)] and the transformation between the adsorption features [B-to-D (solid arrow) and D-to-B (dashed arrow)] were observed. The images were filled-state images taken at $V_s = -1.0$ V.

THE JOURNAL OF PHYSICAL CHEMISTRY C

Article

pubs.acs.org/JPCC

Adsorption of CO Molecules on Si(001) at Room Temperature

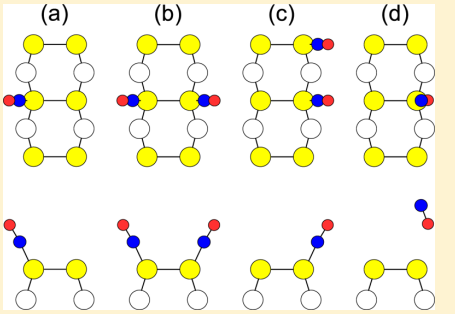
Eonmi Seo,[†,‡] Daejin Eom,[‡] Hanchul Kim,*[,§] and Ja-Yong Koo*[,†,‡]

[†]Korea University of Science and Technology, 217 Gajeong, Yuseong, Daejeon 305-333, Korea

[‡]Korea Research Institute of Standards and Science, Yuseong, Daejeon 305-340, Korea

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

ABSTRACT: Initial adsorption of CO molecules on Si(001) is investigated by using room-temperature (RT) scanning tunneling microscopy (STM) and density functional theory calculations. Theoretical calculations show that only one adsorption configuration of terminal-bound CO (T-CO) is stable and that the bridge-bound CO is unstable. All the abundantly observed STM features due to CO adsorption can be identified as differently configured T-COs. The initial sticking probability of CO molecules on Si(001) at RT is estimated to be as small as ~1×10^{-4} monolayer/Langmuir, which is significantly increased at high-temperature adsorption experiments implying a finite activation barrier for adsorption. Thermal annealing at 900 K for 5 min results in the dissociation of the adsorbed CO molecules with the probability of 60−70% *instead of desorption*, indicating both a strong chemisorption state and an activated dissociation process. The unique adsorption state with a large binding energy, a tiny sticking probability, and a finite adsorption barrier is in stark contrast with the previous low-temperature (below 100 K) observations of a weak binding, a high sticking probability, and a barrierless adsorption. We speculate that the low-temperature results might be a signature of a physisorption state in the condensed phase.



I. INTRODUCTION

The adsorption of a CO molecule on the semiconductor surface is an interesting topic for elucidating basic chemical reactions such as electron-stimulated desorption or photon-stimulated desorption, since CO is a prototypical diatomic molecule and its gas-phase spectroscopy is well understood. Compared to the vast amount of studies on CO reaction with metal surfaces, only limited studies have been reported on the adsorption of CO on the Si(001) surface, and they seem to provide contrasting adsorption behavior of CO molecules.

An early study by high-resolution electron energy loss spectroscopy (HREELS), UV photoelectron spectroscopy (UPS), and temperature-programmed desorption (TPD) reported that CO molecules adsorbed almost perpendicularly on the Si(001)-(2 × 1) surface at 100 K and desorbed at 180 K.[1] The study by X-ray photoelectron spectroscopy (XPS), UPS, and TPD at room temperature (RT) reported that CO adsorbed molecularly on the Si(001)-(2 × 1) surface with an activation energy of ~0.48 eV and with a very small sticking probability.[2] Hu et al. exposed the Si(001) surface to an energetic molecular beam of CO at 85 K and found two possible stable adsorption configurations of the CO molecules in the terminal-bound CO (T-CO) with no energy barrier and in the bridge-bound CO (B-CO) with an energy barrier of 0.9 eV, where T-CO was more stable with the C atom bonding to the down-buckled Si atom of the Si dimer.[3,4] A study by scanning tunneling microscopy (STM) at 70 K and valence-band photoelectron spectroscopy (PES) reported that there existed only one adsorbed state T-CO forming islands near surface defects with no isolated adsorbed CO molecules.[5] Early theoretical studies reported that both T-CO and B-CO were stable,[6,7] which is in contrast to a recent theoretical result that the B-CO is unstable.[8] A model of CO−CO adsorbate interaction on the Si(001)-(2 × 1) surface was also calculated by using cluster models of the surface.[9]

Thus, previous studies on the CO adsorption on Si(001) suggest scattered adsorption behaviors lacking a consensus. Some reported a unique adsorption configuration, while others reported two different configurations. The results from low-temperature (LT) and RT experiments seem to be contradicting to each other, too. For instance, a high sticking probability and a barrierless adsorption process were suggested by the LT experiments, in contrast with a low sticking probability and an activated adsorption at RT. As such, the adsorption of CO molecules on Si(001) needs more detailed investigation for a clear understanding.

In this Article, we investigate the adsorption of CO molecule on the Si(001)-(2 × 1) surface by using STM at RT and density functional theory (DFT) calculations. We find by STM three adsorption features (zigzag, ondimer, and interdimer) that increase linearly with the CO dosage. Our theoretical calculations confirm that the T-CO adsorption configuration is solely stable while the B-CO is unstable. We deduce that the zigzag pattern is the footprint of one CO molecule in the T-CO configuration and that the remaining two adsorption features (ondimer and interdimer) are made up of two adsorbed CO

Received: June 16, 2014
Revised: August 25, 2014
Published: August 27, 2014

molecules. The initial sticking probability is estimated to be ~1 × 10^{-4} monolayer/Langmuir (ML/L) up to the CO dosage of 50 L, whereas the initial sticking probability is significantly increased for the high-temperature adsorption around 450 K. The dissociation probability of the adsorbed CO molecules is 60−70% for the thermal annealing at temperature as high as 900 K, showing a strong chemisorption at RT in contrast to the previously reported weakly bound adsorption state at LT.

II. EXPERIMENTAL AND THEORETICAL METHODS

Experiments were performed using a homemade STM in an ultrahigh vacuum (UHV) chamber with a base pressure less than 1 × 10^{-10} Torr. The samples were phosphorus-doped Si(001) wafers with a resistivity of 1−10 Ω cm. Atomically flat Si(001)-(2 × 1) surfaces were obtained by repeated flashings at temperatures up to 1450 K.[10] The CO molecules were introduced through a leak valve and backfilled the STM chamber for a predetermined time interval under the pressure of 5 × 10^{-8} Torr at RT. The structure of CO/Si(001) was investigated by STM at RT.

The DFT calculations were performed within the generalized gradient approximation,[11] using the Vienna ab initio simulation package.[12] We used a plane-wave cutoff of 400 eV, the theoretical lattice constant of 5.46 Å, the Monkhorst−Pack *k*-point sampling equivalent to a 16 × 16 *k*-mesh in the 1 × 1 surface Brillouin-zone,[13] and the projector-augmented-wave potentials.[14,15] The Si(001) surface was modeled by a six-layer slab separated by 19 Å vacuum with a (4 × 8) surface supercell. The CO molecules were adsorbed on the *c*(4 × 2)-reconstructed surface, and the opposing surface was passivated by hydrogens. All adsorbate atoms and the top four Si layers were relaxed until the Hellman−Feynman forces were smaller than 0.01 eV/Å.

III. RESULTS AND DISCUSSION

Figure 1 shows an STM image of the Si(001)-(2 × 1) surface dosed to 20 L CO molecules. There are three types of features created by CO molecules at RT. (i) The first feature enclosed by a circle is characterized by a strong local zigzag pattern in the filled-state image, where the central dimer is strongly buckled. In the empty-state image, this feature appears as a bright spot at the site of the up-buckled Si atom of the central dimer. (ii) The second one enclosed by a diamond looks darker and more localized than the neighboring clean Si dimers in the filled-state image. It also appears darker than clean Si dimers in the empty-state image. (iii) The third one enclosed by a square looks similar to the C defect, showing the dark depression at one side of two successive Si dimers in the filled-state and the bright protrusion at the other side of the two dimers in the empty-state.[16,17] Figure 1d shows the initial coverages of the three features. Since the third feature is indiscernible from the C defect caused by residual H_2O molecules,[17] we subtracted the number of background C defects (0.03%) from the counted number of the third feature. The three STM image features increase in proportion with the CO dosage, which confirms that these are indeed the footprints of the adsorbed CO molecules on the Si(001)-(2 × 1) surface.

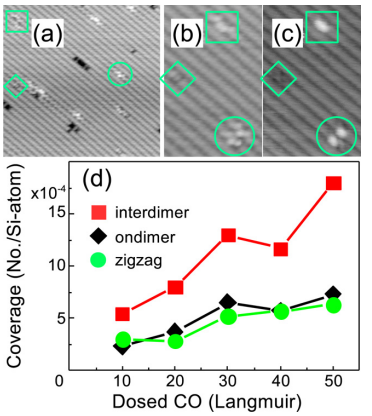


Figure 1. (a) Filled-state STM image (V_s = −2.0 V) of the Si(001)-(2 × 1) surface exposed to 20 L CO molecules at RT. Three types of features appear as the zigzag (○), ondimer (◊), and interdimer (□). (b, c) Three adsorption CO configurations in the enlarged filled-state (V_s = −2.0 V) and empty-state (V_s = 1.5 V) STM images, respectively. (d) Initial coverages of the three adsorption configurations at RT.

To compare with the experimental data, we have calculated the ab initio total energies for four types of *single-molecule* adsorption configurations shown in Figure 2a−d: the T-CO where a CO molecule bound almost vertically to a down-buckled Si, the B-CO where a CO molecule bound to two Si atoms of a Si dimer,[4] an on-dimer configuration where a CO molecule bridge bound to both Si atoms of a dimer (OD-CO), and an end-bridge configuration where a CO molecule bridge bound to two dimers along the dimer row (EB-CO). We found that the B-CO is unstable in contrast to previous theoretical calculations[4,6,7] but in accordance with a recent theoretical study.[8] Both OD-CO and EB-CO were also found to be unstable and relaxed to the T-CO configuration. Resultantly, our calculations for the single-molecule adsorption show that there exists only one stable adsorption configuration, T-CO, shown in Figure 2a. In this unique adsorption configuration, the C atom forms a chemical bond with the down-buckled Si atom of a dimer. We also considered variants of T-CO and B-CO, where the O atom of CO is heading toward Si atoms, and it turned out that the O atom would not bind to a Si atom in any case.

Next, we have examined three types of *two-molecule* adsorption configurations as shown in Figure 2e−g: two CO molecules terminal bound to two Si atoms in an ondimer (2CO-OD), an interdimer (2CO-ID), and a cross-dimer (2CO-XD) configurations, respectively. The calculated adsorption energies of the stable one- and two-molecule adsorption configurations are listed in Table 1. The adsorption energy of T-CO is as large as 0.92 eV, which is too large for the adsorbed CO molecule to desorb at ~180 K,[1] suggesting that the LT adsorption state should be inherently different from the T-CO. All the two-molecule structures have smaller adsorption energies per molecule compared with the single-molecule T-CO structure, suggesting an effective repulsion between adsorbed CO molecules that is in contrast with a LT observation of the adsorbate island formation.[5]

It is interesting to notice that there is a weakly bound locally stable configuration (P-CO in Table 1) with an adsorption energy as small as 0.024 eV. In the P-CO configuration, the CO molecule is oriented almost vertically with the O atom heading toward the down-buckled Si atom of a Si dimer as shown in Figure 2h, and the C−O bond is slightly dilated from the theoretical molecular bond length of 1.143 Å to 1.144 Å. Thus, this P-CO structure can be regarded as a physisorption state.

Journal of the Korean Physical Society, Vol. 64, No. 1, January 2014, pp. 89∼92

Annealing Effects on the properties of Amorphous CoSiB/Pt Multilayer Films with Perpendicular Magnetic Anisotropy

Sol Jung, Jisun Park and Haein Yim*
Department of Physics, Sookmyung Women's University, Seoul 140-742, Korea

Taewan Kim
Department of Advanced Materials Science and Engineering, Sejong University, Seoul 143-747, Korea

(Received 22 August 2013, in final form 6 September 2013)

The perpendicular magnetic anisotropy (PMA) of amorphous CoSiB/Pt multilayer systems was studied as a function of the thickness of the CoSiB/Pt bilayer and the number of repeated CoSiB/Pt bilayers. In this letter, we investigate the thermal property of a CoSiB single layer film annealed at 150 ∼ 350 °C for 3 hours and the perpendicular magnetic anisotropic property of amorphous ferromagnetic Ta(50 Å)/Pt(30 Å)/[CoSiB(2, 3, 4, 5, 6 Å)/Pt(14 Å)]$_5$/Ta(50 Å) multilayer films annealed at 200 ∼ 400 °C for 3 hours. The thermal properties were measured by using a differential scanning calorimeter and an X-ray diffractometer, and the magnetic properties were measured by using a vibrating sample magnetometer. The PMA of the CoSiB/Pt multilayer film disappeared and the multilayer film show isotropy after annealing at a temperature of 350 °C or above.

PACS numbers: 75.70.Cn, 75.70.-I, 75.50.Kj
Keywords: Perpendicular anisotropy, Annealing effect, Amorphous CoSiB, CoSiB/Pt multilayer
DOI: 10.3938/jkps.64.89

I. INTRODUCTION

Magnetic materials with perpendicular magnetic anisotropy (PMA) are applicable to high-density data storage and to spin transfer torque magnetic random access memories (STT-MRAMs) [1–3]. PMA is the phenomenon in which a magnetic thin film is magnetized perpendicular to the film's plane. For the development of STT-MRAMs, the thermal stability of a free layer and a decrease in the switching current become issues. If these problems are to be solved, ferromagnetic materials with large magnetic anisotropy are required for a free layer [4,5].

Materials with PMA have anisotropy energies larger than those of with in-plane anisotropy (IMA). For patterned devices, in addition, the magnetizations of materials with PMA do not suffer from thermal instability due to the magnetization curling observed at the edge of in-plane case [6]. Moreover, reverse of magnetization of a PMA free layer is easier than that of an IMA free layer. Based on these general theories, devices with PMA have lower switching currents than devices with IMA for the same magnetic field [7–9]. Multilayer systems based on crystalline materials, for example, Co, Fe, and CoFe show PMA for various thicknesses of layer and the number of repeated layer with noble metals like Pt, Pd, and Au [10–14]. Also, amorphous multilayer systems based on CoFeSiB, NiFeSiB and CoSiB show PMA. Both crystalline and amorphous multilayers show PMA, but amorphous multilayer systems have two strong advantages compared with crystalline multilayer systems: no grain boundary and the less roughness for the amorphous than that of crystalline [15,16].

In previous study, we investigated the properties of amorphous ferromagnetic material CoSiB(1000 Å) single layer film. The saturated magnetization (M_s) and the magnetic anisotropy constant (K_u) were 407 emu/cm^3 and 1,500 erg/cm^3, respectively. PMA is affected by various factors like the thickness of the single layer (ferromgnetic layer or nonmagnetic layer) and the number of repeated ferromagnetic/nonmagnetic bilayers. In order to investigate the role of these amorphous multilayers, we studied the perpendicular magnetic anisotropic properties of amorphous CoSiB/Pt multilayers as the functions of the thicknesses of CoSiB and Pt layers and of the repetitions of CoSiB/Pt bilayer [17]. The coercivity (H_c) and the saturated magnetization (M_s) were measured by using a vibrating sample magnetometer (VSM).

Based on the previous study, in this research, we investigated the thermal property of CoSiB(1000 Å) single layer films annealed at 150 ∼ 350 °C for 3 hours and the PMA of Ta(50 Å)/Pt(30 Å)/[CoSiB(2, 3, 4, 5, 6 Å)/Pt(14 Å)]$_5$/Ta(50 Å) multilayer films annealed at 200

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

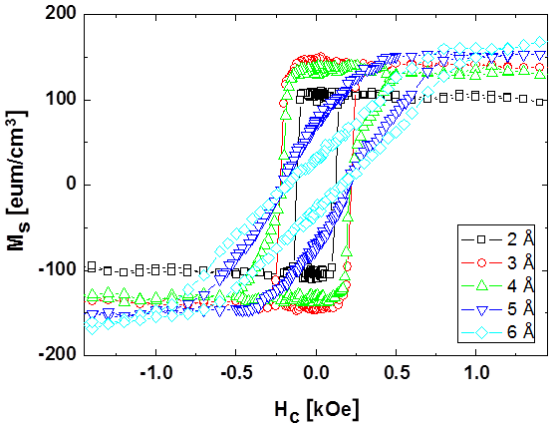


Fig. 1. (Color online) Out-of-plane magnetic hysteresis loops of Ta(50 Å)/Pt(30 Å)/[CoSiB(t_{CoSiB} Å)/Pt(14 Å)]$_5$/Ta(30 Å) multilayer films for values of t_{CoSiB} equal to 1, 2, 3, 4, 5 and 6 Å.

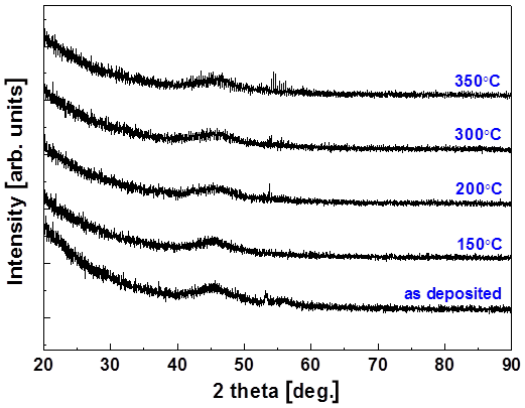


Fig. 2. (Color online) XRD patterns of Ta(50 Å)/Pt(30 Å)/[CoSiB(t_{CoSiB} Å)/Pt(14 Å)]$_5$/Ta(30 Å) multilayer films annealed at room temperature, 150 °C, 200 °C, 300 °C and 350 °C for 3 hours.

$\sim$ 400 °C for 3 hours. The thermal properties were measured by using a differential scanning calorimeter (DSC) and an X-ray diffractometer (XRD), and magnetic properties were measured by using a VSM.

II. EXPERIMENTS AND DISCUSSION

The multilayer is deposited by a 3-inch, and 6-gun-type DC magnetron sputtering system at room temperature. A SiO_2 (100) wafer was placed into an ultrasonic washer with ethanol for 10 minutes, acetone for 10 minutes, and deionize water for 10 minutes in order to the remove impurities on the SiO_2 (100) wafer's surface. The base pressure of the main chamber was up to 2.0 $\times$ 10^{-7} Torr, and the working pressure was 2 $\times$ 10^{-3} Torr with flowing argon. The buffer layer and the capping layer consist of Ta(50 Å)/Pt(30 Å) and Ta(50 Å), respectively. The hysteresis loop is systematically measured by using a VSM to verify the PMA of the amorphous multilayers. The annealing process was conducted in a high-temperature oven flushed with nitrogen. XRD measurements were carried out both at wide-angle and grazing-angle incidence.

The hysteresis loops of the M_s and the H_c for the Ta(50 Å)/Pt(30 Å)/[CoSiB(t_{CoSiB} Å)/Pt(13 Å)]$_5$/Ta(50 Å) multilayers with t_{CoSiB} equal to 1, 2, 3, 4, 5 and 6 Å are summarized in Fig. 1. The magnetic field was applied in the direction perpendicular to the surface of the multilayer film. The M_s increased with increasing t_{CoSiB}. The H_c increased to maximum value at t_{CoSiB}= 3 Å and then decreased for $t_{CoSiB} > 3$ Å. A very thin ferromagnetic CoSiB layer is usually believed to have a large strain and a small lattice mismatch with nonmagnetic Pt layers. Above t_{CoSiB}= 3 Å, the PMA of these system decreased with further increases in the thickness of CoSiB layer because the strain relaxation increased and additional Co did not contribute to the PMA [18]. In order to study the thermal property, we measured the differential scanning calorimeter (DSC) curve of an amorphous material CoSiB(30 μm), by a DSC measurement system, and we found the crystallization temperature (T_x) to be 525 °C. The T_x represents the onset temperature of the first crystallization event. In the other amorphous materials, CoFeSiB and NiFeSiB, these T_x's were around 517 °C and 497 °C, respectively [19]. The thermal stability of CoSiB is observed to be much higher than those of CoFeSiB and NiFeSiB.

Figure 2 shows the XRD patterns of CoSiB(1000 Å) single layer film after annealing at 150 °C, 200 °C, 300 °C and 350 °C for 3 hours [20]. Within the detection limit of the XRD, the patterns show a typical diffuse first nearest neighbor peak associated with amorphous structure. In Fig. 2, all samples show the patterns of amorphous materials. Amorphous ferromagnetic CoSiB(1000 Å) single layer films maintain their amorphous structure up to 350 °C.

We carried out an annealing for 3 hours to investigate the PMA of CoSiB/Pt multilayer films. Fig. 3 and Fig. 4 show the hysteresis loops of Ta(50 Å)/Pt(30 Å)/[CoSiB(2, 3, 4, 5, 6 Å)/Pt(14 Å)]$_5$/Ta(50 Å) multilayer films annealed at 200 $\sim$ 400 °C for 3 hours. The M_s and the H_c of these films were measured in a hard axis (out-of-plane) and an easy axis (in-plane; not presented in this letter). Figure 3(a) and 3(b) show the hysteresis loops when the thickness of CoSiB layer are 2 Å and 3 Å, respectively. In Fig. 3(a), a decrease in the PMA and a squareness (M_s/M_r, M_r: remanent magnetization) occur after annealing at a temperature of 350 °C, and the multilayer film annealed at 400 °C shows an isotropic property. Figure 3(b) shows the same tendency

Journal of the Korean Physical Society, Vol. 64, No. 2, January 2014, pp. 301∼304

Soft Magnetic Properties of $Fe_{(87-x)}Ti_7Zr_6[B_xSi_x]_{0.5}$ Amorphous Metallic Ribbons Prepared by Melt-spinning

Bo-Kyeong Han and Haein Choi-Yim*

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

(Received 5 November 2013, in final form 14 November 2013)

The soft magnetic Fe-based amorphous metallic ribbons $Fe_{(87-x)}Ti_7Zr_6[B_xSi_x]_{0.5}$ (x = 8, 10, 12%) were prepared using the melt-spinning technique to have thickness of 15-22 μm and widths of 1-2 mm. Increasing the amounts of B and Si improved the soft magnetic properties: the coercivity (H_c) decreased, and the saturation magnetization (M_s) increased, the thermal and structural properties remained the same. Due to the high sensitivity of the alloys to the process conditions, variations in the widths of the melt-spun ribbons induced structural modifications, resulting in significant change in the soft magnetic properties. While the ribbons made with 1 mm widths had fully amorphous structures and soft magnetic properties, the ribbons with 2 mm widths showed semi-hard magnetic properties, which was attributed to the formation of nanocrystals inside the amorphous structures.

PACS numbers: 81.05.Kf, 81.40.Rs, 81.70.Pg
Keywords: Soft magnetic property, Fe-based, Amorphous ribbons, Metallic glasses, Ferromagnetic
DOI: 10.3938/jkps.64.301

I. INTRODUCTION

The amorphous phase was first obtained in thin ribbon form for Au-Si binary systems through a rapid solidification process in 1960 [1]. Later on, various systems of amorphous alloys were produced, and a group of alloys with robust stability of the amorphous phase, bulk metallic glasses (BMGs), was developed. BMGs possess high yield, fracture toughness and strength, large elastic strain, excellent corrosion resistance and low plasticity compared with crystalline metals [2–6].

Among the many BMGs, Fe-based amorphous alloys have attracted interest because of their low material cost, ultrahigh strength, high corrosion resistance and good soft magnetic properties [3,7–11]. Thus, since the first Fe-based bulk amorphous alloy, Fe-(Al,Ga)-(P,C,B), was synthesized in 1995 [12], a large number of Fe-based BMGs have been developed.

Particularly, in Fe-based amorphous alloy systems, improving the soft magnetic properties has been tried for the applications to sensors, magnetic heads, magnetic memories and electro-magnetic cores [13–15]. The soft magnetic property is generally defined by a low coercivity (H_c) under 10^3 A/m (12.5 Oe). In practical applications, the coercivity (H_c) is known to be one of the crucial that determine the soft magnetic performance, and it is known to be dominantly affected by the microstructure [16].

Among the Fe-based amorphous alloys, the Fe-Ti-Zr-B system which is known to form an amorphous structure up to a 1 mm size [17], has not been studied in terms of its magnetic properties yet. Therefore, in this work, the Fe-Ti-Zr-B alloy system was modified by partially replacing B by Si and varying the total amount of B and Si at the expense of Fe to study the effects on the soft magnetic properties and the stability of the amorphous structure. X-ray diffraction (XRD), differential scanning calorimetry (DSC), and vibrating sample magnetometer (VSM) were utilized to study the structural, thermal, and magnetic properties, respectively, of $Fe_{(87-x)}Ti_7Zr_6[B_xSi_x]_{0.5}$ (x = 8, 10, 12%) alloy ribbons with widths of 1 and 2 mm prepared by using a melt-spinning process.

II. EXPERIMENTS AND DISCUSSION

Fe-based multi-component ingots (6 g) with compositions $Fe_{(87-x)}Ti_7Zr_6[B_xSi_x]_{0.5}$ (x = 8, 10, 12%) were synthesized from high-purity elements (Fe, Ti, Zr, B, Si > 99.9 wt. %) by using an arc-melting furnace under a Ti-gettered argon atmosphere. Each ingot of high purity was re-melted at least four times to minimize the compositional inhomogeneity within the ingot. Ribbon samples of glassy alloys with thicknesses of 15-22 μm and widths of 1-2 mm were prepared by single copper roller melt-spinning under an argon atmosphere.

The structures of the ribbon samples were identified by using XRD with Cu-Kα. The thermal characteriza-

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

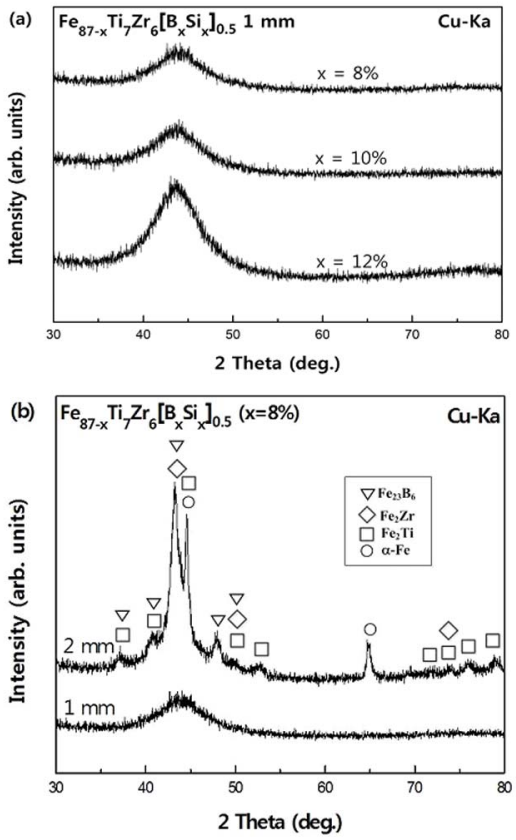


Fig. 1. XRD patterns of (a) $Fe_{(87-x)}Ti_7Zr_6[B_xSi_x]_{0.5}$ (x = 8, 10, 12%) amorphous ribbons with widths of 1 mm and (b) $Fe_{(87-x)}Ti_7Zr_6[B_x\ Si_x]_{0.5}$ (x = 8%) ribbons with widths of 1 & 2 mm.

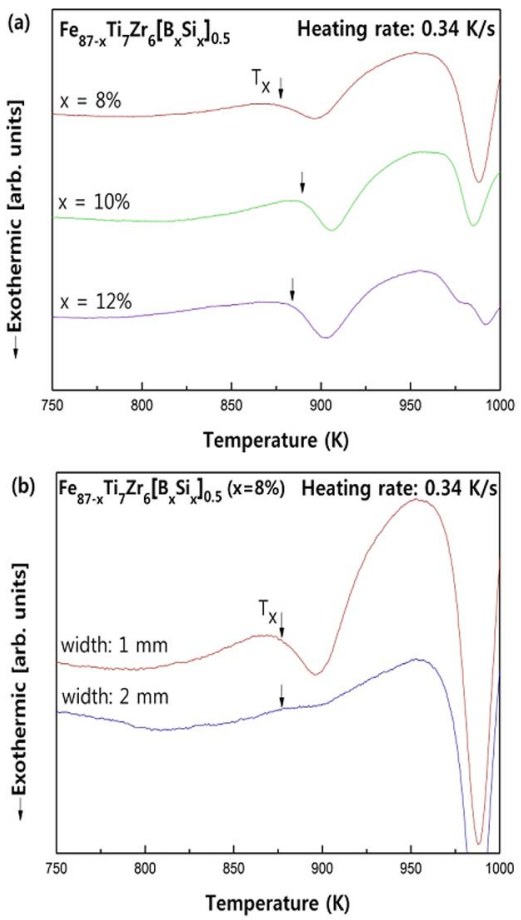


Fig. 2. (Color online) DSC curves of (a) $Fe_{(87-x)}Ti_7Zr_6[B_xSi_x]_{0.5}$ (x = 8, 10, 12%) amorphous ribbons with widths of 1 mm and (b) $Fe_{(87-x)}Ti_7Zr_6[B_xSi_x]_{0.5}$ (x = 8%) ribbons with widths of 1 & 2 mm.

tion was carried out using DSC under constant heating up to 1073 K at a heating rate of 0.34 K/s to evaluate the crystallization temperatures (T_x) of the alloys. The magnetic properties, such as the saturation magnetization (M_s) and the coercivity (H_c), of these amorphous ribbons were measured by using VSM at room temperature. These amorphous ribbons were examined for in-plane magnetic fields in the range from -1.0 Tesla to 1.0 Tesla.

The X-ray diffraction patterns of (a) $Fe_{(87-x)}Ti_7Zr_6[B_xSi_x]_{0.5}$ (x = 8, 10, 12%) amorphous ribbons with widths of 1 mm and (b) $Fe_{(87-x)}Ti_7Zr_6[B_xSi_x]_{0.5}$ (x = 8%) ribbons with widths of 1 & 2 mm are shown in Fig. 1. The ribbons of $Fe_{(87-x)}Ti_7Zr_6[B_xSi_x]_{0.5}$ (x = 8, 10, 12%) with widths of 1 mm have a fully amorphous phase as no sharp Bragg peaks are observed and all diffraction lines show the typical broad halo pattern of an amorphous structure. The broad halo patterns indicate that these ribbons consist of a fully amorphous phase. On the other hand, the ribbon of $Fe_{(87-x)}Ti_7Zr_6[B_xSi_x]_{0.5}$ (x = 8%) with a 2 mm width shows several crystalline peaks over a broad amorphous background, as shown in Fig. 1(b). The peaks due to the crystalline phases were associated with the αFe, $Fe_{23}B_6$, Fe_2Zr and Fe_2Ti phase, which is consistent with the available data [17–19]. A more systematic study will be done in the near future.

The thermal properties of $Fe_{(87-x)}Ti_7Zr_6[B_xSi_x]_{0.5}$ (x = 8, 10, 12%) were examined. The DSC curves for the Fe-Ti-Zr-B-Si ribbons are shown in Fig. 2. The T_x represents the onset temperature of the first exothermal crystallization event. The measured values of T_x for the alloys are summarized in Table 1. The crystallization temperature ranges from 877 to 889 K for the alloys, indicating that the alloys had similar thermal stabilities.

The magnetic properties of the amorphous ribbons

Current Applied Physics 14 (2014) 685–687

Contents lists available at ScienceDirect

Current Applied Physics

journal homepage: www.elsevier.com/locate/cap

Effect of compositional variation on the soft magnetic properties of $Fe_{(87-x-y)}Co_xTi_7Zr_6B_y$ amorphous ribbons

BoKyeong Han [a], Young Keun Kim [b], Haein Choi-Yim [a,*]

[a] Department of Physics, Sookmyung Women's University, Seoul 140-742, Republic of Korea
[b] Department of Materials Science and Engineering, Korea University, Seoul 136-713, Republic of Korea

ARTICLE INFO

Article history:
Received 30 December 2013
Received in revised form
11 February 2014
Accepted 11 February 2014
Available online 4 March 2014

Keywords:
Amorphous ribbons
Fe–Co based
Soft magnetic property
Ferromagnetic

ABSTRACT

The effect of the replacement of Fe by Co or B on the thermal stability and soft magnetic properties of the Fe-based amorphous metallic ribbons with $Fe_{(87-x-y)}Co_xTi_7Zr_6B_y$ ($x = 10$, 20% and $y = 8$, 10, 12%) produced by melt-spinning technique was investigated. For the melt-spun amorphous ribbons, the values of saturation magnetization and coercivity were observed to range from 107.00 to 152.38 emu/g and from 0.012 to 0.446 Oe, respectively. The thermal properties such as T_g, T_x, and ΔT_x were in the range of 796.7–809.6 K, 840.2–853.5 K, and 35.8–54.5 K, respectively. In the Fe–Co–Ti–Zr–B alloys, the Co substitution for Fe improved the soft magnetic properties but decreased the thermal stability. For magnetic properties, the coercivity (H_c) decreased and saturation magnetization (M_s) increased by the addition of Co. However, the supercooled liquid region (ΔT_x) decreased by the addition of Co. Meanwhile, the B substitution for Fe had no meaningful change on the thermal stability and soft magnetic properties. The amorphous ribbon of $Fe_{59}Co_{20}Ti_7Zr_6B_8$ exhibited the best soft magnetic properties such as the low coercivity of 0.025 Oe and the high saturation magnetization of 152.38 emu/g.

1. Introduction

Fe-based amorphous alloys have attracted technological interests due to their excellent mechanical properties, high corrosion resistance, and outstanding soft magnetic properties [1–6]. Thus, since the first amorphous alloy ribbons of Fe–P–C were fabricated by using melt spinning technique in 1967 [7], a large number of Fe-based alloy systems have been developed and investigated to improve the soft magnetic properties and thermal stability in bulk or melt-spun ribbon forms for commercial applications [8,9].

Especially, improving the soft magnetic properties such as very low coercivity, high saturation magnetization, and high permeability has been attempted for various application such as sensors, magnetic delay lines, magnetic heads, magnetic memories and electro-magnetic cores [10–12], In practical applications, the coercivity is known to be one of the crucial properties that determine the soft magnetic performance, and is known to be dominantly affected by the microstructure [13]. The soft magnetic property is generally defined by coercivity (H_c) under 10^3 A/m (12.5 Oe).

Among these Fe-based amorphous alloys, the Fe–Zr–B-based alloys have been investigated in recent years because of their uncommon properties such as double transition behavior, invar effect, resistivity anomalies, and magnetocaloric behavior in the amorphous phase [14–17]. Also, the Fe–Ti–Zr–B system, which is known to form an amorphous structure up to a 1 mm size [18], has been studied in terms of its mechanical property, but not yet in terms of its magnetic properties. In this study, the Fe–Ti–Zr–B alloy system was modified by the replacement of Fe by Co or B to study the effects on the soft magnetic properties and the thermal stability of the amorphous structure. The Fe–Co system is of high interest due to its highest known saturation magnetization (2.45 T) for bulk transition metal alloys for the alloy having 35% Co [19]. Moreover, the partial Co substitution for Fe is known to enhance glass forming ability (GFA). By using differential scanning calorimetry (DSC), and vibrating sample magnetometer (VSM), the effect of the replacement of Fe by Co or B on the thermal stability and magnetic properties were investigated in the $Fe_{(87-x-y)}Co_xTi_7Zr_6B_y$ ($x = 10$, 20% and $y = 8$, 10, 12%) alloy system.

* Corresponding author.
E-mail address: haein@sm.ac.kr (H. Choi-Yim).

http://dx.doi.org/10.1016/j.cap.2014.02.009
1567-1739/

2. Experiments and discussion

The Fe-based multi-component alloys with nominal compositions of $Fe_{(87-x-y)}Co_xTi_7Zr_6B_y$ ($x = 10$, 20% and $y = 8$, 10, 12%) were synthesized by using an arc-melting furnace of high-purity Fe, Co, Ti, Zr, and B into a 6 g ingot under a Ti-gettered argon atmosphere. Each ingot was re-melted at least four times to minimize the compositional inhomogeneity. Rapidly solidified ribbons of amorphous alloys with thickness of $30–35 \times 10^{-6}$ m and width of 1×10^{-3} m were produced by using the single copper roller vacuum melt-spinning technique under a controlled argon gas atmosphere. The spinning speed of the copper wheel of the melt spinner was optimized at 33 m/s.

The amorphous structures of the prepared ribbon samples were confirmed by using X-ray diffraction (XRD) with Cu-Kα radiation. The thermal characterization for the ribbons was carried out using the differential scanning calorimeter (DSC) under an argon atmosphere and constant heating up to 1073 K at a heating rate of 0.34 K/s to evaluate the thermal parameters such as glass transition, crystallization temperatures, and supercooled liquid region of the alloys.

The magnetic properties such as the saturation magnetization and the coercivity of these amorphous ribbons were measured by using the vibrating sample magnetometer (VSM) with a hysteresis loop tracer at room temperature. These amorphous ribbons were examined for in-plane magnetic fields in the range from −1.0 T to 1.0 T.

The X-ray diffraction patterns of the as-spun $Fe_{(87-x-y)}Co_xTi_7Zr_6B_y$ ($x = 10$, 20% and $y = 8$, 10, 12%) amorphous ribbons are shown in Fig. 1. The ribbons of $Fe_{(87-x-y)}Co_xTi_7Zr_6B_y$ ($x = 10$, 20% and $y = 8$, 10, 12%) have a fully amorphous phase as no sharp Bragg peaks are observed and all diffraction lines show the typical broad halo pattern of an amorphous state. The broad halo patterns indicate that these ribbons consist of a fully amorphous phase.

To investigate the thermal properties of the ribbons with the Co or B substitution for Fe, DSC curves of all of the ribbons were obtained. Fig. 2 shows the DSC pattern obtained for the rapidly solidified $Fe_{(87-x-y)}Co_xTi_7Zr_6B_y$ ($x = 10$, 20% and $y = 8$, 10, 12%) amorphous ribbons where T_g, T_x, and ΔT_x are marked by arrows. The thermal stability parameters such as glass transition temperature (T_g), crystallization temperature (T_x), and supercooled liquid region (ΔT_x) are reported in Table 1. The glass transition temperature, the crystallization temperature, and the supercooled liquid region of the $Fe_{(87-x-y)}Co_xTi_7Zr_6B_y$ ($x = 10$, 20% and $y = 8$, 10, 12%) amorphous ribbons ranges from 796.7 to 809.6 K, 840.2 to 853.5 K and 35.8 to 54.5 K, respectively. With partial Co substitution for Fe, the values of T_x, ΔT_x were found to be decreased and T_g was found to be increased. The T_g, T_x and ΔT_x of the $Fe_{(77-y)}Co_{10}Ti_7Zr_6B_y$ ($y = 8$, 10, 12%) ribbons were in the range of 796.7–798.6 K, 850.7–853.5 K and 52.5–54.5 K, respectively. Also, the T_g, T_x and ΔT_x of the amorphous ribbons with $Fe_{(67-y)}Co_{20}Ti_7Zr_6B_y$ ($y = 8$, 10, 12%) are in range of 807.7–809.6 K, 839.7–847.4 K and 35.8–39.7 K, respectively. The ΔT_x is larger for 10% Co than 20% Co in the Fe–Co–Ti–Zr–B system. Therefore, the addition of 10% Co in the Fe–Ti–Zr–B system resulted in better thermal stability more than does adding 20% Co. However, no remarkable change caused by the B addition was observed on thermal stability.

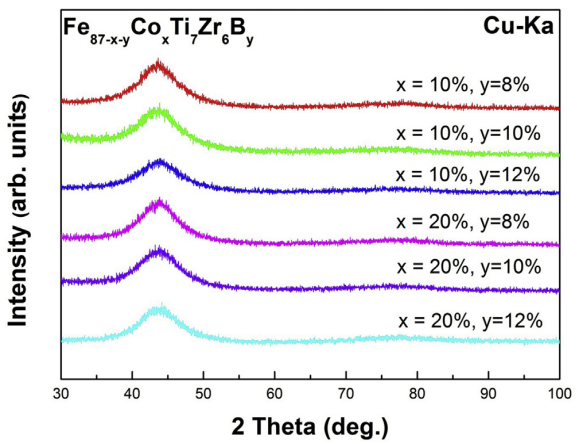


Fig. 1. XRD patterns of $Fe_{(87-x-y)}Co_xTi_7Zr_6B_y$ ($x = 10$, 20% and $y = 8$, 10, 12%) amorphous ribbons.

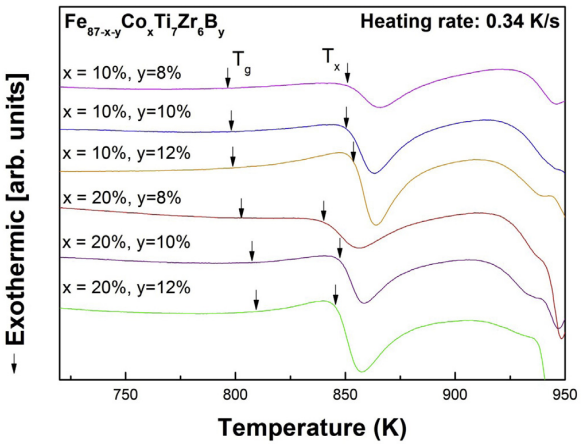


Fig. 2. DSC curves of $Fe_{(87-x-y)}Co_xTi_7Zr_6B_y$ ($x = 10$, 20% and $y = 8$, 10, 12%) amorphous ribbons.

The magnetic properties of the amorphous ribbons were measured by using VSM. The hysteresis $M–H$ loops for the as-spun $Fe_{(87-x-y)}Co_xTi_7Zr_6B_y$ ($x = 10$, 20% and $y = 8$, 10, 12%) ribbons are shown in Fig. 3. The values of the saturation magnetization (M_s) and the coercivity (H_c) are listed in Table 1. The Fe–Co–Ti–Zr–B amorphous ribbons showed typical soft magnetic properties, such as a low coercivity (H_c) for magnetic fields under 1 Oe. The saturation magnetization of amorphous ribbons with $Fe_{(77-y)}Co_{10}Ti_7Zr_6B_y$ ($y = 8$, 10, 12%) is in the range of 107.00–112.22 emu/g and coercivity is in the range of 0.146–0.446 Oe. Also, the saturation magnetization of the $Fe_{(67-y)}Co_{20}Ti_7Zr_6B_y$ ($y = 8$, 10, 12%) amorphous ribbons ranges from 108.61 emu/g to 152.38 emu/g and the coercivity ranges from 0.012 Oe to 0.22 Oe. In the Fe–Co–Ti–Zr–B systems, it turned out that the addition of 20% Co resulted in better soft magnetic properties than the addition of 10% Co. As shown in Fig. 3(a), the addition of B has no meaningful change in the soft magnetic properties of Fe–Co–Ti–Zr–B system. However, the

Table 1
Magnetic and thermal properties of $Fe_{(87-x-y)}Co_xTi_7Zr_6B_y$ ($x = 10$, 20% and $y = 8$, 10, 12%) amorphous ribbons.

Alloys	Thermal properties			Magnetic properties	
	T_g (K)	T_x (K)	ΔT_x (K)	H_c (Oe)	M_s (emu/g)
$Fe_{69}Co_{10}Ti_7Zr_6B_8$	796.7	851.1	54.4	0.316	112.22
$Fe_{67}Co_{10}Ti_7Zr_6B_{10}$	798.2	850.7	52.5	0.146	107.00
$Fe_{65}Co_{10}Ti_7Zr_6B_{12}$	798.6	853.5	54.5	0.446	109.78
$Fe_{59}Co_{20}Ti_7Zr_6B_8$	802.7	840.2	37.5	0.025	152.38
$Fe_{57}Co_{20}Ti_7Zr_6B_{10}$	807.7	847.4	39.7	0.012	108.61
$Fe_{56}Co_{20}Ti_7Zr_6B_{12}$	809.6	845.4	35.8	0.220	113.62

Journal of the Korean Physical Society, Vol. 65, No. 1, July 2014, pp. 65∼69

Effects of Rapid Thermal Annealing on the Magnetic Properties of CoSiB/Pd Multilayers with Perpendicular Anisotropy

Sol Jung, Bo-kyeong Han and Haein Yim*
Department of Physics, Sookmyung Women's University, Seoul 140-742, Korea

Nam-Hui Kim, Jaehun Cho and Chun-Yeol You
Department of Physics, Inha University, Incheon 405-751, Korea

(Received 30 April 2014, in final form 8 May 2014)

The perpendicular magnetic anisotropic property of amorphous CoSiB/Pd multilayers was studied as a function of the thicknesses of the CoSiB and the number of repeated CoSiB/Pd bilayers. In this study, we investigate the effects of thermal annealing on the perpendicular magnetic anisotropic property of the [CoSiB 2, 3, 4, 5 Å/Pd 13 Å]$_5$ multilayer films. We vary the annealing temperature between from 200 to 350°, with the thicknesses of the CoSiB layer being kept constant. We examine the magnetic properties such as the saturation magnetization and the coercivity. The annealing temperature is closely related to the magnetic property of the CoSiB/Pd multilayers, and the coercivity especially is proportional to the annealing temperature.

PACS numbers: 75.70.Cn, 75.70.-I, 75.50.Kj
Keywords: Annealing effects, Perpendicular magnetic anisotropy, Amorphous alloy, CoSiB/Pd multilayer
DOI: 10.3938/jkps.65.65

I. INTRODUCTION

Perpendicular magnetic anisotropy (PMA) is the perpendicular direction dependence of the magnetic properties of a material. Magnetic thin films with PMA are magnetized in a direction normal (perpendicular) to the plane. Experimental studies of the PMA in magnetic thin films and multilayers have been done for almost 40 years since Iwasaki and Takemura first investigated the mechanism of the PMA in Co/Cr thin films [1]. In 1985, Carcia *et al.* established the importance of the interface between the magnetic layer and the nonmagnetic layer as the driving mechanism for the PMA [2]. Subsequently, in studies of magneto resistive random access memories (MRAMs), magnetic tunnel junctions (MTJs) with PMA became a key issue for realizing nextgeneration high-density and non-volatile memory devices, such as a spin-transfer torque (STT) MRAM [3–7].

The STT-MRAMs demand two important elements: thermal stability and low switching current density (J_c) [8]. The thermal stability factor E/k_BT of the free layers of the MTJs needs to be more than 40 for non-volatility. E is the energy barrier and is given by $E = V(M_sH_k/2)$, where V is the volume of the free layer, M_s is the saturation magnetization and H_k is the in-plane magnetic field. This equation affords that $M_sH_k/2$ needs to be high enough to ensure high thermal stability because the V decreases as the size of the MTJ is decreased for the development of high-density MRAMs [5,9]. $M_sH_k/2$ is the anisotropy energy density (K), and it is closely related with the PMA.

The threshold current density for current-induced magnetizing switching (CIMS) is expressed as $J_c = \alpha(2eM_sd/\hbar\eta)(H_k+H_d/2)$, where α, d, η and H_d are the damping constant, the thickness, the spin polarization efficiency factor and the out-of-plane demagnetizing field of the free layer, respectively [10,11]. This equation provides two ways to get low J_c: decrease the M_s of the free layer and the d [12–14].

PMA has two strong advantages for the nextgeneration devices: thermal stability and low J_c. Materials with PMA have better thermal stability and lower J_c than materials with in-plane magnetic anisotropy (IMA) because of the opposite role of the demagnetization energy in the determination of the J_c [15–17]. Therefore, recently, materials with PMA were used as free layers to get high thermal stability and low J_c. Especially, because amorphous materials have good magnetic properties, they have been intensively studied for applications to the free layers of MTJs [18–20].

If the MTJs size is decreased for high-density, the written data are more likely to be lost because of thermal fluctuations. Moreover, the film's roughness is a critical issue because it influences the magnetic property strongly in thin films. If thermal fluctuations are to be

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

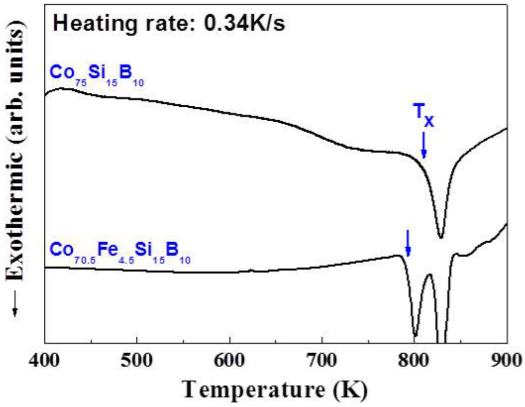


Fig. 1. (Color online) DSC scans of the $Co_{75}Si_{15}B_{10}$ and the $Co_{70.5}Fe_{4.5}Si_{15}B_{10}$ ribbons. The T_x is the onset crystalline temperature of the first crystallization event. The T_x of the $Co_{75}Si_{15}B_{10}$ ribbon is 811 K, and the T_x of the $Co_{70.5}Fe_{4.5}Si_{15}B_{10}$ ribbon is 790 K.

overcome, magnetic films should possess large PMA, a smooth interface and a smooth surface. Therefore, we used amorphous ferromagnetic materials as free layers because it can lead to a decrease in the surface roughness [21,22].

In the present study, we investigated annealing effects in order to examine the thermal stability of the CoSiB/Pd multilayers. We varied the annealing temperature (T_a) or the CoSiB thickness (t_{CoSiB}) to find the magnetic properties for the various values of T_a and t_{CoSiB}.

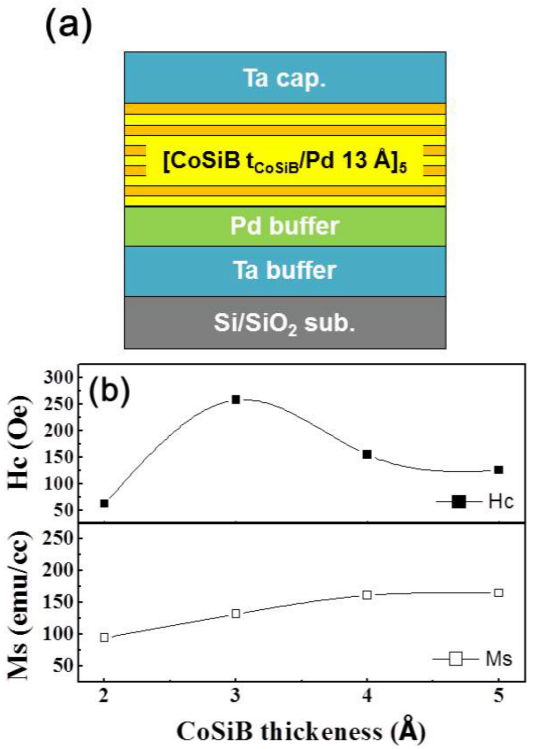


Fig. 2. (Color online) (a) Schematic of the [CoSiB t_{CoSiB}/Pd 13 Å]$_5$ multilayers for t_{CoSiB} values of 2, 3, 4 and 5 Å. (b) The coercivity and the magnetization dependences of the CoSiB/Pd multilayers.

II. EXPERIMENTS AND DISCUSSION

The amorphous ferromagnetic CoSiB/Pd multilayers were prepared with various thicknesses of the CoSiB layer. The CoSiB/Pd multilayers deposited with a Ta capping layer and a Pd buffer layer on a Si/SiO_2 substrate in a DC-magnetron sputtering system at room temperature. The base pressure was lower than 2.0×10^{-7} Torr, and the working pressure was 2×10^{-3} Torr. After deposition, rapid thermal annealing (RTA) was carried out at various temperatures for 1 hour. The magnetic properties of the annealed CoSiB/Pd multilayers were measured by using a vibrating sample magnetometer (VSM) at room temperature. The CoSiB/Pd multilayers were examined for in-plane and out-of-plane magnetic fields in the range from -2.0 and $+2.0$ kOe. For differential scanning calorimeter (DSC) measurements, a rapidly-solidified CoSiB ribbon with a thickness of 30×10^{-6} m and a width of 2×10^{-3} m was produced by using a single copper-roller vacuum melt-spinning technique under a controlled Ar-gas atmosphere. The spinning speed of the copper wheel of the melt spinner was optimized at 33 m/s. The thermal characterization of the CoSiB ribbon was carried out using the DSC under an Ar atmosphere at a heating rate of 0.34 K/s, and the crystalline temperature (T_x) was obtained.

The crystallization and the glass transition temperatures of the amorphous alloys have investigated by using the DSC. Figure 1 presents the DSC scans of the amorphous ferromagnetic $Co_{75}Si_{15}B_{10}$ and $Co_{70.5}Fe_{4.5}Si_{15}B_{10}$ [23] ribbons and presents the T_x's of the two ribbons. The T_x's of the two ribbons are 811 K and 790 K, respectively. The T_x is the onset temperature of first crystallization event, and it is one of which the thermal stability can be observed. Based on this result, the CoSiB is more suitable than the CoFeSiB for use as a free layer.

In a previous study, we investigated the perpendicular magnetic anisotropic properties of CoSiB/Pd multilayers with various thicknesses of the CoSiB and the Pd layers. The H_c's and the M_s's of the CoSiB/Pd multilayers were observed by using the VSM [24]. The CoSiB/Pd multilayers consist of Ta/Pd/[CoSiB t_{CoSiB}/Pd 13 Å]$_5$/Ta (Fig. 2(a)). Figure 2(b) shows the H_c and the M_s of the CoSiB/Pd multilayers. Both the H_c and the M_s increase for CoSiB layers with thicknesses from 2 to 3 Å. After

화학과

http://dx.doi.org/10.5012/bkcs.2014.35.8.2281

Controlled Synthesis of Spherical Polystyrene Beads and Their Template-Assisted Manual Assembly

Seo Young Yoon, Yi-Seul Park, and Jin Seok Lee*

*Department of Chemistry, Sookmyung Women's University, Seoul 140-742, Korea. *E-mail: jinslee@sookmyung.ac.kr*
Received February 21, 2014, Accepted April 8, 2014

Polystyrene beads (PS beads) with narrow size distribution were synthesized, and their diameter was controlled from 1.2 to 5 μm by varying the injection rate of a styrene solution containing initiator and the concentration of reactant, such as initiator and capping material. The diameter of the PS beads increased with increasing in the injection rates and the initiator concentration or decreasing the capping material concentration. Then, we used the PS beads as building block, and organized them into a hexagonally close-packed monolayer on substrate with template-assisted manual assembly. We showed perfect hexagonally close-packed organization of the PS beads with various sizes in large-scale area. And we demonstrated the superiority of the dry manual assembly over the wet self-assembly in terms of simplicity, speed, perfect ordering, and large scale.

Key Words : Polystyrene beads, Template-assisted manual assembly, Capping material, Initiator

Introduction

Colloidal solutions are heterogeneous mixtures in which the dispersed-phase particles are microscopically dispersed throughout another substance. They have long been used as versatile and essential building blocks of various industrial products such as foods, drinks, inks, paints, toners, coatings, papers, cosmetics, photographic films, and rheological fluids.[1] In particular, spherical polymeric colloidal particles with various dimensions have attracted attention because of their potential biomedical and chromatographic applications and applications in preparing display coatings for use as coating materials and spacers in LCD displays.[2-6]

Efforts to synthesize colloidal particles with uniform size, shape, composition, and properties have resulted in a number of recent advances in elucidating and understanding the optical, rheological, and electrokinetic behaviors of these materials.[7] Various chemical methods have now been developed to generate spherical colloids from a range of different materials such as organic polymers and inorganic compounds.[8] On the basis of studies in materials science, chemistry, biology, condensed matter physics, applied optics, and fluid dynamics,[9] a precise control over the properties of spherical colloids has been achieved by changing their intrinsic parameters such as the diameter, chemical composition, dispersion force, crystallinity, and surface functional group.[10]

The organization of these colloidal particles into two-dimensional (2D) arrays on various substrates is an important area of study in modern science and technology.[11,12] The ability to organize spherical colloids into spatially periodic structures has afforded interesting and often useful functionality not only from the colloidal materials but also from the long-range order exhibited by these crystalline lattices.[13] As a result, the assembly of colloidal particles has attracted attention for not only fundamental research[14] but also catalysis,[15] chemical sensing,[16] and investigations on bit-patterned magnetic media.[17-19] In the fabrication of photonic band-gap structures, these organizations resulted in extremely effective gravitational, convective, and electro-hydrodynamic forces.[20-28]

General approaches for the organization of colloidal particles are commonly based on self-assembly processes from the solution to substrates, including sedimentation,[29] evaporation,[30] adsorption,[31] Langmuir–Blodgett,[32] and spin coating.[33] Although these self-assembly methods are useful for organizing colloidal particles into hexagonally ordered arrays by using solvents, the organization of colloidal particles into large and perfect 2D arrays based on the self-assembly in solution has several drawbacks.[20] For instance, defects may be easily formed on the self-assembled arrays as these arrays are fragile, only a part of the array may be organized as required, or the array formation may proceed in an uncontrollable manner. Moreover, formation of close-packed arrays over a large area is difficult, expensive and time consuming. Therefore, high-precision applications in nano-material science and engineering necessitate the organization of colloidal particles into large, perfect, defect-free arrays with precise control of array symmetry and inter-particle distance, which in turn requires the development of effective and practical methods that are different from conventional methods based on self-assembly in solution.

On the basis of our recent finding that rubbing is a highly effective method for the organization of zeolite micro-particles into monolayers on flat substrates,[34] in this study, template with the shape of nanowell arrays was used to organize dry spherical colloidal particles into hexagonally close-packed monolayers by the manual assembly method of rubbing. For precise orientation control, polystyrene (PS) beads were synthesized with narrow size distribution as building blocks and controlled their diameter by varying the reaction temperature and the injection rate of a styrene solution containing 2,2'-azobis(2-methylpropionitrile) (AIBN)

as the initiator. In addition, we demonstrated the superiority of the dry manual assembly over the wet self-assembly in terms of simplicity, speed, perfect ordering, and large scale.

Experimental

Chmicals. Styrene (≥ 99%), 2,2'-azobis(2-methylpropionitrile) (AIBN) and polyethyleneimine (branched) (PEI) were purchsed from Sigam-Aldrich. Polyvinylpyrrolidone (K 30, average Mw 40,000) (PVP K 30) was purchased form Fluka. Silicone elastomer base and silicone elastomer curing agent (SYLGARD® 184) were purchased from Dow-Corning. Ethyl alcohol (99.9%), sulfuric acid and hydrogen peroxide were purchased from Duksan.

Synthesis of the Polystyrene Beads. To make the PS beads as building blocks, styrene, PVP K 30 and AIBN was used as monomer source, capping material and initiator, respectively. First, the PVP K 30 was dissolved in ethanol in three-necked round-bottom flask under argon atmosphere with stirring, then, the reaction temperature was elevated until 70 °C in which AIBN could decompose. After arrival of the reaction temperature, styrene solution containing AIBN was injected to the flask using the syringe and the injection rate was constant. Following injection, polymerization reaction was maintained for over 24 h. The PS beads were obtained by centrifugation and washed to remove residual styrene and PVP K 30 several times. The PS beads were dried at the room temperature.

Preparation of the Patterned Si Substrates. The patterned Si wafer with hexagonal arrays with silicon pillar was fabricated by KrF stepping method. At first, a bottom anti-reflection (BAR) layer was spin-coated with the thickness of 58 nm on a 6 inch Si(100) wafer. And then, a positive photoresists (PR) was spin-coated onto the BARC/Si wafer. The PR/BAR/Si wafer and a photomask was placed and the set up was exposed to the KrF UV light (248 nm) with the non-contact and step, repeatedly. The desired pattern sizes of hole and pitch in the photomask are 3.4 μm/4.8 μm, 8.5 μm/12 μm, and 14.2 μm/20 μm which are four times larger than the desired image sizes on the wafer (4 ×). After the KrF stepper, the patterned Si wafer was obtained with the pattern sizes of hole/pitch which are 0.85 μm/1.2 μm, 2.125 μm/3 μm, and 3.55 μm/5 μm. We can employ the 1.2 μm, 3 μm, and 5 μm, of PS beads. The irradiated PR/BAR/Si wafer was developed on a AZ 300 MIF developer. Deep etching of the PR-free Si areas was carried out by ion coupled plasma (ICP) using Ar-based SF_6. The depths of etching were varied from 700 nm to 2 μm. The remaining BARC and PR layers after etching were removed by a PR asher. The patterned Si wafer was diced into 1.5 cm × 1.5 cm.

Assembly of the Polystyrene Beads. The glasses were coated with poly (ethyleneimine) (PEI) by spin coating. The patterned Si substrates were immersed in a piranha solution (a 3:1 mixture of sulfuric acid and hydrogen peroxide) and washed with deionized (DI) water. On the patterned Si substrates, the poly (dimethylsiloxane) (PDMS) solution which was a 10:1 mixture of silicone elastomer base and curing agent was poured and maintained in the 70 °C for few hours. When the PDMS substrate was prepared as a template, a small amount of the PS beads powder was placed on the template and this powder was rubbed using PDMS slab with uniform direction repeatedly. After rubbing process, the randomly aggregated upper layers of PS beads were removed from the bottom monolayer with hexagonally close-packing (HCP) by using a fresh sticky PDMS slab for a few seconds on top of the PS beads array and subsequently removing the PDMS slab. Finally, this monolayer of the HCP PS beads on the PDMS substrate was transferred to the pre-coated glasses with PEI.

Results and Discussion

PS beads were synthesized with various sizes by controlling the injection rate of the source which was a mixture of styrene monomer and AIBN. The size of the PS beads increased with increasing the injection rate of the source, as shown in Figure 1. As the injection rate of the source was increased from 0.1 to 0.2, and then to 0.3 mL/min, the average particle diameters increased from 2.3, to 3.4, and then to 3.7 μm, respectively; the standard deviations for each particle size were 3.4%, 1.1%, and 0.8%, respectively. The rapid injection rate of the styrene solution resulted in the formation of larger styrene monomers for some time. PVP K 30 in the flask had a role to cap the injectioned source during the synthesis. When the injection rate of the source was fast, PVP K 30 can not embrace a large amount of source, and induced the combination of these styrene monomers with the oligomers or nanoparticles contained in the solution. This difference in the amounts of combined styrene monomers clearly explains the proportional increasing size of the PS beads under the rapid injection rate of the styrene solution.

Previous studies[35-38] demonstrated that the size of PS

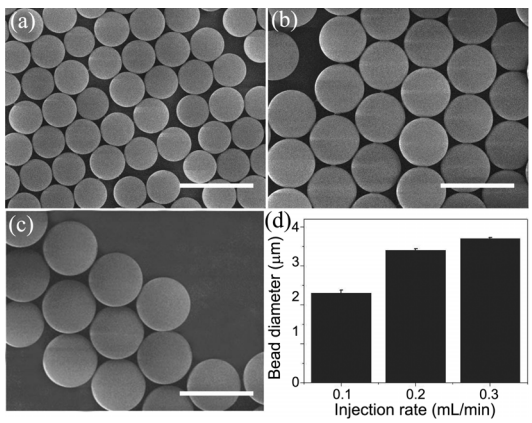


Figure 1. SEM images of the PS beads synthesized under different injection rate of the styrene solution containing AIBN: (a) 0.1, (b) 0.2, and (c) 0.3 mL/min at the polymerization temperature of 70 °C. (d) The each average diameter and standard deviation of PS beads. According to an increase of the injection rate, the size of the PS beads became larger. The scale bars are 5 μm.

Article

Journal of
Nanoscience and Nanotechnology
Vol. 14, 9169–9173, 2014
www.aspbs.com/jnn

Correlating Defect Density with Growth Time in Continuous Graphene Films

Cheong Kang, Da Hee Jung, Ji Eun Nam, and Jin Seok Lee*

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

We report that graphene flakes and films which were synthesized by copper-catalyzed atmospheric pressure chemical vapor deposition (APCVD) method using a mixture of Ar, H_2, and CH_4 gases. It was found that variations in the reaction parameters, such as reaction temperature, annealing time, and growth time, influenced the domain size of as-grown graphene. Besides, the reaction parameters influenced the number of layers, degree of defects and uniformity of the graphene films. The increase in growth temperature and annealing time tends to accelerate the graphene growth rate and increase the diffusion length, respectively, thereby increasing the average size of graphene domains. In addition, we confirmed that the number of pinholes reduced with increase in the growth time. Micro-Raman analysis of the as-grown graphene films confirmed that the continuous graphene monolayer film with low defects and high uniformity could be obtained with prolonged reaction time, under the appropriate annealing time and growth temperature.

Keywords: Graphene, Growth Temperature, Annealing Time, Growth Time.

1. INTRODUCTION

Graphene is a two-dimensional allotrope of carbon comprising of a one-atom-thick planar sheet of sp^2-bonded carbon atoms densely packed in a honeycomb crystal lattice. It has been attracting great interest because of its excellent structural and electrical properties.[1] Moreover, the high electron mobility and tunable band gap of graphene makes it a potential material for electronics and sensing applications.[2–5] Originally, the most commonly adopted technique for the preparation of graphene was based on the mechanical exfoliation of graphite. However, the limitations associated with the scalability of graphene hinder the industrial-scale implementation of this method. To this end, extensive efforts have been made to develop appropriate methods for graphene synthesis, including reduction of graphene oxide, thermal decomposition of SiC, and transition-metal-assisted chemical vapor deposition (CVD) process.[6–13]

Among them, the CVD method using metal catalysts, such as Ir,[8] Ru,[9] Ni,[10–12] and Cu,[13] can facilitate the synthesis of large-area monolayer graphene films, suitable for use in several applications. Recently, copper has been demonstrated to be the most effective catalyst for the growth of high-quality monolayer graphene films. This could be attributed to the low carbon solubility of copper.[14] With the remarkable development of the CVD method, it is possible to obtain graphene films with higher electron mobility, comparable to those of the exfoliated graphene.

Despite the complexity of the CVD process involving different catalysts, carbon sources, and other variables, the physical principles underlying this method are relatively simple. In the CVD process, graphene growth occurs either via surface catalytic reaction[9, 13] in case of the catalyst with low carbon solubility, or via bulk carbon precipitation onto the surface during cooling[11, 12] in case of catalyst with high carbon solubility. In both the cases, graphene nucleation on a catalyst surface is one of the critical steps governing the growth process. In principle, several factors affect the initiation of the nucleation, including the surface microstructure of the metal catalyst,[9, 15] carbon source,[16] carbon segregation from metal-carbon melts,[17] and reaction parameters adopted during the CVD growth.[18–20]

Herein, we have analyzed the role of various CVD reaction parameters on the initiation of graphene nucleation. To this end, we synthesized hexagonal graphene flakes onto Cu foils under various CVD reaction conditions and analyzed the effect of the various reaction parameters on the quality of the graphene films. In particular,

*Author to whom correspondence should be addressed.

we analyzed the role of growth temperature, annealing time, and growth time in the formation of uniform and highly continuous graphene films, by comparing the coverage, density, and the size of the graphene domains using scanning electron microscopy (SEM) and micro-Raman spectroscopy.

2. EXPERIMENTAL DETAILS

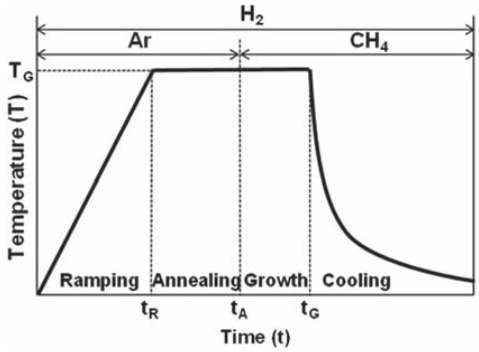
Materials. Copper foils (99.999% purity, 0.025-mm (0.001 in) thickness) that was used for the synthesis of graphene films was purchased from Alfa Aesar. The 300-nm-thick SiO_2/Si substrate, which used for transferring the as-grown graphene films prior to SEM and Raman analysis, was purchased from Hitsan.

Characterization of the as-grown graphene. The as-grown graphene domains and films were characterized by scanning electron microscopy (SEM) and micro-Raman spectroscopy. The morphology of the as-grown graphene domains were observed by using a field-emission scanning electron microscope (FE-SEM, JEOL 7600F) at an acceleration voltage of 15 kV. Prior to SEM observation, the graphene films were transferred onto SiO_2/Si substrates. The micro-Raman mapping of the as-grown graphene films was obtained on an indigenously developed micro-Raman spectroscopy system. For this, 514.5 nm line of an Ar ion laser was used as the excitation source with a power of ~1 mW. The heating effect can be neglected at this power range. The laser beam was focused onto the graphene sample by a 50× microscope objective lens (0.8 N.A.), and the scattered light was collected in the backscattering geometry. The collected scattered light was dispersed by a Shamrock SR 303i spectrometer (1200 grooves/mm) and was detected by using a CCD detector.

3. RESULTS AND DISCUSSION

In this study, graphene flakes were synthesized by copper-catalyzed atmospheric pressure chemical vapor deposition (APCVD) method using a mixture of Ar, H_2, and CH_4. Scheme 1 illustrates the different growth stages associated with the growth of graphene films under the various reaction conditions. In the typical process, Cu foil (99.999% purity) of thickness 0.025 mm was initially cleaned with acetone, methanol, and deionized (DI) water. Subsequently, the Cu foil was placed at the center of a quartz tube (1 inch diameter), positioned inside a horizontal tube furnace. The typical graphene synthesis process involves four steps, namely, ramping, annealing, growth, and cooling. In the scheme, the ramping, annealing, and growth times are denoted as t_R, t_A, and t_G, respectively, and the growth temperature is denoted as T_G. The graphene flakes grown onto the Cu foils were subsequently transferred to a 300-nm-thick SiO_2/Si substrate.

In this study, we varied the growth temperature T_G and analyzed its influence on the size of graphene domains.



Scheme 1. Schematic diagram indicating the four stages involved in the growth of graphene: ramping, annealing, growth and cooling process; In the schematic, t_R, t_A, t_G, and T_G indicate ramping time, annealing time, growth time, and growth temperature, respectively.

Figure 1 shows the SEM images of the graphene flakes synthesized at different growth temperatures: (a) 1030, (b) 1040, and (c) 1050 °C. The growth time was maintained at a constant duration of 15 min for all the samples. In principle, the nucleation and growth of the graphene domains on the Cu foil is initiated via the supersaturation of the active carbon species, which are produced by the decomposition of carbon sources, such as methane, at relatively high temperatures. This implies that the nucleation and growth of the graphene depends on the growth temperature. As can be seen from the SEM image shown in Figure 1(A), the sample grown at 1030 °C does not have graphene domains on the Cu surface. This could possibly be attributed to the undersaturation of carbon sources on the Cu surface.[21] On the other hand, with the realization of certain critical point of supersaturation at a high temperature, graphene nuclei start to form on the Cu catalyst and grow into graphene domains. Besides, it can be observed that higher synthesis temperature favors the growth of larger graphene domains (Figs. 1(B) and (C)). The observed increase in the average size of graphene domains with increase in growth temperature can be attributed to the increase in the growth rate of graphene, due to the higher supersaturation probability of the thermally decomposed carbon species. Therefore, we believe that the increase in the growth temperature accelerates

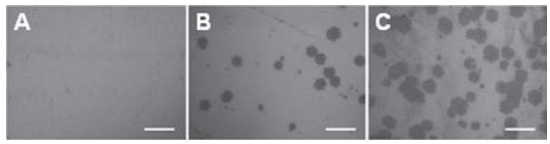


Figure 1. SEM images of the as-grown graphene domains on Cu foil using the APCVD method, obtained with different growth temperature: (A) 1030, (B) 1040, and (C) 1050 °C. All graphene domains were synthesized at a fixed annealing time and growth time of 60 min and 15 min, respectively. Prior to SEM observation, the graphene films were transferred onto SiO_2/Si substrates. Scale bars correspond to 10 μm.

CARBON 80 (2014) 708–715

Available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/carbon

Correlating nucleation density with heating ramp rates in continuous graphene film formation

Da Hee Jung [a,1], Cheong Kang [a,1], Byung Hee Son [b,c], Yeong Hwan Ahn [b,c], Jin Seok Lee [a,*]

[a] Department of Chemistry, Sookmyung Women's University, Seoul 140-742, Republic of Korea
[b] Department of Physics, Ajou University, Suwon 443-749, Republic of Korea
[c] Department of Energy Systems Research, Ajou University, Suwon 443-749, Republic of Korea

ARTICLE INFO

Article history:
Received 29 May 2014
Accepted 5 September 2014
Available online 16 September 2014

ABSTRACT

Graphene films with different nucleation densities were successfully grown on Cu foils using an atmospheric pressure chemical vapor deposition (APCVD) method. We investigated the effect of heating ramp rate on graphene growth, which is critical to the control of both the domain density and the defects density in further synthesized graphene film. Density of graphene domains was reduced with a decrease in ramp rate; heating Cu foil at a lower rate produced flat and smooth surface with larger Cu grains. These surface roughness and grain size-heating rate association were confirmed by atomic force microscope and electron backscatter diffraction analysis. By comparing micro Raman mappings of the graphene films grown with different ramp rates, we found that the number of layers, defect density, and uniformity of as-grown graphene films were strongly dependent on heating ramp rate. With these observations, we demonstrate the growth mechanism of graphene films as a function of the heating ramp rates.

1. Introduction

Graphene, which is a monolayer of sp^2-bonded carbon atoms, is a two-dimensional (2D) material that has been attracting great interest because of its fascinating structural and electrical properties [1]. The extremely high electron mobility of Graphene and its tunable band gap make it potentially attractive for use in innovative electronic and sensing devices [2–5]. Since the first demonstration of its synthesis by mechanical exfoliation of highly oriented pyrolytic graphite (HOPG) [6], a number of alternative synthesis methods have been developed. These include decomposition of SiC surfaces by thermal treatments [7], synthesis using the solution phase [8,9], solvothermal synthesis of graphene [10,11], and chemical vapor deposition (CVD) of graphene on transition metals [12–17]. Of all these approaches, the CVD method is the most effective for device fabrication since it is highly reproducible and yields high-quality films of controllable thickness [18,19]. In particular, graphene films grown on Cu foil by CVD have attracted much interest, because it can produce large area monolayer graphene films due to its low solubility of carbon in Cu, which can be applicable for wafer scale

* *Corresponding author.*
E-mail address: jinslee@sookmyung.ac.kr (J.S. Lee).
[1] These authors contributed equally.
Abbreviations: APCVD, atmospheric pressure chemical vapor deposition; R_r, heating ramp rate; SEM, scanning electron microscopy; AFM, atomic force microscopy; rms, root mean square; EBSD, electron back-scattered diffraction
http://dx.doi.org/10.1016/j.carbon.2014.09.016
0008-6223/

graphene-based devices and circuits [17,20]. Furthermore, atmospheric pressure CVD (APCVD) is highly suited for synthesis of large-scale graphene films as it is a cost-effective, scalable, quick, and high-throughput method. During CVD, various factors affect the initiation of the graphene nucleation process such as the surface microstructure of the metal catalyst [21], the carbon source used [22], surface oxygen [23], and the reaction parameters used during CVD growth [24–26]. However, a number of other factors involved in CVD growth of graphene films have not yet been studied fully and need to be elucidated so that graphene films with low defects density can be fabricated.

In general, it has been found that the presence of domain boundaries in synthesized graphene films has a detrimental effect on their fundamental electronic properties [27,28]. Therefore, it is recently reported to grow graphene domains on liquid copper [29,30] or resolidified copper in order to decrease the number of domain boundaries, increasing each size of graphene domains and hence decrease the defects density of large-area, monolayer graphene films [31–34]. However, graphene grown on liquid or resolidified copper has some drawbacks such as difficulties in transferring on other substrates and high temperature needed to melt the copper and still challenges [34]. Owing to the fact that initial nucleation of graphene domains takes place more readily at step edges, folds, and other imperfections on the Cu foil used as substrate, the existence of graphene domains as well as their number depends strongly on the topography of the Cu foil surface. It is well-known that the density of graphene domains can be suppressed by increasing growth temperature or annealing time, because the number of step edges, folds, and other defects are reduced at higher growth temperatures or by annealing for longer periods [26,35]. Having lower defects density on the Cu foil surface results in a lower density of nucleation sites and thus fewer graphene domains [36].

In this study, we investigated the effects of heating ramp rate on the growth of graphene films by monitoring the characteristics of graphene domains such as nucleation density during the initial stage of graphene nucleation. To elucidate the role of heating ramp rate on graphene growth, we synthesized graphene domains and films on Cu foil using various ramp rates and measured the densities of graphene domains, the roughness of Cu surface, and the size of Cu grains. Characterization techniques, such as scanning electron microscopy (SEM), atomic force microscopy (AFM), and electron back-scatter diffraction (EBSD) analyses were employed to obtain the correlation of nucleation density with heating ramp rates. In addition, we characterized graphene films grown with different ramp rates using ultraviolet–visible (UV–vis) spectroscopy and micro-Raman spectroscopy.

2. Experiment

The graphene films were synthesized using the APCVD method. The Cu foil used as substrate (25 μm thick, 99.999%, Alfa Aesar) was first cleaned with acetone, isopropyl alcohol, and deionized (DI) water. The foil was then placed in a 1 inch-diameter quartz tube, and the tube put inside a horizontal furnace (Lindberg/Blue M, Thermo Scientific). The system was evacuated for 10 min. Next, the pump was isolated, and the growth chamber brought back to atmospheric pressure by introducing a H_2 and Ar. This process was repeated several times, all at room temperature. The Cu foil was then heated to 1050 °C using one of several selected ramp rates (35 °C/min, 18 °C/min, or 9 °C/min) in the H_2 and Ar atmosphere. For the initial Cu cleaning and annealing, the sample was maintained under the same temperature and atmosphere as the annealing process for another 1 h. Then, the supply of Ar was shut off, and methane was introduced for graphene growth. The flow setup for this stage consisted of 300 sccm of diluted methane (50 ppm in Ar), along with 20 sccm of H_2, flowing into the reaction tube. Process times used were 10 min for identification of initial graphene nucleation and 60 min for film growth. Finally, the samples were rapidly cooled to room temperature (20 °C/s) in a protective atmosphere of Ar and H_2. (see the Supplementary information Fig. S1 for the detailed description of the experiment.) The annealing and growth conditions, and the cooling rate were the same for all samples, in order to minimize variations in the characteristics of synthesized films owing to these factors. We only varied heating ramp rates in each experiment.

Scanning electron microscopy (SEM) images of the synthesized graphene flakes were obtained using a field-emission SEM system (JEOL 7600F) operated at acceleration voltages of 10–20 kV. AFM and EBSD analyses were performed to identify the surface roughness and crystallographic nature of the Cu foil samples, respectively. Raman spectra of the synthesized graphene samples were obtained using a laboratory-made micro-Raman spectroscopy system; the system used the 514.5 nm line of an Ar ion laser as excitation source with a power of ~1 mW. The heating effect of the laser could be neglected at this power level. The laser beam was focused onto the graphene sample using a 50× microscope objective lens (Numerical Aperture = 0.8). The collected scattered light was dispersed using a Shamrock SR 303i spectrometer (1200 grooves/mm) and was detected using a charge-coupled device (CCD) detector.

3. Results and discussion

As mentioned previously, we investigated the effects of heating ramp rate on initial nucleation density of graphene domains during APCVD growth of graphene films on Cu foil. Fig. 1(a–c) shows SEM images of the graphene domains grown on Cu foil using different ramp rates, namely, 35 °C/min, 18 °C/min, and 9 °C/min and transferred on SiO_2/Si substrates; these rates are denoted as R_r-°C/min in the figure. The graphene domains synthesized in the case of R_r-35 (i.e., for a ramp rate of 35 °C/min) were densely distributed on the Cu surface (Fig. 1(a)). However, the R_r-18 condition (i.e., a rate of 18 °C/min) resulted in less densely distributed graphene domains (Fig. 1(b)). Finally, R_r-9 (i.e., 9 °C/min) resulted in graphene domains with a significantly lowered nucleation density (Fig. 1(c)). This tendency was also correspond to the results of graphene domains grown for 30 min with different ramp rates. (see the Supplementary information Fig. S2) On the basis of the above-mentioned results, it could be concluded that a low heating ramp rate (R_r-9) has a similar effect on density of graphene domains as

DOI: 10.1002/anie.201400653

Neurochemistry

Cytoskeletal Actin Dynamics are Involved in Pitch-Dependent Neurite Outgrowth on Bead Monolayers**

Kyungtae Kang, Seo Young Yoon, Sung-Eun Choi, Mi-Hee Kim, Matthew Park, Yoonkey Nam, Jin Seok Lee,* and Insung S. Choi**

Abstract: *Neurite outgrowth is an important preceding step for the development of nerve systems. Given that the in vivo environments of neurons consist of numerous hierarchical micro/nanotopographies, there have been many efforts to investigate the relationship between neuronal behaviors and surface topography. The acceleration of neurite outgrowth was recently reported on surfaces with a periodic nanotopography, but the biological mechanism has not yet been elucidated. In this work, the initial neurite development of hippocampal neurons on assembled silica beads with diameters ranging from 700 to 1800 nm was explored. The acceleration of neurite outgrowth increased with the surface-pitch size and leveled off after a pitch of 1 μm. Biochemical analysis indicated that cytoskeletal actin dynamics were primarily responsible for the recognition of surface topography. This work contributes to the emerging research field of topographical neurochemistry, as well as applied fields including neuroregeneration and neuroprosthetics.*

Neurites (axons and dendrites) are the cytoplasmic projections of neurons, and their direction, location, elongation, and even the time point of neuritogenesis directly determine the synaptic connections between neurons for the flow of information through the network. For the proper construction of the neural network, each neuron has the innate machinery to interpret the surrounding environment and determine the correct direction for sprouting neurites. Previous studies have shown that the specific recognition of biochemical cues is intimately connected with the local cytoskeletal dynamics at growth cones, the F-actin-based dynamic tips of neurites, during neurite outgrowth or guidance.[1–7] In addition to biochemical cues, there exists an often ignored but important environmental factor that also controls neuronal behaviors in vivo, namely "topography". The extracellular matrix (ECM) constructs the topographical environment for neurons and contains hierarchical micro/nanostructures composed of protein fibers (e.g., collagens and elastins) embedded with smaller fibril structures.[8]

The connection between neuronal behavior and micro/nanotopographical features has been investigated since contact guidance, which is the aligned growth or migration of cells to the physical shape of substrates, was demonstrated in neurons by using lithographically fabricated microgrooves.[9–11] The fabrication of nanostructures has allowed for the discovery of far more intriguing responses of neurons to nanotopographical cues;[12–21] most importantly, the acceleration of neuritogenesis. The developmental acceleration of neurites was initially reported on electrospun nanofibers,[20] and the presence of a threshold pitch for the acceleration effect was subsequently found on periodic nanostructures such as anodized aluminum oxide (AAO)[22] and assembled silica beads.[23,24] Although these reports strongly suggest that nanotopography is biologically recognized by neurons, the nature of the topographical influence on neuronal behaviors at the intracellular level remains elusive. With the consideration that both neurons and the ECM have heterogeneous, hierarchical structures, particularly at the submicro- and micrometer scales, we investigated the initial development of hippocampal neurons on topographical substrates in this range (pitch from approximately 700 to 1800 nm), and found that the acceleration of neurite outgrowth was saturated at pitches longer than 1000 nm. Biochemical analysis indicated that neuronal recognition of surface topography was governed by cytoskeletal actin.

Surfaces with various pitches were fabricated by organizing silica beads into two-dimensional, monolayer arrays on glass substrates (Figure S1 in the Supporting Information). Beads with diameters of 110, 320, 670, 1000, 1200, 1300, 1480, or 1750 nm were used in the study (SB-110, SB-320, SB-670, SB-1000, SB-1200, SB-1300, SB-1480, and SB-1750, respectively; with SB-110 as the control). Primary hippocampal neurons derived from E-18 Sprague Dawley rats were cultured on the closely packed bead layer (density of 50–200 cells/mm^2; Scheme 1). Before plating, each substrate with a different bead size was tailored to be neuron-adhesive by

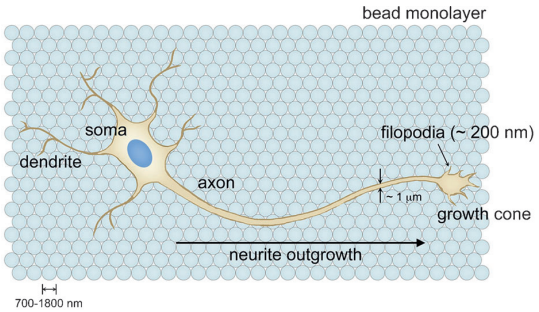
[*] Dr. K. Kang,[+] M.-H. Kim, M. Park, Prof. Dr. I. S. Choi
Center for Cell-Encapsulation Research, Department of Chemistry
KAIST, Daejeon 305-701 (Korea)
E-mail: ischoi@kaist.ac.kr
Homepage: http://cisgroup.kaist.ac.kr
Dr. K. Kang,[+] Prof. Dr. Y. Nam, Prof. Dr. I. S. Choi
Department of Bio and Brain Engineering
KAIST, Daejeon 305-701 (Korea)
E-mail: ynam@kaist.ac.kr
Homepage: http://neuros.kaist.ac.kr
S. Y. Yoon,[+] S.-E. Choi, Prof. Dr. J. S. Lee
Department of Chemistry, Sookmyung Women's University
Seoul 140-742 (Korea)
E-mail: jinslee@sookmyung.ac.kr
Homepage: http://fetns.sookmyung.ac.kr
[+] These authors contributed equally to this work.

[**] This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (2012R1A3A2026403, 2012R1A2A1A01007327, 2012-0009562, 2012R1A1A2022597, and 2012M3A7B4034986).

Supporting information for this article, including experimental details, is available on the WWW under http://dx.doi.org/10.1002/anie.201400653.

coating it with poly-D-lysine (PDL), preceded by O_2 plasma treatment. Neurons adhered and survived well on the bead layers with normal development of neurites and growth cones (Figure S2). The developmental behavior of the cultured neurons was analyzed at 1 day in vitro (DIV) in order to investigate the differences in the neuritogenesis phase. Since the beads made the substrates opaque, the cultured neurons were fixed and immunostained with anti-β-tubulin and phalloidin to target microtubules and F-actin, respectively.

The neuronal morphologies on the submicrometric pitches were remarkably different to those on the controls (coverslip and SB-110; Figure 1). At 1 DIV, the neurons on



Scheme 1. Illustration of a neuron cultured on a silica-bead substrate.

the coverslip or SB-110 had barely sprouted neurites (red staining) and possessed the lamellipodial structures (green staining) that were generally found in the neuron culture on flat surfaces at 1 DIV (Figure 1a,b).[23] In stark contrast, the neurons on SB-670 had already developed the major neurite (a neurite more than double the length of the other neurites), and the formation of a growth cone and lateral filopodial structures was confirmed by the F-actin staining (Figure 1c). Lamellipodial structures surrounding the somas were not observed on SB-670 at 1 DIV. These morphological changes continued for the neurons on bigger beads (SB-1000, SB-1200, SB-1300, SB-1480, and SB-1750) but with significantly longer major neurites (Figure 1d,e and Figure S3). The averaged lengths of the longest neurite for each neuron clearly showed enhanced acceleration on the larger beads (Figure 1f): the length increased with the bead diameter (pitch; e.g., [SB-670: (52.46 ± 2.30) μm; SB-1000: (67.64 ± 4.90) μm]. Beyond SB-1000, however, there was no significant increase in neurite length with increasing bead diameter [SB-1200: (71.24 ± 7.51) μm; SB-1300: (74.93 ± 3.77) μm; SB-1480: (75.59 ± 4.81) μm; SB-1750: (69.09 ± 3.35) μm; there was no statistically significant difference between the values]. This result indicates that there was a saturation point or an upper threshold for the pitch dependency of accelerated neurite outgrowth. Developmental analysis of the neurons on selected bead substrates showed a similar trend of pitch dependency (Figure S4), thus implying that accelerated neuritogenesis contributed to the increased neurite length observed.

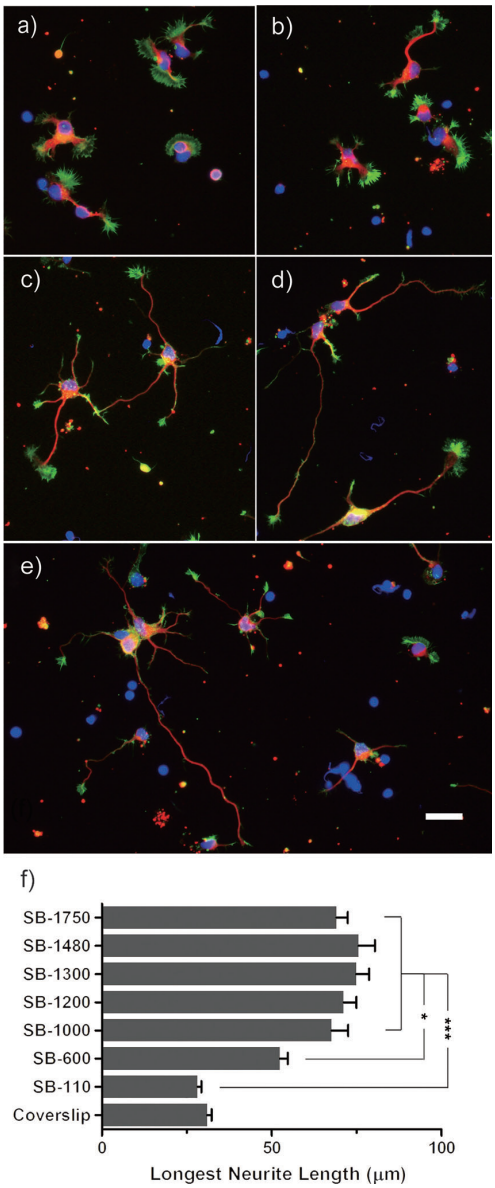


Figure 1. Fluorescence micrographs of hippocampal neurons at 1 DIV. The neurons were immunostained with anti-β-tubulin (red), phalloidin (green), and Hoechst 33342 (blue). The cells were cultured on a) Glass coverslip, b) SB-110, c) SB-670, d) SB-1000, and e) SB-1200. Scale bar: 25 μm. f) Average lengths (mean ± standard error) of the longest neurite on each substrate. All of the results were analyzed by one-way ANOVA at a significance level of 95%, followed by Bonferroni's multiple comparison test ($N = 157, 114, 123, 117, 108, 113, 118$, and 153 for coverslip, SB-110, SB-670, SB-1000, SB-1200, SB-1300, SB-1480, and SB-1750, respectively; $*P < 0.05$; $***P < 0.001$).

Notably, the acceleration effect was not intensified limitlessly as the pitch increased. The existence of an upper threshold found in this study complemented our previous report that showed the presence of a lower threshold of around 200–300 nm, a size comparable with the thickness of

http://dx.doi.org/10.5012/bkcs.2014.35.2.349

Articles

Effect of Aluminum Purity on the Pore Formation of Porous Anodic Alumina

Byeol Kim and Jin Seok Lee*

*Department of Chemistry, Sookmyung Women's University, Seoul 140-742, Korea. *E-mail: jinslee@sookmyung.ac.kr*
Received September 23, 2013, Accepted October 14, 2013

Anodic alumina oxide (AAO), a self-ordered hexagonal array, has various applications in nanofabrication such as the fabrication of nanotemplates and other nanostructures. In order to obtain highly ordered porous alumina membranes, a two-step anodization or prepatterning of aluminum are mainly conducted with straight electric field. Electric field is the main driving force for pore growth during anodization. However, impurities in aluminum can disturb the direction of the electric field. To confirm this, we anodized two different aluminum foil samples with high purity (99.999%) and relatively low purity (99.8%), and compared the differences in the surface morphologies of the respective aluminum oxide membranes produced in different electric fields. Branched pores observed in porous alumina surface which was anodized in low-purity aluminum and the size; dimensions of the pores were found to be usually smaller than those obtained from high-purity aluminum. Moreover, anodization at high voltage proceeds to a significant level of conversion because of the high speed of the directional electric field. Consequently, anodic alumina membrane of a specific morphology, *i.e.*, meshed pore, was produced.

Key Words : Porous anodic alumina, Impurities, Electric field, Branched pore, Meshed pore

Introduction

Recently, porous anodic aluminum oxide (AAO), a highly ordered nano-pore array, has attracted considerable attention as a nanotemplate for the synthesis of various nanostructures,[1] such as nanowires,[2-4] nanotubes,[5] nanodots,[6,7] semiconductors,[4,8] and superconductors.[9] These applications involve electrochemical deposition,[2,3] atomic layer deposition,[6,7] and chemical vapor deposition.[10] Porous alumina membrane formed by the anodization of aluminum consists of a close-packed array of hexagonal cells, each containing a cylindrical central pore. These pores are extended down to the barrier layer, which is a continuous, non-porous dielectric oxide layer between the pore bottom and the aluminum.[11] Moreover, pore diameter (D_p) and interpore distance (D_{int}) can be adjusted by anodization conditions such as the type of the electrolyte,[12-16] anodizing potential,[13,16] current density,[17] and temperature,[18] The most important factors affecting D_{int} are the chemical composition of the electrolyte and anodizing potential. Sulfuric acid,[12,13] oxalic acid,[12,18] and phosphoric acid[12,14] are most commonly used electrolytes, each of which is used with a specific applied potential.[12-18] In general, anodic alumina membranes can be obtained within three well-known growth regimes: sulfuric acid at 25 V for D_{int} of 63 nm,[12,13,19] oxalic acid at 40 V for D_{int} of 100 nm,[12,18,20] and phosphoric acid at 195 V for D_{int} of 500 nm.[12,14] Moreover, D_p can be modified with pore widening process after anodization. The configuration of pores affected porosity and pore density, which are characteristic of anodic alumina membranes.

However, these characteristic factors were distinguished at anodization using a high purity aluminum foil (99.999%). Most reports on the anodizing of aluminum use high-purity aluminum for improving the quality of porous alumina membranes.[1-19] However, few studies have been conducted on the anodization of low-purity aluminum.[21-29] Fabrication of anodic alumina membranes from low-purity aluminum foil is not trivial and often requires specific conditions, different from those typically applied to the anodization of high-purity aluminum.[29]

Herein, we use two different aluminum foils which are high purity (99.999%) and relatively low purity (99.8%) to fabricate a porous alumina membrane. The anodization was preceded by two-step process[20] under same conditions in oxalic acid with 40 V. In phosphoric acid condition, aluminum was anodized at 190 V for the distinct divisions of two types of purity. We think that impurities of aluminum can disturb the direction of the electric field, the main driving force for the pore formation during anodization. We compare surface morphologies of aluminum oxide to demonstrate the effect of changed electric field due to impurity of aluminum, and describe the effect of aluminum purity on the formation of anodic alumina pores during anodization. Furthermore, we provide the possibility of fabricating a novel nanostructure using low-purity aluminum.

Experimental

A high purity aluminum foil (99.999%, Good fellow) and relatively low purity aluminum foil (99.8%, Sigma Aldrich)

were cut into specimens (2 × 3 cm) with the working area of 4 cm^2. Aluminum foil was degreased in acetone and washed in deionized water. Then, the sheet was annealed under argon atmosphere at 500 °C in order to enhance the grain size in the metal and to obtain homogenous conditions for pore growth.[30] After this, the foil was electropolished in a mixture of perchloric acid ($HClO_4$) and ethyl alcohol (C_2H_5OH) (volume ratio 1:4) at constant potential of 10 V for 3 min to diminish the roughness of the aluminum surface.[31] The sample was anodized in a 0.3 M oxalic acid solution at 40 V for 1 h, under constant temperature at 10 °C. The other sample was anodized under 0.1 M phosphoric acid solution at 190 V and 1 °C. After anodization, random and disordered pores were obtained. Formed anodic aluminum oxide layer was chemically removed by a mixture of 6 wt % H_3PO_4 and 1.8 wt % CrO_3 at 50 °C for 30 min. The second anodization was carried out under the same condition as were used during the first anodization step for 30 min. After this, the pores were opened in a same mixture of etching solution at 30 °C for 10 min. During experimental, in order to keep the constant electrolyte temperature, a double walled bath with a refrigerated circuiting system is used. The morphology of porous alumina film was evaluated by a field emission scanning electron microscope (FE-SEM).

Results and Discussion

We proposed that electric field is the main driving force for the pore growth during anodization and impurities in the aluminum disturb the direction of electric field. Therefore, we carefully compared the morphologies of anodized alumina membranes and the changes in the electric field as a function of aluminum purity.

To fabricate a porous alumina membrane, we used two different aluminum foil samples with high purity (99.999%, Al_{high}) and relatively low purity (99.8%, Al_{low}), respectively. The anodization process was preceded by a two-step process[20] in oxalic acid at 40 V and phosphoric acid at 190 V for high- and low-purity aluminum samples, respectively.

Figure 1 shows the scanning electron microscopy (SEM) images of Al_{low} (Figures 1(a) and 1(b)) and Al_{high} (Figures 1(c) and 1(d)) anodized by a two-step process in 0.3 M oxalic acid solution at 10 °C. The applied potential was 40 V and the anodization duration was 1 h. The surface of anodized Al_{low} exhibits a disordered array of nanopores with different pore sizes (Figure 1(a)). Longer first-anodizing duration results in better ordering of the pores array in the two-step anodizing process.[32] Hence, we expected to obtain higher regularity ratio, circularity, and lower concentration of defects in the resulting anodic alumina. However, we observed that the surface of anodized Al_{low} unfavorably influences the regularity ratio, circularity, and defects, which had been also observed by others.[29] Interestingly, the majority of nanopores of anodized Al_{low} were found to consist of branched pores that had additional pores located near the main pore, which we attribute to defects formed during pore formation. Figure 1(b) shows a low-magnified SEM image of anodized Al_{low}, with the branched pores shown in red. Most pores appear highly disordered and branched. Because using Al_{low} for anodization increases the defects, we consider Al_{low} unsuitable for fabricating a highly ordered pore array.

Therefore, we conducted the two-step anodization process of Al_{high} under the conditions used for Al_{low} in order to investigate the effect of aluminum purity on the geometry of anodized pores. As shown in Figure 1(c), the surface of anodized Al_{high} exhibits fewer pore defects and better circularity of pores than the surface of anodized Al_{low}. The majority of pores which are anodized at Al_{high} is straight and single (not double or triple). Accordingly, the degree of branched pore for Al_{high} is less than that of Al_{low} (Figures 1(b) and 1(d)). Based on these observations, we confirmed that the purity of aluminum foil strongly affects the pore geometry.

The effect of aluminum purity on the geometry of anodized pores can be more accurately observed by investigating the size of D_p and D_{int}. Figure 2 shows D_p and D_{int} for pores

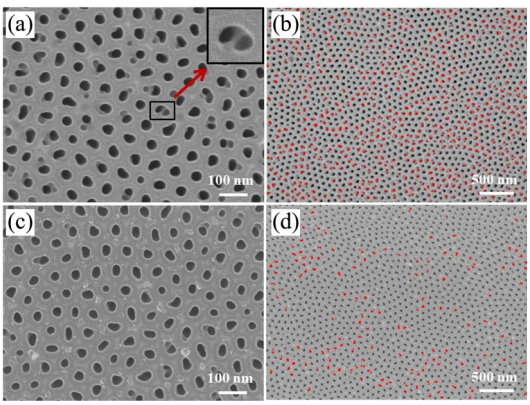


Figure 1. SEM images of Al_{low} (a, b) and Al_{high} (c, d) anodized in a two-step process in 0.3 M oxalic acid solution at 10 °C and 40 V for 1 h. Branched pores are represented as red in (b) and (d).

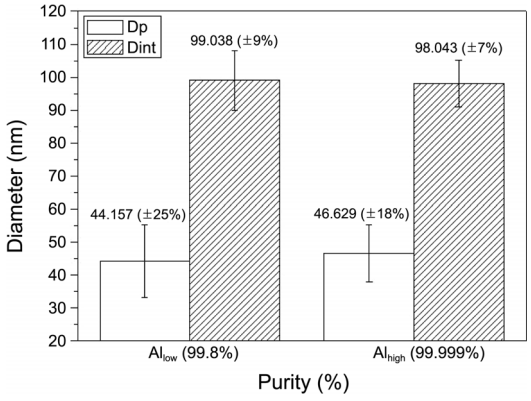


Figure 2. The histogram of average size of pore diameter (D_p) and interpore distance (D_{int}) in anodized Al_{low} and Al_{high}, respectively. Standard deviations are shown in parentheses.

Microporous and Mesoporous Materials 192 (2014) 89–94

Contents lists available at ScienceDirect

Microporous and Mesoporous Materials

journal homepage: www.elsevier.com/locate/micromeso

Energy transfer in dye molecule-containing zeolite monolayers

Hyunjin Lim [a], Sung-Eun Choi [b], Hyeonsik Cheong [a], Jin Seok Lee [b,*]

[a] Department of Physics, Sogang University, Seoul 121-742, South Korea
[b] Department of Chemistry, Sookmyung Women's University, Seoul 140-742, South Korea

ARTICLE INFO

Article history:
Available online 17 August 2013

Keywords:
Energy transfer
Host-guest material
Zeolite
Anisotropic photoluminescence

ABSTRACT

The inter- and intra-molecule energy transfers in dye molecules in zeolite crystals were studied using polarized photoluminescence spectroscopy. We used nanoporous zeolite-L crystal and two kinds of dye molecules, pyronine B (PyB^+) and pyronine Y (PyY^+), as the host and guest materials, respectively. To quantitatively monitor the arrangement of PyB^+ and PyY^+ molecules inserted into the channels of zeolite-L, the polarized fluorescence microscopy of dye-containing zeolite-L were investigated. From the polarized fluorescence microscopy images, it is clear that the incorporated PyB^+ molecules forced to be lined up along the channel direction of zeolite-L crystal and that the various arrangements of PyY^+ molecules were located mostly in the both ends of zeolite-L crystal. Here, we report that the vertically aligned zeolite monolayer can be applied as novel supramolecularly organized systems to study energy transfer dynamics between the internal and external fluorophores, and to develop zeolite-based advanced materials. Based on the polarized fluorescence spectroscopy and the angle-dependent intensity change depending on dye molecules, we discussed the dominant energy transfers between internally incorporated molecules and externally absorbed molecules in dye-containing zeolite-L monolayer.

1. Introduction

In natural photosynthesis, sunlight is absorbed by pigment molecules in a chloroplast. These systems allow directional energy transfer from an electronically excited molecule to an unexcited neighbor molecule in such a way that the excitation energy reaches the reaction center of photosynthesis with high probability and efficiency. Here, the anisotropic arrangement of the pigment molecules is crucial for efficient energy transportation that allows the flow of energy to be controlled and directed to regions for a desired purpose [1–3].

From the efforts to mimic a natural photosynthesis in the form of an optoelectronic device, a number of recent attempts have been made to improve device performance by the arrangement of fluorescence molecules using various methods. Approaches include stretching-drawing [4,5], Langmuir–Blodgett technique [6,7], liquid–crystalline self-organization [8], alignment on specific substrates [9], and chemical vapor deposition technique [10]. A similar approach is also possible by enclosing fluorescence molecules inside a porous material and by choosing conditions such that the cavities are able to accommodate only monomers but not aggregates [11–16]. The wide-ranging tenability of these highly organized materials offers fascinating new possibilities for exploring excitation energy transfer phenomena, and challenges for developing new photonic devices for solar energy conversion and storage [11].

Among these porous materials, zeolite is known the suitable hosts for specific organization of chromophores, which is inorganic crystalline aluminosilicates with pores in nanometer size (typically 0.4–0.8 nm), because of their well-defined internal structure with uniform cages, cavities or channels. Particularly, the aluminosilicate framework of the zeolite-L comprises a channel system of molecular dimensions which is providing a host matrix for adsorbent molecules.

Many investigations have been devoted to the study of the adsorption of organic molecules in zeolites and the modulation of the photochemical and photophysical processes when they are incorporated into the channels or cavities [17–19]. The properties of the organic guest-zeolite supramolecular assembly are clearly different from the molecular properties of the pure organic due to different factors such as the presence of cations, Si/Al atomic ratio, electrostatic field into the zeolite framework and channel and pore size [14,16,20–23]. The presence of Al in the framework is related to high electrostatic potentials in the zeolite medium that have been claimed to be responsible for the stabilization of charge separated transient species [24]. The pore and channel sizes are also crucial to understand the different optical properties under the incorporation of guest molecules into the hosts. Therefore, the extent to which one or more of these effects can affect to the photophysical properties of organic molecule depends on the dimension of guest molecules inserted into the channels of zeolite host.

* Corresponding author. Tel.: +82 220777464; fax: +82 220777321.
E-mail address: jinslee@sookmyung.ac.kr (J.S. Lee).

1387-1811/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.micromeso.2013.08.012

In this study, we explored the dye molecules and dye-containing zeolite crystals using polarized photoluminescence. We used nanoporous zeolite-L crystal and two kinds of dye molecules, pyronine B (PyB^+) and pyronine Y (PyY^+), as the host and guest materials, respectively. To quantitatively monitor the arrangement of PyB^+ and PyY^+ molecules inserted into the channels of zeolite-L, the polarized fluorescence microscopy of dye-containing zeolite-L were investigated. Contrary to the case of randomly oriented dye molecules, the luminescence of dye-containing zeolite crystals has a polarization along the direction of oriented zeolite pore. Because the aggregation can be prevented by locking the dye molecules inside the pores of a zeolite crystal, the interaction among dye molecules can be suppressed.

2. Experimental

2.1. Materials

The cylindrical zeolite-L crystals was prepared according to the procedures described in our previous report [25]. Pyronin B ($PyB^+FeCl_4^-$), Pyronin Y (PyY^+Cl^-), (3-aminopropyl)trimethoxysilane (AP-TMS), (3-chloropropyl)trimethoxysilane (CP-TMS), dimethylsulfoxide (DMSO), and glycerol were purchased from Aldrich and used as received. Cover glasses (18×18 mm^{2-}) were purchased from Marienfeld and treated in a piranha solution (H_2SO_4: 30% H_2O_2 = 3:1) at 95–100 °C for 1 h to remove organic residues on the surface. Toluene was purchased from Junsei and distilled over sodium under argon prior to use.

2.2. Insertion of organic dye ions into the channels of zeolite-L crystals [11,12]

For the synthesis of PyB^+-containing zeolite-L crystals (PyB^+@L), PyB^+ ions were incorporated into the channels of zeolite-L crystals (3 g) by aqueous ion exchange of K^+ ions in zeolite-L with $PyB^+FeCl_4^-$ (180 mL, 0.1 mM) under the reflux condition for 90 min. After rigorous washing with distilled deionized water, the PyB^+-loaded zeolite-L crystals were further washed with DMSO by soxhlet extraction until no further $PyB^+FeCl_4^-$ was detected in the DMSO. PyY^+-containing zeolite-L crystals (PyY^+@L) were synthesized in the same way taking PyY^+Cl^- (180 mL, 0.1 mM) as organic dye ions instead of $PyB^+FeCl_4^-$.

2.3. Preparation of vertically aligned dye-containing zeolite-L monolayer

Clean cover glass plates were treated with the toluene solution of CP-TMS (a mixture of 40 mL of toluene and 1 mL of CP-TMS) at 120 °C for 3 h under argon. After cooling to room temperature, 3-chloropropyl-tethering glass (CP-g) plates were successively washed with copious amounts of freshly distilled toluene and ethanol. The washed CP-g plates were dried with a gentle stream of high purity nitrogen, and kept in desiccators. The vertically aligned dye-containing zeolite-L monolayer on CP-g (v-dye@L-g) was prepared by 'sonication with stacking between the glass plates (SS)' method according to the procedure describe in our previous report [26]. Typically, a comb-shaped Teflon support was placed at the bottom of the 50 mL round-bottomed flask charged with 40 mL of toluene and 50 mg of dye-containing zeolite-L crystals (dye@L). And, a CP-g plate was interposed between two clean bare glass (Bg) plates, and three sets of Bg/CP-g/Bg stack were inserted, stack by stack, into each gap of the comb-shaped Teflon support. The round-bottomed flask was placed directly above an ultrasound generator in the bath, and then the attachment reaction between zeolite-L microcrystals and CP-g was carried out for 2 min.

2.4. Measurements of anisotropic photoluminescence of v-dye@L-g [12]

Three v-dye@L-g plates were placed horizontally on a stack of five bare glass plates placed horizontally in a quartz container ($20 \times 20 \times 10$ mm^3) charged with glycerol as an index-matching fluid. The zero-order $\lambda/2$-waveplate was used to change the angle of linearly polarized incident light to vertical (V) or horizontal (H) direction, and a compensator to change the linearly polarized light from the $\lambda/2$-waveplate to circularly polarized light. A vertically, horizontally or circularly polarized beam (λ = 514.5 nm) generated from an Ar^+ ion laser (Coherent) was introduced into the stack of v-dye@L-g, and the produced fluorescent light was introduced into a spectrometer (Horiba Jobin Yvon, TRIAX 550) equipped with a charge-coupled device (CCD) array (2048×512) detector after passing through an achromatic $\lambda/2$-waveplate to rotate the polarization to 0° (horizontal) and a linear polarizer fixed at 0° (horizontal). Care was taken to ascertain that the detection system does not introduce any systematic anisotropy. This was checked by measuring dyes dissolved in glycerol with circularly polarized laser light and obtaining the ratio of vertically-polarized to horizontally-polarized luminescence intensity of 1.05.

3. Results and discussion

The cylindrical zeolite-L crystals with the sizes of 1 μm (diameter) × 3–4 μm (length) used in this report was prepared according to the procedures described in our previous report [25]. The main channels of zeolite-L consist of unit cells with a length of 7.5 Å in the c-direction, which they are joined by 12-membered ring windows with a diameter of 7.1 Å. The largest free diameter is about 12.6 Å, depending on the charge compensating cations. The dye molecules are positioned at sites along the channels [27]. The length required for the dye molecule inserted into the channels of zeolite-L crystal is equal to a number *S* times the length of primitive unit cell, which has a length of 7.5 A in the c-direction. As an example, a dye of 1.5 nm length requires two primitive unit cells in zeolite-L, and hence *S* is equal to 2. With this reason, the length of dye should be longer than 1.5 nm to be lined up along the direction of the channel in zeolite-L crystal as shown in Fig. 1.

In this studies, as guest molecules for zeolite-L host, two kinds of dye ions having a different molecular dimension, pyronin B (PyB^+) and pyronin Y (PyY^+), were incorporated into the channels of zeolite-L crystals by ion exchange method in water, respectively. And, the external surfaces of pyronin-containing zeolite-L crystals

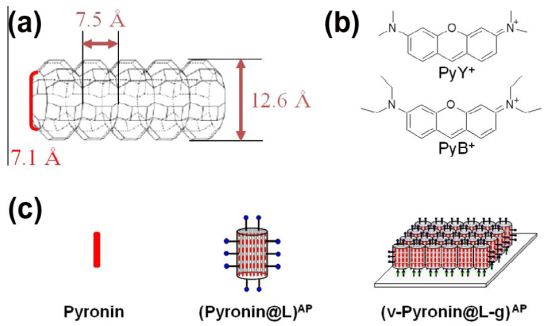


Fig. 1. Structures and dimensions of zeolite-L channel, PyY^+ and PyB^+. (a) The zeolite-L channel consists of 7.5 Å long lobes with the pore opening of 7.1 Å. The internal width of a lobe is 12.6 Å. (b) Molecular structure of PyY^+ and PyB^+ ion, and (c) illustrations of pyronin, aminopropyl-tethering pyronin@L, and aminopropyl-tethering v-pyronin@L-g, denoted as pyronin, $(pyronin@L)^{AP}$ and $(v\text{-}pyronin@L\text{-}g)^{AP}$, respectively.

Journal of the Korean Physical Society, Vol. 64, No. 2, January 2014, pp. 263∼267

High-yield Synthesis of Single-crystalline InSb Nanowires by Using Control of the Source Container

Yi-Seul Park, Eun Ji Kang and Jin Seok Lee*
Department of Chemistry, Sookmyung Women's University, Seoul 140-742, Korea

(Received 23 September 2013, in final form 28 October 2013)

Single-crystalline InSb nanowires were synthesized on a SiO_2wafer through the vapor-liquid-solid mechanism by using the chemical vapor deposition method. We also identified the effect of the source container system, closed- or open-boat, which contained SiO_2 wafers, on the InSb powder, and the single-crystalline InSb nanowires were found to have different growth mechanisms. The structural characterization of the InSb nanowires was performed using a scanning electron microscope. The crystallinity and the composition of the InSb nanowires were investigated using X-ray diffraction and energy dispersive X-ray spectroscopy. We demonstrated that the InSb nanowires were thick and long, when an open-boat system was used, because the open-boat system can carry a greater amount of the vapor-phase InSb precursor than the closed-boat system. A high yield of the InSb nanowires can be obtained using the open-boat system. Additionally, the diameter of the InSb nanowires was quite thick because of the increased the growth time as a result of the dominant vapor-solid mechanism.

PACS numbers: 61.46.Hk, 61.82.Fk, 61.50.-f
Keywords: Nanowire, InSb, Chemical Vapor Deposition, Vapor-Liquid-Solid, Vapor-Solid
DOI: 10.3938/jkps.64.263

I. INTRODUCTION

Indium antimonide (InSb) possesses many unique properties as a result of its outstanding electrical characteristics. For example, it has a narrow band gap of 0.17 eV [1], an effective mass of 0.0135 m_0 [2], a high electron mobility of 7.8×10^4 $cm^2V^{-1}S^{-1}$at room temperature [3], and a large Bohr exciton radius of 54 nm [4]. These unique properties make InSb suitable for a wide range of device applications [3], including high-speed field-effect transistors [5], infrared detectors [6], magnetic sensors, thermoelectric power generators, and cooling devices [7,8]. One-dimensional (1D) structures of InSb, namely InSb nanowires, are being actively investigated as a result of their excellent electronic and optical properties, for their size, the quantum confinement effect, their desirable vectorial transport properties, and their large surface area [9]. Chemical vapor deposition (CVD) is a very important method for in the synthesis of nanowires, because it can achieve single-crystal growth through the vapor-liquid-solid (VLS) mechanism. This growth on various substrates is characterized by high electrical properties [10–12].

In the general CVD method, the nanowire morphology can be controlled by changing experimental parameters such as the temperature, pressure and gas flow rate [13]. Based on the fact that the vapor densities of the precursors increase with increasing temperature and pressure in a CVD reactor, thick nanowires were mostly synthesized at high reaction temperatures and pressures. Also, considering that the transport velocity of the vapor precursors is governed by the gas flow rate, we expect a high gas flow rate to induce thick nanowires. Analogously, we can deduce that the yield and the morphology of the nanowires depend on the amounts of precursors transported into the metal catalyst. For this reason, we believe that the shape of source container also influences the amount of precursors transported in the reactor, resulting in different yields and morphologies for the nanowires.

In this study, we focused on the effect of the container's shape on the yield and the morphology of nanowires, including their growth mechanism. We synthesized a single-crystalline InSb nanowire on a SiO_2 wafer through the VLS mechanism [14] by using the CVD method. The amount of the vapor-phase InSb precursor was adjusted by using the shape of the source container, which can be either a closed- or open-boat system. The InSb nanowires exhibit different features depending on the type of source container, and we confirmed their growth by using the VLS and VS mechanisms [15]. Additionally, the diameter of the InSb nanowires was controlled by adjusting the reaction time.

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

II. EXPERIMENTAL

InSb(99.999%) powder was purchased from Alfa-Aesar. Gold nanoparticles, 10nm, were purchased from Ted Pella. Poly-L-lysine (PLL) solution was purchased from Sigma-Aldrich. SiO_2 wafers were purchased from LG Siltron.

SiO_2 wafers were cut into small pieces a specific size of 10 mm (width) × 10 mm (length). The SiO_2 substrate was dipped in a piranha solution for 1 h to eliminate organic compounds and was ultrasonically cleaned in distilled water for 1 min by using an ultrasound cleaning bath equipped with two 100-W ultrasound generators operating at 28 kHz (Saehan Ultrasonic SH-2100). The water remaining on the SiO_2 wafer was removed by using a N_2 gun. The PLL was dropped on the SiO_2wafer so that gold could be attached. After 3 min, the PLL was removed by using a N_2 gun. PLL and the gold nanoparticles have (+) charge and (−) charge, respectiely, so they have an electrostatic attraction. For this reason, gold nanoparticles can be attached to the SiO_2 wafer. Lastly, a solution of 10-nm gold nanoparticles was dropped on the SiO_2wafer as a catalyst. After 3 min, the gold nanoparticle solution was removed by using a N_2 gun.

The Scanning electron microscopy (SEM) images of the InSb nanowires were obtained using field-emission scanning electron microscopy (FE-SEM) (JEOL JSM-7600F) at an acceleration voltage of 15 kV, which was performed on the as-synthesized product on the substrates. The single-crystalline structure of the nanowires was analyzed by using a JEOL JSM-7600F with energy dispersive X-ray spectroscopy (EDS) mapping capabilities. The X-ray diffraction patterns for the identification of produced nanowires were obtained using a Rigaku diffractometer (D/MAX-1C) with a monochromatic beam of Cu Kα radiation.

III. RESULTS AND DISCUSSION

We synthesized single-crystalline InSb nanowires by using the CVD method, as shown in Fig. 1(a). The synthesis of InSb nanowires was carried out in a tube furnace under a reduced pressure and a hydrogen environment. InSb powders, the precursor for InSb crystal growth, was located at the center of the source container, and the SiO_2 substrate was loaded on the InSb powder in the source container. The source container was placed at the center of the heating zone in the tube furnace (Lindberg/Blue M, 12-in. length), which was equipped with a quartz tube 1-inch in diameter. The reactor tube was evacuated to a base pressure of approximately 0.1 Torr prior to the growth of the InSb nanowires and was flushed several times with high-purity H_2 gas (99.999%) in order to minimize oxygen contamination. The furnace was then heated to a growth temperature of 500 °C at

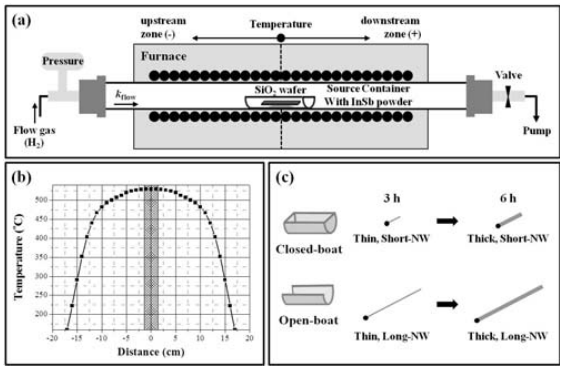


Fig. 1. (a) Schematic illustration of the CVD reactor for the growth of the InSb nanowires. (b) Measured furnace temperature at different locations. The gray color indicates the temperature at which the InSb nanowires were grown. (c) Schematic of the open-quartz boat and the closed-quartz boat filled with InSb powder. (k_{flow} is the flow rate)

20 °C/min and was maintained at 500 °C for 3 h or 6 h. H_2 was flowed at 100 sccm (standard cubic centimeters per minute), and the pressure was maintained at 350 Torr. The vapor-phase InSb precursors were transported by the H_2 carrier gas and condensed onto the SiO_2 substrate. At the end of the growth time, the reactor was slowly cooled to room temperature.

When the temperature of the tube furnace reached 500 °C the growth temperature, the solid-phase InSb powder was vaporized, and the vapor-phase InSb precursors were moved downstream from the center of the furnace (Fig. 1(a)) because of the flow direction of the H_2 gas. At that time, 10-nm gold nanoparticles on the SiO_2 wafer were melted and transformed to the liquid-phase [16]. The vapor-phase InSb precursors carried by H_2 carrier gas were melted in the liquid-phase gold nanoparticles. During the growth period, large amounts of the vapor-phase InSb precursors were accumulated by the liquid-phase gold nanoparticles until they reach a state of supersaturation. As a result, the InSb composition in the liquid-phase gold nanoparticles was precipitated as a solid-phase InSb material that had a one-dimensional structure, a nanowire. The VLS mechanism is the pathway of InSb precursors, which pass through the vapor phase, liquid phase, and solid phase.

Figure 1(b) displays the temperature profile of the tube furnace at a growth temperature of 500 °C. The x-axis and the y-axis indicate the location of the furnace whose zero-point is at the center of the furnace and the temperature at an individual location in the furnace, respectively. In the temperature profile, the temperature of the furnace is higher than other locations. The InSb nanowire was synthesized at approximately 529 – 530 °C, which is marked as the cross-checked pattern in the temperature profile. Figure 1(c) displays the shapes of source container, which are closed- or open-boat. The

Langmuir

Article

pubs.acs.org/Langmuir

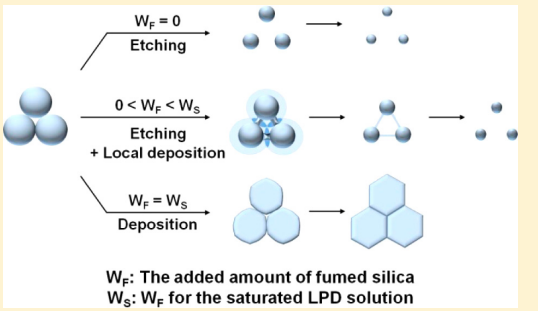
Local Liquid Phase Deposition of Silicon Dioxide on Hexagonally Close-Packed Silica Beads

Seo Young Yoon, Yi-Seul Park, and Jin Seok Lee*

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

Supporting Information

ABSTRACT: Liquid phase deposition (LPD) is a useful method for the production of oxide film with low reaction temperature and production cost. With the report that the LPD of oxide films is conformally processed with uniform thickness and composition, there has been significant attention given to investigating its kinetic controls and growth mechanism on the flat surface. In this work, we explored the LPD of silicon dioxide on the hexagonally close-packed silica beads array as a nanostructured surface. The deposition and etching reactions of SiO_2 occurred locally and simultaneously on silica beads, and were distinguished from the amount of fumed silica added in LPD solution. From locally competitive reactions, we obtained the anisotropic morphology of close-packed silica beads, and proposed a mechanism for the local LPD of SiO_2 driven by nanostructured surfaces. This work contributes highly to improve metal oxide-based engineering, and also provide greater insight into the topography-driven LPD.



INTRODUCTION

Liquid phase deposition (LPD) is a process of producing oxide films under aqueous conditions without the use of electrochemical methods, vacuum systems, or sensitive organometallic precursors.[1] This method has considerable advantages as compared to other conventional methods, such as thermal oxidation,[2] chemical vapor deposition (CVD),[3] and sputtering.[4] In particular, these conventional methods require high reaction temperatures of up to several hundred degrees Celsius, whereas LPD can be performed at room temperature with low production costs and environmental impacts.[1] This lower temperature deposition also does not influence device characteristics and wiring reliability because of decreased thermal stress, and thus it is useful on multilevel interconnection interlayer dielectrics.[5−7]

Silicon dioxide (SiO_2) thin films, which are the most common dielectric materials formed using LPD, are quite versatile and applicable to various fields, such as ultralarge-scale integration (ULSI) technologies,[2] fabrication of integrated circuits (ICs),[8] and interlayer dielectrics and gate oxides in transistors.[9,10] In addition, SiO_2 films could potentially be applied as one of the chief components of optical antireflection coatings and liquid crystal display (LCD) substrates as a mask to prevent alkali ions from diffusing to the indium tin oxide (ITO) films.[3]

Until now, LPD research has focused on the formation of the oxide film, particularly its thickness, density, and composition on the flat surface. The previous studies reported that the growth rate and mechanism of oxide films are strongly affected by reagent concentration and reaction temperature,[5] and their composition and properties are determined by the chemical reactions between reagents and organic functional groups on substrates.[11] Thus, the kinetic controls and growth mechanism of various oxide films were demonstrated on the flat surface by controlling the concentration of reactant or using fluoride scavengers such as boric acid or aluminum metal,[1,12] but the mechanism of the LPD process on the nanostructured surfaces has not yet been conducted.

With the report that the LPD on nanomaterials is conformally processed with uniform thickness and composition over the entire external surface,[13−17] there have been many efforts to form the dielectric oxide film on various nanomaterials for nanoelectronic applications. However, unlike LPD research for nanomaterial coating, there have been few reports investigating the formation of oxide film on nanostructured surfaces. Differing from flat substrates even including the confined surface of a nanomaterial, it is expected that the LPD process on nanostructured surfaces has locally different reaction conditions originating from their surface topography. Considering that this topography-driven LPD conditions can directly affect the conformality and morphology of an as-grown oxide film, it is necessary to understand the local LPD process occurring on nanostructured surfaces.

In this work, we explored the LPD of SiO_2 on a hexagonally close-packed silica beads array as a nanostructured surface. The different amount of fumed silica was used to control the

Received: October 20, 2014
Revised: December 10, 2014
Published: December 10, 2014

DOI: 10.1021/la5041536
Langmuir 2015, 31, 249−253

concentration of silicic acid ($Si(OH)_4$) in LPD solution, consisting of hydrofluorosilicic acid (H_2SiF_6) and water. The deposition and etching reactions of SiO_2 occurred locally and simultaneously on hexagonally close-packed silica beads, and were distinguished from the amount of fumed silica added in LPD solution. With the competition of deposition and etching reactions, we obtained the anisotropic morphology of close-packed silica beads, and proposed a mechanism for the local LPD of SiO_2 on nanostructured surfaces.

RESULTS AND DISCUSSION

To fabricate the nanostructured surface, we first synthesized silica beads with narrow size distribution using the Stöber method,[18,19] and organized them into a hexagonally close-packed monolayer using the Langmuir−Blodgett method (see Supporting Information Figure S1, and the Experimental Section).[20−22] On the basis of the fact that the formation of SiO_2 is governed by the concentration of $Si(OH)_4$, we added different amounts of fumed silica (SiO_2) into LPD solution to vary the concentration of $Si(OH)_4$, as described in eq 1:

$$H_2SiF_6 + SiO_2 + 6H_2O \rightleftarrows 2Si(OH)_4 + 6HF \quad (1)$$

For convenience, the added amount of fumed silica was denoted as W_F, and the amount of fumed silica required for the saturated LPD solution was denoted as W_S. As for the W_S, the average W_F was measured to be 5 g at 25 °C. When the LPD occurred on a silica beads array in the saturated LPD solution containing 5 g of added fumed silica ($W_F = W_S$), this induced morphology changes in their nanostructured surfaces (Figure 1). This phenomenon is attributed to the local deposition of

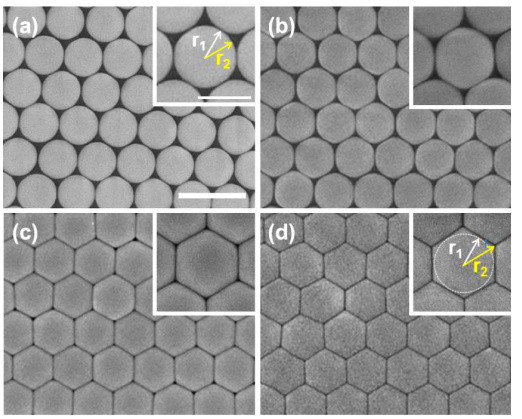


Figure 1. SEM images of the topography of silica beads array monitored over varying reaction times in LPD solution with 5 g of added fumed silica. The reaction time was (a) 0, (b) 0.5, (c) 2, and (d) 6 h, respectively. The deposition was directed toward 6-pinholes (r_2 direction), not toward adjacent silica beads (r_1 direction), and thus the shape of spherical silica beads changed to that of a hexagon. The scale bar is 1 μm, and that of the inset is 500 nm.

$Si(OH)_4$ caused by the dehydrating reaction with hydroxyl groups on the surface of hexagonally close-packed silica beads.[1]

In Figure 1a, r_1 and r_2 are the distances from the center of the silica bead to the adjacent silica bead and pinhole, respectively. Interestingly, the deposition was only directed toward pinholes (r_2) because the deposition toward the r_1 direction was limited by adjacent silica beads. Thus, r_1 was equal to the radius of the silica bead (r_0), and was constant. However, r_2 increased to $(2/\sqrt{3})r_0$ until the pinhole among the silica beads could not decrease further. Because the angle between r_1 and r_2 was 30° and each r_1 and r_2 existed at an interval of 60° for a hexagonally close-packed silica beads array, we obtained an anisotropic morphology for close-packed silica beads. As the reaction progressed, the pinhole size gradually decreased to being barely visible, and the shape of spherical silica beads became angular and ultimately became hexagonal after 6 h through the local LPD process on the nanostructured surface (Figure 1d).

In the close-packed silica beads array, the surface roughness of silica beads and average height between the peak and valley at the silica beads array were 13.25 and 197.80 nm, respectively (Figure 2a). After LPD occurred in the LPD solution with 5 g

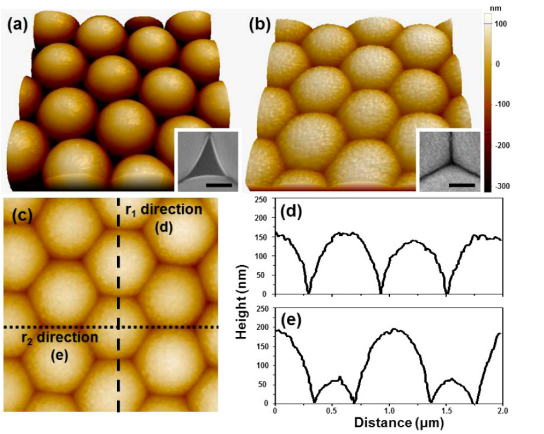


Figure 2. AFM 3D images of the surface topography of silica beads array (a) before and (b) after the LPD process in LPD solution with 5 g of added fumed silica for 6 h. The inset shows a SEM image of the pinholes among the silica beads. The scale bar is 100 nm. (c) AFM 2D image of the surface topography of silica beads array after the LPD process for 6 h. Parts (d) and (e) are scan profiles in (c) along the r_1 and r_2 directions, respectively.

of added fumed silica ($W_F = W_S$) for 6 h, the pinholes among the silica beads became filled, and the surface roughness and average height changed to 13.73 and 149.18 nm, respectively (Figure 2b). The time-dependent AFM data of surface roughness and average height are shown in Supporting Information Figure S2. Although the deposition reaction was followed by absorption of $Si(OH)_4$ on the surface of silica beads, the surface roughness of silica beads remained nearly constant during the LPD process for 6 h, providing strong evidence for good coating conformality. However, this deposition process induced a decrease in the average height between the peak and valley at the silica beads array. Thus, the topography curve of silica beads became fluent curves after the LPD process (Supporting Information Figure S2).

To investigate the topography dependence of the local LPD process on hexagonally close-packed silica beads array, we measured line scanned profiles in the AFM 2D image obtained after LPD process for 6 h (Figure 2c) along the r_1 and r_2 directions, respectively. Figure 2d showed the scan profile along the r_1 direction, and demonstrated that the surface topography of silica beads was retained after LPD. However, seeing the scan profile along the r_2 direction, there were large and small humps

DOI: 10.1021/la5041536
Langmuir 2015, 31, 249−253

J Nanopart Res (2014) 16:2226
DOI 10.1007/s11051-013-2226-9

RESEARCH PAPER

Synthesis of single-crystalline topological insulator Bi_2Se_3 nanomaterials with various morphologies

Yi-Seul Park · Jin Seok Lee

Received: 2 August 2013 / Accepted: 19 December 2013 / Published online: 4 January 2014

Abstract We report the synthesis and characterization of layer-structure Bi_2Se_3 nanomaterials. Bi_2Se_3 nanomaterial has attracted many researchers, because it has a unique three-dimensional topological insulator property characterized by a metallic surface which coexists with an insulating interior. This can be achieved by having large surface-to-volume ratio in the nanomaterial. We synthesized highly single-crystalline topological insulator Bi_2Se_3 nanomaterials with various morphologies, including straight nanowires, zig-zag nanowires, and nanobelts, by adjusting experimental parameters such as the growth temperature, pressure, and carrier gas flow rate. The results show that the width and length of Bi_2Se_3 nanowires increase significantly with increasing values of each parameter. Furthermore, we studied the growth mechanism of individual morphologies based on the layered structure of Bi_2Se_3.

Keywords Layered-structure · Bi_2Se_3 nanomaterial · Topological insulator · Various morphology · Growth mechanism · Spintronics

Y.-S. Park · J. S. Lee (✉)
Department of Chemistry, Sookmyung Women's University, Seoul 140-742, Korea
e-mail: jinslee@sookmyung.ac.kr

Introduction

Recently, research on Bi_2Se_3 and related compounds (Bi_2Te_3 and Sb_2Te_3) has attracted much interest, because they are predicted to be three-dimensional (3D) topological insulators (TIs) (Fu et al. 2007; Moore and Balents 2007; Qi et al. 2008), showing unusual phases of quantum matter with an insulating bulk gap and gapless edges or surface states (Zhang et al. 2009). The inevitable loss and mismatch of periodic lattice structure in topological insulators could encourage atomic reconstruction and dangling bonds. These partial changes lead to an absence of surface states in the bulk properties (Peng et al. 2010). These distinct states in topological insulators result in different electronic properties at their insulating bulk component and the surface protected on their own boundary. Despite the extensive transport experiments on bulk Bi_2Se_3 since the 1970s, there has been no report on the conducting surface state, and the predicted topological features have not been addressed. Because these materials are strongly affected by the contribution of the bulk carriers, it was very hard to distinguish the conductive properties originating from the surface states and those from the insulating bulk.

Interestingly, it was reported that the surface conduction of a topological insulator can be enhanced in the nanostructured-material due to their larger surface-to-volume ratio than the bulk phase. Decreasing their size can suppress the contribution of the bulk

carriers, and the contribution of surface states, which could affect the electrical and optical properties, is enhanced. So, it is essential to control the morphology of Bi_2Se_3 nanomaterials. In previous research, various morphologies of Bi_2Se_3 nanomaterials have been synthesized, such as Bi_2Se_3 nanoplates by solvothermal method (Wang et al. 2003a), and Bi_2Se_3 nanobelts by sonochemical method (Cui et al. 2004). However, these solution-phase synthetic methods cause surface contamination due to organic residues or low crystallinity, which affect the surface states of the topological insulator. Alternatively, vapor-phase synthetic methods by vapor–liquid–solid (VLS) mechanism have been shown to be effective for synthesizing high-crystallinity semiconductor nanomaterials (Gao and Wang 2002; Lieber 2003; Yang 2005).

We tried to synthesize highly single-crystalline Bi_2Se_3 nanomaterials using chemical vapor deposition (CVD) followed by characterization. With different reaction conditions regarding temperature, pressure, and gas flow rates, the nanomaterials are synthesized with various morphologies and widths, including thick nanowires, zig-zag nanowires, and nanobelts, from thin nanowires through a gold-catalyzed vapor–liquid–solid (VLS) growth process.

Experimental

The highly single-crystalline Bi_2Se_3 nanomaterials were synthesized by CVD in a 12-in. horizontal tube furnace (Lindberg/Blue M) equipped with a 1-in.-diameter quartz tube (Fig. 1A). The source material, 0.2 g of bismuth selenide (Bi_2Se_3) powder (99.999 % from Alfa Aesar), was placed at the center of the heating zone. The Bi_2Se_3 nanomaterials were grown on Si(100) substrates with a native oxide functionalized by immersion in poly-L-lysine (0.1 % w/v aqueous solution from Ted Pella) for 3 min, and then in gold colloid solution (10 nm in diameter from Ted Pella). The negatively charged Au nanoparticles adhere to the positively charged poly-L-lysine on the Si(100) substrate. The substrate was placed in the downstream zone of the furnace (7 cm from the center), which is the expected location to accurately set the temperature. The temperature profile of the tube furnace is important for controlling the morphologies of Bi_2Se_3 nanomaterials, and was measured at different set temperatures (Fig. 1B). Ar gas (high purity, 99.999 %) was used as a carrier gas to transport the precursor vapor into the reaction region of the quartz tube. Before the synthetic process, the quartz tube of the furnace was evacuated to below 1 Torr and flushed with the carrier gas repeatedly to minimize oxygen contamination. The furnace was then heated to the reaction temperature of 530 °C at a rate of 20 °C/min, where it was maintained for 1.5 h. During this period, the carrier gas was flowed at 50 sccm, and the pressure was maintained at 50 Torr. The vapor-phase precursors were transported by the Ar carrier gas and deposited onto the substrate as various nanostructures depending on the reaction conditions. The temperature, pressure, and Ar gas flow rate for different nanomaterial morphologies are listed in Fig. 1C. After completing the reaction, the quartz tube was cooled, and then the substrate was removed from the furnace. The substrate had a twinkling gray coating, which indicated the existence of Bi_2Se_3 nanomaterials on the substrate.

The surface morphologies of the synthesized Bi_2Se_3 nanomaterials were characterized using a Hitachi S-4300 field emission scanning electron microscope (FE-SEM) at an acceleration voltage of 20 kV, which was performed on the as-synthesized product on the substrate. A platinum/palladium alloy (in the ratio of 8–2) was deposited with a thickness of about 15 nm on top of the sample. We obtained X-ray diffraction (XRD) patterns of the Bi_2Se_3 material using a Rigaku diffractometer (D/MAX-1C) with a monochromatic beam of Cu Kα radiation. Then, the single-crystalline structure of Bi_2Se_3 nanomaterials was analyzed with a JEOL 2010 transmission electron microscope (TEM) with energy-dispersive X-ray spectroscopy (EDX) for chemical analysis, as shown in Fig. 2. The samples for TEM imaging were made by depositing a hexane solution of nanomaterials prepared by sonication of the as-synthesized substrate in hexane onto holey carbon 300 mesh copper grids (Structure Probe, Inc.).

Results and discussion

Bi_2Se_3 has a layered structure in which each charge-neutral layer consists of a quintuple layer of five covalently bonded atomic planes, Se–Bi–Se–Bi–Se (Fig. 1C). The bonding in the Bi_2Se_3 material is strongly covalent within the layer, but has a weak van

THE JOURNAL OF PHYSICAL CHEMISTRY C

Article

pubs.acs.org/JPCC

Effects of Hydrogen Partial Pressure in the Annealing Process on Graphene Growth

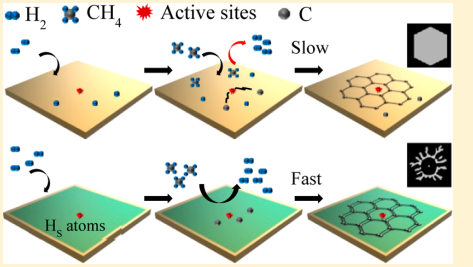
Da Hee Jung,[†] Cheong Kang,[†] Minjung Kim,[‡] Hyeonsik Cheong,[‡] Hangil Lee,*[,†] and Jin Seok Lee*[,†]

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

[‡]Department of Physics, Sogang University, Seoul 121-742, Republic of Korea

Supporting Information

ABSTRACT: Graphene domains with different sizes and densities were successfully grown on Cu foils with use of a chemical vapor deposition method. We investigated the effects of volume ratios of argon to hydrogen during the annealing process on graphene growth, especially as a function of hydrogen partial pressure. The mean size and density of graphene domains increased with an increase in hydrogen partial pressure during the annealing time. In addition, we found that annealing with use of only hydrogen gas resulted in snowflake-shaped carbon aggregates. Energy-dispersive X-ray spectroscopy (EDX) and high-resolution photoemission spectroscopy (HRPES) revealed that the snowflake-shaped carbon aggregates have stacked sp^2 carbon configuration. With these observations, we demonstrate the key reaction details for each growth process and a proposed growth mechanism as a function of the partial pressure of H_2 during the annealing process.



INTRODUCTION

Graphene is a single-atom-thick planar sheet of sp^2-bonded carbon atoms densely packed into a honeycomb crystal lattice. It has attracted great interest recently, mostly due to its unusual physical properties.[1] It has an extremely high electron mobility and a tunable band gap, making graphene potentially useful for innovative approaches to electronics.[2−5] In the efforts to industrialize graphene film as a carbon nanomaterial, a number of recent attempts have been made to improve its large-scale production by mechanical exfoliation of highly orientated polymeric graphite (HOPG), thermal treatment of SiC surface, solution-phase synthesis, and chemical vapor deposition (CVD).[6−11] Of the above approaches, the CVD method is known to be a powerful and reproducible technique for the large-scale production of high-quality graphene films with controlled thicknesses.[7] Recently, CVD growth of graphene on metal catalysts such as Ir,[12] Ru,[13] Ni,[14−16] and Cu[17] foils has been reported. In particular, uniform single-layer deposition of graphene on Cu foil with low-carbon solubility has allowed the production of high-quality graphene for industrial applications.[17]

In spite of the complexity of CVD procedures involving different catalysts, different carbon sources, volume ratios of the gas mixture, and other variables, the physical principles underlying this method mainly rely either on a surface catalytic reaction[13,17] for catalysts with low-carbon solubility or on bulk carbon precipitation onto the surface during cooling[18,19] for catalysts with high-carbon solubility. In both cases, graphene nucleation on the catalyst surface is a critical step in the growth process. Various factors affect the initiation of the graphene nucleation process, including the surface microstructure of the metal catalyst,[13,18] the carbon source,[19] carbon segregation from metal−carbon melts,[20] and the reaction parameters used during CVD growth.[21−23] Recently, it was reported that graphene growth is strongly dependent on the hydrogen (H_2) contribution during growth time, where it seems to act both as an activator for the formation of surface-bound carbon and as an etching reagent affecting the morphology of the graphene domains.[24] Thus, the growth rate and shape of graphene domains rely on the partial pressure of H_2 mixed with Ar as the buffer gas.[24] Although the role of H_2 during growth time was identified, its role during annealing was not stated in detail and remained unclear.

In this research, we focused on understanding the role of H_2 during the annealing process as well as the growth process in order to control the growth rate and shape of graphene domains. We chose atmospheric-pressure CVD, as it was technologically more accessible for the desired production of graphene. To elucidate the role of H_2 in the mixture gas, we synthesized graphene domains on a Cu foil with various volume ratios of H_2 to Ar gas and carefully investigated the size, morphology, and composition of the as-produced graphene domains using scanning electron microscopy (SEM), micro-Raman spectroscopy, energy-dispersive X-ray spectroscopy (EDX) analysis, and high-resolution photoemission spectroscopy (HRPES).[25] Based on our observation and characterization of as-produced graphene domains, we present the key reaction details for each growth process and a proposed growth

Received: November 7, 2013
Revised: January 20, 2014
Published: January 29, 2014

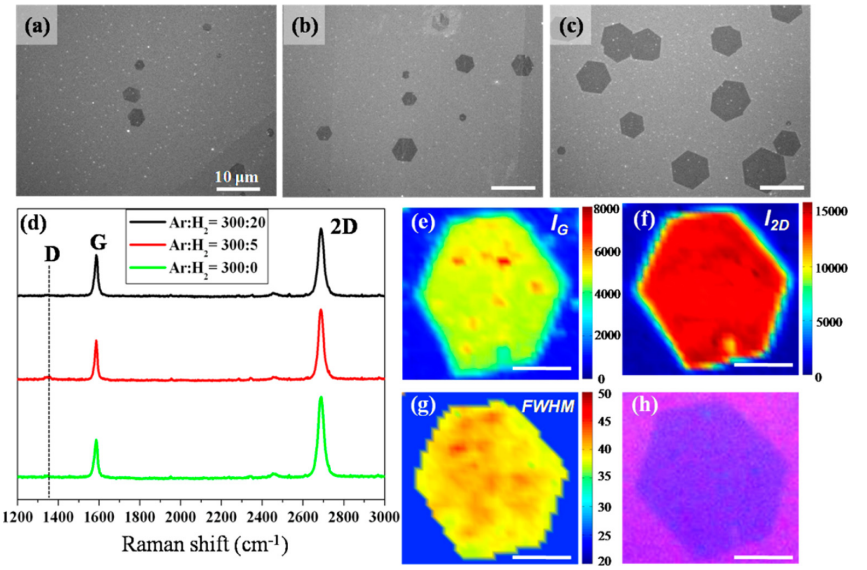


Figure 1. SEM images of graphene domains synthesized with Ar flow fixed at 300 sccm and H_2 flow varied in the annealing process: (a) H_2 = 0, (b) 5, and (c) 20 sccm. Scale bar =10 μm. (d) Raman spectra corresponding to domains shown in panels a, b, and c. (e, f) Intensity mappings of the G and 2D peaks, respectively. (g) Image mapping of the full width at half-maximum (fwhm) of the 2D peak. (h) Optical image of the graphene sample on a SiO_2/Si substrate. Scale bar = 5 μm.

mechanism as a function of the partial pressure of H_2 during the annealing process.

EXPERIMENTAL SECTION

Graphene synthesis was done by using the chemical vapor deposition method. Cu foil (25 μm thick, 99.999%, Alfa Aesar) was first cleaned with acetone, isopropyl alcohol, and deionized (DI) water. Then, the foils were put into a 1 in. quartz tube inside a horizontal furnace (Lindberg/Blue M, Thermo Scientific). The system was evacuated for 10 min. After this, the pump was shut down and the growth chamber was brought back to atmospheric pressure by introducing an argon and hydrogen mixture. This process was repeated several times. The sample was then heated to 1050 °C at a ramping rate of 17.5 deg/min with the desired ratio of argon to hydrogen. For the initial Cu cleaning and annealing, the chamber was maintained at 1050 °C for another 1 h with the desired gas composition. During the annealing process, in order to clarify the effect of hydrogen, Ar flow rate was fixed to 300 standard cubic centimeters per minute (sccm) and H_2 flow was set to 0, 10, and 15 sccm, respectively. In addition, a different Ar flow was used to identify the role of Ar flow in the annealing process. For this experiment, H_2 flow rate was fixed to 20 sccm and Ar flow was set to 100 and 0 sccm. Then, Ar was shut off and 300 sccm of the diluted methane (50 ppm in Ar) was allowed to flow into the tube for the graphene growth with 20 sccm of H_2. The growth time was set to 10 min. Finally, the sample was cooled to room temperature under the H_2/Ar atmosphere but without CH_4 supply. For the study of the effect of H_2 and Ar in the annealing process, we always keep the same growth condition and cooling rates to minimize the effects from the change of them.

The scanning electron microscope (SEM) images of graphene flakes were obtained from a FE-SEM (JEOL 7600F) at an acceleration voltage of 10 to 20 kV, which was performed on the as-synthesized product on substrates. We obtained the Raman spectrum of graphene samples with a homemade micro-Raman spectroscopy system. In micro-Raman spectroscopy, the 514.5 nm line of an Ar ion laser was used as the excitation source with a power of ~1 mW. The heating effect can be neglected at this power range. The laser beam was focused onto the graphene sample by a 50× microscope objective lens (0.8 Numerical Aperture). The collected scattered light was dispersed by a Shamrock SR 303i spectrometer (1200 grooves/mm) and was detected with a CCD detector. EDX spectroscopy was used to perform the elemental analysis of the composites. HRPES experiments were performed at the 8A2 beamline at the Pohang Accelerator Laboratory (PAL), which is equipped with an electron analyzer (SES100, Gamma Data Scienta). The Si 2p, C 1s, and O 1s core level spectra were obtained with use of photon energies of 150, 340, and 600 eV, respectively. Secondary electron emission spectra (−20 V sample bias) 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 4f 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 by using a standard nonlinear least-squares fitting procedure with Voigt functions.[25]

RESULTS AND DISCUSSION

We conducted the atmospheric-pressure CVD growth of graphene domains on Cu foil using various volume ratios of H_2 and Ar during annealing to investigate the influence of the partial pressure of H_2 on the growth rate and shape of the graphene domains. We fixed several factors, such as the flow rate (300 sccm) of Ar for the annealing process and the gas flow rate ratios (300:20 sccm) of diluted CH_4 (50 ppm in Ar) and H_2 for the growth process, respectively, changing only the

수학과

COMMUNICATIONS IN
ANALYSIS AND GEOMETRY
Volume 22, Number 5, 857–879, 2014

Isoperimetric inequalities for soap-film-like surfaces spanning nonclosed curves

Keomkyo Seo

For a soap-film-like surface Σ spanning a nonclosed curve Γ in $\mathbf{R}^n$, it is proved that

$$4\pi \mathrm{Area}(\Sigma) \le \mathrm{Length}(\Gamma)^2.$$

In the upper hemisphere $\mathbf{S}^n_+$ we use a mixed area $M_p(\Sigma)$, which was introduced by Choe and Gulliver [9] to prove an isoperimetric inequality

$$4\pi M_p(\Sigma) \le \mathrm{Length}(\Gamma)^2$$

for a soap-film-like surface Σ with nonclosed boundary $\partial\Sigma$. In a complete simply connected Riemannian manifold with sectional curvature bounded above by a nonpositive constant K, a soap-film-like surface Σ with an embedded, connected, and nonclosed boundary curve Γ satisfies

$$4\pi \mathrm{Area}(\Sigma) - K\mathrm{Area}(\Sigma)^2 \le \mathrm{Length}(\Gamma)^2.$$

Moreover, for a soap-film-like surface Σ with nonclosed boundary $\partial\Sigma$ in a complete simply connected Riemannian manifold M with sectional curvature bounded above by a constant K, we obtain a weak isoperimetric inequality

$$2\pi \mathrm{Area}(\Sigma) - K\mathrm{Area}(\Sigma)^2 \le \mathrm{Length}(\partial\Sigma)^2.$$

Finally, we prove that a soap-film-like surface $\Sigma \subset \mathbf{R}^n$ satisfies

$$2\sqrt{2}\pi \mathrm{Area}(\Sigma) \le \mathrm{Length}(\partial\Sigma)^2.$$

1. Introduction

It has been a long-standing conjecture that the classical isoperimetric inequality

$$4\pi \mathrm{Area}(\Sigma) \le \mathrm{Length}(\partial\Sigma)^2 \tag{1.1}$$

should hold for any compact smooth minimal surface Σ in $\mathbf{R}^n$. In 1921, it was partially proved by Carleman [5] for simply connected minimal surfaces in $\mathbf{R}^n$. Osserman and Schiffer [16] showed that the isoperimetric inequality (1.1) still holds with a strict inequality for doubly connected minimal surfaces in $\mathbf{R}^3$, using the Weierstrass representation formula. Later Feinberg [12] generalized this to doubly connected minimal surfaces in $\mathbf{R}^n$ for all n. In 1984, Li *et al.* [15] proved the inequality (1.1) for minimal surfaces in $\mathbf{R}^3$ with one or two boundary components, introducing the concept of weakly connected boundary. Choe [6] later showed that the inequality (1.1) holds also for any minimal surfaces in $\mathbf{R}^n$ with one or two boundary components. However, this conjecture is still open for minimal surfaces with at least three boundary components.

On the other hand, one might propose the same conjecture for soap-film-like surfaces with singularities. Almgren [1] showed that the conjecture is true for area minimizing flat chains mod k, using geometric measure theory. In [7], Choe proved that the inequality (1.1) holds also for two-dimensional stationary varifolds with connected boundary of multiplicity ≥ 1. However, until now, it was unknown whether soap-film-like surfaces with nonclosed boundary satisfy this isoperimetric inequality. Physically such surfaces may be observed in soap-film experiments, where the nonclosed boundary curve is connected by a singular curve in the interior of surface and three surfaces meet along the singular curve. The mathematical existence of such soap-film-like surfaces was studied by Parks [17] and Drachman and White [10]. In this paper, we study isoperimetric inequalities for soap-film-like surfaces whose boundary contains a nonclosed curve.

It turns out that for a soap-film-like surface Σ spanning a nonclosed curve Γ in $\mathbf{R}^n$ (Theorem 3.4),

$$4\pi \mathrm{Area}(\Sigma) \leq \mathrm{Length}(\Gamma)^2. \tag{1.2}$$

We emphasize that the above inequality does not contain the length of singular curves of Σ. The key idea of the proof of this inequality is the following: firstly, we obtain the area and density comparison between Σ and $p \gg \Gamma$ for a point $p \in \Sigma$. Secondly, we prove the isoperimetric inequality for the cone $p \gg \Gamma$, where $p \gg \Gamma$ is the union of the line segments from p to q over all $q \in \Gamma$. Note that we cannot apply Stokes' theorem to Σ because Σ is singular, hence it is nonorientable in general. For that reason, we use a classical minimal surface $\widetilde{\Sigma}$ which is obtained by cutting the soap-film-like surface Σ along the singular curves, hence the boundary of $\widetilde{\Sigma}$ is the union of the boundary

PACIFIC JOURNAL OF MATHEMATICS
Vol. 268, No. 1, 2014
dx.doi.org/10.2140/pjm.2014.268.205

L^p HARMONIC 1-FORMS AND FIRST EIGENVALUE OF A STABLE MINIMAL HYPERSURFACE

KEOMKYO SEO

We estimate the bottom of the spectrum of the Laplace operator on a stable minimal hypersurface in a negatively curved manifold. We also derive various vanishing theorems for L^p harmonic 1-forms on minimal hypersurfaces in terms of the bottom of the spectrum of the Laplace operator. As consequences, the corresponding Liouville type theorems for harmonic functions with finite L^p energy on minimal hypersurfaces in a Riemannian manifold are obtained.

1. Introduction

Hodge theory plays an important role in the topology of compact Riemannian manifolds. Unfortunately, the Hodge theory does not work anymore in noncompact manifolds. However, the L^2-Hodge theory works well in noncompact cases [Anderson 1988; Dodziuk 1982]. In this direction, there are various results for L^2 harmonic 1-forms on stable minimal hypersurfaces. Recall that a minimal hypersurface in a Riemannian manifold is called *stable* if the second variation of its volume is always nonnegative for any normal variation with compact support. More precisely, an n-dimensional minimal hypersurface M in a Riemannian manifold N is called *stable* if it holds that, for any compactly supported Lipschitz function f on M,

$$\int_M |\nabla f|^2 - (|A|^2 + \overline{\mathrm{Ric}}(\nu, \nu)) f^2 \, dv \geq 0,$$

where ν is the unit normal vector of M, $\overline{\mathrm{Ric}}(\nu, \nu)$ denotes the Ricci curvature of N in the ν direction, $|A|^2$ is the square length of the second fundamental form A, and dv is the volume form for the induced metric on M.

Using the nonexistence of L^2 harmonic 1-forms, Palmer [1991] proved that if there exists a codimension-one cycle on a complete minimal hypersurface M in Euclidean space, which does not separate M, M is unstable. Using Bochner's

This research was supported by the Sookmyung Women's University Research Grants (1-1303-0116).
MSC2010: primary 53C42; secondary 58C40.
Keywords: minimal hypersurface, stability, first eigenvalue, L^p harmonic 1-form, Liouville type theorem.

vanishing technique, Miyaoka [1993] showed that a complete noncompact stable minimal hypersurface in a nonnegatively curved manifold has no nontrivial L^2 harmonic 1-forms. Pigola, Rigoli, and Setti [Pigola et al. 2005] gave general Liouville type results and the corresponding vanishing theorems on the L^2 cohomology of stable minimal hypersurfaces. Refer to [Carron 2002; Pigola et al. 2008] for a survey in this area. While the L^2 theory is quite well understood, in the case $p \neq 2$, the L^p theory is less developed. See [Scott 1995] for general L^p theory of differential forms on a manifold.

The purpose of this paper is twofold. Firstly, we estimate the smallest spectral value of the Laplace operator on a complete noncompact stable minimal hypersurface in a Riemannian manifold under the assumption on L^p norm of the second fundamental form. Secondly, we obtain various vanishing theorems for L^p harmonic 1-forms on minimal hypersurfaces.

Let M be a complete noncompact Riemannian manifold and let Ω be a compact domain in M. Let $\lambda_1(\Omega) > 0$ denote the first eigenvalue of the Dirichlet boundary value problem

$$\begin{cases} \Delta f + \lambda f = 0 & \text{in } \Omega, \\ f = 0 & \text{on } \partial\Omega, \end{cases}$$

where Δ denotes the Laplace operator on M. Then the first eigenvalue $\lambda_1(M)$ is defined by

$$\lambda_1(M) = \inf_{\Omega} \lambda_1(\Omega),$$

where the infimum is taken over all compact domains in M. Cheung and Leung [2001] gave the first eigenvalue estimate for an n-dimensional complete noncompact submanifold M with the norm of its mean curvature vector bounded in the hyperbolic space. In particular, they proved that if M is minimal, the first eigenvalue $\lambda_1(M)$ satisfies

$$\tfrac{1}{4}(n-1)^2 \leq \lambda_1(M).$$

Note that this inequality is sharp because equality holds if M is totally geodesic [McKean 1970]. This result was extended to an n-dimensional complete noncompact submanifold with the norm of its mean curvature vector bounded in a complete simply connected Riemannian manifold with sectional curvature bounded above by a negative constant. More precisely, we have the following theorem.

Theorem [Bessa and Montenegro 2003; Seo 2012]. *Let N be an n-dimensional complete simply connected Riemannian manifold with sectional curvature K_N satisfying $K_N \leq -a^2 < 0$ for a positive constant $a > 0$. Let M be a an m-dimensional complete noncompact submanifold with bounded mean curvature vector H in N satisfying $|H| \leq b < (m-1)a$. Then*

$$\tfrac{1}{4}[(m-1)a - b]^2 \leq \lambda_1(M). \tag{1}$$

J. reine angew. Math. **694** (2014), 203–214
DOI 10.1515/crelle-2012-0119

Journal für die reine und angewandte Mathematik

Optimal isoperimetric inequalities for complete proper minimal submanifolds in hyperbolic space

By *Sung-Hong Min* at Seoul and *Keomkyo Seo* at Seoul

Abstract. Let Σ be a k-dimensional complete proper minimal submanifold in the Poincaré ball model B^n of hyperbolic geometry. If we consider Σ as a subset of the unit ball B^n in Euclidean space, we can measure the Euclidean volumes of the given minimal submanifold Σ and the ideal boundary $\partial_\infty\Sigma$, say $\mathrm{Vol}_{\mathbb{R}}(\Sigma)$ and $\mathrm{Vol}_{\mathbb{R}}(\partial_\infty\Sigma)$, respectively. Using this concept, we prove an optimal linear isoperimetric inequality. We also prove that if $\mathrm{Vol}_{\mathbb{R}}(\partial_\infty\Sigma) \geq \mathrm{Vol}_{\mathbb{R}}(\mathbb{S}^{k-1})$, then Σ satisfies the classical isoperimetric inequality. By proving the monotonicity theorem for such Σ, we further obtain a sharp lower bound for the Euclidean volume $\mathrm{Vol}_{\mathbb{R}}(\Sigma)$, which is an extension of Fraser–Schoen and Brendle's recent results to hyperbolic space. Moreover we introduce the Möbius volume of Σ in B^n to prove an isoperimetric inequality via the Möbius volume for Σ.

1. Introduction

Let $\Sigma \subset \mathbb{R}^k$ be a domain with smooth boundary $\partial\Sigma$. Then the classical isoperimetric inequality says that

$$k^k \omega_k \,\mathrm{Vol}(\Sigma)^{k-1} \leq \mathrm{Vol}(\partial\Sigma)^k, \tag{1}$$

where equality holds if and only if Σ is a ball in $\mathbb{R}^k$. Here $\mathrm{Vol}(\Sigma)$ and $\mathrm{Vol}(\partial\Sigma)$ denote, respectively, the k and $(k-1)$-dimensional Hausdorff measures, and ω_k is the volume of the k-dimensional unit ball B^k. As an extension of this classical isoperimetric inequality for domains in Euclidean space, it is conjectured that inequality (1) holds for any k-dimensional compact minimal submanifold Σ of $\mathbb{R}^n$. In 1921, Carleman [6] gave the first partial proof of the conjecture. He proved the isoperimetric inequality for simply connected minimal surfaces in $\mathbb{R}^n$ by using complex function theory. Osserman and Schiffer [23] proved in 1975 that the isoperimetric inequality holds for doubly connected minimal surfaces in $\mathbb{R}^3$. Two years later Feinberg [13] generalized to doubly connected minimal surfaces in $\mathbb{R}^n$ for any dimension n. In 1984, Li, Schoen, and Yau [16] proved that any minimal surface in $\mathbb{R}^3$ with two boundary components (not necessarily doubly connected) satisfies the isoperimetric inequality. Later, Choe [8] extended this inequality to any minimal surface in $\mathbb{R}^n$ with two boundary components. For higher-dimensional minimal submanifolds, the conjecture was proved only for area-minimizing cases, which was due to Almgren [2].

In hyperbolic space, there is also a sharp isoperimetric inequality for minimal surfaces. In 1992, Choe and Gulliver [10] showed that any minimal surface Σ with two boundary components in hyperbolic space $\mathbb{H}^n$ satisfies the sharp isoperimetric inequality

$$4\pi \operatorname{Area}(\Sigma) \leq \operatorname{Length}(\partial\Sigma)^2 - \operatorname{Area}(\Sigma)^2 \tag{2}$$

with equality if and only if Σ is a geodesic ball in a totally geodesic 2-plane in $\mathbb{H}^n$. However, there are a few results for higher-dimensional minimal submanifolds in hyperbolic space. Yau [28], Choe and Gulliver [9] obtained that if Σ is a domain in $\mathbb{H}^k$ or a k-dimensional minimal submanifold of $\mathbb{H}^n$, then it satisfies the following linear isoperimetric inequality:

$$(k-1)\operatorname{Vol}(\Sigma) \leq \operatorname{Vol}(\partial\Sigma).$$

See [1, 7, 15, 20, 25, 27] for isoperimetric inequalities involving mean curvature in the more general setting of arbitrary submanifolds.

In this paper we obtain optimal isoperimetric inequalities for complete proper minimal submanifolds in hyperbolic space $\mathbb{H}^n$. Recall that a submanifold in $\mathbb{H}^n$ is *proper* if the intersection with any compact set in $\mathbb{H}^n$ is always compact. The existence of complete minimal submanifolds in hyperbolic space was established in [3, 4, 18, 19]. Throughout this paper, $\mathbb{H}^n$ will denote the hyperbolic n-space of constant curvature -1. Among several models of $\mathbb{H}^n$, we will identify $\mathbb{H}^n$ with the unit ball B^n in $\mathbb{R}^n$ using the Poincaré ball model. If $ds^2_{\mathbb{H}}$ denotes the hyperbolic metric on B^n, then we have the following conformal equivalence of $\mathbb{H}^n$ with $\mathbb{R}^n$:

$$ds^2_{\mathbb{H}} = \frac{4}{(1-r^2)^2} ds^2_{\mathbb{R}},$$

where $ds^2_{\mathbb{R}}$ is the Euclidean metric on B^n and r is the Euclidean distance from the origin. We recall that any k-dimensional totally geodesic submanifold in $\mathbb{H}^n$ is a domain on a k-dimensional Euclidean sphere meeting ∂B^n orthogonally. The unit sphere $\mathbb{S}^{n-1} = \partial B^n$ is called the *sphere at infinity* and denoted by $\partial_\infty \mathbb{H}^n$.

In Section 2, we study the isoperimetric inequality for a k-dimensional complete proper minimal submanifold Σ in the Poincaré ball model B^n of hyperbolic space $\mathbb{H}^n$. Obviously, the k-dimensional hyperbolic volume of Σ is infinite. However, if we consider Σ as a subset of the unit ball B^n in Euclidean space, we can measure the k-dimensional Euclidean volume of Σ, which is denoted by $\operatorname{Vol}_{\mathbb{R}}(\Sigma)$. Similarly, we measure the $(k-1)$-dimensional Euclidean volume of the ideal boundary $\partial_\infty\Sigma := \overline{\Sigma} \cap \partial_\infty \mathbb{H}^n$, denoted by $\operatorname{Vol}_{\mathbb{R}}(\partial_\infty\Sigma)$. By using the concept of the Euclidean volume, we obtain an optimal linear isoperimetric inequality for a k-dimensional complete proper minimal submanifold in the Poincaré ball model B^n of $\mathbb{H}^n$.

Theorem. *Let Σ be a k-dimensional complete proper minimal submanifold in the Poincaré ball model B^n. Then*

$$\operatorname{Vol}_{\mathbb{R}}(\Sigma) \leq \frac{1}{k} \operatorname{Vol}_{\mathbb{R}}(\partial_\infty\Sigma),$$

where equality holds if and only if Σ is a k-dimensional unit ball B^k in B^n.

This theorem can be regarded as an extension of the result obtained in [9] and [28]. We further obtain an optimal isoperimetric inequality for a k-dimensional complete proper

JOURNAL OF THE
CHUNGCHEONG MATHEMATICAL SOCIETY
Volume **27**, No. 1, 2014
http://dx.doi.org/10.14403/jcms.2014.27.1.145

PORTFOLIO SELECTION WITH NONNEGATIVE WEALTH CONSTRAINTS: A DYNAMIC PROGRAMMING APPROACH

Yong Hyun Shin*

Abstract. I consider the optimal consumption and portfolio selection problem with nonnegative wealth constraints using the dynamic programming approach. I use the constant relative risk aversion (CRRA) utility function and disutility to derive the closed-form solutions.

1. Introduction

Following the seminal works of Merton [9, 10], there have been many research articles on the continuous-time portfolio optimization problems with various economic constraints. Especially nonnegative wealth constraint, which is sometimes called borrowing constraint, is one of the most important economic constraints on portfolio selection problems. The nonnegative wealth constraint means that the agent cannot borrow against her future labor income, that is, she always supports a nonnegative wealth level (refer to [1, 2, 3, 4, 5, 7, 8]).

In this paper I investigate the optimal consumption and portfolio selection problem with nonnegative wealth constraints based on the dynamic programming framework of Karatzas *et al.* [6]. I use the constant relative risk aversion (CRRA) utility function and disutility to obtain the closed-form solutions.

2. The economy

It is assumed that only two assets are traded in the financial market. One is a riskless asset and the other is a risky asset. The constant interest

Received January 08, 2014; Accepted January 27, 2014.
2010 Mathematics Subject Classification: Primary 91G10.
Key words and phrases: nonnegative wealth constraints, dynamic programming approach, CRRA utility, portfolio selection.

rate is $r > 0$, and the risky asset follows the geometric Brownian motion $dS_t/S_t = \mu dt + \sigma dB_t$, where μ and σ are constants, and B_t is a standard Brownian motion on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

The market price of risk is defined by $\theta := (\mu - r)/\sigma$. Let π_t be the $\mathcal{F}_t$-progressively measurable portfolio process at time t and c_t be the nonnegative $\mathcal{F}_t$-progressively measurable consumption rate process at time t. It is assumed that the following mathematical conditions hold:

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

The agent receives constant labor income $y_0 > 0$. So the agent's wealth process X_t at time t follows the stochastic differential equation (SDE)

$$dX_t = [rX_t + \pi_t(\mu - r) - c_t + y_0]\, dt + \sigma \pi_t dB_t, \; X_0 = x \geq 0.$$

It is assumed that the agent in this model always faces nonnegative wealth constraints such that

$$X_t \geq 0, \text{ for all } t \geq 0. \tag{2.1}$$

3. The optimization problem

The agent's optimization problem is to maximize her expected utility

$$V(x) = \sup_{(c,\pi)\in\mathcal{A}(x)} \mathbb{E}\left[\int_0^\infty e^{-\rho t}\left(\frac{c_t^{1-\gamma}}{1-\gamma} - l\right) dt\right], \tag{3.1}$$

where $\gamma > 0 (\gamma \neq 1)$ is the agent's coefficient of relative risk aversion, $\rho > 0$ is a subjective discount rate, $l > 0$ is constant disutility due to labor, and $\mathcal{A}(x)$ is an admissible set of pairs (c, π). The following assumption always holds without further comments:

Assumption 3.1.

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

Remark 3.2. For later use, I consider a quadratic equation

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

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

Next theorem implies my main results.

Operations Research Letters 42 (2014) 145–149

Contents lists available at ScienceDirect

Operations Research Letters

journal homepage: www.elsevier.com/locate/orl

An optimal job, consumption/leisure, and investment policy

Gyoocheol Shim [a], Yong Hyun Shin [b,*]

[a] Department of Financial Engineering, Ajou University, Suwon 443749, Republic of Korea
[b] Department of Mathematics, Sookmyung Women's University, Seoul 140742, Republic of Korea

ARTICLE INFO

Article history:
Received 3 September 2013
Received in revised form
27 January 2014
Accepted 27 January 2014
Available online 2 February 2014

Keywords:
Job choice
Consumption
Leisure
Portfolio selection
Martingale method

ABSTRACT

In this paper we investigate an optimal job, consumption, and investment policy of an economic agent in a continuous and infinite time horizon. The agent's preference is characterized by the Cobb–Douglas utility function whose arguments are consumption and leisure. We use the martingale method to obtain the closed-form solution for the optimal job, consumption, and portfolio policy. We compare the optimal consumption and investment policy with that in the absence of job choice opportunities.

1. Introduction

We study an optimal job, consumption, and investment policy of an infinitely-lived economic agent whose preference is characterized by the Cobb–Douglas utility function of consumption and leisure. We consider two kinds of jobs one of which provides higher income but lower leisure than the other. We provide the closed-form solution for the optimal job, consumption, and investment policy by using the martingale and duality approach. We show that there is a threshold wealth level below which the optimally behaving agent chooses the job providing higher income, but above which he chooses the other job providing higher leisure. This is intuitively appealing since leisure is more important than income as the agent's wealth level gets higher. We show that the agent in our model consumes less (resp. more) when the agent's wealth is below (resp. above) the threshold level than he would if he did not have such job choice opportunities. We also show that the agent in our model takes more risk than he would without the job choice options.

There have been many extensive research studies on continuous-time portfolio selection after Merton's pioneering study (Merton [10,11]). Bodie, Merton, and Samuelson [1] have studied the effect of the labor–leisure choice on portfolio choice of an economic agent who has flexibility in his labor supply, by using the dynamic programming method. However they did not derive the closed-form solution. In this paper we use the martingale method to derive the closed-form solution. Many papers have considered portfolio selection with a retirement option: for example, Choi and Shim [2], Choi, Shim, and Shin [3], Dybvig and Liu [4], Farhi and Panageas [5], Lim and Shin [9], etc. The retirement in these papers is irreversible in that the agent cannot come back to his job after retirement, while the job choices in our model are reversible in that the agent can change the current job at any state and time.

The rest of the paper proceeds as follows. Section 2 sets up the optimization problem. Section 3 provides a solution to the problem and Section 4 investigates properties of the optimal policy.

2. The model

We consider the continuous-time financial market in an infinite-time horizon. We assume that there are two financial assets in the market: one is a riskless asset and the other is a risky asset. The risk-free interest rate $r > 0$ is assumed to be a constant and the price S_t of the risky asset is governed by the geometric Brownian motion $dS_t/S_t = \mu dt + \sigma dB_t$ for $t \geq 0$, where $(B_t)_{t=0}^{\infty}$ is a standard Brownian motion on the underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and the parameters μ and $\sigma > 0$ are assumed to be constants. We let $\{\mathcal{F}_t\}_{t\geq 0}$ be the augmentation under $\mathbb{P}$ of the natural filtration generated by the standard Brownian motion $(B_t)_{t=0}^{\infty}$.

Let Θ_t denote the job of an economic agent at time t. The job process $\boldsymbol{\Theta} \triangleq (\Theta_t)_{t=0}^{\infty}$ is $\mathcal{F}_t$-adapted. For simplicity, we assume

* Corresponding author.
E-mail addresses: gshim@ajou.ac.kr (G. Shim), yhshin@sookmyung.ac.kr (Y.H. Shin).

http://dx.doi.org/10.1016/j.orl.2014.01.009
0167-6377/

that there are two kinds of jobs, A_0 and A_1. The agent receives constant labor income $Y_i > 0$ and have a leisure rate L_i at each job A_i, $i = 0, 1$, where

$$0 \le Y_0 < Y_1 \quad \text{and} \quad 0 < L_1 < L_0.$$

Let $c_t \ge 0$ and π_t denote the consumption rate and the amount of money invested in the risky asset, respectively, at time t. The consumption rate process $\mathbf{c} \triangleq (c_t)_{t=0}^{\infty}$ and the portfolio process $\boldsymbol{\pi} \triangleq (\pi_t)_{t=0}^{\infty}$ are $\mathcal{F}_t$-progressively measurable, $\int_0^t c_s ds < \infty$ for all $t \ge 0$ almost surely (a.s.), and $\int_0^t \pi_s^2 ds < \infty$ for all $t \ge 0$ a.s.

Thus the agent's wealth process $(X_t)_{t=0}^{\infty}$ with $X_0 = x$ evolves according to

$$dX_t = \left[rX_t + (\mu - r)\pi_t - c_t + Y_0 \mathbf{1}_{\{\Theta_t = A_0\}} + Y_1 \mathbf{1}_{\{\Theta_t = A_1\}}\right] dt + \sigma \pi_t dB_t. \tag{2.1}$$

The present value of the future labor income stream is Y_i/r for $\Theta_t = A_i$ where $i = 0, 1$. Since $Y_1/r > Y_0/r$ and the job state process $\boldsymbol{\Theta}$ is chosen endogenously by the agent, we let $X_0 = x > -Y_1/r$ and the agent faces the following wealth constraint:

$$X_t \ge -\frac{Y_1}{r}, \quad \text{for all } t \ge 0 \text{ a.s.} \tag{2.2}$$

We call a triple of control $(\boldsymbol{\Theta}, \mathbf{c}, \boldsymbol{\pi})$ satisfying the above conditions including (2.2) with $X_0 = x > -Y_1/r$ admissible at x. Let $\mathcal{A}(x)$ be the set of all admissible policies.

We assume that the agent has the Cobb–Douglas utility function $u(c_t, l_t)$, as in Farhi and Panageas [5]:

$$u(c_t, l_t) \triangleq \frac{1}{\alpha} \frac{\left(c_t^{\alpha} l_t^{1-\alpha}\right)^{1-\gamma}}{1-\gamma}, \quad 0 < \alpha < 1 \text{ and } 0 < \gamma \ne 1, \tag{2.3}$$

where γ is the agent's coefficient of relative risk aversion, α is a constant, and l_t is the leisure rate at time t. Let $\gamma_1 \triangleq 1 - \alpha(1-\gamma)$, and then $0 < \gamma_1 \ne 1$ and the Cobb–Douglas utility function $u(\cdot,\cdot)$ in (2.3) can be rewritten as

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

Remark 2.1. If $\gamma > 1$, then $\gamma > \gamma_1 > 1$, $\frac{\gamma_1}{1-\gamma_1} < 0$ and $L_0^{\frac{\gamma_1-\gamma}{\gamma_1}} - L_1^{\frac{\gamma_1-\gamma}{\gamma_1}} < 0$. If $0 < \gamma < 1$, then $0 < \gamma < \gamma_1 < 1$, $\frac{\gamma_1}{1-\gamma_1} > 0$ and $L_0^{\frac{\gamma_1-\gamma}{\gamma_1}} - L_1^{\frac{\gamma_1-\gamma}{\gamma_1}} > 0$. Thus the following inequality always holds:

$$\frac{\gamma_1}{1-\gamma_1}\left(L_0^{\frac{\gamma_1-\gamma}{\gamma_1}} - L_1^{\frac{\gamma_1-\gamma}{\gamma_1}}\right) > 0.$$

Problem 2.1. The agent's optimization problem is to maximize the expected utility

$$J(x; \boldsymbol{\Theta}, \mathbf{c}, \boldsymbol{\pi}) = \mathbb{E}\left[\int_0^{\infty} e^{-\rho t}\left(L_0^{\gamma_1-\gamma}\frac{c_t^{1-\gamma_1}}{1-\gamma_1}\mathbf{1}_{\{\Theta_t=A_0\}} + L_1^{\gamma_1-\gamma}\frac{c_t^{1-\gamma_1}}{1-\gamma_1}\mathbf{1}_{\{\Theta_t=A_1\}}\right) dt\right],$$

over $(\boldsymbol{\Theta}, \mathbf{c}, \boldsymbol{\pi}) \in \mathcal{A}(x)$, where $\rho > 0$ is a subjective discount factor.

Thus the value function $V(x)$ is given by

$$V(x) = \sup_{(\boldsymbol{\Theta}, \mathbf{c}, \boldsymbol{\pi}) \in \mathcal{A}(x)} J(x; \boldsymbol{\Theta}, \mathbf{c}, \boldsymbol{\pi}).$$

Assumption 2.1. We assume, as in Farhi and Panageas [5], that

$$K_1 \triangleq r + \frac{\rho - r}{\gamma_1} + \frac{\gamma_1 - 1}{2\gamma_1^2}\theta^2 > 0,$$

where $\theta \triangleq (\mu - r)/\sigma$, called the market price of risk.

3. The solution to the optimization problem

We denote the state price density by H_t:

$$H_t \triangleq e^{-\left(r + \frac{1}{2}\theta^2\right)t - \theta B_t}.$$

For any fixed $T \in [0, \infty)$, we denote the equivalent martingale measure by $\widetilde{\mathbb{P}}^T$:

$$\widetilde{\mathbb{P}}^T(A) = \mathbb{E}\left[e^{-\frac{1}{2}\theta^2 T - \theta B_T}\mathbf{1}_A\right], \quad \text{for } A \in \mathcal{F}_T.$$

By the Girsanov theorem, the new process $\widetilde{B}_t = B_t + \theta t$ is a standard Brownian motion for $t \in [0, T]$ under the measure $\widetilde{\mathbb{P}}^T$. As shown in Proposition 7.4 in Section 1.7 of Karatzas and Shreve [7], there exists a unique probability measure $\widetilde{\mathbb{P}}$ on $\mathcal{F}_{\infty}$ which agrees with $\widetilde{\mathbb{P}}^T$ on $\mathcal{F}_T$, for $T \in [0, \infty)$, and $\widetilde{B}_t$ is a standard Brownian motion for $t \in [0, \infty)$ under $\widetilde{\mathbb{P}}$. Thus Eq. (2.1) can be rewritten as

$$dX_t = \left[rX_t - c_t + Y_0 \mathbf{1}_{\{\Theta_t = A_0\}} + Y_1 \mathbf{1}_{\{\Theta_t = A_1\}}\right] dt + \sigma \pi_t d\widetilde{B}_t. \tag{3.1}$$

By (2.2) and (3.1), we derive, similarly to Lim and Shin [9], the following budget constraint:

$$\mathbb{E}\left[\int_0^{\infty} \left(c_t - Y_0 \mathbf{1}_{\{\Theta_t = A_0\}} - Y_1 \mathbf{1}_{\{\Theta_t = A_1\}}\right) H_t dt\right] \le x.$$

For a Lagrange multiplier $\lambda > 0$, a dual value function is

$$\begin{aligned}
&\widetilde{V}(\lambda) + \lambda x \\
&= \sup_{(\boldsymbol{\Theta}, \mathbf{c}, \boldsymbol{\pi}) \in \mathcal{A}(x)} \mathbb{E}\left[\int_0^{\infty} e^{-\rho t}\left(L_0^{\gamma_1-\gamma}\frac{c_t^{1-\gamma_1}}{1-\gamma_1}\mathbf{1}_{\{\Theta_t=A_0\}} + L_1^{\gamma_1-\gamma}\frac{c_t^{1-\gamma_1}}{1-\gamma_1}\mathbf{1}_{\{\Theta_t=A_1\}}\right) dt \right.\\
&\qquad \left. - \lambda \int_0^{\infty} \left(c_t - Y_0 \mathbf{1}_{\{\Theta_t = A_0\}} - Y_1 \mathbf{1}_{\{\Theta_t = A_1\}}\right) H_t dt\right] + \lambda x \\
&= \mathbb{E}\left[\int_0^{\infty} e^{-\rho t}\left(\widetilde{u}_0(z_t)\mathbf{1}_{\{0 < z_t \le \bar{z}\}} + \widetilde{u}_1(z_t)\mathbf{1}_{\{z_t > \bar{z}\}}\right) dt\right] + \lambda x,
\end{aligned} \tag{3.2}$$

if the job and consumption strategy $(\Theta_t^{\lambda}, c_t^{\lambda})$ is given by

$$\Theta_t^{\lambda} = \begin{cases} A_0, & \text{if } 0 < z_t \le \bar{z}, \\ A_1, & \text{if } z_t > \bar{z}, \end{cases}$$

$$c_t^{\lambda} = \begin{cases} L_0^{\frac{\gamma_1-\gamma}{\gamma_1}}(z_t)^{-\frac{1}{\gamma_1}}, & \text{if } 0 < z_t \le \bar{z}, \\ L_1^{\frac{\gamma_1-\gamma}{\gamma_1}}(z_t)^{-\frac{1}{\gamma_1}}, & \text{if } z_t > \bar{z}, \end{cases}$$

where

$$\begin{aligned}
\widetilde{u}_i(z) &= \sup_{c \ge 0}\left(L_i^{\gamma_1-\gamma}\frac{c^{1-\gamma_1}}{1-\gamma_1} - cz\right) + Y_i z \\
&= L_i^{\frac{\gamma_1-\gamma}{\gamma_1}}\frac{\gamma_1}{1-\gamma_1}z^{-\frac{1-\gamma_1}{\gamma_1}} + Y_i z, \quad i = 0, 1,
\end{aligned}$$

$$z_t \triangleq \lambda e^{\rho t} H_t = \lambda e^{\left(\rho - r - \frac{1}{2}\theta^2\right)t - \theta B_t}, \tag{3.3}$$

and $\bar{z}$ is the solution to the algebraic equation $\widetilde{u}_0(z) = \widetilde{u}_1(z)$:

$$\bar{z} = \left(\frac{\frac{\gamma_1}{1-\gamma_1}\left(L_0^{\frac{\gamma_1-\gamma}{\gamma_1}} - L_1^{\frac{\gamma_1-\gamma}{\gamma_1}}\right)}{Y_1 - Y_0}\right)^{\gamma_1} > 0, \tag{3.4}$$

which is positive by Remark 2.1. Similarly to Proposition 6.5 in Karatzas and Wang [8], the Lagrange multiplier λ is chosen in (3.12) so that

$$\mathbb{E}\left[\int_0^{\infty} \left(c_t^{\lambda} - Y_0 \mathbf{1}_{\{0 < z_t \le \bar{z}\}} - Y_1 \mathbf{1}_{\{z_t > \bar{z}\}}\right) H_t dt\right] = x \tag{3.5}$$

JOURNAL OF THE
CHUNGCHEONG MATHEMATICAL SOCIETY
Volume **27**, No. 2, May 2014
http://dx.doi.org/10.14403/jcms.2014.27.2.219

AN OPTIMAL CONSUMPTION AND INVESTMENT PROBLEM WITH LABOR INCOME AND REGIME SWITCHING

Yong Hyun Shin*

Abstract. I use the dynamic programming approach to study the optimal consumption and investment problem with regime-switching and constant labor income. I derive the optimal solutions in closed-form with constant absolute risk aversion (CARA) utility and constant disutility.

1. Introduction

Following the seminal works of Merton [3, 4], the field of continuous-time portfolio selection is one of the most important areas in mathematical finance. Also recently regime-switching technique is widely used in mathematical finance (see [1, 7, 5, 6]).

In this work I study the optimal consumption and investment problem with two-state regime-switching and constant labor income under the dynamic programming framework based on Karatzas *et al.* [2]. I use the constant absolute risk aversion (CARA) utility function and constant disutility to derive the optimal solutions in closed-form.

2. The financial market

On a proper probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a standard Brownian motion B_t and a continuous-time two-state Markov chain ϵ_t are defined and it is assumed that they are independent. The filtration $\{\mathcal{F}_t\}_{t\geq 0}$ is generated by both the Brownian motion B_t and the Markov chain ϵ_t.

Received January 10, 2014; Accepted April 07, 2014.
2010 Mathematics Subject Classification: Primary 91G10.
Key words and phrases: regime-switching, labor income, CARA utility, portfolio selection.

In the financial market, it is assumed that only two assets are traded: One is a riskless asset with constant interest rate $r > 0$ and the other is a risky asset (or stock). It is also assumed that there are two regime states, 1, 2 in the market and regime i switches into regime j at the first jump time of an independent Poisson process with intensity λ_i, for $i, j \in \{1, 2\}$. 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 $\theta_i := (\mu_i - r)/\sigma_i,\ i = 1,\ 2$. Let π_t be the $\mathcal{F}_t$-progressively measurable portfolio process at time t and c_t be the nonnegative $\mathcal{F}_t$-progressively measurable consumption rate process at time t. I assume that the portfolio process π_t and the consumption rate process c_t satisfy the following conditions:

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

In regime $i \in \{1, 2\}$, the agent receives constant labor income $y_i > 0$. The agent's wealth process X_t at time t follows

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

3. The optimization problem

The agent's expected utility maximization problem with CARA utility $u(c) := -e^{-\gamma c}/\gamma$ is given by

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

where τ_i is the first jump time from i-th state to j-th state, $\rho > 0$ is a subjective discount factor, $\gamma > 0$ is the coefficient of absolute risk aversion, $l > 0$ is constant disutility because of labor, and $\mathcal{A}(x)$ is an admissible class of pairs (c, π) at x, where $i, j \in \{1, 2\}$ and $i \neq j$. It is assumed that the following inequality always holds without further comments:

Assumption 3.1.

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

Assumption 3.2. It is assumed that the value function $V_i(x)$ for this optimization problem (3.1) is an increasing function, that is,

$$V_i'(x) > 0, \text{ for } i = 1, 2.$$

Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2014, Article ID 153793, 6 pages
http://dx.doi.org/10.1155/2014/153793

Research Article

Portfolio Selection with Subsistence Consumption Constraints and CARA Utility

Gyoocheol Shim[1] and Yong Hyun Shin[2]

[1] *Department of Financial Engineering, Ajou University, Suwon 443-749, Republic of Korea*
[2] *Department of Mathematics, Sookmyung Women's University, Seoul 140-742, Republic of Korea*

Correspondence should be addressed to Yong Hyun Shin; yhshin@sookmyung.ac.kr

Received 2 January 2014; Revised 25 March 2014; Accepted 6 April 2014; Published 16 April 2014

Academic Editor: Pankaj Gupta

We consider the optimal consumption and portfolio choice problem with constant absolute risk aversion (CARA) utility and a subsistence consumption constraint. A subsistence consumption constraint means there exists a positive constant minimum level for the agent's optimal consumption. We use the dynamic programming approach to solve the optimization problem and also give the verification theorem. We illustrate the effects of the subsistence consumption constraint on the optimal consumption and portfolio choice rules by the numerical results.

1. Introduction

Following the seminal research works of Merton [1, 2], various problems of continuous-time optimal consumption and portfolio selection have been considered under various financial/economic constraints. One of the interesting research topics in a continuous-time portfolio selection problem is the optimization problem subject to a subsistence consumption constraint (or a downside consumption constraint). A subsistence constraint means that there exists a positive lower bound level for the agent's optimal consumption rate. Thus this constraint affects the agent's financial decision including her optimal portfolio.

Lakner and Nygren [3] have studied the portfolio optimization problem subject to a downside constraint for consumption and an insurance constraint for terminal wealth with a martingale approach. Gong and Li [4] have investigated the role of index bonds in the optimal consumption and portfolio selection problem with constant relative risk aversion (CRRA) utility and a real subsistence consumption constraint using the dynamic programming approach. Shin et al. [5] have also considered a similar problem to that of Gong and Li [4]. They have studied the portfolio selection problem with a general utility function and a downside consumption constraint using the martingale approach. Yuan and Hu [6] have investigated the optimal consumption and portfolio selection problem with a consumption habit constraint and a terminal wealth downside constraint using the martingale approach. In this paper we use the dynamic programming method based on Karatzas et al. [7] to derive the value function and the optimal policies in closed-form with a constant absolute risk aversion (CARA) utility function and a subsistence consumption constraint. Lim et al. [8] have considered a similar portfolio optimization problem combined with the voluntary retirement choice problem. Shin and Lim [9] have analyzed the effects of the subsistence consumption constraint for behavior of investment in the risky asset.

The rest of this paper proceeds as follows. Section 2 introduces the financial market. In Section 3 we consider the main optimization problem. We use the dynamic programming principle to derive the solutions in closed form with CARA utility and a subsistence consumption constraint. We also give some numerical results and the solutions derived by the martingale method. Section 4 concludes.

2. The Financial Market Setup

We assume that there are two assets in the financial market: one is a riskless asset with constant interest rate $r > 0$, and the

other is a stock whose price process $\{S_t\}_{t\geq 0}$ evolves according to the stochastic differential equation (SDE)

$$\frac{dS_t}{S_t} = \mu dt + \sigma dB_t, \quad \text{for } t \geq 0, \tag{1}$$

where μ is the constant expected rate of return of the stock, $\sigma > 0$ is the constant volatility of the stock, and B_t is a standard Brownian motion on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ endowed with the filtration $\{\mathcal{F}_t\}_{t\geq 0}$ which is the augmentation under $\mathbb{P}$ of the natural filtration generated by the standard Brownian motion $\{B_t\}_{t\geq 0}$. We assume that $\mu \neq r$ so that the market price of risk θ is not zero:

$$\theta \triangleq \frac{\mu - r}{\sigma} \neq 0. \tag{2}$$

Let X_t be an economic agent's wealth at time t, π_t the amount of money invested in the stock at time t, and c_t the consumption rate at time t. The portfolio process $\{\pi_t\}_{t\geq 0}$ is adapted to $\{\mathcal{F}_t\}_{t\geq 0}$ and satisfies, for all $t \geq 0$, almost surely (a.s.),

$$\int_0^t \pi_s^2 ds < \infty, \tag{3}$$

and the consumption rate process $\{c_t\}_{t\geq 0}$ is a nonnegative process adapted to $\{\mathcal{F}_t\}_{t\geq 0}$ such that, for all $t \geq 0$, a.s.,

$$\int_0^t c_s ds < \infty. \tag{4}$$

We assume that there is a subsistence consumption constraint which restricts the minimum consumption level. That is, the consumption process should satisfy

$$c_t \geq R, \quad \forall t \geq 0, \tag{5}$$

where $R > 0$ is a constant lower bound for the consumption rates. Thus the agent's wealth process $\{X_t\}_{t\geq 0}$ follows the SDE

$$dX_t = \left[rX_t + \pi_t(\mu - r) - c_t\right] dt + \sigma \pi_t dB_t, \tag{6}$$

with an initial endowment $X_0 = x > R/r$. (We need this restriction on the initial endowment for the positive consumption rate. See Lemma 3.1 of Gong and Li [4]). A consumption-portfolio plan $(\mathbf{c}, \boldsymbol{\pi}) := (\{c_t\}_{t\geq 0}, \{\pi_t\}_{t\geq 0})$ satisfying the above conditions is called admissible at $x > R/r$ if $X_t \geq R/r$, for all $t \geq 0$. We let $\mathcal{A}(x)$ denote the class of admissible controls at $x > R/r$.

3. The Optimization Problem

Now the agent's optimization problem with initial wealth $X_0 = x > R/r$ is to choose $(\mathbf{c}, \boldsymbol{\pi}) \in \mathcal{A}(x)$ to maximize the following expected life-time utility:

$$\mathbb{E}\left[\int_0^\infty e^{-\beta t} u(c_t)\, dt\right]. \tag{7}$$

Here, $\beta > 0$ is the subjective discount factor and $u(\cdot)$ is a constant absolute risk aversion (CARA) utility function defined by

$$u(c) \triangleq -\frac{e^{-\gamma c}}{\gamma}, \tag{8}$$

where $\gamma > 0$ is the agent's coefficient of absolute risk aversion. Thus the agent's value function is given by

$$V^*(x) \triangleq \sup_{(\mathbf{c},\boldsymbol{\pi})\in\mathcal{A}(x)} \mathbb{E}\left[-\int_0^\infty \frac{e^{-\beta t - \gamma c_t}}{\gamma} dt\right]. \tag{9}$$

Bellman equation corresponding to the optimization problem for $x > R/r$ is

$$\max_{c\geq R, \pi}\left[\left\{rx + \pi(\mu - r) - c\right\} V'(x) + \frac{1}{2}\sigma^2\pi^2 V''(x) - \beta V(x) - \frac{e^{-\gamma c}}{\gamma}\right] = 0. \tag{10}$$

We assume that the wealth process X_t must satisfy a transversality condition

$$\lim_{t\to\infty} e^{-\beta t} V(X_t) = 0. \tag{11}$$

We will find the solution $V(x)$, as the candidate value function, to Bellman equation (10) under the conditions that $V'(x) > 0$ and $V''(x) < 0$ for $x > R/r$ and $V'(\overline{x}) = u'(R) = e^{-\gamma R}$ for a real number $\overline{x} > R/r$. After obtaining the solution, we can check these conditions. Under these conditions, in particular, the first-order condition (FOC), $-V'(x) + u'(c) = 0$ with respect to $c \geq R$, is binding if $R/r < x < \overline{x}$ so that the maximizing $c \geq R$ in Bellman equation (10) is R in this case.

Thus, from the first-order conditions (FOCs) of Bellman equation (10), we derive the candidate optimal consumption and portfolio

$$c^* = \begin{cases} R, & \text{if } \dfrac{R}{r} < x < \overline{x} \\ -\dfrac{\log\left\{V'(x)\right\}}{\gamma}, & \text{if } x \geq \overline{x}, \end{cases} \qquad \pi^* = -\frac{\theta V'(x)}{\sigma V''(x)}. \tag{12}$$

Remark 1. For later use, we consider two quadratic algebraic equations:

$$rm^2 - \left(r + \beta + \frac{\theta^2}{2}\right) m + \beta = 0, \tag{13}$$

with two roots m_1 $(0 < m_1 < 1)$ and $m_2 > 1$ and

$$\frac{\theta^2}{2\gamma} n^2 + \left(r - \beta - \frac{\theta^2}{2}\right) n - r\gamma = 0, \tag{14}$$

with two roots $n_1 < 0$ and $n_2 > \gamma$.

JOURNAL OF THE
CHUNGCHEONG MATHEMATICAL SOCIETY
Volume **27**, No. x, xxxx 201x

SOME REMARKS ON TYPES OF NOETHERIAN LOCAL RINGS

Kisuk Lee*

Abstract. We study some results which concern the types of Noetherian local rings, and improve slightly the previous result: For a complete unmixed (or quasi-unmixed) Noetherian local ring A, we prove that if either $A_{\mathfrak{p}}$ is Cohen-Macaulay, or $r(A_{\mathfrak{p}}) \leq \operatorname{depth} A_{\mathfrak{p}} + 1$ for every prime ideal $\mathfrak{p}$ in A, then A is Cohen-Macaulay. Also, some analogous results for modules are considered.

1. Backgrounds on types

Throughout this paper, we assume that $(A, \mathfrak{m})$ is a commutative Noetherian local ring of dimension d, and M is a finitely generated A-module. We also assume that all modules are unitary.

In this section, we introduce some previously known results about types of rings. Although most of the backgrounds on types quoted here are already present in [7,8], we have included them for the readers' convenience.

We first define the type of a ring. The i-th Bass number of M at a prime $\mathfrak{p}$, denoted $\mu_i(\mathfrak{p}, M)$, is defined to be $\dim_{k(\mathfrak{p})}\operatorname{Ext}^i_{A_{\mathfrak{p}}}(k(\mathfrak{p}), M_{\mathfrak{p}})$, where $k(\mathfrak{p}) = A_{\mathfrak{p}}/\mathfrak{p}A_{\mathfrak{p}}$: we set $\mu_i(A)=\mu_i(\mathfrak{m}, A)$ for brevity. The *type* of A, denoted by $r(A)$, is defined to be $\mu_d(A)$. Many mathematicians have been interested on the conditions related on types, which make a ring A Cohen-Macaulay ([1,4,5,6,9], etc).

Gorenstein rings were characterized as Cohen-Macaulay rings A with $r(A) = 1$ by Bass ([2]), and then it was conjectured by Vasconcelos that the condition $r(A) = 1$ is sufficient for A to be Gorenstein ([14]). In [5], Foxby proved this conjecture for essentially equicharacteristic rings using a version of the Intersection Theorem. The conjecture was proven

Received August 13, 2014; Accepted October 10, 2014.
2010 Mathematics Subject Classification: Primary 13C14, 13C15, 13D07, 13D45, 13H10.
Key words and phrases: Cohen-Macaulay ring, type of a ring, Gorenstein ring.

in general by Roberts ([13]): he showed that a Noetherian local ring of type one is Cohen-Macaulay, (and hence Gorenstein).

THEOREM 1.1. [13] *Let A be a commutative Noetherian local ring, let n be the Krull dimension of A, and suppose that $\mu_n(A) = 1$. Then A is Gorenstein.*

In [9], the above theorem was slightly developed using the behavior of maps in minimal injective resolution of a module as follows:

THEOREM 1.2. [9] *Let $(A, \mathfrak{m}, k)$ be a Noetherian local ring of dimension n and M a finitely generated A-module of dimension d. If $\mu_t(\mathfrak{q}, M) = 1$ for some $\mathfrak{q} \in \mathrm{Supp}(M)$ and some $t \leq \dim M_{\mathfrak{q}}$, then $\mu_j(\mathfrak{q}, M) = 0$ for all $j < t$.*

Costa, Huneke and Miller showed the following:

THEOREM 1.3. [4] *Let A be a local ring whose completion is a domain. If $r(A) = 2$, then A is Cohen-Macaulay.*

After they gave two examples: a complete equidimensional local ring of type two that is not Cohen-Macaulay, and a complete reduced local ring of type two that is not Cohen-Macaulay, they posed a question: *Does there exist a complete, equidimensional, reduced local ring A with $r(A) = 2$ that is not Cohen-Macaulay?* However, Marley answered this question in negative by proving the following theorem:

THEOREM 1.4. [10] *Let A be an unmixed local ring of type two. Then A is Cohen-Macaulay.*

Marley also asked that if a complete local ring of type n satisfies Serre's condition (S_{n-1}), then it is Cohen-Macaulay: we say that a finite A-module M satisfies a Serre's condition (S_t) if $\mathrm{depth}\, M_{\mathfrak{p}} \geq \min\{t, \dim M_{\mathfrak{p}}\}$ for every prime $\mathfrak{p}$ in $Supp(M)$. Kawasaki answered this question in the affirmative when rings contain a field and $n \geq 3$ ([6]), and later Aoyama gave a general proof.

THEOREM 1.5. [1] *Let $n \geq 3$ be an integer. If $r(A) \leq n$ and $\hat{A}$ is (S_{n-1}), then A is Cohen-Macaulay.*

Kawasaki conjectured the following which is still open:

CONJECTURE 1.6. [6] *Let A be a complete unmixed local ring of type n. If $A_{\mathfrak{p}}$ is Cohen-Macaulay for all $\mathfrak{p}$ in $Spec(A)$ such that $ht(\mathfrak{p}) < n$, then A is Cohen-Macaulay.*

J. Chosun Natural Sci.
Vol. 7, No. 4 (2014) pp. 278 – 282
http://dx.doi.org/10.13160/ricns.2014.7.4.278

Various Row Invariants on Cohen-Macaulay Rings

Kisuk Lee†

Abstract

We define a numerical invariant $row_j^*(A)$ over Cohen-Macaulay local ring A, which is related to the presenting matrices of the j-th syzygy module (with or without free summands). We show that $row_d(A) = row_{CM}(A)$ and $row_d^*(A) = row_{CM}^*(A)$ for a Cohen-Macaulay local ring A of dimension d.

Key words : Row Invariants, Cohen-Macaulay Local Ring, Maximal Cohen-Macaulay Module, Free Summands

1. Introduction

Throughout this paper, we assume that $(A,\boldsymbol{m})$ is a Noetherian local ring, and all modules are unitary.

It is proved[1] that there are certain restrictions on the entries of the maps in the minimal free resolutions of finitely generated modules of infinite projective dimension over Noetherian local rings. This fact provides not only a new way to understand some previously known results in commutative ring theory (see for instance Corollary 2.8[1], or Proposition 2.2[1]), but also new interesting invariants of local rings. These invariants have turned out to be quite useful; for example, the Auslander index of A can be described as a column invariant when A is Gorenstein[1], and the multiplicity of A can be also explained by these invariants when A is hypersurface. (For the further background of these invariants, we refer the reader to some papers[1-5].)

This paper particularly deals with invariants $row_d(A)$, $row_{CM}(A)$, $row_d^*(A)$, and $row_{CM}^*(A)$. We will eventually show that all of those are equal to $row(A)$ for a Cohen-Macaulay local ring of dimension d.

We recall that $row(A)$ is defined with the rows of the maps in infinite minimal resolutions: $row(A)$ is the smallest integer $t \geq 1$ such that for each finitely generated A-module M of infinite projective dimension, each row of φ_i contains an element outside $\boldsymbol{m}^t$ for all $i > depth(A)$, where φ_i is the i-th map of a minimal free resolution of M. It was shown[6] that $row(A)$ can be described in terms of the rows of presenting matrices of maximal Cohen-Macaulay modules (not necessarily without free summand) when A is Cohen-Macaulay (not necessarily Gorenstein). In this paper, we define $row_j^*(A)$ as the smallest positive integer $t \geq 1$ such that for each j-th syzygy module (not necessarily without free summand) of infinite projective dimension, each row of its presenting matrix contains an element outside $\boldsymbol{m}^t$, and show that $row_j^*(A) = row_j(A)$ for each $j \geq d$.

In Section 2, we give the definitions of $row_j(A)$, $row_{CM}(A)$, $row_j^*(A)$, and $row_{CM}^*(A)$ in detail, and study some basic properties of these invariants.

In Section 3, we prove that $row_d^*(A) = row_d(A)$ for a Cohen-Macaulay local ring A of dimension by showing that $row_d(A) = row_{CM}(A)$ and $row_d^*(A) = row_{CM}^*(A)$ which explains $row(A)$ can be described in terms of the rows of presenting matrices of the d-th syzygy modules with or without free summands. We also prove that if A is a Gorenstein local ring of dimension d, then $row(A) = row_j^*(A)$ for all $j \geq d$.

2. Basic Definitions and Facts

Let M be a finitely generated A-module and φ_i denote an i-th map of a minimal free resolution of M: $\cdots \to A^{n_2} \xrightarrow{\varphi_2} A^{n_1} \xrightarrow{\varphi_1} A^{n_0} \to M \to 0$. A map φ_i is

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

†Corresponding author : kilee@sookmyung.ac.kr
(Received : October 16, 2014, Revised : December 15, 2014, Accepted : December 25, 2014)

represented by a matrix of size $n_{i+1}\times n_i$. By a (minimal) presenting matrix of M, we mean the representation matrix of φ_i. Then we note that φ_i is represented by the presenting matrix of the $(i-1)$-st syzygy module, which is the image of φ_i.

Definition 2.1. Let $(A,\boldsymbol{m})$ be a Cohen-Macaulay (non-regular) local ring, and M a (nonfree) maximal Cohen-Macaulay module. We define $row_{CM}(A)$ to be the smallest positive integer c such that each row of the presenting matrix of M contains an element outside $\boldsymbol{m}^c$. We now define $row_{CM}(A) = \sup\{row_{CM}(M) : M : M$ is a maximal Cohen-Macaulay module without free summands$\}$, and $row^*_{CM}(A) = \sup\{row_{CM}(M) : M : M$ is a maximal Cohen-Macaulay module$\}$.

Using the syzygy modules, we also define the following invariants:

Definition 2.2. Let $(A,\boldsymbol{m})$ be a Cohen-Macaulay local ring. For a non negative integer j, we define $row_j(A) = \inf\{ t \geq 1$: for each j-th syzygy module without free summands of infinite projective dimension, each row of its presenting matrix contains an element outside $\boldsymbol{m}^t\}$, and $row^*_j(A) = \inf\{ t \geq 1$: for each j-th syzygy module of infinite projective dimension, each row of its presenting matrix contains an element outside $\boldsymbol{m}^t\}$.

If A is regular, then we define $row_{CM}(A) = 1$, and $row_j(A) = row^*_j(A) = 1$ for each non-negative integer j.

We here remark that if A is not Cohen-Macaulay and the Canonical Element Conjecture holds for A, then $row^*_j(A) = \infty$ for $0 \leq j < \dim A$. We first recall the Canonical Element Conjecture studied[1] for reader's convenience.

Canonical Element Conjecture (Hochster). Let $\boldsymbol{x} = x_1,\cdots,x_n$ be any system of parameters of a Noetherian local ring $(A,\boldsymbol{m})$ of dimension n. Let $\phi_\cdot$ be a map of complexes from the Koszul complex $K_\cdot(\boldsymbol{x};A)$ to a free resolution $F_\cdot$ of the residue field $A/\boldsymbol{m}$ lifting the identity map in degree 0. Then $\phi_n \neq 0$.

It is known[7,8] that this conjecture holds for rings containing a field and also in dimension ≤ 3. P. Roberts also pointed out the following statement, which was proved[1]:

Fact. (P. Roberts) Let $(A,\boldsymbol{m})$ be a local ring of dimension n. Let $\boldsymbol{x} = x_1,\cdots,x_n$ be a system of parameters. Let $(J_\cdot, d_\cdot)$ be a free complex such that $H_0(J_\cdot) \cong A/(\boldsymbol{x})$. Let $(K_\cdot(n),\delta_\cdot)$ denote the Koszul complex $K_\cdot(\boldsymbol{x};A)$. Let $\lambda_\cdot$ be a map of complexes from $(K_\cdot(n),\delta_\cdot)$ to $J_\cdot$ which lifts the identity map on $A/(\boldsymbol{x})$. Suppose that the Canonical Element Conjecture holds for all local rings of dimension less than equal to q. Then λ_i splits for all $0 \leq i \leq q$, i.e., we may assume that $J_i = K_{i(n)} \oplus J_i'$, and $d_i = \begin{pmatrix} \delta_i & 0 \\ \alpha & \beta_i \end{pmatrix}$.

Therefore, if $\boldsymbol{x}$ is a system of parameters of a non Cohen-Macaulay ring A, and the Canonical Element Conjecture holds for A, then for each $t \geq 1$, j-th map of the minimal resolution of $A/(\boldsymbol{x}^t)$ contains a row all of whose entries are in $\boldsymbol{m}^t$ for $1 \leq j \leq \dim A$, and hence $row^*_j(A) = \infty$ for $0 \leq j < \dim A$.

The following proposition is obtained immediately from the definitions:

Proposition 2.3. Let $(A,\boldsymbol{m})$ be a Cohen-Macaulay local ring of dimension d. Then

(1) $row^*_j(A) \leq row(A) < \infty$ for all $j \geq d$, and $row^*_{(j+1)}(A) \leq row^*_j(A)$ for all $j \geq 0$.

(2) There is some $j_0 > d$ such that $row^*_j(A) = row(A)$ for all j with $d \leq j < j_0$. In particular, $row(A) = row^*_d(A)$.

Proof. (1) For a fixed $j \geq d$, suppose that $row^*_j(A)$ is finite, and let $t = row^*_j(A)$. Then there is a j-th syzygy module N such that every entry of some row of the presenting matrix of N is contained in $\boldsymbol{m}^{t-1}$. Since the presenting matrix of N is $(j+1)$-th differential map of a minimal resolution of certain module, we know $row(A) > t-1$ by the definition of $row(A)$, and thus $row(A) \geq row^*_j(A)$. If $row^*_j(A) = \infty$, then using the above reasoning, we arrive at $row(A) = \infty$, which is a contradiction.

The second part of (1) follows immediately since a $(j+1)$-th syzygy module is also a j-th syzygy module for any non-negative integer j.

(2) Let $t = row(A)$. Then there is a finitely generated A-module M of infinite projective dimension such that for some $j_0 > d$, every entry of some row of φ_{j_0} in $\boldsymbol{m}^t$,

통계학과

Journal of the Korean Statistical Society 43 (2014) 631–641

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

journal homepage: www.elsevier.com/locate/jkss

Quasilikelihood and quasi-maximum likelihood for GARCH-type processes: Estimating function approach

S.Y. Hwang [a,*], M.S. Choi [b], I.K. Yeo [a]

[a] Department of Statistics, Sookmyung Women's University, Seoul 140-742, Republic of Korea
[b] Korea Energy Economics Institute, Uiwang-si 437-713, Republic of Korea

ARTICLE INFO

Article history:
Received 2 October 2012
Accepted 26 January 2014
Available online 13 February 2014

AMS 2000 subject classifications:
primary 62M10
62P10

Keywords:
GARCH-type
Quasilikelihood
Quasi-maximum likelihood
Skewed t-distribution

ABSTRACT

When modeling conditionally heteroscedastic processes, it is usually the case that exact likelihood is not known to researchers or it is too complicated for practical purposes. Consequently, a quasi-maximum likelihood (QML) is frequently employed rather than the exact maximum likelihood (cf. Gourieroux (1997, Ch. 4), Straumann (2005, Ch. 5)). When a GARCH process is partially specified only through the first and second order conditional moments, a systematic and unified approach for inference is via the so called quasilikelihood (QL, see Godambe (1985)). This article aims to discriminate between the QL and QML for general GARCH-type processes. It is verified that the QL differs with the QML in terms of conditional skewness and kurtosis. Asymptotics for the QL and QML are obtained and then compared with each other. A class of skewed t-distributions (cf. Fernandez and Steel (1998)) is considered to illustrate the difference between the QL and QML.

1. Introduction

Conditionally heteroscedastic processes have been useful for modeling volatilities in the field of financial time series. The time series $\{\epsilon_t\}$ is referred to as the first order GARCH, viz., GARCH(1, 1) when $\{\epsilon_t\}$ is recursively obtained by the equation

$$\epsilon_t = \sqrt{h_t} e_t \quad \text{and} \quad h_t = \alpha_0 + \alpha_1 \epsilon_{t-1}^2 + \beta_1 h_{t-1}. \tag{1}$$

The term GARCH stands for the Generalized ARCH (AutoRegressive Conditional Heteroscedasticity). Here and in the sequel, $\{e_t\}$ is a sequence of iid random innovations with mean zero and variance unity. We define an increasing sequence of the sigma fields $\{F_t\}$ generated by $\epsilon_t, \epsilon_{t-1}, \ldots$. Let h_t denote the conditional variance of ϵ_t given the past values $\epsilon_{t-1}, \epsilon_{t-2}, \ldots$, that is, $h_t = \mathrm{Var}(\epsilon_t | F_{t-1})$. Note that h_t is F_{t-1} measurable. Higher order GARCH model, for instance, a GARCH(p, q) process is readily obtained by adding additional lags of $\epsilon_{t-1}^2, \ldots, \epsilon_{t-p}^2$ and $h_{t-1}, \ldots, h_{t-q}$ in the defining equation (1). Making some variations to (1), a number of GARCH-type models have been suggested and investigated in the literature. We refer to the monograph by Straumann (2005) and Part I of the handbook edited by Andersen, Davis, Kreiss, and Mikosch (2009) for various nonlinear and asymmetric GARCH processes. See also a recent paper of Choi, Park, and Hwang (2012) and references therein for diverse GARCH-type models.

* Corresponding author. Tel.: +82 1091809078.
E-mail addresses: shwang@sm.ac.kr, shwang@sookmyung.ac.kr (S.Y. Hwang), stat.mschoi@keei.re.kr (M.S. Choi), inkwon@sookmyung.ac.kr (I.K. Yeo).

http://dx.doi.org/10.1016/j.jkss.2014.01.005
1226-3192/

Extending the class of GARCH-type models under consideration in the paper, we are willing to introduce a conditional mean function in GARCH-type processes. Consider the observable time series $\{X_t\}$ generated by the GARCH-type process $\{\epsilon_t\}$ such that

$$X_t = \mu_t + \epsilon_t, \tag{2}$$

where μ_t denotes the conditional mean function of X_t given the past, i.e., μ_t is F_{t-1} measurable, as defined by $\mu_t = E(X_t|F_{t-1})$. It follows from (2) that there is a one to one correspondence between $\{X_t\}$ and $\{\epsilon_t\}$ and therefore F_t denotes also the sigma field generated by observations $X_t, X_{t-1}, \ldots$. If we choose $\mu_t = \theta X_{t-1}$ and $h_t = \alpha_0 + \alpha_1\epsilon_{t-1}^2 + \beta_1 h_{t-1}$, $\{X_t\}$ is referred to as an AR(1)–GARCH(1, 1) model, indicating a combination of AR(1) in the conditional mean μ_t and the GARCH(1, 1) in the conditional variance h_t. In this article, we do not specify functional form of μ_t and h_t and thus we are dealing with a general GARCH-type process.

In the parameter estimation for general GARCH-type processes defined in (2), it is usually the case that the exact likelihood is hardly known to us and/or it is too complicated for practical purposes. Equivalently, no specific distributional assumptions are made on the iid innovation process $\{\epsilon_t\}$ other than zero mean and unit variance. Then, the quasi-maximum likelihood (QML) estimator is obtained by maximizing an objective function of pseudo-likelihood which is possibly falsely specified likelihood. Often, the pseudo-likelihood is taken via Gaussian innovations, standardized t-distributions with unknown degrees of freedom, and generalized error distributions (refer to, e.g. Tsay (2010, Ch. 10)). Asymmetric (and leptokurtic) distributions for the innovation process $\{e_t\}$ have recently been paid attention including skewed t-distributions (cf., Fernandez & Steel, 1998). It is obvious that the QML estimator reduces to the maximum likelihood (ML) estimator provided that the pseudo-likelihood coincides with the (unknown) true likelihood.

When the general GARCH-type process $\{X_t\}$ is partially specified only through the first and second order conditional moments (i.e., μ_t and h_t), a systematic and unified approach for a partially specified model is via the so-called quasilikelihood (QL). We refer to, among others, Godambe (1985), Heyde (1997) and Hwang and Basawa (2011a) for a background on quasilikelihood in the context of stochastic processes. It is shown that provided that $\mu_t = 0$, the QML based on the Gaussian innovation is essentially the same as the QL (see Proposition 1) and thus two terms QML and QL are typically used interchangeably in the GARCH context. The question that arises naturally is as to the difference between the QML and QL when the QML is based on non-Gaussian innovations. In this paper, we clarify the difference (if any) between the QML and QL for general GARCH-type processes. It will be verified that the QML differs with the QL in terms of conditional skewness and kurtosis (see Theorem 2). Asymptotics for the QML and QL are obtained and then compared with each other (see Theorems 3 and 4). A simulation study is conducted to see finite sample properties of the QL and QML under various innovation distributions including a class of skewed t-distributions. The main goal of the paper is to discriminate and compare QL and QML for general GARCH-type processes.

2. The model description and preliminary results

We are concerned with the general GARCH-type process defined by

$$X_t = \mu_t(\theta) + \epsilon_t \tag{3}$$

where $\mu_t(\theta) = E(X_t|F_{t-1})$ and $\{\epsilon_t\}$ denotes a conditionally heteroscedastic time series. Here, θ is a parameter (vector) indexing the conditional mean function $\mu_t(\theta)$. Baek, Park, and Hwang (2012) discussed a test for fit and evaluated misspecification effects for the model (3). For modeling the conditional mean function $\mu_t(\theta)$, one may include the standard AR (autoregressive) model, threshold AR process and exponential AR model. See for instance Tong (1990) for various nonlinear autoregressive processes. Although higher order models are allowed in our model (3), we consider first order models only, for simplicity of presentation.

- *Threshold autoregressive (TAR) process*: the first-order threshold autoregressive process (TAR(1)) is defined by

$$\mu_t(\theta) = \theta_1 X_{t-1}^+ + \theta_2 X_{t-1}^-, \tag{4}$$

  where $x^+ = \max(x, 0)$ and $x^- = \max(-x, 0)$ so that $x = x^+ - x^-$. We refer to Tong (1990) for a comprehensive treatment of TAR models including higher order processes. It is noted that $\theta_1 = -\theta_2$ $(= \theta$, say) in TAR(1) reduces to a standard AR(1) process.
- *Exponential autoregression (EAR)*: one may consider

$$\mu_t(\theta) = \left[\theta_1 + \theta_2 \exp\left(-\theta_3 X_{t-1}^2\right)\right] X_{t-1} \tag{5}$$

  which is referred to as a first-order exponential autoregressive process (EAR(1)). Note that $\theta_2 = 0$ in EAR(1) gives AR(1) model.

For various models for conditional variance h_t, we refer to, among others, Choi et al. (2012) and Andersen et al. (2009) with references therein. See, e.g., Eq. (1.3) of Straumann (2005) for accommodating broader class of models for h_t. Refer also to Table 5 of Zivot (2009) for various asymmetric GARCH models. One may generate diverse GARCH type processes by combining the mean function $\mu_t(\theta)$ and variance function h_t.

Journal of the Korean Statistical Society 43 (2014) 475–482

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

journal homepage: www.elsevier.com/locate/jkss

Non-stationary quasi-likelihood and asymptotic optimality

S.Y. Hwang [a,*], Tae Yoon Kim [b]

[a] Department of Statistics, Sookmyung Women's University, Seoul, Republic of Korea
[b] Department of Statistics, Keimyung University, Daegu, Republic of Korea

ARTICLE INFO

Article history:
Received 10 October 2013
Accepted 15 February 2014
Available online 5 March 2014

AMS 2000 subject classifications:
primary 62M10
secondary 62P10

Keywords:
Asymptotic optimality
Non-stationary quasi-likelihood
Non-stationary ARCH
Branching Markov processes
Non-stationary random-coefficient AR

ABSTRACT

This article is concerned with non-stationary time series which does not require the full knowledge of the likelihood function. Consequently, a quasi-likelihood is employed for estimating parameters instead of the maximum (exact) likelihood. For stationary cases, Wefelmeyer (1996) and Hwang and Basawa (2011a,b), among others, discussed the issue of asymptotic optimality of the quasi-likelihood within a restricted class of estimators. For non-stationary cases, however, the asymptotic optimality property of the quasi-likelihood has not yet been adequately addressed in the literature. This article presents the asymptotic optimal property of the non-stationary quasi-likelihood within certain estimating functions. We use a random norm instead of a constant norm to get limit distributions of estimates. To illustrate main results, the non-stationary ARCH model, branching Markov process, and non-stationary random-coefficient AR process are discussed.

1. Introduction

When the likelihood of the data from a stochastic process is known to us, the likelihood-based methods are readily available for estimating parameters. A unified approach for asymptotic optimal inference based on the likelihood (from a stationary process) is to use the general framework of local asymptotic normality (LAN). Refer to, among others, Hall and Mathiason (1990) and Wefelmeyer (1996) for an excellent review of the LAN. The LAN is extended to local asymptotic mixed normality (LAMN) to suit non-stationary processes. Asymptotic optimality properties of the maximum (exact) likelihood estimates under various criteria are well documented in the literature via the LAN and the LAMN for the stationary case and non-stationary case, respectively. See, for instance, Basawa and Scott (1983), Hwang and Basawa (2011a) and Hwang, Basawa, Choi, and Lee (2013) for the various asymptotic optimal properties of the maximum (exact) likelihood estimates via the LAN and LAMN approaches.

In this article, we suppose that the likelihood function is *unknown*. For instance, in a time series, the error distribution may not be known and/or the time series is specified only through a first few conditional moments. A quasi-likelihood (QL, for short) method in the context of estimating functions is suited to cases of unknown likelihood (cf. Heyde, 1997). Introducing the Godambe information criterion, Godambe (1985) established finite sample optimality of the QL within a certain class of estimating functions. To establish asymptotic optimality properties of the QL, one needs to impose constraints on the class of estimators among which the quasi-likelihood performs "best". Readers refer to, e.g., Wefelmeyer (1996) and Hwang and Basawa (2011b, 2014) for the stationary QL case. However, for the non-stationary QL case, the issue of asymptotic optimality has not been adequately pursued in the literature. Our main goal is to establish certain asymptotic optimality of

* Corresponding author. Tel.: +82 1091809078.
E-mail address: shwang@sm.ac.kr (S.Y. Hwang).

http://dx.doi.org/10.1016/j.jkss.2014.02.004
1226-3192/

the non-stationary QL within a restricted class of estimating functions (see Section 2 for details). To illustrate main results, the non-stationary ARCH model, branching Markov process and non-stationary random coefficient AR model are discussed.

2. Non-stationary quasi-likelihood: main results

We are concerned with the non-stationary stochastic process $\{X_t,\ t = 0, 1, 2, \ldots\}$ with the starting value $X_0 = x_0$. The probability measure associated with $\{X_t\}$ involves a $(k \times 1)$ vector parameter θ which belongs to an open subset Θ of R^k, the k-dimensional Euclidean space. It is noted that θ denotes a parameter of interest, allowing nuisance parameter η in addition to θ. Semiparametric cases can also be included by treating the error distribution as η. This will be illustrated via examples in Section 3. Let $\{X_1, X_2, \ldots, X_n\}$ denote the sample. Suppose that the likelihood of the data is *unknown* and we are interested in estimating θ based on the sample. We are then led to the theory of estimating functions instead of the likelihood-based methods. Throughout the paper, the estimating function is shortened as EF. We regard quasi-likelihood (QL) as a member of certain class of EFs within which the asymptotic optimality property of the QL can be derived. To proceed, one needs to restrict the class of EFs under consideration in order to discuss the optimality of the non-stationary QL.

Godambe (1985) introduced a "linear" class G of EFs $G_n(\theta) : (k \times 1)$ defined by

$$G = \left\{ G_n(\theta) = \sum_{t=1}^{n} w_t(\theta) g_t(\theta) \right\} \tag{2.1}$$

where $\{g_t(\theta)\}$ is a sequence of martingale differences with respect to F_t (here and in what follows, F_t denotes the σ-field generated by $X_t, X_{t-1}, \ldots, X_1$ for each $t \geq 1$). In (2.1), $g_t(\theta)$ is a pre-specified martingale difference which is referred to as an innovation at time t, that is, $E(g_t(\theta)|F_{t-1}) = 0$. And $w_t(\theta)$ is an F_{t-1} measurable weight vector of order $(k \times 1)$. The class G is generated by varying the "coefficient" $w_t(\theta)$ while the given innovation $g_t(\theta)$ being fixed. We use the notation E_{t-1} for denoting the conditional expectation given F_{t-1}, viz., $E_{t-1}(\cdot) = E(\cdot|F_{t-1})$ so that $E_{t-1}g_t(\theta) = 0$. It is assumed that $E_{t-1}[g_t^2(\theta)] < \infty$ and $E_{t-1}[\partial g_t(\theta)/\partial\theta] \neq 0$. As a special member of G, the quasi-likelihood (QL) $G_n^*(\theta)$ is defined by

$$G_n^*(\theta) = \sum_{t=1}^{n} E_{t-1}[\partial g_t(\theta)/\partial\theta](E_{t-1}g_t^2(\theta))^{-1} g_t(\theta). \tag{2.2}$$

Although it would be more precise to use the term QL-EF for $G_n^*(\theta)$, we shall refer to $G_n^*(\theta)$ as QL for simplicity. In this paper, the superscript $*$ is reserved for denoting terms related to the quasi-likelihood $G_n^*(\theta)$. Due to Godambe (1985), the QL $G_n^*(\theta)$ enjoys maximum of the so-called Godambe's information matrix among the class G of EFs $G_n(\theta)$. Refer also to, e.g., Hwang and Basawa (2011b) and Thavaneswaran, Liang, and Frank (2012) for the expression of Godambe's information matrix. As is noted in, e.g., Hwang and Basawa (2011b), the optimality of QL $G_n^*(\theta)$ established by Godambe (1985) is for the estimating function itself and does not lead to any finite or asymptotic optimality of the estimator derived from the QL equation $G_n^*(\theta) = 0$. Dealing with the estimators directly (rather than EFs themselves), let $\hat{\theta}_{\mathrm{QL}}$ denote a consistent solution of the QL equation $G_n^*(\theta) = 0$. For stationary cases, Chandra and Taniguchi (2001), Hwang and Basawa (2011b, 2014) and Wefelmeyer (1996) showed that the asymptotic variance of $\sqrt{n}(\hat{\theta}_{\mathrm{QL}} - \theta)$ achieves the "minimum" among all the estimators derived from EFs in the class G, under appropriate conditions.

To complement the literature, this article covers non-stationary cases. To do this, we use a random norm for $\hat{\theta}_{\mathrm{QL}}$ instead of a constant norm $\sqrt{n}$. Consider $G_n(\theta) = \sum_{t=1}^{n} w_t(\theta) g_t(\theta) \in G$ for which the sum of conditional covariance matrices $V_n(\theta)$ is defined by

$$V_n(\theta) = \sum_{t=1}^{n} w_t(\theta) w_t^T(\theta) E_{t-1} g_t^2(\theta) : (k \times k) \tag{2.3}$$

where the superscript T denotes "transpose". We shall regard two EFs in G as being identical if one is a constant multiple of the other. It will be assumed that $\|V_n(\theta)\| \to \infty$ almost surely as the sample size n tends to infinity. Here, $\|\cdot\|$ is used to denote the matrix (or vector) norm, e.g., $\|A\|^2 = \mathrm{trace}(A^T A)$. The half matrix of a symmetric positive definite matrix A is denoted by $A^{1/2}$ obtained from the spectral decomposition of A. The inverse of $A^{1/2}$ is denoted by $A^{-1/2}$. Denote by I_k the identity matrix of order k. Fix $\theta \in \Theta$ and define the random local neighborhood $N_\delta(\theta)$ about θ.

$$N_\delta(\theta) = \{\bar{\theta};\ \|V_n^{1/2}(\theta)(\bar{\theta} - \theta)\| < \delta\}, \quad \text{for some } \delta > 0 \tag{2.4}$$

Note that $N_\delta(\theta)$ reduces almost surely to θ as $n \to \infty$. From now on, we confine ourselves to "Regular EFs" in the class G.

Definition. Any EF $G_n(\theta) \in G$ is called regular if it satisfies the following three conditions.
(C1) For each fixed $\theta \in \Theta$,

$$\sup \|V_n^{-1/2}(\theta)[\partial G_n(\bar{\theta})/\partial\theta^T - \partial G_n(\theta)/\partial\theta^T]V_n^{-1/2}(\theta)\| = o_p(1)$$

where the "sup" is taken over $\bar{\theta} \in N_\delta(\theta)$ and $o_p(1)$ stands for a term converging to zero in probability.

Statistics, 2014
Vol. 48, No. 3, 487–507, http://dx.doi.org/10.1080/02331888.2012.748772

Taylor & Francis
Taylor & Francis Group

Non-ergodic martingale estimating functions and related asymptotics

S.Y. Hwang[a]*, I.V. Basawa[b], M.S. Choi[a] and S.D. Lee[c]

[a]*Department of Statistics, Sookmyung Women's University, Seoul, South Korea;* [b]*Department of Statistics, University of Georgia, Athens, GA, USA;* [c]*Department of Statistics, Chungbuk National University, Cheongju, South Korea*

(*Received 17 September 2011; final version received 24 September 2012*)

Well-known estimation methods such as conditional least squares, quasilikelihood and maximum likelihood (ML) can be unified via a single framework of martingale estimating functions (MEFs). Asymptotic distributions of estimates for ergodic processes use constant norm (e.g. square root of the sample size) for asymptotic normality. For certain non-ergodic-type applications, however, such as explosive autoregression and super-critical branching processes, one needs a *random norm* in order to get normal limit distributions. In this paper, we are concerned with non-ergodic processes and investigate limit distributions for a broad class of MEFs. Asymptotic optimality (within a certain class of non-ergodic MEFs) of the ML estimate is deduced via establishing a convolution theorem using a random norm. Applications to non-ergodic autoregressive processes, generalized autoregressive conditional heteroscedastic-type processes, and super-critical branching processes are discussed. Asymptotic optimality in terms of the maximum *random limiting power* regarding large sample tests is briefly discussed.

Keywords: non-ergodic process; martingale estimating functions; branching process; asymptotic optimality; explosive autoregressive model

2000 MSC Classification: Primary 62M10

1. Introduction

Estimating functions corresponding to standard estimation methods such as conditional least squares (CLS), quasilikelihood (QL), and maximum likelihood (ML) are special cases of martingale estimating functions (MEF, hereafter) (refer to Section 2 to follow; see, for instance, [1]). Consistent and asymptotically normal estimators are usually obtained by equating MEFs to zero, that is, by solving the martingale estimating equations (MEEs, for short). Most of the literature on MEF (or MEE) asymptotics has been directed to ergodic processes for which a *non-random norm* is used to get asymptotic normal distributions of the estimates. For ergodic stationary processes, a constant norm $\sqrt{n}$ (throughout the paper, n denotes the sample size) is often used to get asymptotic normal distributions. For a comprehensive discussion of estimating functions from ergodic processes, see among others [1–5]. In particular, refer to [6] for a rigorous treatment of MEFs from a general ergodic Markov processes. On the other hand, for non-ergodic-type processes (refer to [7, Preface] for a definition of non-ergodic process), limit distributions of standard

*Corresponding author. Email: shwang@sm.ac.kr

estimators (including CLS, QL, and ML) are mixed normal when a non-random norm is used. Instead, a *random norm* is required to get normal limit distributions for non-ergodic processes. Asymptotics of various statistics normalized by random norms in a broad context have recently been discussed by Pena *et al.* [8]. The non-ergodic process is distinguished from ergodic process by the random norm (other than non-random norm) required to get normal limits. Refer to [7] for various non-ergodic processes including normal mixture models, explosive autoregressive processes, birth processes, branching processes, and non-stationary diffusion processes. Some specific examples of non-ergodic processes are discussed in detail in Section 4.

As commented in [3], rigorous treatments of asymptotics on MEFs in the context of non-ergodic processes have not yet been adequately explored in the literature. In this paper, we consider large sample inference based on MEFs for a class of non-ergodic processes. Under a mild set of regularity conditions, asymptotic distributions of estimators obtained by solving MEEs are derived. Furthermore, we show that the maximum-likelihood estimator (MLE) is asymptotically optimum in a sense of 'minimum variance' within a class of estimators obtained from MEEs. We use a natural random norm in our asymptotics. A convolution theorem (using a random norm) is established which yields the desired optimality (refer to Theorems 3 and 4). These results are illustrated with specific applications to non-ergodic processes. Section 2 introduces a general class of MEFs and derives the limit distributions of the estimators therefrom. The convolution theorem for non-ergodic MEFs is established in Section 3. Three specific applications (one ergodic and two non-ergodic models) are discussed in Section 4. Section 5 addresses large sample tests based on MEFs and asymptotic optimality of tests is discussed in terms of the 'maximum' *random limiting power*. One of the main goals is to provide a simplified but a broad treatment of asymptotic theory on MEFs from non-ergodic processes.

2. MEFs and asymptotic distributions

Let $\{X_t, t = 0, 1, 2, \ldots\}$ denote a possibly non-ergodic stochastic process with initial value $X_0 = x_0$. Suppose that the probability measure associated with $\{X_t\}$ is indexed by a $(k \times 1)$ vector parameter θ taking values in Θ which is an open subset of the k-dimensional Euclidian space R^k. Based on a sample of observations $X_1, X_2, \ldots, X_n$, we are interested in estimating the parameter vector θ. Let F_t denote the σ-field generated by $X_t, X_{t-1}, \ldots, X_1$. Use $p_t(\theta)$ to denote the conditional density of X_t given F_{t-1} and consider the following MEF $U_n(\theta)$ given by

$$U_n(\theta) = \sum_{t=1}^{n} u_t(\theta) : (k \times 1) \text{ vector}, \tag{1}$$

where $\{u_t(\theta)\}$ is a sequence of martingale differences satisfying $E(u_t(\theta)|F_{t-1}) = 0$. It will be assumed throughout that the $(k \times 1)$ vector $u_t(\theta)$ is differentiable (with respect to θ) and differentiability under integral sign is valid. Denote by ξ_n the sum of conditional covariance matrices corresponding to $u_t(\theta)$, that is,

$$\xi_n = \sum_{t=1}^{n} \mathrm{Var}(u_t(\theta)|F_{t-1}) = \sum_{t=1}^{n} E(u_t(\theta)u_t^{\mathrm{T}}(\theta)|F_{t-1}) : (k \times k) \text{ matrix}. \tag{2}$$

We suppress the dependence of ξ_n on θ whenever necessary, that is, $\xi_n = \xi_n(\theta)$, for notational simplicity. In addition, ξ_n is assumed to be positive definite almost surely. Throughout, the matrix (or vector) norm will be denoted by $\|\cdot\|$, viz., for any matrix A, $\|A\|^2 = \mathrm{tr}(A^{\mathrm{T}}A) = \mathrm{tr}(AA^{\mathrm{T}})$. Here, A^{T} is the transpose of A. For simplicity, for any non-singular matrix A, $A^{-\mathrm{T}}$ is used for denoting $A^{-\mathrm{T}} = (A^{-1})^{\mathrm{T}} = (A^{\mathrm{T}})^{-1}$. The identity matrix of order k is denoted by I_k. In addition, for a positive

Journal of the Korean Data & Information Science Society
2014, **25**(5), 1137–1149

http://dx.doi.org/10.7465/jkdi.2014.25.5.1137
한국데이터정보과학회지

Contemporary review on the bifurcating autoregressive models : Overview and perspectives[†]

S. Y. Hwang[1]

[1]Department of Statistics, Sookmyung Women's University
Received 5 July 2014, revised 1 August 2014, accepted 13 August 2014

Abstract

Since the bifurcating autoregressive (BAR) model was developed by Cowan and Staudte (1986) to analyze cell lineage data, a lot of research has been directed to BAR and its generalizations. Based mainly on the author's works, this paper is concerned with a contemporary review on the BAR in terms of an overview and perspectives. Specifically, bifurcating structure is extended to multi-cast tree and to branching tree structure. The AR(1) time series model of Cowan and Staudte (1986) is generalized to tree structured random processes. Branching correlations between individuals sharing the same parent are introduced and discussed. Various methods for estimating parameters and related asymptotics are also reviewed. Consequently, the paper aims to give a contemporary overview on the BAR model, providing some perspectives to the future works in this area.

Keywords: Bifurcating autoregression, branching correlation, branching tree, estimating function, multi-cast model.

1. Introduction

Cowan and Staudte (1986) suggested bifurcating autoregressive (BAR) models to analyze binary splitting data where each individual (mother) gives rise to exactly two offspring (daughters) in the next generation. Powell's data, G15 + C data and cell lineage data are typically of this kind (cf. Huggins and Basawa, 1999). Figure 1.1 shows a binary splitting (that is, bifurcating) data consisting of 7 observations. See, for instance, Hwang and Basawa (2009). Let X_i be an observation of individual i. For example, X_i may represent blood pressure, cholesterol level, dose of a drug and protein content of a cell i, etc. In a bifurcating model, i produces $(2i, 2i+1)$. Cowan and Staudte (1986) introduced the first order BAR(1) model recursively defined by, with the initial observation $X_1 = x_1$,

$$\begin{aligned} X_{2i} &= \theta X_i + \epsilon_{2i} \\ X_{2i+1} &= \theta X_i + \epsilon_{2i+1} \end{aligned} \tag{1.1}$$

where $|\theta| < 1$ and $(\epsilon_{2i}, \epsilon_{2i+1})$ is a sequence of iid bivariate random vectors with common mean zero, common variance σ^2 and $Corr(\epsilon_{2i}, \epsilon_{2i+1}) = \rho$. It is noted that the BAR(1) model views each line of descendants as a standard AR(1) time series.

[†] This research was supported by a grant (2013) from Sookmyung Women's University.
[1] Professor, Department of Statistics, Sookmyung Women's University, Seoul 140-742, Korea.
E-mail: shwang@sm.ac.kr

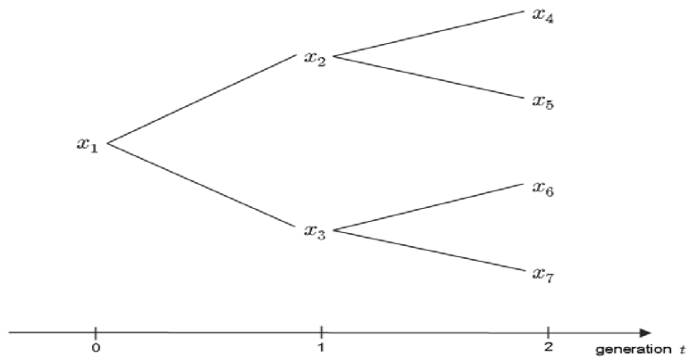


Figure 1.1 A binary splitting (bifurcating) data consisting of 7 observations

Three issues are discussed to generalize BAR(1) time series (1.1). First issue is on the nature of trees (as in Figure 1.1) indexing the data. Specification of the time series models defined on the tree may be the second issue. The correlation (conditionally on their mother) between sisters is referred to as the branching correlation (cf. Hwang, 2011). It is noted in the BAR(1) formulation (1.1) that ρ represents a correlation between sisters, conditionally on their mother. That is, the branching correlation of the BAR(1) models is given by

$$\rho = Corr(X_{2i}, X_{2i+1}|X_i). \tag{1.2}$$

Mathematical modeling of the branching correlation is discussed as the third issue in generalizing BAR(1) time series. Details of the three issues will be provided next.

Regarding the parameter estimation for BAR(1) and its generalizations, it is resonable to assume that the exact likelihood is not known. Equivalently, no specific distributional assumptions are made on $(\epsilon_{2i}, \epsilon_{2i+1})$. Then, quasi-maximum likelihood (QML) estimator is obtained by maximizing the pseudo-objective function which is taken as Gaussian bivariate innovations $(\epsilon_{2i}, \epsilon_{2i+1})$. When BAR(1) process X_i is partially specified only through the first and second order conditional moments, a systematic approach for a partially specified model is via the so called quasilikelihood (QL). Due to the wideness of the generalized structures discussed in this paper, it will be appropriate to use a QL estimation. We refer to, for instance, Heyde (1997) and Hwang *et al.* (2014b) for a background on QL and QML in a broader context of stochastic processes.

2. From the bifurcating tree to multicast and branching tree

Many applications require a multicast tree in which each node may have m branches where m may be larger than 2. For instance, in an internet traffic data, a package (signal) may be sent along m links and delay times X over each link may be of interest. One may then extend bifurcating structure to a multicast tree where each individual splits exactly into m-offspring ($m \geq 1$). A tri-casting ($m = 3$) data is illustrated in Figure 2.1. See for instance, Hwang and Choi (2009) and Hwang and Kang (2012).

JOURNAL OF MEDICINAL FOOD
J Med Food 17 (1) 2014, 76–82

DOI: 10.1089/jmf.2013.3054

Laminaria japonica Combined with Probiotics Improves Intestinal Microbiota: A Randomized Clinical Trial

Seok-Jae Ko,[1] Jinsung Kim,[1] Gajin Han,[1] Seul-Ki Kim,[1] Hong-Geol Kim,[1] Inkwon Yeo,[2] Bongha Ryu,[1] and Jae-Woo Park[1]

[1]*Department of Internal Medicine, College of Korean Medicine, Kyung Hee University, Seoul, Republic of Korea.*
[2]*Department of Statistics, College of Sookmyung Women's University, Seoul, Republic of Korea.*

ABSTRACT *Laminaria japonica*—a widely used ingredient in seaweed kimchi—and lactic acid bacteria (LAB)—a main component of traditional fermented Korean food—may alter human intestinal microbiota composition and have a positive effect on various digestive problems. However, few clinical trials have investigated the potential benefits of *L. japonica* when combined with LAB for human intestinal microbiota. Therefore, this study was designed to evaluate the effects of *L. japonica* and representative LAB on the human intestine. Forty participants with no known digestive diseases were randomly assigned to one of the two combination groups: (1) *L. japonica* with LAB and (2) *L. japonica* with placebo LAB. The study agents were administered for 4 weeks with a 2-week follow-up period. The primary outcome measure was the number of each of the seven LAB species in the human intestine, and the secondary outcome measures included the Korean version of the Gastrointestinal Symptom Rating Scale, the World Health Organization Quality of Life, and bowel functions. The primary outcome was evaluated before and after administration of the study agents (0 and 4 weeks), and the secondary outcomes were evaluated at 0, 4, and 6 weeks. Four of the seven LAB species were found to be significantly increased in the *L. japonica* with the LAB group and five species were significantly different from those of the placebo group. The secondary outcome measures did not change significantly. In conclusion, *L. japonica* with LAB facilitated the proliferation of beneficial human intestinal microbiota. (Trial number: ClinicalTrials.gov NCT01651741).

KEY WORDS: • ***health functional food*** • ***intestine*** • ***probiotics*** • ***seaweeds***

INTRODUCTION

KOREAN NATIONAL CUISINE has evolved based on the perspective that ''the root of medicine and food is equal,'' and is largely founded on fermented food such as kimchi, gochujang (fermented red chili paste), and *doenjang* (fermented bean paste). Korean food is known internationally for its high nutrition, diversity, and taste.[1] Kimchi, the representative traditional Korean fermented food, is made with vegetables such as Chinese cabbage. It is seasoned with a variety of ingredients, including garlic, green onion, and red pepper. The main vegetable ingredients and the degree of fermentation affect the taste of kimchi and determine the different kimchi varieties. *Laminaria japonica*, an edible seaweed, has been used as one of the main ingredients in kimchi for many years. It contains various factors, including alginate and fucoidans, which are known for their dietary benefits and therapeutic effects against constipation[2] and inflammation.[3] Because of the beneficial effect of *L. japonica*, its consumption has increased annually, reaching a mean 5 kg/year for an ordinary Korean adult.[4,5]

Lactic acid bacteria (LAB) are primarily responsible for the fermentation of kimchi and more than 10^8 colony-forming units of LAB are involved.[6] During the early stages of kimchi fermentation, the LAB profile is comparatively diverse; however, in the later stages, the LAB profile becomes simpler and consists mostly of *Lactobacillus plantarum*.[7] LAB have been the focus of international research for their potentially beneficial roles in stabilizing the gut microflora.[8] LAB are also effective for preventing various digestive problems such as colon cancer[9] and chronic diarrhea.[10]

Recent research on *L. japonica* and LAB has shown that *L. japonica* with LAB possesses a high antioxidant activity and protects against hepatic damage, obesity, hypertension, stress, insomnia, and alcoholic liver damage.[11,12] However, although the beneficial effects of *L. japonica* with LAB are well known, few clinical trials have investigated its effect in human intestine.

This study aimed to evaluate the effect of *L. japonica* with LAB as representative compounds of seaweed kimchi. We examined the microbial communities in the healthy human intestine before and after the administration of *L. japonica*

Manuscript received 2 September 2013. Revision accepted 27 November 2013.

Address correspondence to: Jae-Woo Park, PhD, KMD, Department of Internal Medicine, College of Korean Medicine, Kyung Hee University, Seoul 130-701, Republic of Korea, E-mail: pjw2907@khu.ac.kr

with LAB. We also compared the number of intestinal microbiota between the two groups (*L. japonica* with LAB and *L. japonica* with the placebo LAB group). Additionally, questionnaires on gastrointestinal symptoms and quality of life were used to investigate the combined effect of *L. japonica* with LAB.

MATERIALS AND METHODS

Subjects and study design

Forty-three volunteers with no history of disease, aged 18–75 years were recruited from the local community through bulletin board advertising, brochures, and banners. Three subjects failed in the screening test and the remaining 40 participants were enrolled in the study. The current study was reviewed by the Institutional Review Board and Ethics Committee of the Kyung Hee University Hospital at Gangdong (KHNMC-OH-IRB 2012-005). The trial followed the standards of the International Committee on Harmonization of Good Clinical Practice and the revised version of the Declaration of Helsinki. All participants were asked to fill out informed consent forms, and they were allowed adequate time to decide whether they would participate in the trial before signing the consent.

The study was performed as a placebo-controlled and double-blinded trial with 40 participants randomly allocated to two groups: (1) *L. japonica* with Duolac7S (DUO) and (2) *L. japonica* with placebo DUO. Participants were required to complete a 4-week treatment period (weeks 0–4) and a 2-week follow-up period (weeks 4–6). The flowchart of the entire trial is shown in Figure 1.

All eligible participants were free of diseases and clinical symptoms, according to the Korean version of Gastrointestinal Symptom Rating Scale (KGSRS) with scores of < 3 points and no history of abdominal surgery related to the digestive system. Subjects who took over-the-counter medications that affected gastrointestinal motility or antibiotics, herbal medicine, or probiotics 2 weeks before the start of the trial were excluded. Pregnant women or women who planned to become pregnant during the trial period were also excluded. The study participants were asked not to take any medication that could affect the gastrointestinal tract during the trial, and if medications were taken, the drug name, usage, and duration were recorded on the case report form.

An independent statistician was responsible for the assignment of random numbers that were generated through a creation program. An opaque sealed envelope containing the table of random serial numbers was kept in a locked cabinet. The participants, investigator, and clinical research coordinator (CRC) were blinded to the randomization, and the blinding procedure was monitored by an authorized clinical research associate.

All participants were asked to report any adverse events to the principal investigator or CRC during the entire trial. The values for serum complete blood count, blood urea nitrogen, creatinine, aspartate aminotransferase, alanine aminotransferase, gamma-glutamyltransferase levels, and erythrocyte sedimentation rate were assessed to confirm the safety of the study agents after the treatment period. The safety test analyses were performed at an accredited laboratory, and any adverse events and safety issues were documented on the case report form.

Interventions

The agents used in this trial were water-extracted *L. japonica* (LJE) and DUO. The LJE tablet was an oval-shaped,

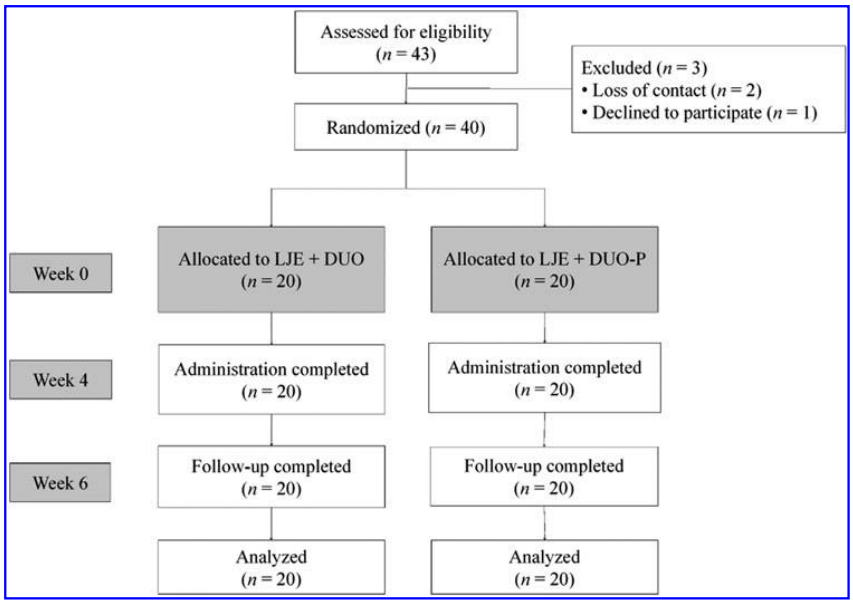


FIG. 1. Flow chart of the trial. LJE, *Laminaria japonica* extract; DUO, Duolac7S; DUO-P, Placebo of Duolac7S.

Using a bimodal kernel for a nonparametric regression specification test

Cheolyong Park, Tae Yoon Kim, Jeongcheol Ha, Zhi-Ming Luo and Sun Young Hwang

Keimyung University and Sookmyung Womens University

Abstract: For a nonparametric regression model with a fixed design, we consider the model specification test based on a kernel. We find that a bimodal kernel is useful for the model specification test with a correlated error, whereas a conventional unimodal kernel is useful only for an iid error. Another finding is that the model specification test suffers from a convergence rate change depending on whether the errors are correlated or not. These results are verified by deriving an asymptotic null distribution and asymptotic (local) power, and by performing a simulation. The validity of the bimodal kernel for testing is demonstrated with the "drum roller" data (see Laslett(1994) and Altman (1994)).

Key words and phrases: Nonparametric specification test, correlated error, bimodal kernel, convergence rate change.

1. Introduction

Suppose that we are concerned with the nonparametric regression function specification test given

$$Y_i = m(x_i) + \eta_i \quad (i = 1, \ldots, n), \tag{1.1}$$

where m is a smooth function defined on [0,1], $x_i = i/n$, and $\{\eta_i\}$ is a zero-mean, covariance stationary process. Here, the design points grow closer as the sample size increases; while, the error process remains the same. Refer to Kim *et al.* (2009) and references therein for detailed discussions about this model. The testing problem under consideration is whether $m(x)$ belongs to a specific parametric family. This can be described as

$$H_0 : m(x) = g(x, \gamma_0) \text{ for all } x \in [0, 1] \text{ with some } \gamma_0 \in \mathcal{B} \subset R^q$$

versus

$$H_1 : m(x) \neq g(x, \gamma) \text{ for some } x \in [0, 1] \text{ with all } \gamma \in \mathcal{B} \subset R^q.$$

For testing H_0 nonparametrically, we consider a kernel-based test statistic

$$T_n = (n^2h)^{-1}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} K((x_i - x_j)/h)\hat{\epsilon}_i\hat{\epsilon}_j, \tag{1.2}$$

where K is a kernel satisfying (C1) below, $\epsilon_i = Y_i - g(x_i, \gamma)$ and $\hat{\epsilon}_i = Y_i - g(x_i, \hat{\gamma})$, with a consistent estimator $\hat{\gamma}$ of γ_0. Here, T_n is based on the average squared error

$$d_A(\hat{m}, m) = n^{-1}\sum_{i=1}^{n}(\hat{m}(x_i) - m(x_i))^2, \tag{1.3}$$

where $\hat{m}(x) = (nh)^{-1}\sum_{i=1}^{n} K((x - x_i)/h)Y_i$, In recent years, much research has been performed on applying T_n to the model specification test for a random design regression model. See, for example, Fan and Li (1996), Zheng (1996), Luo et al. (2011), and Khmaladze and Koul (2004). For the model specification test for the fixed design regression model, not many results are available. See, e.g., Eubank and Spiegelman (1990) or the monograph by Hart (1997), which studies nonparametric lack-of-fit test with iid errors. We study T_n as a nonparametric regression specification test for the fixed design regression model, particularly when errors are correlated. The major strengths of T_n is its consistency, since the existing parametric tests fail to be consistent against all deviations from the null.

It is well known that when a nonparametric method such as $\hat{m}$ is used to recover m, correlated errors cause trouble. See Opsomer et al. (2001) for a detailed discussion of this. We demonstrate that an analogous size distortion problem arises for a nonparametric specification test when errors are correlated. As a possible solution, we recommend the use of bimodal kernel K with $K(0) = 0$. In addition, we find that T_n shows a power rate change, that yields a continuous but non-monotonic power function over the hypothesis domain. We proceed as follows. Section 2 proposes the specification test with bimodal kernel and demonstrates its usefulness with the "drum roller" data (Laslett (1994) and Altman (1994)). Section 3 discusses the power rate change of T_n and its impact. Some other tests are discussed there. Section 4 reports on simulations that check our theoretical results. All proofs are deferred to the Appendix.

2. Size distortion and bimodal kernel

The Korean Journal of Applied Statistics (2014) DOI: http://dx.doi.org/10.5351/KJAS.2014.27.1.115
27(1), 115–122

A Test for Randomness of the Binary Random Sequence

In-Kwon Yeo[a,1]

[a]Department of Statistics, Sookmyung Women's University

(Received November 18, 2013; Revised December 16, 2013; Accepted January 9, 2014)

Abstract

A test for randomness of the binary random sequence is proposed in this paper. The proposed test statistic is based on the mean length of runs distributed with truncated geometric distribution and asymptotically χ^2_2-distributed when the size of the sequences is large. A small Monte Carlo simulation compared the size of the test with a significant level as well as evaluated the test power. We applied the proposed method to the sequence of yes or no numbers in Lotto 6/45 and concluded that the randomness of Lotto is retained.

Keywords: Length of runs, Lotto 6/45, truncated geometric distribution.

1. 서론

로또 6/45는 사회적으로 사행성, 한탕주의와 같은 많은 부작용도 낳았지만 저소득층이나 소외 계층을 위한 각종 복지사업이나 임대주택 건설 등 다양한 사업의 재원으로 사용되고 있다. 2002년 12월 7일 처음 추첨을 실시된 이후 현재까지 각종 사건과 이슈를 만들어 왔고 많은 사람들에게 인생역전이라는 환상을 주며 운영되고 있다. 엄청난 금액의 누적당첨금으로 로토광풍이 불기도 했으며 당첨번호를 예측하여 회원에게 제공해 주는 회사도 우우죽순처럼 생겨나고 있다. 또한 당첨번호의 조작이나 추첨 후에도 해당 회차의 복권을 발행 할 수 있다는 의혹들이 인터넷이나 언론을 통해 제기되기도 했다.

당첨번호에 대한 균일성(uniformity)이나 무작위성(randomness)에 대한 연구는 Genest 등 (2002), Helman (2005) 등에 의해 연구되었다. 특히 Johnson과 Klotz (1993)는 Lotto America Megabucks Lottery의 200회의 당첨번호를 사용하여 54개 각 번호가 당첨번호에 속할 확률을 최대가능도추정법을 통해 추정하고 이를 이용하여 균일성 검정을 실시하였다. 그 결과 p-값이 0.084로 확실히 균일하다고 보기 어려우며 낮은 번호의 공이 조금 많이 선택되는 것을 보아 이는 매 회 공을 넣을 때 크기순서대로 넣었기 때문일 것으로 추측했다. 우리나라의 경우 Kim (2004)는 71회까지의 당첨번호에서 가장 많이 나온 번호의 표집분포를 몬테칼로 모의실험을 통해 유도하여 균일성에 위배되지 않은 것을 보였으며 Lim과 Baek (2009)는 331회까지의 당첨번호에 대해 순서통계량 평균을 이용한 Coronel-Brizio 등 (2008)의 방법을 이용하여 로또 추첨이 무작위적인 것을 확인했다.

이 논문에서는 임의의 어떤 번호가 매 회에서 전 회에서의 추첨여부와 관계없이 독립적으로 추첨되는지를 검정하는 방법에 대해 알아본다. 이 번호가 어떤 회에서 추첨되었으면 1, 아니면 0이라고 표시하

This Research was supported by the Sookmyung Women's University Research Grants 2012.
[1]Professor, Department of Statistics, Sookmyung Women's University, Chongpa-dong 2ga, Yongsan-gu, Seoul 140-742, Korea. E-mail: inkwon@sm.ac.kr

면 자료는 이진확률수열(binary random sequence) 구조를 가진다. 이러한 수열에서의 독립성은 연검정(run test)을 통해 확인할 수 있으나 연검정의 경우 1이 발생할 확률 π가 0.5일 때 사용된다. 이 논문에서는 π가 임의의 값을 가질 때에도 이진확률수열의 무작위성을 검정하기 위한 검정통계량을 제안하고 이 통계량의 점근적 성질에 대해 알아본다. 이 방법은 로또뿐만 아니라 각종 컴퓨터 프로그램에서 제공하는 난수의 무작위성 여부를 확인하는데 사용될 수 있다. Rukhin 등 (2010)에는 적합도검정, 연검정, 최대연길이 등 15가지의 난수검정방법이 제시되어 있으나 우리가 알아보고자 하는 형태를 가진 이진확률수열의 무작위성을 검정하는 방법을 제시하지 않고 있다.

2. 이진확률수열의 무작위성 검정통계량

$\{X_i\}_{i=1}^n$는 0과 1로 이루어진 이진확률수열이라고 하자. 이 수열이 베르누이 시행에 의한 수열이라고 하면 각각의 X_i는 독립이고 1일 확률이 π가 된다. 로또에서 어떤 번호가 i번째 회에서 당첨번호로 뽑히면 1 아니면 0이라고 하자. 이 때 1이 추출될 확률이 6/45인 것처럼 이 논문에서는 π의 값을 알고 있다고 가정한다. 확률 π가 1/2이라고 하면 연(run)의 개수를 이용하여 독립성을 검정하는 연검정을 실시할 수 있으나 π가 1/2이 아닌 경우에는 연검정으로 독립성을 확인하는데 무리가 있다. 이 논문에서는 연의 개수 대신 연의 길이를 이용하여 π가 1/2이 아닌 경우에도 적용할 수 있는 검정통계량을 소개하고자 한다.

통계량 Y_j^+는 j번째 1로 이루어진 연의 길이, Y_j^-는 j번째 0으로 이루어진 연의 길이라고 하자. 예를 들어, 이진확률수열이 다음과 같이 이루어졌다고 하면

$$111\ 00\ 1\ 00\ 11\ 0\ 1\ 000\ 1$$

$Y_1^+ = 3, Y_2^+ = 1, Y_3^+ = 2, Y_4^+ = 1, Y_5^+ = 1$이 되고 $Y_1^- = 2, Y_2^- = 2, Y_3^- = 1, Y_4^- = 3$이 된다. 이때 n^+와 n^-를 각각 1과 0으로 이루어진 런의 개수라고 하면 위의 예제의 경우 $n^+ = 5$, $n^- = 4$가 된다. 연은 순환구조를 가지기 때문에 $\left|n^+ - n^-\right| \le 1$이 된다. 만약 $\{X_i\}_{i=1}^n$가 베르누이 확률변수라고 하면 각각의 Y_j^+는 0이 나올 때까지 베르누이를 반복할 때의 1의 개수이고 Y_j^-는 1이 나올 때까지의 0의 개수로 n이 무한이 커진다면 기하분포를 이용하여 모형화 할 수 있다. 즉 Y_j^+는 절사된 기하분포로

$$P\left(Y_j^+ = y | Y_j^+ \ge 1\right) = \frac{P\left(Y_j^+ = y\right)}{1 - P\left(Y_j^+ = 0\right)} = \frac{\pi^y(1-\pi)}{1-(1-\pi)} = \pi^{y-1}(1-\pi), \quad y = 1, 2, \ldots$$

가 된다. 마찬가지로 Y_j^-의 분포는 다음과 같다.

$$P\left(Y_j^- = y | Y_j^- \ge 1\right) = \pi(1-\pi)^{y-1}, \quad y = 1, 2, \ldots,$$

여기서 Y_j^+와 Y_j^-의 평균은 각각 $1/(1-\pi)$과 $1/\pi$이고 분산은 $\pi/(1-\pi)^2$와 $(1-\pi)/\pi^2$가 된다. 만약 n^+와 n^-가 커지면 중심극한정리에 의해

$$Z_+ = \frac{\bar{Y}^+ - 1/(1-\pi)}{\sqrt{\pi}/(\sqrt{n^+}(1-\pi))} = \frac{(1-\pi)\bar{Y}^+ - 1}{\sqrt{\pi/n^+}}, \qquad Z_- = \frac{\bar{Y}^- - 1/\pi}{\sqrt{1-\pi}/(\sqrt{n^-}\pi)} = \frac{\pi\bar{Y^-} + 1}{\sqrt{(1-\pi)/n^-}}$$

는 점근적으로 표준정규분포를 따른다. 여기서 $\bar{Y}^+ = \sum_j Y_j^+/n^+$이고 $\bar{Y}^- = \sum_j Y_j^-/n^-$이다.

표본크기 n이 유한하더라도 매우 크다면 근사적으로 위의 결과들을 활용할 수 있다. 여기서 한 가지 주의해야 할 것은 실제분석에 있어는 n이 유한하기 때문에 마지막 연의 길이는 더 길어질 수 있으나 n에 의해 강제로 잘릴 수 있다. 그러므로 위의 예제에서 Y_5^+의 성질은 앞의 Y_j^+와는 통계적 성질이 다르다

Journal of the Korean Statistical Society 43 (2014) 631–641

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

journal homepage: www.elsevier.com/locate/jkss

Quasilikelihood and quasi-maximum likelihood for GARCH-type processes: Estimating function approach

S.Y. Hwang [a,*], M.S. Choi [b], I.K. Yeo [a]

[a] Department of Statistics, Sookmyung Women's University, Seoul 140-742, Republic of Korea
[b] Korea Energy Economics Institute, Uiwang-si 437-713, Republic of Korea

ARTICLE INFO

Article history:
Received 2 October 2012
Accepted 26 January 2014
Available online 13 February 2014

AMS 2000 subject classifications:
primary 62M10
62P10

Keywords:
GARCH-type
Quasilikelihood
Quasi-maximum likelihood
Skewed t-distribution

ABSTRACT

When modeling conditionally heteroscedastic processes, it is usually the case that exact likelihood is not known to researchers or it is too complicated for practical purposes. Consequently, a quasi-maximum likelihood (QML) is frequently employed rather than the exact maximum likelihood (cf. Gourieroux (1997, Ch. 4), Straumann (2005, Ch. 5)). When a GARCH process is partially specified only through the first and second order conditional moments, a systematic and unified approach for inference is via the so called quasilikelihood (QL, see Godambe (1985)). This article aims to discriminate between the QL and QML for general GARCH-type processes. It is verified that the QL differs with the QML in terms of conditional skewness and kurtosis. Asymptotics for the QL and QML are obtained and then compared with each other. A class of skewed t-distributions (cf. Fernandez and Steel (1998)) is considered to illustrate the difference between the QL and QML.

1. Introduction

Conditionally heteroscedastic processes have been useful for modeling volatilities in the field of financial time series. The time series $\{\epsilon_t\}$ is referred to as the first order GARCH, viz., GARCH(1, 1) when $\{\epsilon_t\}$ is recursively obtained by the equation

$$\epsilon_t = \sqrt{h_t} e_t \quad \text{and} \quad h_t = \alpha_0 + \alpha_1 \epsilon_{t-1}^2 + \beta_1 h_{t-1}. \tag{1}$$

The term GARCH stands for the Generalized ARCH (AutoRegressive Conditional Heteroscedasticity). Here and in the sequel, $\{e_t\}$ is a sequence of iid random innovations with mean zero and variance unity. We define an increasing sequence of the sigma fields $\{F_t\}$ generated by $\epsilon_t, \epsilon_{t-1}, \ldots$. Let h_t denote the conditional variance of ϵ_t given the past values $\epsilon_{t-1}, \epsilon_{t-2}, \ldots$, that is, $h_t = \mathrm{Var}(\epsilon_t | F_{t-1})$. Note that h_t is F_{t-1} measurable. Higher order GARCH model, for instance, a GARCH(p, q) process is readily obtained by adding additional lags of $\epsilon_{t-1}^2, \ldots, \epsilon_{t-p}^2$ and $h_{t-1}, \ldots, h_{t-q}$ in the defining equation (1). Making some variations to (1), a number of GARCH-type models have been suggested and investigated in the literature. We refer to the monograph by Straumann (2005) and Part I of the handbook edited by Andersen, Davis, Kreiss, and Mikosch (2009) for various nonlinear and asymmetric GARCH processes. See also a recent paper of Choi, Park, and Hwang (2012) and references therein for diverse GARCH-type models.

* Corresponding author. Tel.: +82 1091809078.
E-mail addresses: shwang@sm.ac.kr, shwang@sookmyung.ac.kr (S.Y. Hwang), stat.mschoi@keei.re.kr (M.S. Choi), inkwon@sookmyung.ac.kr (I.K. Yeo).

http://dx.doi.org/10.1016/j.jkss.2014.01.005
1226-3192/

Extending the class of GARCH-type models under consideration in the paper, we are willing to introduce a conditional mean function in GARCH-type processes. Consider the observable time series $\{X_t\}$ generated by the GARCH-type process $\{\epsilon_t\}$ such that

$$X_t = \mu_t + \epsilon_t, \tag{2}$$

where μ_t denotes the conditional mean function of X_t given the past, i.e., μ_t is F_{t-1} measurable, as defined by $\mu_t = E(X_t|F_{t-1})$. It follows from (2) that there is a one to one correspondence between $\{X_t\}$ and $\{\epsilon_t\}$ and therefore F_t denotes also the sigma field generated by observations $X_t, X_{t-1}, \ldots$. If we choose $\mu_t = \theta X_{t-1}$ and $h_t = \alpha_0 + \alpha_1\epsilon_{t-1}^2 + \beta_1 h_{t-1}$, $\{X_t\}$ is referred to as an AR(1)–GARCH(1, 1) model, indicating a combination of AR(1) in the conditional mean μ_t and the GARCH(1, 1) in the conditional variance h_t. In this article, we do not specify functional form of μ_t and h_t and thus we are dealing with a general GARCH-type process.

In the parameter estimation for general GARCH-type processes defined in (2), it is usually the case that the exact likelihood is hardly known to us and/or it is too complicated for practical purposes. Equivalently, no specific distributional assumptions are made on the iid innovation process $\{\epsilon_t\}$ other than zero mean and unit variance. Then, the quasi-maximum likelihood (QML) estimator is obtained by maximizing an objective function of pseudo-likelihood which is possibly falsely specified likelihood. Often, the pseudo-likelihood is taken via Gaussian innovations, standardized t-distributions with unknown degrees of freedom, and generalized error distributions (refer to, e.g. Tsay (2010, Ch. 10)). Asymmetric (and leptokurtic) distributions for the innovation process $\{e_t\}$ have recently been paid attention including skewed t-distributions (cf., Fernandez & Steel, 1998). It is obvious that the QML estimator reduces to the maximum likelihood (ML) estimator provided that the pseudo-likelihood coincides with the (unknown) true likelihood.

When the general GARCH-type process $\{X_t\}$ is partially specified only through the first and second order conditional moments (i.e., μ_t and h_t), a systematic and unified approach for a partially specified model is via the so-called quasilikelihood (QL). We refer to, among others, Godambe (1985), Heyde (1997) and Hwang and Basawa (2011a) for a background on quasilikelihood in the context of stochastic processes. It is shown that provided that $\mu_t = 0$, the QML based on the Gaussian innovation is essentially the same as the QL (see Proposition 1) and thus two terms QML and QL are typically used interchangeably in the GARCH context. The question that arises naturally is as to the difference between the QML and QL when the QML is based on non-Gaussian innovations. In this paper, we clarify the difference (if any) between the QML and QL for general GARCH-type processes. It will be verified that the QML differs with the QL in terms of conditional skewness and kurtosis (see Theorem 2). Asymptotics for the QML and QL are obtained and then compared with each other (see Theorems 3 and 4). A simulation study is conducted to see finite sample properties of the QL and QML under various innovation distributions including a class of skewed t-distributions. The main goal of the paper is to discriminate and compare QL and QML for general GARCH-type processes.

2. The model description and preliminary results

We are concerned with the general GARCH-type process defined by

$$X_t = \mu_t(\theta) + \epsilon_t \tag{3}$$

where $\mu_t(\theta) = E(X_t|F_{t-1})$ and $\{\epsilon_t\}$ denotes a conditionally heteroscedastic time series. Here, θ is a parameter (vector) indexing the conditional mean function $\mu_t(\theta)$. Baek, Park, and Hwang (2012) discussed a test for fit and evaluated misspecification effects for the model (3). For modeling the conditional mean function $\mu_t(\theta)$, one may include the standard AR (autoregressive) model, threshold AR process and exponential AR model. See for instance Tong (1990) for various nonlinear autoregressive processes. Although higher order models are allowed in our model (3), we consider first order models only, for simplicity of presentation.

- *Threshold autoregressive (TAR) process*: the first-order threshold autoregressive process (TAR(1)) is defined by

$$\mu_t(\theta) = \theta_1 X_{t-1}^+ + \theta_2 X_{t-1}^-, \tag{4}$$

  where $x^+ = \max(x, 0)$ and $x^- = \max(-x, 0)$ so that $x = x^+ - x^-$. We refer to Tong (1990) for a comprehensive treatment of TAR models including higher order processes. It is noted that $\theta_1 = -\theta_2$ $(= \theta$, say) in TAR(1) reduces to a standard AR(1) process.
- *Exponential autoregression (EAR)*: one may consider

$$\mu_t(\theta) = \left[\theta_1 + \theta_2 \exp\left(-\theta_3 X_{t-1}^2\right)\right] X_{t-1} \tag{5}$$

  which is referred to as a first-order exponential autoregressive process (EAR(1)). Note that $\theta_2 = 0$ in EAR(1) gives AR(1) model.

For various models for conditional variance h_t, we refer to, among others, Choi et al. (2012) and Andersen et al. (2009) with references therein. See, e.g., Eq. (1.3) of Straumann (2005) for accommodating broader class of models for h_t. Refer also to Table 5 of Zivot (2009) for various asymmetric GARCH models. One may generate diverse GARCH type processes by combining the mean function $\mu_t(\theta)$ and variance function h_t.

Communications for Statistical Applications and Methods
2014, Vol. 21, No. 3, 213–224

DOI: http://dx.doi.org/10.5351/CSAM.2014.21.3.213
ISSN 2287-7843

An Empirical Characteristic Function Approach to Selecting a Transformation to Normality

In-Kwon Yeo[1,a], Richard A. Johnson[b], XinWei Deng[c]

[a]Department of Statistics, Sookmyung Women's University, Korea
[b]Department of Statistics, University of Wisconsin-Madison, USA
[c]Department of Statistics, Virginia Tech, USA

Abstract

In this paper, we study the problem of transforming to normality. 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 the normal distribution. Our approach also allows for other symmetric target characteristic functions. Asymptotics are established for a random sample selected from an unknown distribution. The proofs show that the weight function t^{-2} needs to be modified to have thinner tails. We also propose the method to compute the influence function for M-equation taking the form of U-statistics. The influence function calculations and a small Monte Carlo simulation show that our estimates are less sensitive to a few outliers than the maximum likelihood estimates.

Keywords: Box-Cox transformation, influence function, Yeo-Johnson transformation.

1. Introduction

Transformation of data is a useful tool that permits the use of an assumption such as normality, when the observed data seriously violate this condition. 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). There are two ways to circumvent this problem. The first approach is to construct diagnostics which attempt to identify influential observations and then to remove these observations before estimating the transformation parameter. Cook and Wang (1983), Hinkley and Wang (1988), Tsai and Wu (1990) and Kim *et al.* (1996) studied case deletion diagnostics for the Box-Cox transformation. However, for multiple outliers, these diagnostic procedures are quite complicated and require an extensive computation. The second approach is to perform a robust estimation procedure that is not strongly affected by outliers. Carroll (1980) proposed a robust method for selecting a power transformation to achieve approximate normality in a linear model. Hinkley (1975) and Taylor (1985) suggested methods to estimate the transformation parameter in the Box-Cox transformation when the goal is to obtain approximate symmetry rather than normality. Yeo and Johnson (2001) and Yeo (2001) introduced an M-estimator obtained by minimizing the integrated square of the imaginary part of the empirical characteristic function of Yeo-Johnson transformed data.

In this paper, we propose a robust method to estimate a transformation parameter as well as the mean and the variance of the target distribution. The estimators are obtained by minimizing a squared

This research was supported by the Sookmyung Women's University Research Grants 2013.
[1] Corresponding author: Department of Statistics, Sookmyung Women's University, Seoul 140-742, Korea.
E-mail: inkwon@sookmyung.ac.kr

Published online 31 May 2014 / journal homepage: http://csam.or.kr

distance between the empirical characteristic function of the transformed data and the target characteristic function which is often that of a normal distribution. Specifically, we minimize the integral of the squared modulus of the difference of the two characteristic functions multiplied by a weight function. It is assumed that, for some interval about zero, the weight function equals t^{-2} which is used to compute the distance covariance by Szekely *et al.* (2007).

Many authors such as Koutrouvelis (1980), Koutrouvelis and Kellermeier (1981), Fan (1997), Klar and Meintanis (2005), and Jimenez-Gamero *et al.* (2009) have proposed goodness-of-fit test statistics based on measuring differences between the empirical characteristic function and the characteristic function in the null hypothesis. Conversely, we employ this statistic as a measurement to estimate the transformation parameter as well as the mean and the variance. Our estimation procedure 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.

2. Estimation

Let $\psi(x, \lambda)$ be a family of transformations indexed by the transformation parameter λ, for instances, Box and Cox (1964), John and Draper (1980), and Yeo and Johnson (2000). Generally, a main goal of transforming data is to enhance the normality of data. According to the definition of the relative skewness by van Zwet (1964), a convex transformation reduces left-skewness or vice versa. The Box-Cox transformation and the Yeo-Johnson transformation are either convex or concave and can be applied to skewed data to improve symmetry. By contrast, the modulus transformation by John and Draper (1980) is convex-concave and can be applied to symmetric data to reduce kurtosis.

Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with distribution function $F(\cdot)$. It is assumed that there exists a transformation parameter λ for which $\psi(X, \lambda)$ is normally distributed with mean μ and variance σ^2. This assumption can be extended to any symmetric distribution with the location and the scale parameter, μ and σ.

Let $\phi(t)$ be the characteristic function of the standard normal distribution, that is $\phi(t) = e^{-t^2/2}$, and let $\phi_n(\boldsymbol{\theta}, t)$ be the empirical characteristic function of standardized transformed variables $Z_j(\boldsymbol{\theta}) = \{\psi(X_j, \lambda) - \mu\}/\sigma$, $j = 1, \ldots, n$, where $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)^T = (\lambda, \mu, \sigma^2)^T$. Then,

$$\phi_n(\boldsymbol{\theta}, t) = n^{-1} \sum_{j=1}^{n} \exp\left(itZ_j(\boldsymbol{\theta})\right) = \phi_{cn}(\boldsymbol{\theta}, t) + i\phi_{sn}(\boldsymbol{\theta}, t),$$

where

$$\phi_{cn}(\boldsymbol{\theta}, t) = n^{-1} \sum_{j=1}^{n} \cos\left(tZ_j(\boldsymbol{\theta})\right) \quad \text{and} \quad \phi_{sn}(\boldsymbol{\theta}, t) = n^{-1} \sum_{j=1}^{n} \sin\left(tZ_j(\boldsymbol{\theta})\right).$$

Here, the parameter space $\boldsymbol{\Theta}$ is assumed to be a compact set of the form

$$\boldsymbol{\Theta} = \{\boldsymbol{\theta} \mid a_i \le \theta_i \le b_i, \text{ where } 0 < a_3, \ |a_i|, |b_i| < \infty, \text{ for } i = 1, 2, 3\}. \tag{2.1}$$

The goodness-of-fit test statistics based on measuring differences between the empirical characteristic function and the characteristic function in the null hypothesis have been extensively investigated

TRIALS

STUDY PROTOCOL Open Access

Acupuncture for functional dyspepsia: study protocol for a two-center, randomized controlled trial

Gajin Han[1], Seok-Jae Ko[1], Jae-Woo Park[1], Jinsung Kim[1], Inkwon Yeo[2], Hyejung Lee[3], Song-Yi Kim[3] and Hyangsook Lee[3*]

Abstract

Background: Functional dyspepsia (FD) is a common health problem currently without any optimal treatments. Acupuncture has been traditionally sought as a treatment for FD. The aim of this study is to investigate whether acupuncture treatment helps improve symptoms of FD.

Methods/design: A two-center, randomized, waitlist-controlled trial will be carried out to evaluate whether acupuncture treatment improves FD symptoms. Seventy six participants aged 18 to 75 years with FD as diagnosed by Rome III criteria will be recruited from August 2013 to January 2014 at two Korean Medicine hospitals. They will be randomly allocated either into eight sessions of partially individualized acupuncture treatment over 4 weeks or a waitlist group. The acupuncture group will then be followed-up for 3 weeks with six telephone visits and a final visit will be paid at 8 weeks. The waitlist group will receive the identical acupuncture treatment after a 4-week waiting period. The primary outcome is the proportion of responders with adequate symptom relief and the secondary outcomes include Nepean dyspepsia index, EQ-5D, FD-related quality of life, Beck's depression inventory, state-trait anxiety inventory questionnaire, and level of ghrelin hormone. The protocol was approved by the participating centers' Institutional Review Boards.

Discussion: Results of this trial will help clarify not only whether the acupuncture treatment is beneficial for symptom improvement in FD patients but also to elucidate the related mechanisms of how acupuncture might work.

Trial registration: ClinicalTrials.gov Identifier: NCT01921504.

Keywords: Acupuncture, Functional dyspepsia, Ghrelin, Randomized controlled trial, Waitlist

Background

Functional dyspepsia (FD) is a functional gastrointestinal disorder without any structural lesion [1]. It is reported that the prevalence of FD ranges from 8% to 23% in Asia [2], and in particular, FD is claimed to affect 25% of the South Korean population [3]. Though FD is not a life-threatening disease, FD patients suffer from a poor quality of life (QoL), which is regarded as an economic burden on society [4].

Although various treatments for FD, such as proton pump inhibitors, prokinetic agents, tricyclic antidepressants, and antinociceptive agents are available [5-9], many FD patients turn to other complementary and alternative therapies largely due to a lack of satisfactory relief by these treatments [10].

Acupuncture treatment is one of the most sought-after therapeutic modalities in complementary and alternative medicine. It has been frequently used to treat symptoms of FD [11] largely based on the following rationales: firstly, acupuncture is well known to relieve pain in various conditions [12] and may help reduce epigastric pain or burning sensation [4]; secondly, previous studies have shown that manual acupuncture and electro-acupuncture modulates gastric/duodenal motility through the activation of sympathetic efferent nerve fibers or vagal nerve fibers [13]; finally, since patients with FD have been reported to have higher levels of psychological distress or depression than healthy subjects [14,15], the anxiolytic or antidepressant

* Correspondence: erc633@khu.ac.kr
[3]Acupuncture and Meridian Science Research Center, College of Korean Medicine, Kyung Hee University, Kyung Hee dae-ro 26, Dongdaemun-gu, Seoul 130-701, South Korea
Full list of author information is available at the end of the article

effects of acupuncture have been utilized to minimize worsening of FD symptoms [16-18].

However, previous studies have reported inconsistent findings of the clinical benefit of acupuncture treatment for FD [4,19]. Nevertheless, these studies have been criticized for adopting relatively sub-optimal acupuncture treatment, i.e., three times a week for 2 weeks [19,20], or testing a less generalizable treatment regimen, i.e., daily acupuncture treatment for 4 weeks [4]. According to a systematic review by Zhu [21], the acupuncture treatment for FD lacks high-quality evidence. Hence, it was addressed that randomized clinical trials corresponding with the CONSORT statement and STandards for Reporting Interventions in Clinical Trials of Acupuncture (STRICTA) recommendations have to be implemented to evaluate the efficacy of acupuncture for FD. In addition, considering the questions which have been raised so far on the adequacy of various sham acupuncture controls [22,23], care should be taken before simply adopting any available sham devices or procedures in the early phase.

Given that FD is now regarded as a complex disorder of which the pathophysiological mechanisms are inconclusive [24], various factors are likely to come into play. Among them, ghrelin, an orexigenic peptide, has been shown to play an important role in gastric motility, food intake, and potential gastroprotection against acute gastric mucosal injury [25,26]. However, it is yet to be studied whether ghrelin is associated with the mechanisms of acupuncture treatment in patients with FD [27,28].

Taken together, we felt the need to properly test whether acupuncture treatment would help symptom improvement in FD patients and, subsequently, the related mechanism. Therefore, we propose a two-center, randomized, waitlist-controlled trial investigating the effectiveness of acupuncture for symptom improvement in patients with FD and to assess whether the change in ghrelin secretion is associated with acupuncture treatment.

Methods/design

Objectives

The aims of this study are to: i) assess the effect of acupuncture treatment on patients with FD in comparison with waitlist group with respect to adequate symptom relief and ii) to elucidate whether ghrelin level is associated with clinical effect of acupuncture.

Hypothesis

We hypothesize that eight sessions of acupuncture treatment over 4 weeks will improve symptoms of FD.

Design

A two-center, randomized, waitlist-controlled trial will be conducted at the Kyung Hee University Korean Medicine Hospital and Kyung Hee University Hospital at Gangdong in Seoul, Korea from August 2013 to January 2014. The flow of the entire trial is shown in Figure 1.

Inclusion and exclusion criteria

Inclusion criteria

Participants who meet the following criteria will be included:

i) Individuals between the ages of 18 and 75 years, able to read and write Korean language
ii) Individuals who meet the Rome III FD criteria
iii) Individuals who check more than four points on the visual analogue scale (VAS; 0, no symptom disturbance at all; 10, very severe) for dyspeptic symptoms
iv) Individuals who have normal esophagogastroduodenoscopy results within a year and have been diagnosed with FD by a specialist consultation
v) Individuals who receive no other treatments during the study
vi) Individuals who voluntarily agree with a study protocol and sign a written informed consent

Exclusion criteria

Participants who report any of the following will be excluded:

i) Individuals who have peptic ulcer or gastroesophageal reflux disease confirmed by esophagogastroduodenoscopy
ii) Individuals who have obvious signs of irritable bowel syndrome

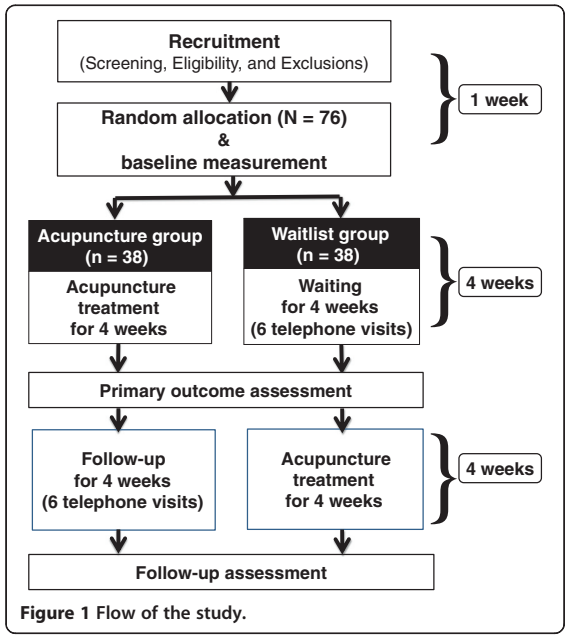


Figure 1 Flow of the study.

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 consider the selection of weight functions that yield a closed form for the distance function. 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 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

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

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 (1981), 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). According to Yeo et al. (2013), the estimator by the empirical characteristic function approach is still sensitive, but less sensitive than the maximum likelihood estimate, to an outlier when the target distribution is normal.

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 $x > 0$,

$$\psi(x,\lambda) = \begin{cases} \left(x^{\lambda}-1\right)/\lambda, & \lambda \neq 0 \\ \log(x), & \lambda = 0 \end{cases}$$

and the Yeo-Johnson transformation

The Korean Journal of Applied Statistics (2014) DOI: http://dx.doi.org/10.5351/KJAS.2014.27.5.681
27(5), 681–691

An Empirical Comparison of Predictability of Ranking-based and Choice-based Conjoint Analysis

Bu-Yong Kim[a,1]

[a]Department of Statistics, Sookmyung Women's University

(Received June 18, 2014; Revised August 9, 2014; Accepted August 13, 2014)

Abstract

Ranking-based conjoint analysis(RBCA) and choice-based conjoint analysis(CBCA) have attracted significant interest in various fields such as marketing research. When conducting research, the researcher has to select one suitable approach in consideration of strengths and weaknesses. This article performs an empirical comparison of the predictability of RBCA and CBCA in order to provide criterion for the selection. A new concept of measurement set is developed by combining the ranking set and choice set. The measurement set enables us to apply two approaches separately on the same consumer group that allows a fair comparison of predictability. RBCA and CBCA are conducted on consumer preferences for RTD-coffee; subsequently, the predicted values of market shares and hit rates are compared. The study result reveals that their predictabilities are not significantly different. Further, the result indicates that RBCA is recommended if the researcher wants to improve data quality by filtering out poor responses or to implement the market segmentation. In contrast, CBCA is recommended if the researcher wants to lessen the burden on the respondents or to measure preferences under similar conditions with the actual marketplace.

Keywords: Predictability, ranking-based conjoint, choice-based conjoint, measurement set, RTD-coffee.

1. 서론

마케팅조사를 비롯하여 다양한 학문 및 산업분야에서 활용되는 컨조인트분석은 소비자들이 제품의 각 속성을 어느 정도 중요시 하는지 측정하고, 선호하는 제품프로파일이 어느 것인지 파악하고, 경쟁 제품들의 시장점유율을 예측하기 위하여 실행된다. 컨조인트분석 결과는 신제품개발을 위한 제품특성의 최적화, 소비자 선호도분석 및 구매행동분석, 제품포지셔닝과 판매촉진을 위한 시장세분화 등에 광범위하게 활용된다. 컨조인트분석은 제품에 대한 소비자의 선호도를 측정하는 방식에 따라 등급기반 컨조인트분석(rating-based conjoint analysis; RTCA), 순위기반 컨조인트분석(ranking-based conjoint analysis; RBCA), 선택기반 컨조인트분석(choice-based conjoint analysis; CBCA)으로 구분되는데, 컨조인트분석가는 각 기법의 특성 및 장단점과 예측력을 고려하여 조사환경에 가장 적절한 기법을 선택하게 된다. 컨조인트분석 기법들의 예측력이나 타당성을 비교연구한 기존의 연구들이 있는데, Elrod 등(1992)은 RTCA와 CBCA의 예측력을 실증적으로 비교하고 두 기법의 예측력에는 차이가 없음을 밝혔

This research was supported by Sookmyung Women's University Research Grants(2013).
[1]Department of Statistics, Sookmyung Women's University, 100 Cheongpa-ro 47-gil, Yongsan-gu, Seoul 140-742, Korea. E-mail: buykim@sookmyung.ac.kr

다. Boyle 등 (2001)은 등급형으로 수집한 선호도 측정 자료를 순위형과 선택형으로 변환하여 해당 모형의 타당성을 비교하였는데, 선호도 자료를 다른 유형의 자료로 변환하는 것은 권할 수 없다고 하였다. Chakraborty 등 (2002)은 RTCA와 CBCA의 시장점유율 예측력을 비교하였는데, 컨조인트분석 기법을 선정하기 위해서는 소비자 선호도의 이질성이나 제품의 유사성의 정도를 고려해야 한다고 하였다. Moore (2004)는 RTCA와 CBCA의 교차타당성을 비교하였는데, 어느 분석기법도 우위에 있지 않다고 결론지었다. Sayadi 등 (2005)은 선호도를 순위로 측정하는 것이 등급으로 측정하는 것보다 제품 간의 선호도 차이를 잘 반영한다고 하였다. Lim 등 (2006)은 의료서비스의 선택에 관한 실증적 비교연구에서 RTCA보다 CBCA의 예측력이 우수함을 입증하였다. Karniouchina 등 (2009)은 RTCA와 CBCA의 적중률을 비교하고 CBCA를 선택할 것을 제안하였다.

본 연구는 분석가들이 컨조인트분석 기법을 선택하는 데 도움이 될 수 있는 준거를 제공하고자 한다. 최근에 많이 사용되고 있는 기법이 RBCA와 CBCA인데도 불구하고 기존의 연구들은 RTCA와 CBCA를 비교하였으며 RBCA와 CBCA를 비교분석한 연구는 아직 없다. 그러므로 RBCA와 CBCA의 예측력을 비교하여 어느 것이 우위에 있는지 확인하고자 한다. 그런데 기존 연구들은 모의실험을 통하여 예측력을 비교하거나, 집단별로 다른 기법을 적용하여 분석한 결과를 비교하는 접근방법을 채택하였다. 이러한 방법들에 의하면 현실성이 결여되거나 외부요인들의 영향이 통제되지 않은 분석결과를 얻게 되므로 기법들의 예측력을 비교하는데 적절하지 않다고 판단된다. 따라서 모의실험이 아니고 실제 상황인 RTD(ready to drink)커피에 대한 선호도 측정을 바탕으로 실증적 비교를 시도하며, 두 집단이 아닌 동일집단에 RBCA와 CBCA을 동시에 적용할 수 있도록 컨조인트조사를 설계하여 예측력을 비교하고자 한다. 이를 위해 RBCA을 위한 순위집합과 CBCA을 위한 선택집합을 통합한 측정집합 개념을 컨조인트조사 설계에 새롭게 도입한다.

2. 컨조인트분석 기법의 비교

컨조인트분석가가 실제 사례에 적용할 기법을 선정하기 위해서는 각 기법의 특성과 장단점을 이해해야 하므로 우선 기법들의 실행과정을 비교하여 설명한다. 컨조인트분석의 실행에서 각 기법에 공통적으로 해당되는 과정은 제품의 속성과 수준을 결정하는 것과 선호도측정 대상 제품프로파일들을 선정하는 것이며, 선호도 측정과정은 기법마다 다른 방식으로 실행한다. 선호도 자료를 수집한 후에 RTCA와 RBCA에서는 응답자별로 실행한 컨조인트분석 결과를 전체 응답자에 대해 종합하지만, CBCA에서는 응답자별 선호도 평가결과를 전체로 종합하고 컨조인트분석을 실행한다는 점이 다르다. 각 기법에서 수집하는 자료의 유형이 상이하므로 컨조인트모형의 계수를 추정하는 방법은 기법에 따라 다른 것을 적용한다. 최종적으로 계수추정치를 바탕으로 각 속성의 상대적 중요도와 수준별 부분가치를 계산하는 과정은 모든 기법에서 동일하며 Kim (2014)에 소개된 방법을 따른다.

2.1. RTCA

RTCA에서는 제품프로파일들에 대한 응답자들의 선호도를 등급으로 측정한다. 이 기법에서는 부분가치를 추정하기 위하여 최소자승추정법을 적용하기 때문에 분석가들이 추정과정을 쉽게 이해할 수 있다는 점, 컴퓨터 뿐만 아니라 조사지를 사용하여 선호도를 측정할 수 있다는 점, 응답자가 프로파일들에 대한 선호도를 등급으로 평가하는 작업을 비교적 쉽게 할 수 있다는 점을 장점으로 꼽을 수 있다. 그러나 선호도 평가과정에서 응답자들이 프로파일들에 대해 충분한 변별력을 갖기 어렵기 때문에 수집된 자료의 품질이 낮다는 심각한 단점을 가지고 있다. 최근에 RTCA를 적용한 사례연구로는 Kim 등 (2010), Mesias 등 (2010), Ryu와 Roh (2010), Chung 등 (2011), Endrizzi 등 (2011), Adanacioglu와

The Korean Journal of Applied Statistics (2014) DOI: http://dx.doi.org/10.5351/KJAS.2014.27.2.185
27(2), 185–195

New Method for Preference Measurement in Ranking-based Conjoint Analysis

Bu-Yong Kim[a,1]

[a]Department of Statistics, Sookmyung Women's University

(Received October 20, 2013; Revised December 2, 2013; Accepted December 2, 2013)

Abstract

Ranking-based conjoint analysis is widely used in various fields such as marketing research. While the ranking-based conjoint affords several advantages over the rating-based or choice-based conjoint, it has a serious shortcoming that respondents have much difficulty in ranking the product profiles in order of preference when many profiles are involved. This article suggests a new method for the preference measurement to improve the response efficiency. The method employs the concept of ranking sets that let the respondent evaluate a small number of profiles at a time. Through the proposed method, preference rankings of profiles obtained from each ranking set are aggregated to generate overall rankings. The balanced incomplete block design is expanded and transformed to the dual design in order to construct well-balanced ranking sets that can accommodate a large number of profiles. The proposed method is applied to the analysis of consumer preferences for perfume-for-women.

Keywords: Ranking-based conjoint, ranking set, balanced incomplete block design, perfume-for-women.

1. 서론

마케팅조사 분야에서 발전해 온 컨조인트분석은 다양한 학문 및 산업에서 광범위하게 활용되고 있다. 이 분석기법은 제품(서비스, 전략, 정책, 디자인, 프로그램 등도 포함됨)의 주요 속성과 수준들에 요인설계를 적용하여 프로파일들을 구성하고 각 프로파일에 대한 소비자들의 선호도를 측정하여, 수준별 부분가치를 추정하고 속성들의 상대적 중요도를 평가하는데 사용된다. 컨조인트분석 결과는 소비자 선호도 분석, 구매행동분석, 신제품개발, 시장점유율 예측, 시장세분화, 제품특성의 최적화, 판매가격 결정, 포지셔닝과 판매촉진 등에 활용되고 있다.

컨조인트분석은 선호도 측정이나 부분가치 추정을 위해 채택된 방법에 따라 몇 가지 기법으로 분류된다. 선호도 측정방법에 따른 자료의 유형을 기준으로 등급기반, 순위기반, 선택기반 컨조인트분석으로 분류되는데, 각 기법들은 상이한 특성과 장단점을 가지고 있다. 등급기반 컨조인트분석에서는 응답자가 프로파일들에 대한 선호도 평가를 비교적 쉽게 할 수 있고, 부분가치 추정을 위하여 최소자승추정법을 적용할 수 있다는 장점을 가지고 있다. 그러나 선호도 평가과정에서 응답자들이 프로파일들에 대해 충분한 변별력을 갖기 어렵기 때문에 측정된 자료의 품질이 낮다는 심각한 단점을 가지고 있다. 반면에 순위

[1]Department of Statistics, Sookmyung Women's University, 100 Cheongpa-ro 47-gil, Yongsan-gu, Seoul 140-742, Korea. E-mail: buykim@sookmyung.ac.kr

기반 컨조인트분석에서는 응답자가 프로파일들에 대한 선호도를 변별력을 가지고 평가할 수 있다는 장점을 가지고 있다. Sayadi 등 (2005)은 순위에 의한 선호도 측정이 등급에 의한 측정보다 제품 간의 차이를 잘 반영한다고 하였다. 그러나 다수의 프로파일에 대해 선호도 순위를 매기는 작업이 응답자에게 큰 부담이 되고 정확한 평가가 이루어지기 어렵다는 단점을 가지고 있다. 선택기반 컨조인트분석에서는 소비자의 실제 구매행동과 매우 유사한 상황에서 현실적인 선호도 평가가 실행된다는 장점을 가지고 있으나, 부분가치 추정과정이 비교적 복잡하다는 단점을 가지고 있다. 한편, 등급기반과 순위기반 컨조인트분석에서는 각 응답자별로 분석한 결과를 전체로 종합하기 때문에 개인별 부분가치에 바탕을 둔 시장세분화 작업을 할 수 있지만, 선택기반 컨조인트분석에서는 응답자 전체를 통합한 자료에 대하여 분석을 하기 때문에 시장세분화 작업이 불가능하다. 이러한 장단점과 특성들이 고려되어 순위기반 컨조인트분석(ranking-based conjoint analysis; RBCA)이 다양한 산업분야(제조, 디자인, 정보보안, 여행, 관광, 요식, 식품, 의료, 보건행정, 교통, 의류 등)에서 널리 사용되고 있다. RBCA를 적용한 사례연구로는 Min 등 (2000), Shin 등 (2004), Lee 등 (2004), Kim (2005), Kim 등 (2005), Lee 등 (2005), Choi (2006), Kim과 Park (2007), Shin 등 (2007), Park 등 (2008), Yoon과 Cho (2009), Lee 등 (2010), Jo (2010), Chen 등 (2010), Jin 등 (2010), Ong 등 (2010), Park과 Lee (2011), Jung (2012), Choi 등 (2012), Hong 등 (2012)이 있다.

본 논문에서는 RBCA를 위한 선호도 순위를 효율적으로 측정할 수 있는 방법에 관하여 연구한다. RBCA의 초기 단계에서는 분석대상 제품의 속성과 수준들을 파악하고, 요인설계에 의해 제품프로파일들을 구성하여 응답자들에게 제시하고 선호도 순위를 측정하는 작업을 한다. 선호도 순위를 측정하기 위한 기존의 방법들이 몇 가지 있지만, 그 방법들은 응답자에게 과중한 부담과 피곤을 느끼게 하고, 응답자가 비교해야 할 프로파일 짝의 개수가 지나치게 많아서 평가에 과도한 시간이 요구되고, 조사지를 사용하는 선호도 측정에서는 채택하기 곤란하고, 응답자가 선호도를 정확하게 평가하기 어렵다는 단점들을 가지고 있다. 따라서 기존의 선호도 측정방법을 개선하기 위하여 순위집합 개념을 제안한다. 즉, 응답자가 순위를 매기기에 어렵지 않을 정도로 소수의 프로파일들을 순위집합으로 묶되, 프로파일 전체의 순위를 결정하는데 충분한 개수의 순위집합들을 구성하여 응답자에게 제시하는 방법이다. 순위집합을 체계적으로 구성하기 위하여, 균형불완비블록설계(balanced incomplete block design; BIBD)를 두 배 또는 세 배 확장하고 쌍체설계로 전환하여 순위집합을 구성하는 방법을 제시한다. 그리고 제안된 선호도 측정방법을 실제로 적용하여 여성용 향수제품에 대한 소비자 선호도분석을 실행하고자 한다.

2. 순위집합을 이용한 선호도 측정방법

2.1. 순위집합 개념의 도입

RBCA에서 선호도 측정을 위한 조사를 설계할 때 중요시해야 할 사항은 통계적 효율성과 응답효율성인데, 통계적 효율성에 관한 연구는 이미 많이 이루어졌기 때문에 본 연구에서는 응답효율성에 초점을 맞춘다. 프로파일들에 대한 선호도 순위를 측정하는 기존의 방법들은 다음과 같다. ① 응답자에게 분석대상 프로파일 전체를 동시에 제시하고 각 프로파일들에 대한 선호도를 순위로 응답하게 하는 방법이다. 제1장에 소개한 사례연구들의 대부분은 이 측정방법을 채택하고 있다. 그런데 요인설계에 의해 구성된 프로파일의 개수는 상당히 많은 경우가 대부분이고, 부분요인설계법에 따라 프로파일의 일부를 선정하더라도 일정 수준의 통계적 효율성을 유지하기 위해서 상당히 많은 개수의 프로파일을 포함시키는 것이 불가피하다. 이러한 상황에서 전체 프로파일에 대한 선호도를 순위로 평가하게 하면, 응답자는 과중한 부담을 느끼고 선호도를 정확하게 평가하기 어렵다. ② 전체 프로파일을 두 개씩의 모든 가능한 짝으로 묶어서 응답자에게 제시하고 각 짝에서 선호하는 프로파일을 선택하거나 프로파일에 대한 상대적인 선

연구논문

2014년 지방선거 여론조사 전화조사방법에 따른 예측오차 및 편향*

Forecasting Error and Bias of Telephone Survey
for 2014 Local Elections

김영원[a] · 황다솜[b]
Young-Won Kim · Dasom Hwang

선거여론조사가 얼마나 정확한지 그리고 어떤 조사가 다른 조사에 비해 왜 더 정확한지 등에 대해서 여론조사 전문가들은 많은 관심을 갖고 있다. 본 연구에서는 이러한 질문에 대한 해답을 찾기 위해 2014년 제6회 전국동시지방선거에 대한 전화여론조사자료를 토대로 전화조사방법에 따른 편향과 예측오차를 분석하였다. 이를 위해 중앙선거여론조사공정심의위원회 홈페이지에 등록된 선거여론조사자료를 활용하였으며, Martin et al. (2005)이 제안한 정확성 척도 A를 예측오차 및 편향의 판단 기준으로 사용하였다. 2014년 지방선거 선거여론조사 중 광역단체장 선거에 관한 150건의 전화조사를 분석해 본 결과, ARS전화조사는 전화면접조사에 비해 새누리당을 과대 추정하는 편향을 발생시키는 경향이 있다는 것을 확인할 수 있었고, 조사방법과 함께 가중치 변동 폭, 조사기간, 그리고 응답률 등도 전화여론조사에서 편향이나 오차 발생에 영향을 주는 요인이라는 결론을 내릴 수 있었다.

주제어 : 선거여론조사, 전화조사, 예측오차, 편향, 2014년 지방선거

* 본 연구는 숙명여자대학교 2012학년도 교내연구비 지원에 의해 수행되었음.

a) 교신저자(corresponding author): 숙명여자대학교 통계학과 교수 김영원.
E-Mail: ywkim@sookmyung.ac.kr

b) 숙명여자대학교 통계학과 대학원생

If late-campaign polls are to be used as forecast, it is important to ask, how well do the polls do and why are some polls better forecast than others? To answer the questions, we investigate the forecasting error and bias of telephone survey conducted for 2014 local election of Korea. To resolve the problem, we use the telephone survey para-data released by National Fair Election Survey Deliberation Commission. The aim of this study is to find main cause of the forecasting error and bias of telephone survey. By analyzing the 150 election telephone survey and using the measure of predictive accuracy(A) proposed by Martin et al.(2005), we found telephone survey method, period of poll, variation level of weight, and response rate are the factors to cause the forecasting error and bias in telephone survey. Also we found that ARS telephone survey as well as large variation of weight bring overestimation of candidate of Saenuri Party.

Key words: election poll, telephone survey, forecasting error, bias, 2014 local election of Korea

Ⅰ. 서론

여론조사는 각종 정치, 사회적 문제나 국가 정책 등에 대한 국민들의 의견을 파악할 수 있는 매우 유효한 방법이며, 특히 선거에서는 유권자들이 자신들을 대표해 줄 수 있는 사람을 선택하는 데 필요한 정보를 얻고 난립하는 후보자를 거를 수 있는 순기능도 갖고 있다. 하지만 같은 시기에 실시된 여론조사의 경우에도 후보자의 지지율이 크게 차이가 나서 여론조사에 대한 신뢰

Journal of the Korean Statistical Society 43 (2014) 439–452

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

journal homepage: www.elsevier.com/locate/jkss

Component selection in additive quantile regression models

Hohsuk Noh [a], Eun Ryung Lee [b,*]

[a] Institut de Statistique, Biostatistique et Sciences Actuarielles (ISBA), Université catholique de Louvain, Belgium
[b] University of Mannheim, Germany

ARTICLE INFO

Article history:
Received 24 November 2013
Accepted 19 January 2014
Available online 1 February 2014

AMS 2000 subject classifications:
primary 62H99
secondary 62G08

Keywords:
Additive model
Covariate selection
Nonparametric quantile regression
Second order cone programming
Penalized estimation

ABSTRACT

Nonparametric additive models are powerful techniques for multivariate data analysis. Although many procedures have been developed for estimating additive components both in mean regression and quantile regression, the problem of selecting relevant components has not been addressed much especially in quantile regression. We present a doubly-penalized estimation procedure for component selection in additive quantile regression models that combines basis function approximation with a ridge-type penalty and a variant of the smoothly clipped absolute deviation penalty. We show that the proposed estimator identifies relevant and irrelevant components consistently and achieves the nonparametric optimal rate of convergence for the relevant components. We also provide an accurate and efficient computation algorithm to implement the estimator and demonstrate its performance through simulation studies. Finally, we illustrate our method via a real data example to identify important body measurements to predict percentage of body fat of an individual.

1. Introduction

Suppose that Y is a response of interest which depends on a vector of random covariates $\mathbf{X} = (X^1, \ldots, X^p) \in \mathbb{R}^p$, $p \geq 2$. We are interested in the conditional quantile of Y given $\mathbf{X}$. The relationship between Y and $\mathbf{X}$ can be modeled as

$$Y = Q_\tau(\mathbf{X}) + U_\tau,$$

where $Q_\tau(\cdot)$ is an unknown real-valued function and U_τ is an unobserved random variable whose τth quantile conditional on $\mathbf{X} = \mathbf{x}$ is 0 for almost every $\mathbf{x}$. Many nonparametric methods are available to estimate $Q_\tau(\cdot)$ and yet they are all severely affected by the so-called curse of dimensionality when p is large. To tackle this issue, several authors have proposed dimension reduction methods, a popular one of which is to consider the following model based on the additive structure of $Q_\tau(\mathbf{X})$:

$$Y_i = Q_\tau(\mathbf{X}_i) + U_{i,\tau} = \mu_\tau + \sum_{k=1}^{p} m_{k,\tau}(X_i^k) + U_{i,\tau}, \tag{1.1}$$

where $(Y_i, \mathbf{X}_i, U_{i,\tau}) \in \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}$ for $1 \leq i \leq n$ are independent and identically distributed and $P(U_{i,\tau} \leq 0 | \mathbf{X}_i = \mathbf{x}_i) = \tau$ for almost every $\mathbf{x}_i$. To fit this additive quantile regression model (1.1), there have been many proposals including Cheng, De Gooijer, and Zerom (2011), De Gooijer and Zerom (2003), Horowitz and Lee (2005), Lee, Mammen, and Park (2010) and Yu and Lu (2004). Most of them have more or less similar implication that it is possible to estimate the additive component

* Corresponding author. Tel.: +49 621 181 1777.
E-mail address: silverryuee@gmail.com (E.R. Lee).

1226-3192/$ – see front matter
http://dx.doi.org/10.1016/j.jkss.2014.01.002

functions in multi-dimension settings at the same accuracy as in the univariate cases. However, all these works are based on the assumption that the given covariates are all relevant to the conditional quantile of interest, which is not always the case in practice. More interestingly, when the error variance is heteroscedastic or other forms of non-location-scale covariate effects exist, the sets of relevant covariates in additive models may differ when we consider different quantiles of the conditional distribution. This is the characteristic feature of covariate selection in quantile regression, which differentiate itself from that in mean regression. Motivated by this concern, we consider the problem of selecting the relevant covariates in the model (1.1).

Although there have been a few works in the literature which deal with the problem of covariate selection in additive mean regression framework such as Fan, Feng, and Song (2011), Huang and Yang (2004) and Xue (2009), not many works have been done in the context of quantile regression. Fenske, Kneib, and Hothorn (2011) adapted an idea of Bühlmann and Yu (2003) (in framework of boosting with squared error loss) for covariate selection in additive mean regression models to their additive quantile regression models in order to identify the relevant factors (covariates) in the data, but did not discuss its statistical properties. Contrary to them, we not only propose a method to identify the relevant covariates in the model (1.1) and estimate the corresponding additive component functions, but also derive its theoretical properties.

The main idea of component selection in our work is that the magnitude of the L_2-norm of m_k is closely related to relevance of the covariate X^k. Motivated by this observation, we propose a two-stage procedure for estimating $m_{k,\tau}(\cdot)$ simultaneously with the identification of irrelevant covariates. The initial estimator is computed for estimating $\|m_k\|_{L_2}$ with $\|f\|_{L_2} = (\int f^2)^{1/2}$. Then, we obtain the final estimator which automatically discards presumably irrelevant component functions by referencing to the estimates of the L_2 norm of each component, $\|\tilde{m}_k\|_{L_2}$, $k = 1, \ldots, p$, in a lasso-type penalized framework with basis approximation. This work can be seen as an extension of the idea of Zou and Li (2008) for parametric linear models to nonparametric additive quantile regression models. Independently, a recent work by Lian (2012) obtained related results based on similar idea. However, our work has several advantages over Lian (2012). First, showing that the problem for our estimator can be formulated into one typical convex optimization problem so called the second order cone programming (SOCP) problem, we provide an accurate and efficient computation algorithm to implement the estimator. Compared to ours, Lian (2012) used the majorization–minimization algorithm twice and hence needed two additional tuning parameter for sparsity of the solution and approximation of the check loss function, which may be hard to choose in practice. Second, in order to mitigate instability of the estimator at the boundaries, which is a typical problem of series estimators when the correlations between the covariates are high, we introduce an extra penalty that controls the L_2-norms of component functions using the idea of P-spline apart from the penalty for component selection. Finally, we derive the asymptotic properties of the estimator under more reasonable assumptions. See the discussion in Section 4.

The rest of the paper is organized as follows. Section 2 describes our initial and final estimators. Section 3 shows how the proposed estimators can be implemented. In Section 4, the asymptotic properties of the estimators are established including the consistency of the final estimator in component selection. We provide the finite sample properties of the estimators via a simulation study and illustrate the use of the estimator while analyzing real data on body fat percentage in Section 5. Technical details are given in the Appendix. Regarding the notation, we use subscripts to represent observations of random variables and superscripts to index components of vectors. Additionally we omit the subscript "τ" whenever no confusion is caused.

2. Description of the estimators

In this section, first we propose an estimator for $\|m_k\|_{L_2}$ based on the idea of P-spline in Eilers and Marx (1996). Then we will explain how to use it in component selection when constructing the final estimator. We assume the support $\mathcal{X}_k$ of X^k is compact for $k = 1, \ldots, p$ and simply let $\mathcal{X}_k = [0, 1]$. Additionally, we assume that $m_k(\cdot)$ is centered in the sense that

$$\int_{\mathcal{X}_k} m_k(x)\, dx = 0, \quad k = 1, \ldots, p. \tag{2.1}$$

This assumption guarantees identifiability of the model (1.1) (in L_2 sense) because there exists an universal constant $c > 0$ such that for any $\mu \in \mathbb{R}$ and any p functions $m_k : [0, 1] \to \mathbb{R}, k = 1, \ldots, p$ satisfying (2.1), $\|\mu + \sum_{k=1}^{p} m_k\|_{L_2} \geq c(|\mu| + \sum_{k=1}^{p} \|m_k\|_{L_2})$. These identifiability constraints (2.1) were also considered for additive models by Horowitz and Lee (2005), Horowitz and Mammen (2006) and Xue, Qu, and Zhou (2010), and for a more complex model which have additive structures inside by Lee, Mammen, and Park (2012). One can use another identifiability constraints $Em_k(X^k) = 0, k = 1, \ldots, p$, which were used in Lee et al. (2010) and Xue (2009). In this case, we have to data-dependently modify a given basis for approximating m_k to make the resulting estimator satisfy the constraint $\sum_{i=1}^{n} \hat{m}_k(X_i^k) = 0$.

Since we are focusing on situations where the problem of component selection should be considered, we assume that only s covariates among the X_i^k's are relevant in the model (1.1). It is unknown which s covariates are relevant, nor what the value of s is. Without loss of generality, we let $m_k(\cdot)$, $k = 1, \ldots, s$ be the nonzero component functions and $m_k(\cdot)$, $k = s+1, \ldots, p$ be identically zero.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA

Model Selection via Bayesian Information Criterion for Quantile Regression Models

Eun Ryung LEE, Hohsuk NOH, and Byeong U. PARK

Bayesian information criterion (BIC) is known to identify the true model consistently as long as the predictor dimension is finite. Recently, its moderate modifications have been shown to be consistent in model selection even when the number of variables diverges. Those works have been done mostly in mean regression, but rarely in quantile regression. The best-known results about BIC for quantile regression are for linear models with a fixed number of variables. In this article, we investigate how BIC can be adapted to high-dimensional linear quantile regression and show that a modified BIC is consistent in model selection when the number of variables diverges as the sample size increases. We also discuss how it can be used for choosing the regularization parameters of penalized approaches that are designed to conduct variable selection and shrinkage estimation simultaneously. Moreover, we extend the results to structured nonparametric quantile models with a diverging number of covariates. We illustrate our theoretical results via some simulated examples and a real data analysis on human eye disease. Supplementary materials for this article are available online.

KEY WORDS: High dimension; Linear quantile regression; Model selection consistency; Nonparametric quantile regression; Regularization parameter selection; Shrinkage method.

1. INTRODUCTION

In regression analysis, model selection is necessary in that an underfitted model brings out seriously biased results, whereas an overfitted model leads to substantial loss in estimation efficiency. Hence, finding a parsimonious model with good prediction ability is an important task.

For a traditional linear mean regression model, Bayesian information criterion (BIC) has been used for model selection because it has been well understood that best subset selection with the BIC identifies the true model consistently (Nishii 1984; Shao 1997). Recently, Wang, Li, and Leng (2009) and Chen and Chen (2008) found that the ordinary BIC is too liberal for large model spaces and proposed its modifications that work when the number of variables increases with the sample size. The selection consistency of BIC has been extended to the M-estimation setting by several authors such as Machado (1993) and Wu and Zen (1999), particularly when the predictor dimension is finite.

However, such methods are computationally expensive, especially when the number of variables is large. To tackle this problem, various shrinkage methods such as the least absolute shrinkage and selection operator (LASSO) and smoothly clipped absolute deviation (SCAD) have been proposed and recognized as a promising way of doing estimation and variable selection simultaneously. An interesting thing is that, along with the popularity of such shrinkage methods, BIC has gained attention as a useful way of determining the amount of shrinkage. The seminal work of Wang and Leng (2007) followed by Wang, Li, and Tsai (2007b) and Wang, Li, and Leng (2009) suggested to use a BIC-type criterion for choosing the shrinkage parameter in penalized least squares and established the model selection consistency of the resulting procedure. Zhang, Li, and Tsai (2010) showed that BIC can play the same role in penalized likelihood settings.

As shrinkage methods gained popularity in mean regression, many researchers have turned their interest to variable selection in quantile regression. Wang, Li, and Jiang (2007a) and Wu and Liu (2009) considered shrinkage estimators for variable selection in linear quantile regression models, and Li and Zhu (2008) proposed an efficient algorithm to compute the entire solution path of L_1-norm-regularized quantile regression. Recently, Li, Peng, and Zhu (2011) and Wang, Wu, and Li (2012) investigated SCAD penalizations for variable selection in high-dimensional median and quantile regression, respectively. In the context of nonparametric models, Noh, Chung, and Van Keilegom (2012) and Noh and Lee (2012) proposed variable selection methods for varying coefficient models and additive models in quantile regression setting, respectively. Although some of these works considered BIC-type criteria to select the shrinkage parameter, apparently there has been no sound theory about why we should consider BIC-type criteria for model selection in quantile regression models. This is the main motivation of the present work.

Model selection in quantile regression has two interesting features. One is that considering median regression it can be seen as a way of achieving robustness in model selection. Along this direction, Machado (1993) proposed a BIC for median regression and showed that it identifies the true model consistently if the number of variables is finite. The other feature is that when

Eun Ryung Lee is postdoctoral researcher, Department of Economics, University of Mannheim, 68131 Mannheim, Germany (E-mail: *silverryuee@gmail.com*). Hohsuk Noh is Assistant Professor, Department of Statistics, Sookmyung Women's University, Seoul 140-742, South Korea (E-mail: *word5810@gmail.com*). Byeong U. Park is Professor, Department of Statistics, Seoul National University, Seoul 151-747, South Korea (E-mail: *bupark@stats.snu.ac.kr*). E. R. Lee acknowledges financial support from the Collaborative Research Center SFB 884 "Political Economy of Reforms," funded by the German Research Foundation (DFG). H. Noh acknowledges financial support from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement no. 203650 and IAP research network P7/06 of the Belgian Government (Belgian Science Policy). Research of B. U. Park was supported by the NRF grant funded by the Korean Government (MEST) no. 2010-0017437. The authors thank two referees, the Associate Editor, and the Coeditor for their valuable suggestions, which have significantly improved the article. This research was partly done while the second author was visiting the first author in University of Mannheim. The first and second authors appreciate the support from Prof. Enno Mammen and Prof. Ingrid Van Keilegom regarding their research visits.

Journal of the American Statistical Association
March 2014, Vol. 109, No. 505, Theory and Methods
DOI: 10.1080/01621459.2013.836975

heterogeneity exists due to either heteroscedastic variance or other types of nonlocation-scale covariate effects, the sets of relevant covariates can vary depending on the segment of interest in the conditional distribution. Wang, Wu, and Li (2012) showed that the SCAD-penalized quantile regression estimator is able to detect this heterogeneity of the relevance pattern in high dimension. However, to the best of our knowledge, no theoretical work about BIC has been done in a quantile regression context for the case where the number of variables is diverging, which is encountered in many applications. From this observation, we are motivated to establish a more general and systematic theory of BIC in quantile regression including high-dimensional cases.

The rest of this article is organized as follows. Section 2 briefly describes how a BIC for linear quantile regression is derived and proposes its modification in the case of a diverging number of variables. We also show the selection consistency of the modified BIC there. In Section 3, we propose BICs for covariate selection in some structured nonparametric models and show that they identify the true model consistently when the number of covariates diverges. Finally, we illustrate the theoretical results with some simulated examples and a real data analysis in Sections 4 and 5, respectively. All the theoretical details are given in the Appendix.

2. BIC FOR LINEAR QUANTILE REGRESSION

In this section, we describe how a BIC for a linear quantile regression model can be derived and then propose its extension to the cases where the number of variables diverges as the sample size increases.

2.1 Motivation of BIC for Linear Quantile Regression

Consider the linear quantile regression model

$$Y^i = \mathbf{X}^{i\top}\boldsymbol{\beta}^* + U^i, i = 1, \ldots, n, \tag{2.1}$$

where $(Y^i, \mathbf{X}^i, U^i)$ are independent and identically distributed as $(Y, \mathbf{X}, U) \in \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}$ and $P(U \le 0|\mathbf{X} = \mathbf{x}) = \tau$ for almost every $\mathbf{x}$. Let $\mathbf{X}^i = (X_1^i, \ldots, X_p^i)^\top$ and $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)^\top$. The number of variables $p = p_n$ is allowed to increase with the sample size n. We omit such dependence on n in notation for simplicity whenever it has no confusion. We consider the situations where only d^* covariates among the X_j^i's are relevant, that is, the $p - d^*$ coefficients are zero in the model (2.1). This brings out the statistical problem of selecting the relevant variables.

To derive a BIC for the model (2.1), suppose that the error U^i follows an asymmetric Laplace distribution whose density is given by

$$f(u) = \frac{\tau(1-\tau)}{\sigma} \exp\left\{-\frac{\rho_\tau(u)}{2\sigma}\right\} \tag{2.2}$$

with the check loss $\rho_\tau(u) = u(2\tau - 2I(u < 0))$, and U^i is independent of $\mathbf{X}^i$. Since $P(U^i < 0) = \tau$, the conditional τ-quantile of Y^i given $\mathbf{X}^i = \mathbf{x}^i$ is $\mathbf{x}^{i\top}\boldsymbol{\beta}^*$. Before deriving a BIC, we introduce some notations. The generic notation $\mathcal{S} = \{j_1, \ldots, j_d\} \subset \{1, \ldots, p\}$ denotes an arbitrary candidate model, which includes $X_{j_1}, \ldots, X_{j_d}$ as relevant predictors and let $\mathbf{X}_\mathcal{S} = (X_{j_1}, \ldots, X_{j_d})^\top$. Additionally, we define $|\mathcal{S}|$ to be the cardinality d of $\mathcal{S}$. Based on these notations and setup, we first note that the maximum likelihood estimator of $(\boldsymbol{\beta}_\mathcal{S}, \sigma)$ for a model $\mathcal{S}$ is given as $(\hat{\boldsymbol{\beta}}_\mathcal{S}, \hat{\sigma})$, where $\hat{\boldsymbol{\beta}}_\mathcal{S} = \operatorname{argmin}_{\boldsymbol{\beta}_\mathcal{S} \in \mathbb{R}^{|\mathcal{S}|}} \sum_{i=1}^n \rho_\tau(Y^i - \mathbf{X}_\mathcal{S}^{i\top}\boldsymbol{\beta}_\mathcal{S})$ and $\hat{\sigma} = (2n)^{-1}\sum_{i=1}^n \rho_\tau(Y^i - \mathbf{X}_\mathcal{S}^{i\top}\hat{\boldsymbol{\beta}}_\mathcal{S})$. From the definition of BIC in a form of a penalized log-likelihood (Schwartz 1978), we obtain the following BIC for linear quantile regression:

$$\mathrm{BIC}_\mathrm{L}^\mathrm{O}(\mathcal{S}) = \log\left(\sum_{i=1}^n \rho_\tau\left(Y^i - \mathbf{X}_\mathcal{S}^{i\top}\hat{\boldsymbol{\beta}}_\mathcal{S}\right)\right) + |\mathcal{S}|\,\frac{\log n}{2n}. \tag{2.3}$$

The above BIC for $\tau = 1/2$ has been already considered as a robust alternative to the traditional BIC by Machado (1993), who showed its robustness in consistent model selection over a range of the error distributions. Actually, Wu and Zen (1999) considered another type of BIC for median regression in the context of M-estimation. If we set $\sigma = 1$ in (2.2) when deriving the BIC, it leads us to one of the BICs proposed by Wu and Zen (1999), which is

$$\sum_{i=1}^n \rho_\tau\left(Y^i - \mathbf{X}_\mathcal{S}^{i\top}\hat{\boldsymbol{\beta}}_\mathcal{S}\right) + |\mathcal{S}|\log n. \tag{2.4}$$

Wu and Zen (1999) showed that the BIC with $\tau = 1/2$ at (2.4) is strongly consistent under certain conditions. The criterion (2.3) is invariant under scale change of the response variable. To see this, note that changing Y^i to cY^i for some constant $c > 0$ results in changing $\hat{\boldsymbol{\beta}}_\mathcal{S}$ to $c\hat{\boldsymbol{\beta}}_\mathcal{S}$ for each $\mathcal{S}$ and thus adding $\log c$ to the criterion. This means the criterion (2.3) chooses the same model regardless of the dependent variable's underlying scale of units, whereas (2.4) does not. For this reason, we concentrate on the criterion (2.3) in this article although it is also possible to extend (2.4) to quantile regression.

2.2 Modified BIC for High Dimension

Regarding the BIC in (2.3), recently Lian (2012) proved its consistency in model selection when the number of variables p is finite. However, when p is diverging with the sample size it does not guarantee a consistent result. The reason is that the ordinary BIC is too liberal for model selection and tends to seriously overfit in such a situation, as observed in our simulations. A recent attempt to remedy this problem is to put more penalty on the complexity of the model in the BIC, so that it can be strengthened in overfit resistance. This type of modification has been shown to be successful in the context of mean regression in Wang, Li, and Leng (2009) and Chen and Chen (2008). Adopting this idea, we consider the following modification of the BIC in (2.3) for the case, where p is diverging:

$$\mathrm{BIC}_\mathrm{L}^\mathrm{H}(\mathcal{S}) = \log\left(\sum_{i=1}^n \rho_\tau\left(Y^i - \mathbf{X}_\mathcal{S}^{i\top}\hat{\boldsymbol{\beta}}_\mathcal{S}\right)\right) + |\mathcal{S}|\,\frac{\log n}{2n}C_n, \tag{2.5}$$

where C_n is some positive constant, which diverges to infinity as n increases.

Additionally, to facilitate the theoretical analysis of the BIC in (2.5) in high-dimensional setting where the number of variables exceeds the sample size, we consider setting an upper bound on the cardinality of $\mathcal{S}$, say s_n, and searching for the best model among those submodels of which the cardinality is smaller than or equal to s_n. This idea of restricting model spaces was

Journal of the Korean Statistical Society 43 (2014) 503–512

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

journal homepage: www.elsevier.com/locate/jkss

Frontier estimation using kernel smoothing estimators with data transformation

Hohsuk Noh

Sookmyung Women's University, Republic of Korea

ARTICLE INFO

Article history:
Received 5 March 2014
Accepted 23 July 2014
Available online 13 August 2014

AMS 2000 subject classifications:
primary 62G05
secondary 62P20

Keywords:
Bandwidth selection
Concavity
Constrained kernel smoothing estimator
Priestly–Chao estimator
Monotonicity
Transformed data
Linear programming

ABSTRACT

In economics, a production frontier function is a graph that shows the maximum output of production units such as firms, industries, or economies, as a function of their inputs. Practically, estimating production frontiers often requires imposition of constraints such as monotonicity or monotone concavity. However, few constrained estimators of production frontier have been proposed in the literature. They are based on simple envelopment techniques which often suffer from lack of precision and smoothness. Motivated by this observation, we propose a smooth constrained nonparametric frontier estimator respecting constraints by considering kernel smoothing estimators from a transformed data. It is particularly appealing to practitioners who would like to use smooth estimates that, in addition, satisfy theoretical axioms of production. The utility of this method is illustrated through application to one real dataset and simulation evidences are also presented to show its superiority over the most known methods.

1. Introduction

In economics, a production frontier specifies the maximum possible production level of output for a given input. Hence, it naturally defines production efficiency in the context of that production set: a point on the frontier implies efficient use of the available input, whereas a point below the curve implies inefficiency. Due to importance and necessity of evaluating efficiencies of production units in various areas, various statistical techniques of estimating frontier functions have been proposed so far.

Practically, estimating frontier functions often requires imposition of constraints such as monotonicity or monotone concavity. Popular constrained nonparametric frontier estimators are the free disposal hull (FDH), the linearized FDH (LFDH) and the data envelopment analysis (DEA) estimators. The FDH and DEA estimators are the (upper) boundary curve of the smallest set that envelops all the data with the restriction of free disposability (FDH and DEA) and convexity (DEA) (see, Gijbels, Mammen, Park, & Simar, 1999 and Korostelev, Simar, & Tsybakov, 1995). The linearized FDH (LFDH) estimator is a linearized version of the FDH estimator, which is obtained by linear interpolation of appropriate FDH-efficient points (see, *e.g.*, Hall & Park, 2002, and Jeong & Simar, 2006). Although the FDH and DEA estimators have easy implementation without the need of choosing smoothing parameters and respect desirable constraints of a frontier curve such as monotonicity or monotone concavity, all of these lack smoothness of an estimated curve. To our knowledge, in many situations it is quite natural to assume that frontier curves change smoothly with respect to inputs except at some input points with special meaning. However, estimated curves by FDH and DEA have many unnecessary nonsmooth points. For example, if we see an

E-mail address: word5810@gmail.com.

http://dx.doi.org/10.1016/j.jkss.2014.07.005
1226-3192/

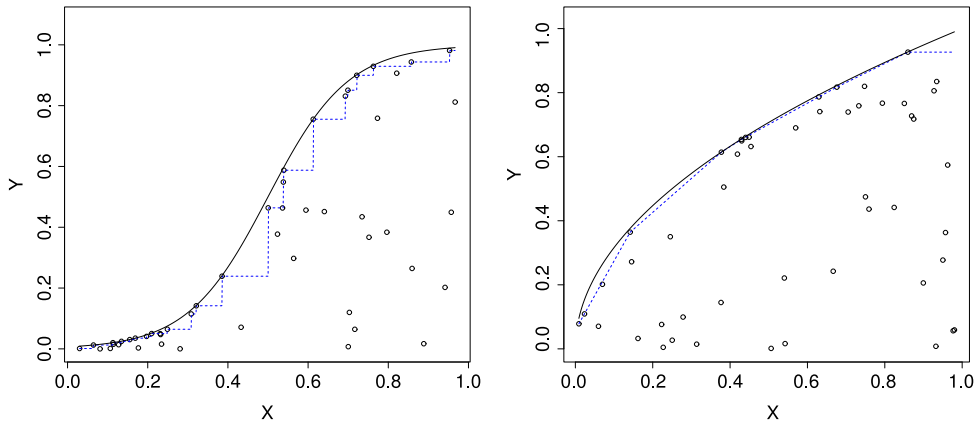


Fig. 1. Examples of FDH and DEA estimates—the true frontier curve (solid black), FDH estimate (dotted blue in the left panel) and DEA estimate (dotted blue in the right panel).

FDH estimate in Fig. 1, the estimated curve repeatedly stays flat and jumps and the location of jumps is just X-locations of FDH points with no appropriate interpretation. A similar problem occurs with the DEA estimator in Fig. 1. At every X-location of DEA points, we have sudden changes of the slope of the estimated curve, for which it is very hard to provide justification. Considering that the unknown frontier function is often assumed to be continuously differentiable and seems to be so in practice, we are motivated to develop a smooth constrained nonparametric frontier estimator less suffering from non-smoothness and underestimation problems.

The main idea of our proposal stems from the works of Hall and Huang (2001), who reweighted the data to impose monotonicity on kernel regression methods and Du, Parmeter, and Racine (2013), who extended the idea to the multivariate and multi-constraint setting. Their common idea is to consider a transformed data by introducing weights for each response and compute the ordinary kernel regression estimator from the transformed data as the final estimator. The weights are decided to make the resulting estimator from the transformed data obey the given constraints (on an unspecified conditional mean), and fit well the given data. Inspired by this insight, we propose a smooth frontier function estimator by developing an appropriate weight-choosing scheme for frontier estimation. Specifically, we choose the weights to make the resulting estimator obey typical constraints in frontier estimation such as monotonicity or monotone concavity, envelop all the data and minimize the area under the frontier curve. We follow Hall, Park, and Stern's (1998) idea that finding a curve beyond the data points under which the area is minimized gives a good estimator in frontier analysis, and some theoretical justification of the idea can be found in Hall, Park, and Stern (1998). As typical in kernel smoothing methods, performance of the estimator depends on the choice of the bandwidth. We propose that an appropriate bandwidth can be determined by analogy to the popular Bayesian information criterion (BIC). This criterion is shown to work well in our simulation studies in Section 3. Actually, the idea of adapting the mean regression method of Du et al. (2013) and Hall and Huang (2001) to frontier estimation was previously considered in Parmeter and Racine (2013). However, our method is distinctively different from theirs in the minimization criterion and the strategy of bandwidth selection. The details about the difference are discussed in Section 2.3.

The rest of the paper is organized as follows. Section 2 describes in detail the proposed smooth frontier estimator, including computation via linear programming as well as bandwidth selection method. Section 3 provides comparison with the most known frontier estimation methods in productivity analysis such as the FDH, LFDH and DEA estimators through Monte Carlo experiments. Section 4 illustrates the utility of our method for the performance analysis of electric utility companies. Finally, the conclusion is given in Section 5.

2. Methodology

Our interest is to estimate from the data the boundary curve of the support of a bivariate distribution with density $f(x, y)$ in $\mathbb{R}^2$. Before introducing our estimator, we will describe some details of the setting for it. We assume that the support Ψ of f is of the form

$$\Psi = \{(x, y)|y \le \varphi(x)\} \supseteq \{(x, y)|f(x, y) > 0\}, \{(x, y)|y > \varphi(x)\} \subseteq \{(x, y)|f(x, y) = 0\}, \tag{1}$$

where φ is an increasing or concavely increasing function whose graph corresponds to the locus of the curve above which the density f is zero. Further, we assume that the data $(x_i, y_i),\ i = 1, \ldots, n$ from the density $f(x, y)$ is obtained from the following data generating process:

$$y_i = \varphi(x_i) - u_i,$$

Journal of the Korean Data & Information Science Society
2014, **25**(4), 915–920

http://dx.doi.org/10.7465/jkdi.2014.25.4.915
한국데이터정보과학회지

Stationary analysis of the surplus process in a risk model with investments[†]

Eui Yong Lee[1]

[1]Department of Statistics, Sookmyung Women's University

Received 15 May 2014, revised 5 June 2014, accepted 24 June 2014

Abstract

We consider a continuous time surplus process with investments the sizes of which are independent and identically distributed. It is assumed that an investment of the surplus to other business is made, if and only if the surplus reaches a given sufficient level. We establish an integro-differential equation for the distribution function of the surplus and solve the equation to obtain the moment generating function for the stationary distribution of the surplus. As a consequence, we obtain the first and second moments of the level of the surplus in an infinite horizon.

Keywords: Integro-differential equation, moment generating function, risk model, stationary distribution, surplus process.

1. Introduction

In this paper, we consider a continuous time surplus process in a risk model with investments. The surplus, with initial level of $u > 0$, linearly increases at a constant rate $c > 0$ due to the incoming premium. Meanwhile, claims arrive according to a Poisson process of rate $\lambda > 0$ and decrease the level of the surplus jump-wise by random amounts which are independent and identically distributed with distribution function G of mean $\mu > 0$. The premium rate c is usually assumed to be larger than $\lambda\mu$ which is the expected amount of claims per unit time. Hence, the surplus process goes eventually to infinity with probability 1 in the classical risk model.

In the present risk model, it is assumed that an investment of the surplus to other business is made whenever the surplus reaches a sufficient level $V > u$. The amounts of investments are independent and identically distributed with distribution function H. To analyze stochastically the level of the surplus in an infinite horizon, we also assume that the surplus process continues to move even though the level of the surplus becomes negative. In the classical risk model, we stop the surplus process and say that a ruin occurs if the level of the surplus goes below 0. However, in practice, the insurance company never lets the surplus of a policy get

[†] This work was done while the author was on sabbatical leave from September 2013 to August 2014 and supported by the National Research Foundation of Korea (NRF) Grant funded by the Korean Government (MOE) (NRF-2013R1A1A2006750).

[1] Professor, Department of Statistics, Sookmyung Women's University, Seoul 100-741, Korea. E-mail: eylee@sookmyung.ac.kr

infinitely large and/or still manages to operate the policy though the surplus of the policy is exhausted. Hence, our assumptions are very practical.

The concept of the investment of the surplus was introduced by Jeong *et al.* (2009) and Jeong and Lee (2010). They studied some optimal policies related to managing the surplus in the risk model. Cho *et al.* (2013) also studied some transient and stationary behaviors of the surplus in the risk model with investments. However, in all previous models, the amount of the investment was a constant.

The classical 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 interesting 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 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. Kwon *et al.* (2014) studied some approximations of the ruin probability of the surplus in the classical risk model.

Dufresne and Gerber (1991) generalized the classical risk model by assuming that the surplus is perturbed by diffusion between the time points of occurrence of claim and studied the ruin probabilities of the surplus. Won *et al.* (2013) generalized the risk model of Dufresne and Gerber (1991) by assuming that there are two types of claim, where one occurs more frequently but the size is stochastically less than the other, and studied the ruin probabilities of the surplus. Song *et al.* (2012) generalized the classical risk model by assuming that the premium rate depends on the current level of the surplus and obtained the joint distribution of the time of ruin, the surplus before ruin and the deficit at ruin.

However, until now, most works have been concentrated on the ruin probability of the surplus and its related characteristics. In this paper, by assuming that the surplus process continues though the level of it goes below 0 and an investment to other business is made when the level of it reaches a sufficient level V, we study the long-term behavior of the surplus in our modified risk model. In Section 2, we establish an integro-differential equation for the distribution function of the surplus and solve the equation to obtain the moment generating function for the stationary distribution of the surplus. In Section 3, by differentiating the moment generating function, we obtain the first and second moments of the level of the surplus in the long-run (in an infinite horizon).

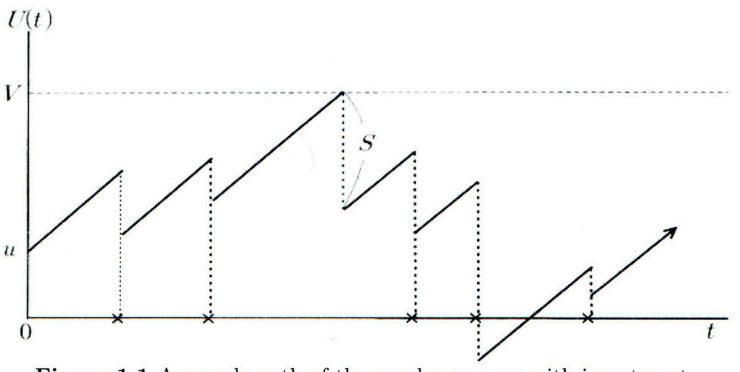


Figure 1.1 A sample path of the surplus process with investment

컴퓨터과학과

Energy-Efficient and High Performance CGRA-based Multi-Core Architecture

Yoonjin Kim and Heesun Kim

Abstract—**Coarse-grained reconfigurable architecture (CGRA)-based multi-core architecture aims at achieving high performance by kernel level parallelism (KLP). However, the existing CGRA-based multi-core architectures suffer from much energy and performance bottleneck when trying to exploit the KLP because of poor resource utilization caused by insufficient flexibility. In this work, we propose a new ring-based sharing fabric (RSF) to boost their flexibility level for the efficient resource utilization focusing on the kernel-stream type of the KLP. In addition, based on the RSF, we introduce a novel inter-CGRA reconfiguration technique for the efficient pipelining of kernel-stream on CGRA-based multi-core architectures. Experimental results show that the proposed approaches improve performance by up to 50.62 times and reduce energy by up to 50.16% when compared with the conventional CGRA-based multi-core architectures.**

Index Terms—**Embedded systems, coarse-grained reconfigurable architecture (CGRA), multi-core, kernel level parallelism (KLP)**

I. INTRODUCTION

The flexibility of a system is very important to accommodate the short time-to-market requirements for embedded systems. On the other hand, application-specific optimization of embedded system becomes inevitable to satisfy the market demand for designers to meet tighter constraints on cost, performance and power. To compromise these incompatible demands, coarse-grained reconfigurable architecture (CGRA) has emerged as a suitable solution for embedded systems [1]. It can boost the performance by adopting multiple processing elements while it can be reconfigured to adapt to evolving characteristics of the embedded applications like audio, video and graphics processing. However, there is a limit when a CGRA is expected to improve the performance of an entire application. This is because single CGRA is sequentially optimized for the parallelized computations in a kernel at a time whereas the overall speedup of the entire application can be achieved by kernel level parallelism (KLP) that several kernels concurrently run. Therefore, such a limitation of single CGRA has resulted in the appearance of CGRA-based multi-core architecture which allows for the multi-CGRA to support diverse KLPs – running separate kernels or inter-dependent kernels (kernel-stream) in parallel.

However, the existing CGRA-based multi-core architectures suffer from much energy and performance bottleneck when trying to achieve the KLP. This is because the existing multi-CGRA structures are not flexible enough to adaptively support various cases of the KLP. It means that the resources in the multi-CGRAs cannot be efficiently utilized under monotonous aggregation of several CGRAs. Therefore, boosting their flexibility level for the efficient resource utilization is considered as a serious concern. For improving their flexibility, this paper provides a new multi-CGRA fabric with a novel reconfiguration technique focusing on the kernel-stream type of the KLP and its hardware

Manuscript received Nov. 24, 2013; accepted Apr. 7, 2014
Dept. of Computer Science, Sookmyung Women's University, Korea
E-mail : {ykim, khs9101}@sookmyung.ac.kr
Yoonjin Kim is the corresponding author for this paper.

implementation.

The paper has following contributions:

- A new ring-based sharing fabric (RSF) has been proposed for raising the flexibility level of the CGRA-based multi-core architectures.
- A novel inter-CGRA reconfiguration technique on the RSF has been introduced for efficient pipelining of kernel-stream on CGRA-based multi-core architectures.
- RT-level design and synthesis have been carried out with varying the number of CGRAs to demonstrate the cost-effectiveness of the proposed approaches in reducing energy while enhancing its performance compared with the existing architecture model.

This paper is organized as follows. After the related work in Section II, we briefly describe CGRA-based multi-core architecture and its representation as preliminaries in Section III. In Section IV, we present the motivation of our approaches. Then we propose the ring-based sharing fabric (RSF) and the inter-CGRA reconfiguration technique in section V. Section VI illustrates the inter-CGRA control mechanism and the experimental results are given in Section VII. Finally we conclude the paper in the Section VIII.

II. Related Works

Until now, there have been a few multi-core architecture projects based on CGRAs for kernel-level parallelism [2-5]. However, most of them are monotonous aggregation of several CGRAs. For example, Samsung reconfigurable processor (SRP)-based multiprocessors have been presented in [6, 7]. In [6], the multiprocessor consists of sixteen reconfigurable processors through networks-on-chip (NoC) interconnection with mesh type topology for exploiting data parallelism of volume rendering. However, the experimental result shows modest performance improvement compared with CPU/GPU improvement. It is because there is performance limitation with general NoC-based multi-core architecture. Another SRP-based multi-core architecture is shown in [7] - it is composed of ARM9 processor, two CGRAs, and AHB bus which couples them. Even though they have demonstrated the software implementation of DVB-T2 on dual CGRAs running at 400 MHz, this work also shows performance limitations because of inefficient resource utilization in the multi-core architecture. In addition, power/energy evaluation of the multiprocessors are not shown in both cases [6, 7].

The Xentium tiles [8] is another example of CGRA-based multi-core architecture. The Xentium is a programmable digital signal processing tile and the different tile processors are connected to a router on NoC. It turns out that the hyperspectral image compression algorithm can indeed be efficiently mapped on this multi-tiled architecture and adding more tiles give a close to linear speedup. However, it is unclear whether other applications may be also successfully mapped on to the architecture already specialized for the compression algorithm. In addition, adding more tiles means increase of power consumption as well as speedup but such a power issue has not been dealt with in [8].

Multi-core architecture with dynamically reconfigurable array processors [10] is more flexible than [6-9] because the shared data-memory banks are connected to all processing cores through crossbar switches unlike communication among the CGRAs is only restricted by NoC or on-chip bus in [6-9]. However the centralized shared data-memory banks may cause performance bottleneck with much power consumption when the number of cores increases. In addition, there are no quantitative evaluation and analysis about power, area, and timing with increasing the number of cores in [10].

III. Preliminaries

1. CGRA-Based Multi-Core Architecture

Typically, a coarse-grained reconfigurable architecture (CGRA)-based multi-core architecture includes general purpose processors (GPP), multi-CGRA, and their interface. Fig. 1 shows such an example of the CGRA-based multi-core architecture – it is composed of a GPP, a DMA, four CGRAs, and on-chip communication architecture like networks-on-chip (NoC) or on-chip bus which couples them. The GPP executes control intensive, irregular code segments and the multi-CGRA performs data-intensive kernel code segments – in this paper, we make use of the multi-CGRA in Fig. 1 as base architecture for comparison with proposed architecture.

Each CGRA consists of PE array (PA), data buffer

Fault-tolerant CGRA-based Multi-Core Architecture[1]

Seungyun Sohn, Heesun Kim, and Yoonjin Kim
Department of Computer Science, Sookmyung Women's University, Seoul 140-742, Korea
E-mail : {kjkj35, khs9101, ykim}@sookmyung.ac.kr

With the increasing defect rate in semiconductor technology, there has been growing need for fault-tolerant systems that can maintain functionality in the presence of multiple failed components. Specially, such a fault-tolerance is necessary for embedded systems to be severely constrained by the harsh environment. To guarantee reliability in the embedded system, the computing engines such as embedded processsors or accelerators should be tolerant to prevent decommissioning the entire systems from a few failures. In order to implement fault-tolerant computing engines, most of researches have taken redundant modules into account – the extra modules enable system to keep its running even though the main module meets fault. Therefore, in terms of redundancy, coase-grained reconfigurable architecture (CGRA)-based multi-core architecture has been expected as a suitable solution because CGRAs include multiple processing elements and storage units that can be reconfigured. Until now, there have been a few research projects based on fault-tolerant CGRA with voter [1] or undo-retry [2]. However, they only show implementations applied with general fault-tolerant design schemes as well as their works are limited to single CGRA. Therefore, in this paper, we propose a new reliable multi-CGRA fabric for the efficient replacement of the modules driving malfuction. The proposed fabric enables exploiting the inherent redundancy and reconfigurability of the CGRA for fault-recovery.

We make use of the multi-CGRA in Fig 1 as base architecture for comparison with proposed architecture. It is composed of a general purpose processor (GPP), a DMA, four CGRAs, and

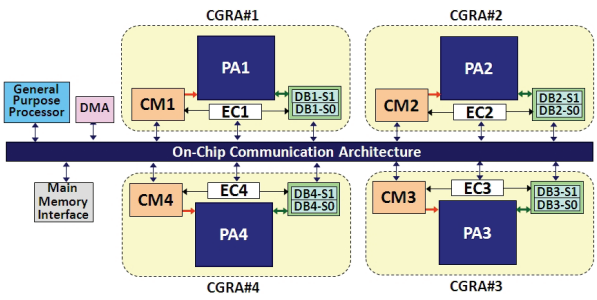


Fig 1. CGRA-based multi-core architecture.

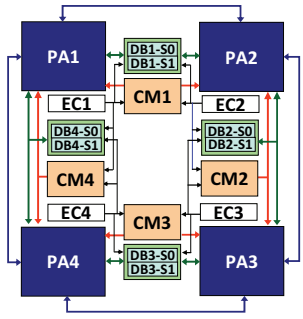


Fig 2. Ring-based sharing fabric (RSF).

[1] This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2011-0013282) and by IC Design Education Center (IDEC).

on-chip communication architecture which couples them. Single CGRA consists of PE array (PA), data buffer (DB), configuration memory (CM), and execution controller (EC). The proposed multi-CGRA fabric based on four CGRAs is shown in Fig. 2 – it is called ring-based sharing fabric (RSF). The RSF connects all of the PAs through single-cycle interconnections and a DB (or a CM) is shared by two adjacent PAs as Fig 2. Such interconnections are utilized for inter-CGRA reconfiguration that enables efficient resource utilization when faults occur - inter-CGRA reconfiguration means that use of each component is not limited to a CGRA.

In order to show the process of the fault-recovery on the RSF, we assume an example that some faults occur in the base architecture as Fig 3. In this case, only one component is broken in each CGRA but no CGRA can normally operate despite most of components are available – it is very typical case that entire system is decomissioned by a few failures. However, Fig 4 shows that the same faults occur in the RSF but it makes 3 CGRAs alive by efficient replacement of the broken modules whereas the base architecture loses all of CGRAs. Such a fault-recovery can be achived by inter-CGRA reconfiguration exploiting the inherent redundancy of the multi-CGRA.

Experimental results show that the proposed RSF shows area/delay/power overhead of up to 19%/10%/21.73% with increasing the number of CGRAs (4~16) when compared with the conventional CGRA-based multi-core architectures.

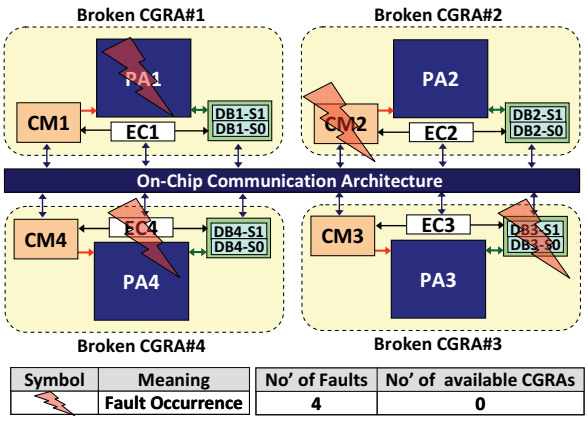


Fig 3. An example when 4 CGRAs have faults.

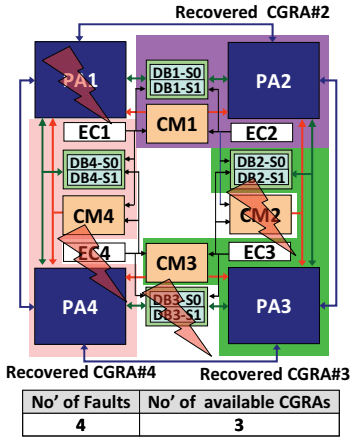


Fig 4. Fault-recovery on RSF.

[1] Dawood Alnajjar, et al, "Implementing Flexible Reliability in a Coarse-Grained Reconfigurable Architecture," *IEEE Trans. on Very Large Scale Integration Systems*, vol. 21, no. 12, pp. 2165-2178, December 2013.

[2] Muhammad Moazam Azeem, et al, "Error Recovery Technique for Coarse-Grained Reconfigurable Architectures," in *Proc. of IEEE Int. Symp. on Design and Diagnostics of Electronic Circuits & Systems*, pp. 441-446, April 2011.

Intra/Inter-CGRA Co-Reconfiguration for Efficient CGRA-based Multi-Core Architecture[1]

Heesun Kim, Seungyun Sohn, and Yoonjin Kim
Department of Computer Science, Sookmyung Women's University, Seoul 140-742, Korea
E-mail : {khs9101, kjkj35, ykim}@sookmyung.ac.kr

Coarse-grained reconfigurable architecture (CGRA) has emerged as a suitable solution for embedded systems but there is a limit as to what a CGRA can improve performance. This is because single CGRA is sequentially optimized for the parallelized computations in a kernel at a time whereas the entire speedup of the applications can be achieved by kernel level parallelism (KLP) that several kernels concurrently run [1]~[4]. Therefore, CGRA-based multi-core architectures have appeared to support diverse KLPs. However, the existing CGRA-based multi-core architectures suffer from much energy and performance bottleneck because the existing multi-CGRA structures are not flexible enough to adaptively support various cases of the KLP. It means that the resources in the multi-CGRAs cannot be efficiently utilized under monotonous aggregation of several CGRAs. For improving their flexibility, we introduce a novel intra/inter-CGRA co-reconfiguration technique based on a ring-based sharing fabric (RSF) focusing on the kernel-stream type of the KLP.

We make use of the multi-CGRA in Fig 1 as base architecture for comparison with proposed architecture. It is composed of a general purpose processor (GPP), a DMA, four CGRAs, and on-chip communication which couples them. Single CGRA consists of PE array (PA), data buffer (DB), configuration memory (CM), and execution controller (EC). The proposed RSF based on four CGRAs connects all of the PAs through single-cycle interconnection and a DB (or a CM) is

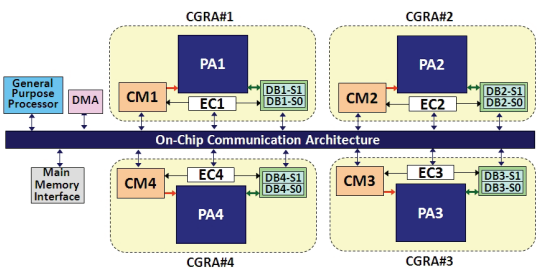


Fig 1. CGRA-based multi-core architecture.

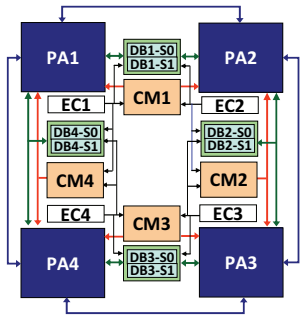


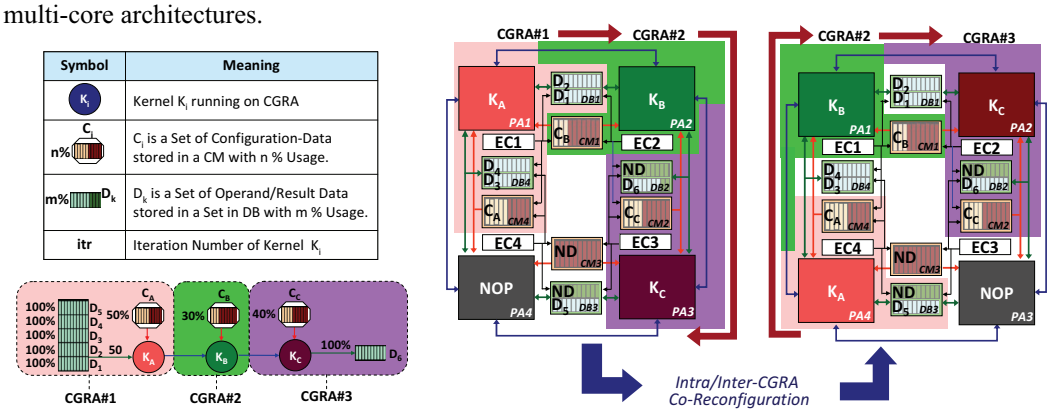
Fig 2. Ring-based sharing fabric.

[1] This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2011-0013282) and by IC Design Education Center (IDEC).

shared by two adjacent PAs as Fig 2. Such connectivity shows minimal interconnection overhead even though the number of CGRAs increases because the design overhead is only interconnections and switching logics between two adjacent PAs.

In order to show the process of the proposed co-reconfiguration technique, we assume an example that the pipelining of kernel-stream requires three DBs (500%) and it iteratively runs on the RSF at 50 times iterations as Fig 3 (a) – each data-set (100%) includes operand-data for the iterative running at 10 times. In this example, base architecture should be sequentially performed with DMA transfer because of insufficient DB capacity. It causes much energy and performance bottleneck. The lack of DB resources may be also exposed on the RSF but we can alleviate the limitation of the RSF by shifting configuration of kernel-stream on the RSF. Fig 3 (b) shows initial configuration of the kernel-stream that PA1 utilizes DB1 (D_1 and D_2) and DB4 (D_3 and D_4) for the running of 40 iterations. Then the RSF can be configured as Fig 3 (c) that shows the utilization of one more DB (DB3) for the remaining 10 iterations. The utilization of DB3 (D_5) can be achieved by shifting the configurations of PAs from 'PA1->PA2->PA3' to 'PA4->PA1->PA2'. In this case, the intra-CGRA reconfiguration means that PA1 and PA2 are reconfigured twice in order to perform K_A/K_B and K_B/K_C. On the other hand, the inter-CGRA reconfiguration enables that three CGRAs are configured with different number of CMs/DBs and connected through the direct interconnections. Such a co-reconfiguration can start immediately because each CM is shared by two adjacent PAs that are dynamically reconfigurable. It means that the pipelining of the kernel-stream continually runs on the RSF without stall as shown in Fig 4.

Experimental results show that the proposed approaches improve performance by up to 50.62 times and reduce energy by up to 50.16% when compared with the conventional CGRA-based multi-core architectures.



(a)Pipelining of kernel-stream with 3 DBs (b) Configuration#1 (c) Configuration#2

Fig 3. Intra/inter CGRA co-reconfiguration for the kernel-stream with three DBs on the RSF.

Ring-based Sharing Fabric for Efficient Pipelining of Kernel-Stream on CGRA-based Multi-Core Architecture

Heesun Kim, Seungyun Sohn, Yoonjin Kim
Embedded Systems Laboratory, Department of Computer Science
Sookmyung Women's University, Seoul, 140-1742, South Korea
{ khs9101, kjkj35, ykim}@sookmyung.ac.kr

Abstract

Coarse-grained reconfigurable architecture (CGRA)-based multi-core architecture aims at achieving high performance by kernel level parallelism (KLP). However, the existing CGRA-based multi-core architectures suffer from much energy and performance bottleneck when trying to exploit the KLP because of poor resource utilization caused by insufficient flexibility. In this work, we propose *a new ring-based sharing fabric (RSF)* to boost their flexibility level for the efficient resource utilization focusing on the kernel-stream type of the KLP. In addition, we introduce *a novel inter-CGRA reconfiguration technique* for the efficient pipelining of kernel-stream based on the RSF. Experimental results show that the proposed approaches improve performance by up to 50.62 times and reduce power by up to 44.03% when compared with the conventional CGRA-based multi-core architectures.

Keywords

Embedded systems, coarse-grained reconfigurable architecture (CGRA), multi-core, kernel level parallelism (KLP).

1. Introduction

The flexibility of a system is very important to accommodate the short time-to-market requirements for embedded systems. On the other hand, application-specific optimization of embedded system becomes inevitable to satisfy the market demand for designers to meet tighter constraints on cost, performance and power. To compromise these incompatible demands, coarse-grained reconfigurable architecture (CGRA) has emerged as a suitable solution for embedded systems [1]. It can boost the performance by adopting multiple processing elements while it can be reconfigured to adapt to evolving characteristics of the embedded applications like audio, video and graphics processing. However, there is a limit when a CGRA is expected to improve the performance of an entire application. This is because single CGRA is sequentially optimized for the parallelized computations in a kernel at a time whereas the overall speedup of the entire application can be achieved by kernel level parallelism (KLP) that several kernels concurrently run. Therefore, such a limitation of single CGRA has resulted in the appearance of CGRA-based multi-core architecture which allows for the multi-CGRA to support diverse KLPs – running separate kernels or inter-dependent kernels (kernel-stream) in parallel.

However, the existing CGRA-based multi-core architectures suffer from much energy and performance bottleneck when trying to achieve the KLP. This is because the existing multi-CGRA structures are not flexible enough to adaptively support various cases of the KLP. It means that the resources in the multi-CGRAs cannot be efficiently utilized under monotonous aggregation of several CGRAs. Therefore, boosting their flexibility level for the efficient resource utilization is considered as a serious concern. For improving their flexibility, this paper provides a new multi-CGRA fabric with a novel reconfiguration technique focusing on the kernel-stream type of the KLP and its hardware implementation.

The paper has following contributions:

- A new ring-based sharing fabric (RSF) has been proposed for raising the flexibility level of the CGRA-based multi-core architectures.
- A novel inter-CGRA reconfiguration technique has been introduced for efficient pipelining of kernel-stream on the RSF.
- RT-level design and synthesis have been carried out with varying the number of CGRAs to demonstrate the cost-effectiveness of the proposed approaches in reducing power while enhancing its performance compared with the existing architecture model.

This paper is organized as follows. After the related work in Section 2, we briefly describe CGRA-based multi-core architecture and its representation as preliminaries in Section 3. In Section 4, we present the motivation of our approaches. Then we propose the ring-based sharing fabric (RSF) and the inter-CGRA reconfiguration technique in section 5 and the experimental results are given in Section 6. Finally we conclude the paper in the Section 7.

2. Related works

Until now, there have been a few multi-core architecture projects based on CGRAs for kernel-level parallelism [2]-[5]. However, most of them are monotonous aggregation of several CGRAs. For example, Samsung reconfigurable processor (SRP)-based multi-processors have been presented in [6][7]. In [6], the multiprocessor consists of sixteen reconfigurable processors through networks-on-chip (NoC) interconnection with mesh type topology for exploiting data parallelism of volume rendering. However, the experimental result shows modest performance improvement compared with CPU/GPU improvement. It is because there is performance limitation with general NoC-based multi-core architecture. Another SRP-based multi-core architecture is shown in [7] - it is composed of ARM9 processor, two CGRAs, and AHB bus which couples them. Even though they have demonstrated the software implementation of DVB-T2 on dual CGRAs running at 400MHz, this work also shows performance limitations because of inefficient resource utilization in the multi-core architecture. In addition, power/energy evaluation of the multiprocessors are not shown in both cases [6][7].

The Xentium tiles [8] is another example of CGRA-based multi-core architecture. The Xentium is a programmable digital signal processing tile and the different tile processors are connected to a router on NoC. It turns out that the hyperspectral image compression algorithm can indeed be efficiently mapped on this multi-tiled architecture and adding more tiles give a close to linear speedup. However, it is unclear whether other applications may be also successfully mapped on to the architecture already specialized for the compression algorithm. In addition, adding more tiles means increase of power consumption as well as speedup but such a power issue has not been dealt with in [8].

Multi-core architecture with dynamically reconfigurable array processors [10] is more flexible than [6]-[9] because the shared data-memory banks are connected to all processing cores through crossbar switches unlike communication among the CGRAs is

only restricted by NoC or on-chip bus in [6]-[9]. However the centralized shared data-memory banks may cause performance bottleneck with much power consumption when the number of cores increases. In addition, there are no quantitative evaluation and analysis about power, area, and timing with increasing the number of cores in [10].

3. Preliminaries

3.1. CGRA-based multi-core architecture

Typically, a coarse-grained reconfigurable architecture (CGRA)-based multi-core architecture includes general purpose processors (GPP), multi-CGRA, and their interface. Figure 1 shows such an example of the CGRA-based multi-core architecture – it is composed of a GPP, a DMA, four CGRAs, and on-chip communication architecture like networks-on-chip (NoC) or on-chip bus which couples them. The GPP executes control intensive, irregular code segments and the multi-CGRA performs data-intensive kernel code segments – in this paper, we make use of the multi-CGRA in Figure 1 as base architecture for comparison with proposed architecture.

Each CGRA consists of PE array (PA), data buffer (DB), configuration memory (CM), and execution controller (EC). The PA has identical processing elements (PEs) containing functional units and a few storage units. The PA has reconfigurable interconnections between PEs for efficient data-transfer. The DB provides operand data to PA through a high-bandwidth data bus. The CM is composed of configuration elements (CEs) and each CE provides context word to configure each PE. The EC has control data that contains execution cycles, read/write mode and addresses of the DB and the CE for correct operations of the PA.

3.2. Symbolic representation

In this paper, we bring up the problems of resource utilization in the conventional multi-CGRA and propose new approaches to overcome such issues. Therefore, panoptic illustration of resource utilization in multi-CGRA is necessary for intelligible explanation of our approaches. In this section, we define an efficient way expressed in symbols to show such a utilization status as Figure 2.

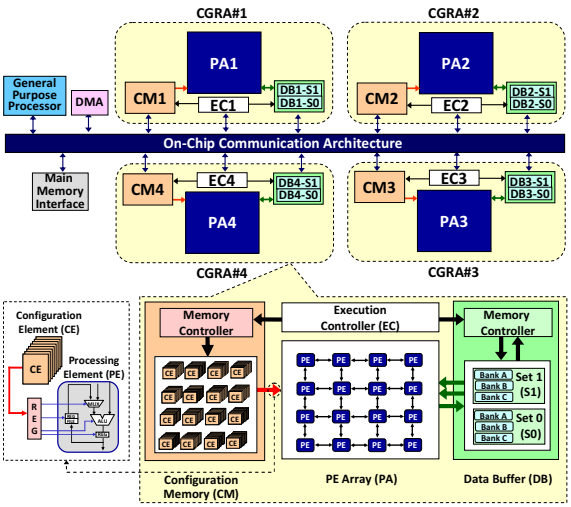
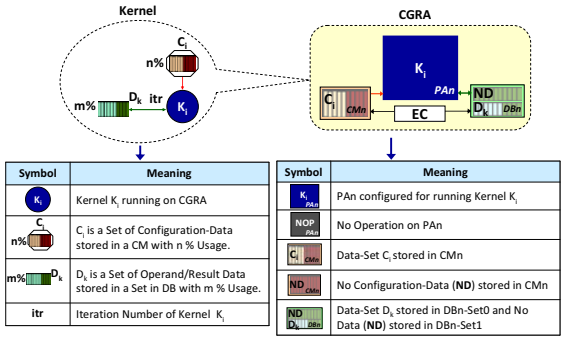


Figure 1: CGRA-based multi-core architecture.



(a) Symbols for kernel **(b)** Symbols for CGRA

Figure 2: Symbolic representation of kernel running on CGRA.

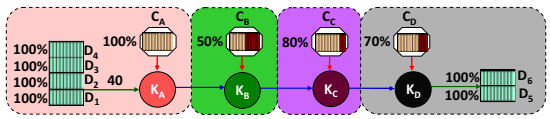
It shows the symbolic representation of the resource utilization with CM/DB usage when kernel K_i run on a CGRA. The meaning of the symbols for kernel and CGRA are defined in Figure 2 (a) and Figure 2 (b) respectively.

4. Motivation

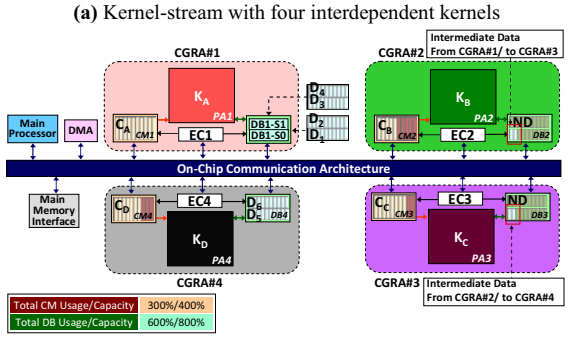
In this section, we present the motivation of our approaches. The main motivation comes from the resource utilization problems when trying to exploit kernel level parallelism (KLP) on multi-CGRA. Even though various cases of the KLP can be considered, we focus on the pipelining of kernel-stream that is the most complex and ever-changing case.

4.1. Pipelining of kernel-stream

Pipelining of kernel-stream is a type of the KLP and means that interdependent kernels (kernel-stream) iteratively run on multi-CGRA in the manner of pipelining. In this case, each kernel may be mapped on each CGRA without any problems of resource utilization if each kernel requires CM/DB usage less than CM/DB capacity of each CGRA. However, there may be more cases causing poor utilization of resources as Example#1 in Figure 3 - four interdependent kernels (kernel-stream) iteratively run on the base



(a) Kernel-stream with four interdependent kernels



(b) Kernel-stream iteratively runs on the base multi-CGRA

Figure 3: Example#1 – pipelining of the kernel-stream composed of four interdependent kernels on the base multi-CGRA.

소셜 네트워크에서의 Hadoop 기반 실시간 광고 효과 분석 시스템 설계

방지선, 이아름, 옥윤청, 김윤희
숙명여자대학교 컴퓨터과학부
e-mail : nin23bix@gmail.com, arsbsp@gmail.com, oyc1216@gmail.com, yulan@sm.ac.kr

A Hadoop-based System of Analyzing Real-time Advertisement Effectiveness in Social Network

Jiseon Bang, A-Reum Lee, YoonJung Ock, Yoonhee Kim
Dept. of Computer Science, Sookmyung Women's University

요 약

소셜 네트워크 서비스의 증가로 인해 개인의 관심분야의 수집과 분석이 용이해졌을 뿐만 아니라, 많은 양의 정보를 활용할 수 있게 되었다. 이에 빅 데이터를 이용한 분석이 여러 분야에서 제안되고 있다. 한편, 광고효과 측정 방법에 있어서 빅 데이터 분석은 많은 부분 정확도가 떨어지고, 시간이 오래 걸린다는 단점이 있었다. 때문에 본 시스템에서는 소셜 네트워크에서의 데이터를 파싱하여 TV 광고에 대한 사람들의 반응을 분석하고 그 효과를 그래프로 보여주도록 제작하였다. 본 시스템을 통해 광고효과 분석이 기존보다 빨라졌으며 다양한 방식의 분석이 가능해졌다.[1]

1. 서론

SNS(Social Network Service)의 확산과 동향[1]에 따르면 한국인터넷 진흥원이 실시한 2010년 인터넷 이용실태 조사보고서에서 SNS이용자가 65.7%로 높은 부분을 차지하고 있다고 한다. 또한 SNS 이용자 수는 현재까지 지속적으로 증가해왔으며 앞으로도 증가해 갈 것으로 추정된다. 이와 같이 점차 확산되고 있는 SNS는 신속성, 개인성, 정보의 개방성과 같은 특성을 지닌다. 이에, SNS를 분석했을 때 보다 개인적이고 진실한 정보를 얻을 수 있을 뿐만 아니라 실시간으로 빠르게 갱신되는 정보들을 분석할 수 있다고 추측할 수 있다. 따라서 SNS를 분석함으로써 실시간으로 유효한 데이터를 통해 광고 효과를 분석 할 수 있을 것이다.

SNS는 개인의 관심분야를 수집·분석하는 데에 많이 이용되고 있고 Social big data, 금융혁신의 새로운 도구[2]에서도 국내 금융회사가 보유 고객에 대한 내부 정보와 SNS를 통한 외부의 데이터를 통합 분석하여 고객 맞춤형 마케팅 전략에 적극적으로 활용할 필요성이 있다고 밝혔다. SNS의 빅 데이터를 실시간으로 분석하기 위해서는 SNS의 정보를 얻어오고, 빅 데이터를 분석하기 위한 별도의 시스템이 구성되어있어야 한다. 따라서 기존에 존재하지 않던 실질적인 시스템이 개발되어야 하며, 해당 시스템은 기존의 광고 분석 기법에 의거하여 SNS의 정보를 분석하고 처리하여 타당한 결론을 낼 수 있어야 한다. 이를 위해서, 본 논문에서 제안하는 시스템은 하둡을 이용하여 빅 데이터를 빠르게 분석한다. 빅 데이터는 그 방대한 양으로 인하여 기존 시스템으로는 분석에 시간 소요가 매우 많이 든다는 단점이 있다. 하지만 하둡을 이용하면 빠르게 빅 데이터를 분석할 수 있기 때문에, 이를 이용하여 광고 효과 분석에 이용하면 분석 시간을 크게 단축할 수 있다.

한편, 제안하는 시스템은 광고효과 측정방법에 있어서도 새로운 접근법을 제시하고 있다. 기존의 광고 효과 측정방법은 일일 후 회상 조사(DAR)와 같은 설문조사나 광고 인지율 조사가 대부분이었다[3]. 이는 정확도가 떨어지고 실제 광고 직후 반응을 얻기가 힘들다. 이에, 광고에 대한 반응을 실시간으로 얻을 수 있는 프로그램의 필요성이 대두되었다. 본 논문에서 제안하는 시스템의 목적은 실시간으로 광고 전후의 소셜 네트워크에서의 언급도를 비교하여 광고 반응을 알 수 있도록 분석하고 광고효과를 판단하는 데 있다. 이 시스템은 SNS 빅데이터를 이용한 프로그램이라는 점에서 광고 분석 기법의 새로운 전환점이 될 수 있을 것이라 기대한다. 기존의 방식보다 더 빠르면서도 더 정확한 처리를 가능하게 하는 것이 제안하는 시스템 설계의 목적이다.

기존의 광고 효과는 개별 광고에 대해서만 조사를

[1] 본 연구는 2013 년도 정부(미래창조과학부)재원으로 한국과학창의재단 청년 과학융합 창업 아이디어 창출활동 지원사업으로 수행되었음. 과제번호: 2013AAC0016

할 수 있었다. 하지만 본 논문에서 제안하는 시스템은 개별 광고뿐만이 아니라 같은 주제에 대한 여러 광고를 함께 분석할 수 있다. 기존에 한 번의 조사를 통해 한 개의 광고만을 조사할 수 있었던 것에 비하여 제안하는 시스템이 갖는 장점이라고 할 수 있다. 이 기능을 통해 특정 주제에 대해서 어떤 기법의 광고가 가장 높은 인지도를 내는지, 많은 긍정적 반응을 이끌어 내는지를 알아낼 수 있다. 또한 특정 광고에 대해서 동종광고를 비교 분석할 수도 있다. 이를 통해 타 광고에 비해서 분석하려는 광고가 얼마나 많이 시청되는지 또한 알 수 있다.

제안하는 시스템에서는 트위터와 광고의 데이터를 수집하여, 이를 토대로 세 가지 시나리오에 대해 그래프를 보여준다. 또, 이용자가 광고주일 때에 한하여 새로운 광고를 만드는 데 도움을 주고, 루머가 발생했을 때 알려주는 시스템을 소개한다. 본 논문의 구성은 다음과 같다. 2장에서는 관련 연구에 대하여 설명하고, 3장에서는 시스템 설계와 구현에 대해 설명한다. 4장에서는 분석 결과를 기술하며 5장에서 결론을 맺는다.

2. 관련연구

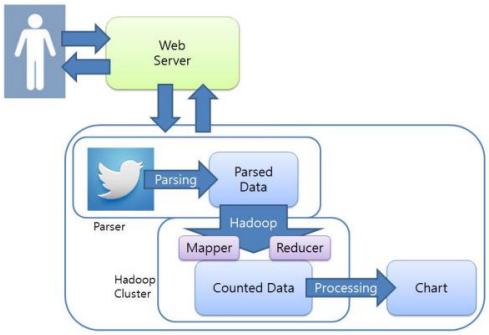
본 논문에서 제안하는 시스템에서 다루는 TV 광고에 대한 광고효과 분석 기법으로는 시청률조사가 있는데, 시청률조사기법은 각 가정에 셋톱박스를 설치하는 방식으로 측정한다. 하지만 조사 기관이 민간기업이고, 그 수가 너무 적어 독점에 대한 문제제기가 있어왔다. 또한 조사 기준이나 절차에 있어서도 문제점이 나타났는데, 구체적으로, 대표성이 미흡하고 비일관적이라는 것이다. 또한, 통계적 오차 범위까지 벗어나, 신뢰도에 있어서도 의심을 받고 있다[4].

한편, 현재 SNS 정보를 수집, 분석하여 보여주는 사이트가 존재하는데, 트윗트렌드[5]와 소셜메트릭스[6]가 그것이다. 하지만, 트윗트렌드의 경우 분석결과가 매우 적어 정확한 분석이 어려웠으며, 소셜 메트릭스의 경우 다양하게 분석이 가능했으나, 여러 광고에 대한 비교는 불가능했다. 또한 두 사이트 모두 분석에 어느 정도의 시간 소요가 있었다.

하둡을 이용한 소셜 네트워크 분석에 대한 논문은 2012 년부터 발표되었고 지난 해인 2013 년에도 발표되었다[7], [8]. [7]번 논문은 하둡을 이용하여 소셜 네트워크 데이터를 이용한 분석이 가능함을 제시한 논문으로, 하둡 시스템과 한글 형태소 분석 프로그램을 도입한 시스템 모델을 설계함으로써 하둡이 SNS 정보를 빠르게 처리할 수 있다는 가능성을 열어 주었다. 하지만 하둡으로 데이터를 처리했을 때, 속도 개선 이외의 다른 성능에 대한 설명이 없었고, SNS 분석에 있어 구체적인 결과에 대한 설명도 미흡했다. [8]번 논문은 한 층 더 발전된 형태로, 직접 크롤링을 통해 데이터를 전달받고, 하둡 시스템을 이용하여 광고 효과를 그래프와 수치로 분석하였다. 이 시스템은 하둡을 이용하여 광고효과를 분석하는 프로그램을 처음 만들었다는 점에서 의의가 있었다. 본 논문은 이 논문의 후속논문으로, 광고 효과를 언급도 이외의 여러 항목에 대해 분석하고, 단일 광고뿐만 아니라 여러 광고를 비교할 수 있도록 하였다.

3. 설계 및 구현

본 논문에서 제안하는 시스템의 형식은 (그림 1)과 같다. 전체 시스템은, 트위터에서 사용자들이 작성한 글 내용을 시간정보와 함께 수집하는 파서, 이 정보를 토대로 조건에 맞는 글의 개수를 계산하는 하둡 프로그램, 이렇게 계산된 정보를 바탕으로 그래프를 보여주는 웹 인터페이스로 이루어져 있다. 각 기능의 자세한 설명은 아래와 같다.



(그림 1) 시스템 구조도

시스템은 크게 데이터 파서(Parser), 하둡 클러스터(Hadoop Cluster), 그리고 웹 서버(Web Server)로 구성되어 있다. 파서는 트위터 글과 광고정보 사이트 데이터를 수집한다. 두 정보는 각각 다른 방식으로 수집되고 저장된다. 자세한 설명은 3.1절에서 다룬다.

하둡 클러스터는 트위터 데이터의 파싱 결과를 세 가지 시나리오 별로 조건에 맞는 항목의 개수를 계산한다. 여기에서 세 가지 시나리오에 대한 설명은 다음과 같다.

- 해당 광고가 어떤 시간대에 방송되는 것이 가장 효과적인지를 알고 이후 광고 방송 시 활용하기 위해 특정 광고에 대한 시간대별 분석을 실시한다.
- 해당 광고가 효과적인 광고인지 여부를 측정하기 위하여 특정 광고에 대한 긍정적/부정적 반응 정도를 분석한다.
- 동종 광고 중에서 어떤 광고가 가장 효과적인지를 측정하기 위하여 동종 광고의 긍정적/부정적 반응 정도를 분석한다.

한편, 이용자가 광고주인 경우에는 다른 두 가지 기능을 제공하는데, 광고 제작 시 활용할 수 있도록 좋은 평가를 받은 광고를 추천해 주는 기능, 루머 발생시 초기 대응을 위해 미리 알려주는 기능이 그것이

클라우드 컴퓨팅에서 Bag-of-tasks 응용을 위한 전력량 고려 가상 머신 할당 기법

(Power-aware Provisioning of Virtual Machines for Bag-of-tasks Applications in Cloud Computing)

최지은 O, 안윤선, 김윤희

숙명여자대학교 컴퓨터과학과

jechoi1205@gmail.com, ahnysun1@gmail.com, yulan@sm.ac.kr

요 약

최근 자원을 가상화하고 기업 또는 개인이 컴퓨터 시스템을 유지 보수하는 비용과 시간 및 인력을 줄일 수 있는 클라우드 컴퓨팅에 관한 연구가 활발히 진행되고 있다. 또한 그린 컴퓨팅이 새로운 트렌드로 자리잡으면서 높은 컴퓨팅 수준을 보장하면서 전력량을 줄일 수 있는 연구가 이뤄지고 있다. Green500 과 같이 전세계의 슈퍼컴퓨터를 에너지 효율성이 높은 순으로 리스트를 제공하는 등 전세계의 관심이 증가하고 있어 에너지 효율적인 컴퓨팅 기법의 필요성이 대두되고 있다. 본 논문에서는 클라우드 컴퓨팅 환경에서 대량의 Bag-of-tasks 응용을 처리하는데 있어서 전력 소모량을 고려한 가상 머신 할당하는 기법에 대해 제안한다. 선행 연구의 경우 빠른 수행시간 내에 작업을 끝낼 수 있는 자원 프로비저닝 기법을 제안하고 있지만 전력 소모량은 고려하고 있지 않다. 따라서 본 논문에서는 사용자의 데드라인 요구사항을 충족하면서 플랫폼 제공자의 전력 소모량을 감소시킬 수 있는 기법에 따라 실험을 진행하였다. 전산유체역학(CFD) 응용에 대해 데드라인 내에 작업을 완료할 수 있는 가장 낮은 주파수 가상 머신을 선택하는 알고리즘에 따라 줄어든 전력 소모량을 계산하여 선행연구와 비교하였다.

키워드: 프로비저닝, Bag-of-tasks, 클라우드 컴퓨팅, 가상 머신, 전력 소모량, 전산유체역학

1. 서론

클라우드 컴퓨팅은 사용자가 지불한 만큼 컴퓨팅 서비스를 사용할 수 있도록 하는 웹 기반의 컴퓨팅 패러다임이다. 클라우드 제공자가 제공하는 서비스의 형태에 따라 Infrastructure as a Service (IaaS), Platform as a Service (PaaS), Software as a Service (SaaS)로 구분된다. Service Level Agreements (SLAs)에 따라 사용자와 클라우드 제공자 사이의 컴퓨팅 성능에 대한 요구사항 및 성능 보장이 이루어진다[1]. 기존에는 컴퓨팅 성능에 중점을 둔 슈퍼컴퓨터가 존재했지만, 최근 Green500[6]과 같이 전세계의 슈퍼컴퓨터를 에너지 효율성이 높은 순으로 평가하여 리스트를 제공하는 등의 에너지 소모량에 관한 세계적인 관심이 높아지고 있는 수준이다[2]. 이처럼 그린 컴퓨팅이 새로운 트렌드로 자리하면서 클라우드 컴퓨팅 환경에서 데이터 센터의 전력량을 고려하는 연구의 중요성이 강조되고 있다.

클라우드 컴퓨팅에서 전력량을 고려한 기법은 Static Power Management (SPM)과 Dynamic Power Management (DPM)로 크게 두 가지로 나눌 수 있다. SPM 기법은 저전력 에너지 효율적인 하드웨어를 사용하여 최대 전력 소비량과 에너지 사용량을 줄이는 기법이다. DPM 기법은 현재 자원 이용률과 응용의 작업배치를 기반으로 에너지 사용량을 줄이는 기법이다. DPM 기법의 대표적인 기술로는 CMOS 회로의 동적 전력을 줄여 전체 전력 소비량을 줄이는 Dynamic Voltage and Frequency Scaling (DVFS) 기술이 있다. DVFS 는 CPU 주파수나 공급전압을 조절하여 CPU 동작 속도를 줄여 전체 에너지 사용량을 줄이는 기술이다[2]. 따라서 본 논문에서는 저전력 가상 머신 할당 기법을 위해 낮은 CPU 주파수의 VM 을 스케줄링 하는 알고리즘을 제안한다.

한편 선행 연구에서는 데드라인 기반으로 작업을 스케줄링하고, 일정한 주기마다 자원 및 작업에 대한 모니터링을 통해 오토 스케일링을 수행하는 알고리즘을 제안하였다. 또한 성능과 비용 지향 정책에 따라 자원을 적절히 프로비저닝 하여 자원 활용의 효율 극대화 및 효율적인 비용 산출을 목표로 하고 있다.

본 논문에서는 기존의 오토 스케일링 기법 연구에서는 고려하지 않았던 수행 응용에 대한 전력 소비량을 고려하여 적정 자원에 작업을 수행시키는 가상 머신 할당 기법을 개발하여 시뮬레이션을 진행하였으며, 각각의 전력량을 비교 실험하였다.

본 논문의 구성은 다음과 같다. 2 장에서는 관련 연구에 대해 설명하며, 3 장에서는 제안하는 저전력 스케줄링 알고리즘을 소개한다. 4 장에서는 알고리즘을 클라우드 컴퓨팅 환경에서 구현하여 실험한 내용을 다루고, 5 장에서 결론을 맺는다.

2. 관련 연구

클라우드 컴퓨팅이 새로운 패러다임으로 화두가 되면서, 데이터센터에서 시스템 신뢰도와 운영비용을 절감시키는 저전력 기법이 중요해지고 있다. DVFS 는 자원의 동작 전압과 동작 주파수를 조절하여 전력량을 감소시키는 기법으로 관련 연구가 활발히 진행되고 있다.

이와 관련하여 [1]연구에서는 실시간 클라우드 서비스에 대한 전력량을 고려한 가상 자원 프로비저닝 기법을 소개한다. 이 연구에서는 클라우드 환경의 실시간 서비스를 데드라인에 관점으로 구분하였으며, 데드라인에 따른 Hard Real-Time (HRT) 서비스와 Soft Real-Time (SRT) 각각에 적합한 DVFS 기법을 소개하고 있다. HRT 서비스에 적용 가능한 DVFS 기법은 또한 3 가지로 구분된다. 가상 머신 요청에 대해 가장 낮은 프로세서 속도를 가진 가상 머신을 할당하는 Lowest-DVFS 기법, 현재의 요구되는 MIPS 율을 δ만큼 오버 스케일링 하여 가상 머신을 할당하는 δ-Advanced-DVFS 기법, 평균 응답시간과 평균 데드라인을 이용한 Adaptive-DVFS 기법이 있다. 이 경우 데드라인을 위반하는 서비스에 대해서는 거절 함으로서 정책을 유지한다. 반면 SRT 서비스는 데드라인을 넘어서는 서비스에 대해서 지연에 따라 패널티를 부과하는 정책을 소개한다. 이 기법은 태스크의 요구성능에 따라 가상 머신을 제공하여 사용자에게 불필요한 비용 부담을 안길 수 있다.

[3]연구는 기존의 DVFS 연구들이 하나의 주파수를 사용하는 것을 보완하여 최대 주파수와 최소주파수 2 개를 사용하여 동적으로 주파수를 변화시키는 DVFS 기법을 제안한다. 그러나 이 기법은 최대 주파수에서 최소 주파수로 주파수를 조절하는 시점을 계산하는 복잡한 계산과정이 존재하기 때문에 이에 따른 오버헤드가 발생할 수 있다.

[4]연구는 데드라인을 고려하고 Bag-of-tasks 응용에 대해 Dynamic Voltage Scaling (DVS) 기법을 적용한 스케줄링 알고리즘을 제안한다. 하지만 이는 클라우드 환경이 아닌 그리드 환경에서 진행되었다. 이와 관련하여 [8]연구에서는 보다 클라우드 환경에서 복잡한 응용인 워크플로우 응용을 대상으로 하여 사용자의 데드라인 요구 조건을 위반하지 않으면서 오토 스케일링 기술을 통해 비용을 감소시키는 연구를 진행하였다. 하지만 이 연구는 간단한 워크플로우만을 이용한 한계가 있으며 실험을 진행하지 않아 기법의 타당성 및 효율성을 검증할 수 없다.

[5]연구에서는 데드라인과 사용자 비용을 고려하여 자원을 할당하고, 사용자가 정의한 규칙을 기반으로 자원 규모의 확장과 축소를 결정하는 기법을 제안한지만 자원의 전력 소모를 고려하지 않는다.

본 논문에서는 클라우드 컴퓨팅 환경에서 전력량을 고려한 가상 자원 할당 기법을 제안한다. 고려되는 클라우드 서비스는 사용자가 요구하는 데드라인을 지켜야 하는 HRT 서비스이다. 제안한 알고리즘은 시뮬레이션을 통해 검증한다. 시뮬레이션에서 사용되는 응용은 많은 양의 Bag-of-tasks 응용으로 실험하였다.

3. Power-aware Scheduling

본 논문에서는 클라우드 컴퓨팅 환경에서 Bag-of-tasks 응용에 대한 가상 자원 할당 기법을 제안한다. 본 장에서는 제안하는 스케줄링 알고리즘을 소개한다.

1) 전력량 고려 스케줄링 알고리즘

본 논문에서 제안하는 전력량을 고려한 가상 머신 스케줄링 알고리즘은 그림 1 과 같다. 먼저 응용이 SLA 와 함께 사용자에 의해 제출된다. 이후 제출된 응용은 작업의 형태로 기다리고 있고, 알고리즘에 따라 작업에 대한 가상 머신(VM)의 스케줄링이 제공된다.

Algorithm – Power-aware Scheduling
Input – Waiting jobs of the applications,
SLA = {a deadline *D*},
Output – Scheduling decision *S* = {jobs -> *VMs*}

```
1: Sort waiting jobs in decreasing order of execution length;
2: VM <- null;
3: for each job j do
4:     for each private vm do
5:         If vm_AvailableTime < CurrentTime then
6:             EST_vm = CurrentTime;
7:         Else
8:             EST_vm = vm_AvailableTime;
9:         VM <- find a vm on which j can start the lowest frequency
                                                    within the D;
10:        EFT_vm <- EST_vm + ET_vm;
11:        schedule j to vm;
12:        continue with the next job;
13:    end for
14: end for
15: return S;
```

그림 1. 저전력 스케줄링 알고리즘

과학 계산 실험을 위한 클라우드 자원을 활용한 실험 프로비넌스 모델 설계

안윤선 김윤희☆

숙명여자대학교 컴퓨터과학부

{ahnysun, yulan}@sookmyung.ac.kr

A Design of Provenance Model for Scientific Computation Experiments over Cloud Environment

Younsun Ahn Yoonhee Kim☆

Dept. of Computer Scinence, Sookmyung Women's University

요 약

과학 응용 실험 중 다양한 파라미터 값을 실행 파일에 적용하고 반복적으로 수행하여 결과를 얻어 파라미터 값에 따른 실험 결과를 비교 분석하는 파라메트릭 스터디(Parametric Study) 응용들이 다수 존재한다. 이러한 응용의 경우 복잡한 계산으로 인해 장시간의 고성능 컴퓨팅 환경을 독점적으로 사용해야 하는 등 실험에 많은 비용이 요구된다. 따라서 실험의 결과를 정규화하고 체계적으로 관리를 할 수 있다면 같은 실험에 대한 반복 실험을 배제하고 유사실험에 대한 효율적인 실험 환경 설정이 가능하다. 실험의 환경과 결과를 정규화하여 데이터를 재사용 할 수 있도록 구성하기 위해 실험 데이터에 대한 프로비넌스가 필요하다. 저장된 프로비넌스 데이터를 이용하여 같은 실험을 생략할 수 있고 비슷한 조건에서 수행수 해야할 때 요구되는 자원 환경의 예측도 가능하다. 본 논문은 프로비넌스 데이터 모델에 대해서 제안하고 실험 환경에 대한 응용 요구사항과 필요 자원, 실험결과를 효율적으로 관리 하도록 한다. 또한 프로비넌스 데이터에 대한 자원 온톨로지 정보를 활용하여 유사실험을 실시할 시 요구되는 자원에 대한 예측과 실험결과를 사전에 예측하여 효율적인 실험이 가능하도록 하고, 특히 장시간 실험에 치명적인 실패율이 적은 자원환경을 고려 할 수 있도록 하여 실험의 성공률을 높이도록 하였다.

1. 서 론

과학 응용 실험 중 다양한 파라미터 값을 실행 파일에 적용하고 반복적으로 수행하여 결과를 얻어 파라미터 값에 따른 실험 결과를 비교하고 분석하는 파라메트릭 스터디(Parametric Study) 응용들이 존재한다. 이러한 응용의 경우 효율적인 수행을 위해 실험의 결과를 정규화하여 기존의 실험을 재사용하도록 구성하는 프로비넌스가 필요하다.

프로비넌스는 실험 결과의 근원정보를 기록하는 것이며 이전 실험 결과의 출처를 이용하여 같은 결과를 얻기 위해서 사용한다[1]. 프로비넌스 데이터에 작업의 환경, 작업의 이력, 작업 수행 동안의 정보를 기록한다.

저장된 프로비넌스 데이터를 이용하여 같은 실험을 다시 수행할 때 비슷한 조건에서 수행할 수 있도록 하며 다음 실험을 효율적으로 수행 할 수 있게 한다.

클라우드 환경에서는 다양한 클라우드 자원을 필요한 만큼 빌려 쓸 수 있다. 기존에 수행한 실험의 프로비넌스 데이터를 이용하여 응용에 적합하고 효율적인 클라우드 자원을 선택하여 작업을 수행 할 수 있다.

이를 위해 작업 수행 환경인 클라우드 자원의 온톨로지를 구축하고 프로비넌스 데이터에서 사용된 클라우드 자원에 대한 부분을 가지고 온톨로지를 생성한다. 온톨로지는 지식 기반이 구축될 수 있는 기본 구조를 제공하는 도메인 개념들의 명확한 표현으로 정의된다[2]. 온톨로지를 기반으로 자원을 표현하면 자원의 특성을 개념화 하고, 의미 있게 그들 사이의 관계를 나타낼 수 있다.

생성한 온톨로지로 다음 같은 실험을 수행할 때 클라우드 자원의 온톨로지와 비교하여 유사도가 높은 자원을 선택하여 작업을 보다 빠른 시간 내에 작업의 실패율을 낮추어 효율적으로 수행 할 수 있다.

본 논문의 구성은 다음과 같다. 1장의 서론에 이어 2장에서는 관련 연구들을 살펴보고, 3장에서는 본 논문에서 제안하는 프로비넌스 데이터 모델에 대해 설명한다. 4장에서는 실험에 대해 살펴본다. 마지막으로 5장에서 결론을 맺는다.

2. 관련 연구

실험을 효율적으로 수행하기 위해서는 이전 실험의 결과를 가지고 실험을 모든 사람에게 의미 있게 일정한 형태로 만드는 것이 필요하다. 실험의 정규화를 위해

☆ 교신저자

"이 논문은 2013년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임"(NRF-2013R1A1A3007866)

필요한 프로비넌스에 대한 연구들이 이루어졌다.

[1]에서는 프로비넌스를 두 가지의 포맷으로 나누어 제시하였다. 프로스펙티브 프로비넌스(Prospective provenance)는 작업의 명세와 실험이 수행되는 단계를 나타내며, 레트로스펙티브 프로비넌스(Retrospective provenance)는 수행 중인 작업의 자세한 정보와 수행 환경에 대한 정보를 나타내고 있고, [3]에서는 프로비넌스 모델을 워크플로우 형태와 노드와 엣지 형태로 나타내었다. 하지만 정확히 어떠한 정보를 프로비넌스 데이터로 정의할지 나타내지 않았다.

온톨로지를 기반으로 자원을 선택한 연구는 [4]가 존재한다. [4]는 그리드 환경에서 온톨로지를 기반으로 자원 명세를 표현하고 온톨로지간 유사도를 비교하여 사용자가 요구하는 환경과 유사한 자원을 찾는 방법을 제안한다. 이를 바탕으로 클라우드 환경에서 온톨로지를 구축하고자 한다.

본 논문에서는 클라우드 환경에서 프로비넌스 데이터를 이용하여 실험을 정규화하고 자원에 대한 온톨로지를 바탕으로 이전 실험의 환경과 비슷한 자원을 선택하여 실험 결과를 예측 할 수 있고, 실험의 신뢰도를 높이고자 한다.

3. 제안한 연구

본 논문에서 제안한 프로비넌스 모델은 그림 1과 같다. 프로비넌스 모델은 반복적인 과학 응용의 효율적인 수행을 위해 필요한 데이터들을 나타내고 있다. 실험을 수행할 때마다 프로비넌스 데이터가 저장되어 축적된다.

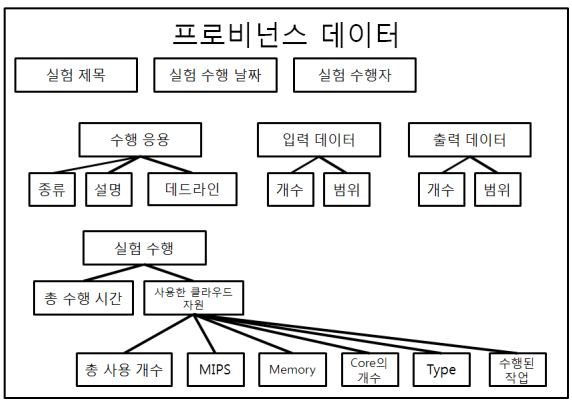


그림1 프로비넌스 데이터 모델

그림 1은 프로비넌스 데이터 모델을 나타내며, 프로비넌스 데이터에 저장되는 실험의 정규화에 필요한 정보들이 나타나있다. 실험에 대한 기본적인 제목, 날짜, 수행자의 정보가 필요하며, 수행 응용에 대한 종류, 설명, 데드라인 정보가 들어간다. 입력 데이터와 출력 데이터의 개수와 범위를 저장하며 실험 수행에서 필요한 총 수행 시간과 사용한 클라우드 자원에 대한 상세한 정보가 저장된다. 클라우드 자원의 온톨로지와 비교하기 위해 필요한, 클라우드 자원의 사용 개수, MIPS, Memory, Core의 개수, Type, 각각의 특정 클라우드 자원에서 수행된 작업 정보를 저장한다.

저장된 프로비넌스 데이터 중 사용한 클라우드 자원 부분의 정보를 가지고 온톨로지를 생성한다. 온톨로지를 생성하여 자원의 특성과 관계를 나타낼 수 있다. 클라우드 자원에 대한 명세를 온톨로지를 기반으로 표현하면 정확히 일치하지 않는 자원이라도 의미상 유사한 자원을 선택할 수 있다.

작업을 클라우드 자원에 스케쥴링 할 경우 성능이 좋고, 비싼 자원을 선택하지 않고, 프로비넌스 데이터를 가지고 구축한 온톨로지와 클라우드 자원의 온톨로지를 비교하여 자원을 선택한다. 이전에 수행한 실험의 환경과 비슷한 패턴, 비슷한 스펙, 유사한 결과를 얻을 수 있는 자원을 선택하여 작업을 클라우드 자원에 할당한다.

4. 실험

본 논문에서의 실험은 파라메트릭 스터디 응용 중 하나인 전산유체역학 (CFD, Computational Fluid Dynamics) [5] 응용을 대상으로 실험을 진행하였다. 전산 유체 역학은 복잡한 유동 현상에 대해 수치 해석 기법을 사용하여 유체 유동 문제를 풀고 해석하는 응용으로 같은 계산을 파라미터 또는 입력 파일을 바꿔가면서 반복적으로 수행한다. 전산 유체 역학 응용은 e-AIRS 2.0[6]을 이용하여 분석한 응용의 평균 길이를 실험에 사용하였다. 응용의 작업의 평균 길이는 660,000MI(Million Instruction)로, 작업의 길이를 정규분포를 이용하여 랜덤 하게 생성하여 수행하였다.

본 실험에서는 하이브리드 클라우드 환경에서 MIPS의 범위가 100~2000인 클라우드 자원을 사용하여 5000개의 전산유체역학 응용 작업들을 여러 번 수행하여 결과를 얻었다.

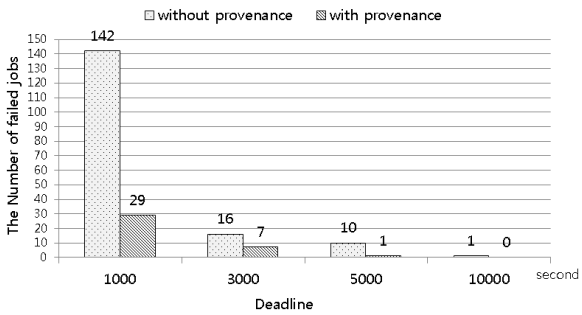


그림 2 프로비넌스 데이터를 이용한 결과와 프로비넌스 데이터를 이용하지 않았을 때 수행 실패한 작업 개수 비교

그림2는 전산유체역학 응용을 클라우드 자원에 할당하여 수행한 결과이다. 프로비넌스 데이터를 이용하지 않았을 때와 프로비넌스 데이터를 이용하여 여러 번 수행한 결과를 비교하였다. 데드라인의 변화에 따라 수행에 실패한 작업들의 평균 개수를 비교하였다.

프로비넌스 데이터를 이용하지 않았을 경우와 프로비넌스 데이터를 이용한 경우 모두 데드라인이 길어짐에 따라 작업의 실패 개수가 줄어든다.

SNS 환경에서의 지능형 실시간 광고 효과 분석 시스템 설계

이아름○ 방지선 안윤선 김윤희
숙명여자대학교 컴퓨터과학과
arsbsp@gmail.com, nin23bix@gmail.com, ahnysun1@gmail.com, yulan@sm.ac.kr

A Design of Analysis System for Intelligent and Real-time Advertisement Effectiveness in SNS Environments

A-Reum Lee○ Jiseon Bang Younsun Ahn Yoonhee Kim
School of Computer Science, Sookmyung Women's University

요 약

SNS(social networking service)에서의 광고에 대한 개인의 반응은 실시간 생성된다. 이를 통해 개인의 관심 분야의 수집과 분석이 용이해졌으며, 많은 양의 정보를 활용할 수 있게 되었다. 본 시스템에서는 특정 광고에 대한 소셜 네트워킹 서비스에서의 개인의 반응을 데이터로 수집하고, 방대한 양의 반응 데이터를 효율적으로 분석하기 위하여 Hadoop을 통한 분산 처리를 하여 실시간에 가까운 분석을 한다. 또한, 보다 신뢰성 높은 분석을 위해 온톨로지를 이용한 지능형 추론 방법을 도입하여 유사 반응 효과 분석을 가능토록 한다. 시스템이 수집한 데이터를 통해 사용자로부터 학습하여 점차 지능화되는 방향으로 확장되어 정확도와 신뢰도가 높아지고 추론이 가능하도록 설계한다.[1)]

1. 서 론

SNS의 사용자 증가로 인해 사이버 스페이스상의 의사소통과 교류가 확대됨에 따라 SNS의 방대한 데이터를 분석하여 다양한 분야에 활용하는 연구가 활발하다. SNS의 확산과 동향[1]에 따르면 한국 인터넷진흥원이 실시한 2010년 인터넷 이용실태 조사에서 SNS이용자가 65.7%로 높은 부분을 차지하고 있다고 한다. SNS 이용자 수는 지속적으로 증가해왔으며 앞으로도 증가해 갈 것으로 추정된다. 이와 같이 점차 확산되고 있는 SNS는 신속성, 개인성, 정보의 개방성과 같은 특성을 지닌다. 이에 근거하여 SNS를 분석했을 때보다 개인적이고 진실한 정보를 얻을 수 있을 거라고 추측할 수 있으며 실시간으로 빠르게 갱신되는 정보들을 분석할 수 있다는 장점을 가진다. 따라서 SNS를 분석함으로써 실시간으로 유효한 데이터를 통해 광고 효과를 분석 할 수 있을 것으로 판단하였다.

SNS의 빅데이터를 실시간으로 분석하기 위해서는 SNS의 정보를 수집하고, 빅데이터를 분석하기 위한 별도의 시스템이 구성되어있어야 한다. 따라서 기존에 존재하지 않던 실질적인 시스템이 개발되어야 하며, 해당 시스템은 기존의 광고 분석 기법에 의거하여 SNS의 정보를 분석하고 처리하여 타당한 결론을 낼 수 있어야 한다. 이를 위해서, 본 논문에서 제안하는 시스템은 하둡을 이용하여 빅 데이터를 빠르게 분석한다. 빅데이터는 그 방대한 양으로 인하여 기존 시스템으로는 분석에 시간 소요가 매우 많이 든다는 단점이 있다. 반면, 하둡을 이용하면 빅 데이터를 짧은 시간 내에 분석할 수 있기 때문에, 이를 이용하여 광고 효과 분석에 이용하면 분석 시간을 크게 단축할 수 있다.

한편, 제안하는 시스템은 광고효과 측정방법에 있어서도 새로운 접근법을 제시하고 있다. 기존의 광고효과 측정방법은 일일 후 회상 조사(DAR)와 같은 설문조사나 광고 인지율 조사가 대부분이었다[2]. 이는 정확도가 떨어지고 실제 광고 직후 반응을 얻기가 힘들다. 이에, 광고에 대한 반응을 실시간으로 얻을 수 있는 프로그램의 필요성이 대두되었다.

따라서 본 논문에서는 광고 반응 분석과 광고 효과 판단에 있어 신뢰성 높고 정확한 처리가 가능한 시스템을 제작하기 위해, 이전 논문[3]에서 구현했던 실시간 광고 효과 분석 시스템에 온톨로지를 적용하여 의미 데이터의 관계를 구성하고, 의미론적 추론 방법을 활용하여 기존 정보를 통해 새로운 정보를 자동으로 생성시키도록 한다. 이로 인해 기존의 방식보다 더 빠르면서 정확한 처리를 가능하게 하며, SNS 데이터를 단순 매칭이 아닌 의미에 따른 추론 방법을 활용하여 광고에 대한 반응을 분석할 수 있는 실시간 광고 효과 분석 시스템을 설계하고자 한다. 본 논문의 구성은 다음과 같다. 1장의 서론에 이어 2장에서는 관련 연구들을 살펴보고, 3장에서는 제안하는 시스템의 기능 설계에 대해 설명하며 마지막으로 4장에서 결론을 맺는다.

2. 관련연구

1) 이 논문은 2014년 정부재원(미래창조과학부 여대학(원)생 공학연구팀제 사업)으로 한국연구재단과 한국여성과학기술인지원센터의 지원을 받아 연구되었습니다.

TV광고에 대한 광고 효과 분석 기법으로는 셋톱박스를 설치하여 실시하는 시청률 조사가 있다. 하지만 조사기관이 민간 기업이고, 그 수가 너무 적어 독점에 대한 문제제기가 있어왔다. 또한, 조사 기준이나 절차에 있어서도 대표성이 미흡하고 비일관적이라는 단점이 있었으며, 통계적 오차 범위에서 벗어나 신뢰도에 있어서도 의심을 받고 있다[4].

현재 SNS 정보를 수집, 분석하여 보여주는 사이트로써 트윗트렌드[5]와 소셜메트릭스[6]가 존재한다. 하지만, 트윗트렌드의 경우 분석결과가 매우 적어 정확한 분석이 어려웠으며, 소셜 메트릭스의 경우 다양하게 분석이 가능했으나, 여러 광고에 대한 비교는 불가능했다. 또한 두 사이트 모두 분석에 일정시간의 소요가 있었다.

하둡을 이용한 소셜 네트워크 분석에 대한 논문은 2012 년부터 발표되어왔다. Hadoop을 이용한 트위터 메시지 분석 시스템 설계[7]에서는 하둡을 이용하여 소셜 네트워크 데이터를 이용한 분석이 가능함을 제시하고 하둡 시스템과 한글 형태소 분석 프로그램을 도입한 시스템 모델을 설계함으로써 하둡이 SNS 정보를 빠르게 처리할 수 있다는 가능성을 열어 주었다. 하지만 속도 개선 이외의 다른 성능에 대한 언급이 없었고, SNS 분석에 있어 구체적인 결과 분석도 미흡했다. 하둡을 이용한 소셜 네트워킹의 TV광고효과 분석 시스템 설계[8]에서는 한 층 더 발전된 형태로, 직접 파싱을 통해 데이터를 전달받고, 하둡 시스템을 이용하여 광고 효과를 그래프와 수치로 분석하였다. 이 시스템은 하둡을 이용하여 광고효과를 분석하는 프로그램을 처음 만들었다는 점에서 의의가 있었지만 다양한 형태의 분석이 불가능했다.

한편, 온톨로지는 지식을 기반으로 하여 구축할 수 있는 기본 구조를 제공하는 도메인 개념들의 명확한 표현이라고 정의할 수 있다[9]. 온톨로지는 또한 정보시스템의 대상이 되는 자원의 개념을 명확하게 정의하고 상세하게 기술하여 보다 정확한 정보를 찾을 수 있도록 하는 데 목적이 있다. 이를 이용하면 자원에 대한 정보의 질을 향상시킬 수 있다. 때문에, 광고 효과 분석에 있어 더 정확한 정보와 예측을 위해서는 온톨로지 기술 접목이 필요하다.

이에 본 논문은 광고 효과를 언급도 이외의 여러 항목에 대해 분석하고, 단일 광고가 아닌 여러 광고를 비교할 수 있도록 한다. 뿐만 아니라, 실시간 광고 효과 분석 시스템에 온톨로지를 적용하여 광고 반응 분석과 광고 효과 판단에 있어 신속하고 정확한 처리가 가능한 분석 프로그램을 만들고자 한다.

3. 기능 설계

본 논문에서 제안하는 시스템의 구조는 그림 1과 같다. 전체 시스템은 이전 연구[3]에서 구현하였던, 트위터에서 사용자들의 글 내용을 시간정보와 함께 수집하는 파서, 조건에 맞는 글의 개수를 계산하여 분석결과를 얻는 하둡 프로그램, 이렇게 계산된 정보를 바탕으로 그래프를 보여주는 웹 인터페이스로 이루어지며 프로세싱 과정에서 온톨로지를 도입하여 시스템을 확장한다.

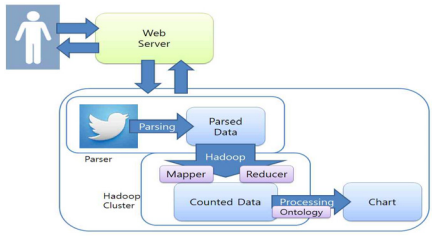


그림 1 시스템 구조도

파서는 트위터 글과 광고정보 사이트 데이터를 수집한다. 파서로는 Jericho, Jsoup HTML Parser를 이용하여 정보를 각각 다른 방식으로 수집, 저장한다. 제안하는 시스템에서 TV광고 정보를 수집하는 사이트는 TVCF라는 사이트로, 1950년부터 현재까지 국내외 TV광고를 카테고리로 분류해 놓았다[10]. 각 데이터는 빠른 업데이트를 위해 한 시간에 한 번씩 crontab을 이용하여 파싱을 수행하며 트위터에서는 글쓴이와 내용, 게시시각을, TVCF에서는 광고명과 카테고리, 광고시각을 얻어와 트위터 데이터 키워드와 매핑 한다.

하둡 클러스터는 트위터 데이터의 파싱 결과를 세 가지 시나리오 별로 조건에 맞는 항목의 개수를 계산한다. 여기에서 세 가지 시나리오에 대한 설명은 다음과 같다.

1) 가장 효과적인 방송 시간을 알고 이후 활용하기 위해 특정 광고에 대한 시간대별 분석을 실시한다.
2) 효과적인 광고인지를 측정하기 위해 광고에 대해 온톨로지 기반 키워드 분석을 통한 긍정적/부정적 반응 정도를 분석한다.
3) 동종 광고 중에서 가장 효과적인 광고를 알기 위해 동종 광고의 긍정적/부정적 반응 정도를 분석한다.

한편, 이용자가 광고주인 경우에는 광고 제작 시 활용할 수 있도록 광고를 추천해 주는 기능과, 루머 발생 시 초기 대응을 위해 이를 미리 알려주는 루머알림 기능의 두 가지 기능을 추가적으로 제공한다. 여기에서 광고 추천 기능은, 동종 광고 중 이전까지의 분석 결과가 가장 긍정적이었던 광고를 추천해 주는 형식으로 구현하였으며, 루머 알림 기능의 경우는 온톨로지 딕셔너리에 설정된 특정 키워드가 단기간 안에 일정 횟수 이상 언급되었을 시 이를 미리 알려주는 형식으로 확장하여 구현한다.

다섯 가지 기능을 구현하기 위해서 제안하는 시스템에서는 하둡을 이용한다. 하둡은 파싱을 통해 얻은 정보 중에서 조건에 맞는 항목의 개수를 세어준다. 하둡 프로그램은 맵과 리듀스 과정을 통해 이루어지는데, 트위터 데이터를 입력 값으로 넣으면, 맵에서는 시나리오에 따라 시간이나 조건 별 항목의 값이 value가 되고, 조건의 이름이 key값이 되도록 한다. 리듀스 과정에서는 이렇게 생성된 값을 통해 실제 개수의 계산이 이루어진다. 맵리듀스 과정을 통해 얻어진 정보는 key, value 쌍으로 정렬되어 지정된 경로에 저장되도록 하였다.

첫 번째 기능은 시간대별로 광고의 언급도를 분석하여 도표화하는 것이다. 이는 얼마나 많은 사람들이 해당 광

Auto-Scaling of Virtual Resources for Scientific Workflows on Hybrid Clouds

Younsun Ahn
Dept. of Computer Science
Sookmyung Women's University
Seoul, Korea
ahnysun@sookmyung.ac.kr

Yoonhee Kim*
Dept. of Computer Science
Sookmyung Women's University
Seoul, Korea
yulan@sookmyung.ac.kr

ABSTRACT

Cloud computing technology enables applications to employ scalable resources dynamically. Scientists can promote large-scale scientific computational experiments over cloud environment. It is essential for many-task-computing (MTC) to certificate stable executions of applications even rapid changes of vital status of physical resources and furnish high performance resources in a long period. Auto-scaling with virtualization provides efficient and integrated cloud resource utilization. Auto-scaling issues have been actively studied as effective resource management in order to utilize large-scale data center in a good shape but most of the auto-scaling methods just easily support performance metrics such as CPU utilization and data transfer latency but seldom consider execution deadline or characteristics of an application. We propose an auto-scaling method that finishes all tasks by user specified deadline. We accomplish our goal by dynamically allocating VMs to maximize resource utilization while meeting a deadline and considering task dependency and data transfer time in workflow application. We have evaluated our auto-scaling method with protein annotation workflow application which tasks are specified as a workflow in hybrid cloud environment. The results of a simulation show the method performs automatically resource allocation actually needed satisfying deadline constraints.

Categories and Subject Descriptors

C.2.4 [Computer Systems Organization]: Distributed Systems

Keywords

Cloud computing, workflow, auto-scaling, hybrid.

1. INTRODUCTION

The appearance of Science Clouds allows scientists to facilitate large-scale scientific computational experiments over cloud environment. Cloud computing enables applications to employ scalable resources dynamically. It is essential for many task computing (MTC) to certificate stable executions of applications even rapid [1] changes of vital status of physical resources and furnish high performance resources in a long period. Thus, it is getting significant to research computational problem solving environment that gives the management of task executions or resources in the large scale of computation. Auto-scaling with virtualization can use efficient integrative utilization of cloud resources for the computational problem solving environment. Technically, auto-scaling can dynamically change the number of Virtual Machine (VM) during execution of an application.

Our previous paper [1] proposed an auto-scaling method to provide efficient resource utilization in a hybrid cloud computing environment. However, proposed auto-scaling algorithm needs to be extended in order to support various patterns of task execution, which are Bag-of-Tasks [2] and workflows. Tasks in Bag-of-Tasks [2] can be scheduled on resources separately from each other while tasks in workflow can be performed in order of dependency pattern. A workflow is commonly represented by a directed acyclic graph (DAG).

This paper proposes an extended version of the auto-scaling method, reflected in patterns of tasks and the requirements of an application based on cloud computing infrastructure. Especially, it automatically allocates resources depending on tasks in workflow on hybrid cloud. We propose an auto-scaling method that finishes all tasks by user-specified deadline. We accomplish our goal by dynamically allocating VMs to maximize resource utilization while meeting a deadline and considering task dependency and data transfer time in workflow application. We have evaluated our auto-scaling method with a specialized computation and data analysis application such as protein annotation workflow application [3] which tasks are specified as a workflow in hybrid cloud environment. The results of a simulation show the method performs automatically resource allocation actually needed satisfying deadline constraints.

The rest of the paper is organized with the sections as follows; we introduce an overview of related works in Section 2. Section 3 explains auto-scaling algorithm, and Section 4 contains contents about a scenario of using auto-scaling algorithm and experiment results are discussed. Finally, we conclude the paper and discuss future work in final section.

2. RELATED WORK

Cloud computing offers endless resources by virtualization technology and facilitates extension and reduction of resources. Auto-scaling issues are currently being discussed and studied as effective resource management. On a cloud service provider side, "Auto-scaling" of AWS [4], Paraleap [5] for Windows Azure [6], and Scalr [7] use a rule-based auto-scaling method which flexibly increases or decreases resources to meet user-defined metrics. Meanwhile, rule-based auto-scaling methods are uncomplicated enough to allocate resources dynamically, it may not be possible

* Corresponding Author

ScienceCloud 2014, June 23, 2014, Vancouver, BC, Canada.

http://dx.doi.org/10.1145/2608029.2608036

to satisfy an amount of resources actually needed, then it could possibly cause violation of deadline and cost.

From this reason, on a user side, [8], [9], [10], [11], [12] are the studies of auto-scaling in consideration of deadline of applications or cost for resource usage. [8] proposes an auto-scaling method minimizing resource usage cost for Bag-of-Tasks [2] jobs. It is connected with horizontal scaling which adds or removes the number of VMs and vertical scaling which expands or reduces the size of a VM. However, it is still deficient for accomplished resource requirements of dynamic workloads because it is lack of the consideration of resource usage during execution of an application. There are few studies considering workflow scheduling. [9] proposes an algorithm to minimize execution costs while meeting deadline for a workflow. Reference [10] respects to finish the execution of jobs within deadlines at minimum financial cost using their auto scaling method. However, they use only public cloud, which means that they just consider simple environment. [11], [12] facilitate the execution of workflow applications using their auto-scaling methods on Grids, which may be changed to adapt to the Cloud computing environment. [11] proposes an algorithm that minimizes execution cost, that is processing cost and data transmission cost within deadline. [11] divides the workflow into partitions and assigned each partition a sub-deadline, thus [11] can minimize execution time for the entire workflow. [12] proposes an efficient scheduling algorithm for dependent jobs. Also, [12] considers communication cost about data transfer time. We propose an algorithm for workflow referring to Reference [11]'s and [12]'s workflow scheduling algorithms.

In this paper, we propose an extended version of an auto-scaling method based on our previous research [1] which enables dynamic resource allocation considering the types of jobs from Bag-of-Tasks [2] as well as the workflow and the characteristics of an application. Auto-scaling method can automatically allocate cloud resources considering task dependency and data transfer time in workflow application.

3. AUTO-SCALING ALGORITHM

We extend [1]'s auto-scaling algorithm which can perform only tasks in Bag-of-Tasks. Our auto-scaling algorithm can schedule tasks in workflow. Auto-scaling technique can discover delay and deadline violation to comparing actual start time and estimated start time of running tasks. Moreover, monitoring perform every regular term. Algorithm's assumption and notation are referred to [1]. Some notations for the workflow scheduling algorithms are defined at the below:

- EFT_{VM}: Estimated finish time of a VM.
- EST_{VM} : Earliest start time of a VM.
- ET_{VM}: Execution time of a task on a VM.

Algorithm 1, Run-time Scaling is extended based on [1]. We propose algorithm 2, Workflow Scheduling algorithm to deal with tasks in workflow. In the reference [1], Run-time Scaling Algorithm has two polices and depicts our general auto-scaling method. Additionally, we extend three policies by adding workflow scheduling. Tasks are scheduled with applying one of the three policies such as Performance-oriented Scheduling (line 5), cost-oriented Scheduling (line 8) and Workflow Scheduling (line 11) of SLA (Service Level Agreements). Tasks in Bag-of-Tasks [2] are sorted as descending order based on their execution time, while tasks in workflow are sorted as sequential order.

Algorithm 1 – **Run-time Scaling**
Input – An application,
SLA={a policy *P*, a deadline *D* [, minimum performance requirement *minPM*]}
Output – Scaling decision *S* = { *toStartUp*, *toShutDown* }
Scheduling decision *S* = { *tasks* → *VMs*}

```
1:  SCALING ← TRUE;
2:  while (true)
3:      if SCALING is TRUE
4:          switch P
5:          case Performance:
6:              Sort waiting tasks in decreasing order of execution length;
7:                  S ← PerformanceOrientedScheduling(
                                sortedTasks, D, minPM);
8:          case Cost:
9:              Sort waiting tasks in decreasing order of execution length;
10:                 S ← CostOrientedScheduling(sortedTasks, D);
11:         case Workflow:
                Sort waiting tasks in sequential order;
12:                 S ← WorkflowScheduling(sortedTasks, D);
13:         each vm where status is running
14:         if no running/waiting tasks on vm then
15:             destroy the vm
16:         send scaling decisions to DRMS
17:         send scheduling decisions to JES
18:         waitForNextInterval();
19:         SCALING ← SLAMonitoring(runningTasks, D);
```

Algorithm 2 defines our Workflow Scheduling algorithm. In this algorithm, DTT (Data Transfer Time) means data transfer time. Proposed Workflow Scheduling algorithm is based on a PCH algorithm [12]. When our scaling method tries to allocate cloud resources, it followed the performance-oriented policy. We could get tasks on a critical path and group of tasks using PCH algorithm [12] and then each group can be scheduled. The total execution time of critical path in a private cloud resource is calculated and set to a deadline value, also additional time is add to the deadline value.

First, tasks on a critical path are scheduled on a same resource, which can execute all tasks in a critical path (line 2). Every time we try to schedule tasks, it is a rule to choose private cloud resource prior to public cloud. It is important to consider task dependencies and data transfer time in workflow. Each task could get an EFT (Estimated Finish Time) of a parent task and set an EST (Earliest Start Time) value to EFT of a parent task in order to consider order of tasks (line 4-5). With an executing on a same resource of tasks on the critical path, a communication overhead is reduced. Parent task and child task are performed on a same VM, data transfer time is zero.

When tasks which are not on a critical path are scheduled to VMs, tasks check a parent task's state. If parent tasks are allocated to cloud resources, child tasks could be allocated (line 6). Unless parent tasks are scheduled cloud resources, child tasks must wait to be allocated. As mentioned above, each task could get an EFT of a parent task and calculate an EST value considering EFT of a parent task and data transfer time. In case there are a number of parent tasks, a task chooses latest EFT among them and calculates an EST value with EFT of a parent task and data transfer time. (line 7-10 (private), line 12-16 (public)).

클라우드 컴퓨팅 응용을 위한 효율적 전력 사용을 고려한 VM 관리 (Power-aware VM Management for Cloud Computing Applications)[1]

최지은, 안윤선, 김윤희

숙명여자대학교 컴퓨터과학과

jechoi1205@gmail.com , ahnysun@sm.ac.kr , yulan@sm.ac.kr

요 약

최근 자원을 가상화하고, 기업 또는 개인이 컴퓨터 시스템을 유지 보수하는데 드는 비용과 시간 및 인력을 줄일 수 있는 클라우드 컴퓨팅이 많은 관심을 받으면서 관련 연구가 활발히 진행되고 있다. 한편 지구 온난화 등의 문제가 대두되면서 그린 컴퓨팅이 새로운 트렌드로 부각되고 있다. Green500 은 전세계의 슈퍼컴퓨터를 에너지 효율성이 높은 순으로 리스트를 제공하는 것으로 에너지 효율적인 컴퓨팅 기법의 전세계적인 관심이 높아졌다는 것을 보여준다. 이러한 관심을 바탕으로 높은 컴퓨팅 수준을 보장하고 슈퍼컴퓨팅에서 전력량을 고려한 기법을 제안하는 연구가 이뤄지고 있다. 본 논문에서는 클라우드 컴퓨팅에서 전력 소비량을 고려한 가상 머신 할당 기법의 동향을 Bag-of-tasks 응용을 중심으로 분석하였다. 선행 연구의 경우 빠른 수행시간 내에 작업을 끝낼 수 있는 자원 프로비저닝 기법을 제안하고 있지만 전력 소모량은 고려하고 있지 않다. 본 논문에서는 클라우드 컴퓨팅 환경에서 대량의 Bag-of-tasks 응용을 처리하는데 있어서 전력 소모량을 고려한 가상 머신 할당하는 기법에 대해 제안하고 사용자의 데드라인 요구사항을 충족하면서 플랫폼 제공자의 전력 소모량을 감소시키는 것을 실험을 통해 분석하였다. 또한 전산유체역학(CFD) 응용을 대상으로 데드라인 내에 작업을 완료할 수 있는 가장 낮은 주파수 가상 머신을 선택하는 알고리즘에 따라 줄어든 전력 소모량을 계산하여 선행연구와 비교하였다.

Keywords: Bag-of-tasks, Cloud Computing, Virtual Machines, Power consumption, CFD

1. 서론

클라우드 컴퓨팅은 사용자가 지불한 만큼 컴퓨팅 서비스를 사용할 수 있도록 하는 웹 기반의 컴퓨팅 패러다임이다. 클라우드 제공자가 제공하는 서비스의 형태에 따라 Infrastructure as a Service(IaaS), Platform as a Service(PaaS), Software as a Service(SaaS)로 구분된다. Service Level Agreements(SLAs)에 따라 사용자와 클라우드 제공자 사이의 컴퓨팅 성능에 대한 요구사항 및 성능 보장이 이루어진다[1]. 한편, 지구 온난화 등의 문제가 대두되면서 그린 컴퓨팅이 새로운 트렌드로 부각되고 있다. 기존에는 컴퓨팅 성능에 중점을 둔 슈퍼컴퓨터가 존재했지

이 논문은 2013 년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF-2013R1A1A3007866)

만, 최근 Green500[6]과 같이 전세계의 슈퍼컴퓨터를 에너지 효율성이 높은 순으로 평가하여 리스트를 제공하는 등의 에너지 소모량에 관한 세계적인 관심이 높아지고 있는 수준이다[2]. 이처럼 지구 온난화와 같은 환경적 문제를 고려한 그린 컴퓨팅이 새로운 트렌드로 자리하면서 클라우드 컴퓨팅 환경에서 데이터 센터의 전력량을 고려하는 연구의 중요성이 강조되고 있다.

클라우드 컴퓨팅에서 전력량을 고려한 기법으로는 크게 Static Power Management(SPM)과 Dynamic Power Management(DPM)의 두 가지로 나눌 수 있다. SPM 기법은 저전력 에너지 효율적인 하드웨어를 사용하여 최대 전력 소비량과 에너지 사용량을 줄이는 기법이다. DPM 기법은 현재 자원 이용률과 응용의 작업배치를 기반으로 에너지 사용량을 줄이는 기법이다. DPM 기법의 대표적인 기술로는 CMOS 회로의 동적 전력을 줄여 전체 전력 소비량을 줄이는 Dynamic Voltage and Frequency Scaling(DVFS) 기술이 있다. DVFS 는 CPU 주파수나 공급전압을 조절하여 CPU 동작 속도를 줄여 전체 에너지 사용량을 줄이는 기술이다[2]. 따라서 본 논문에서는 저전력 가상 머신 할당 기법을 위해 낮은 CPU 주파수의 가상 머신을 스케줄링 하는 알고리즘을 제안한다.

한편 선행 연구[5]에서는 작업의 종료시간인 데드라인을 기반으로 작업을 스케줄링하고, 일정한 주기마다 자원 및 작업에 대한 모니터링을 통해 오토 스케일링을 수행하는 알고리즘을 제안하였다. 또한 성능과 비용 지향 정책에 따라 자원을 적절히 프로비저닝하여 자원 활용의 효율 극대화 및 효율적인 비용 산출을 목표로 하는 연구를 진행하였다.

본 논문에서는 기존의 오토 스케일링 기법 연구에서는 고려하지 않았던 수행 응용에 대한 전력 소비량을 고려하여 적정 자원에 작업을 수행시키는 가상 머신 할당 기법을 제안하였다. 또한 Bag-of-tasks 형태의 전산유체역학(CFD)[9] 응용을 대상으로 CloudSim[7]을 이용하여 시뮬레이션을 진행하였다. Bag-of-tasks 형태의 응용은 작업간의 순서가 정해지지 않은 응용을 말한다. 시뮬레이션을 통해 선행연구와 비교 실험을 진행하여 본 논문에서 제안한 기법의 효율성을 검증하였다. 또한 데드라인을 변화시키면서 전력 소모량을 분석하는 실험을 진행하여 제안된 기법과 데드라인의 관계를 분석하였다.

본 논문의 구성은 다음과 같다. 2 장에서는 관련 연구에 대해 설명하며, 3 장에서는 제안하는 저전력 스케줄링 알고리즘을 소개한다. 4 장에서는 알고리즘을 클라우드 컴퓨팅 환경에서 구현하여 실험한 내용을 다루고, 5 장에서 결론을 맺는다.

2. 관련 연구

클라우드 컴퓨팅이 새로운 패러다임으로 화두가 되면서, 데이터센터에서 시스템 신뢰도와 운영비용을 절감시키는 저전력 자원 프로비저닝 기법이 중요해지고 있다. DVFS(Dynamic Voltage and Frequency Scaling) 기법은 자원의 동작 전압과 동작 주파수를 조절하여 CMOS 회로의 동적 전력을 감소시키는 대표적인 에너지를 고려한 기법으로 관련 연구가 활발히 진행되고 있다.

이와 관련하여 [1]연구에서는 실시간 클라우드 서비스에 대한 전력량을 고려한 가상 자원 프로비저닝 기법을 소개한다. 이 연구에서는 클라우드 환경의 실시간 서비스를 데드라인에 관점으로 Hard Real-Time (HRT) 서비스와 Soft Real-Time (SRT) 서비스로 구분하고 각각에 적합한 DVFS 기법을 소개하고 있다. HRT 서비스에 적용 가능한 DVFS 기법은 다시 Lowest-DVFS, δ-Advanced-DVFS, Adaptive-DVFS 기법의 3 가지로 구분된다. Lowest-DVFS 기법은 가상 머신 요청에 대해 가장 낮은 프로세서 속도를 가진 가상 머신을 할당하는 기법이다. HRT 서비스에 Lowest-DVFS 기법을 적용할 경우 작업의 수락률이 낮다. 따라서 낮은 수락률을 극복하고자 δ-Advanced-DVFS 기법은 현재의 요구되는 MIPS 율을 δ만큼 오버 스케일링 하여 가상 머신을 할당한다. 마지막으로 HRT 서비스에 대한 도착률과 작업 시간이 사전에 알려진 경우 계산식을 사용해 최적의 값으로 스케일링 하는 Adaptive-DVFS 기법이 있다. 이 세가지 기법의 경우 HRT 서비스를 제공하기 위해 데드라인을 위반하는 서비스에 대해서는 거절함으로써 정책을 유지한다. 반면 SRT 서비스는 데드라인을 넘어서는 서비스에 대해서도 지연에 따라 패널티를 부과하는 정책을 소개한다. 이 기법은 태스크의 요구성능에 따라 가상 머신을 제공하여 사용자에게 불필요한 비용 부담을 안길 수 있다.

기존의 DVFS 연구들이 하나의 주파수만을 사용하는 것을 보완하여 [3] 연구에서는 최대 주파

하이브리드 클라우드상의 과학응용을 위한 가상 자원 오토 스케일링 기법

(An Auto-Scaling Technic of Virtual Resources on Hybrid Clouds for Scientific Applications)

안 윤 선 † 정 솔 † 김 윤 희 ††
(Younsun Ahn) (Sol Jeong) (Yoonhee Kim)

요 약 계산 과학 분야에서 자원을 필요한 만큼 빌려 쓸 수 있는 클라우드 기술을 적용하여 과학 클라우드(Science Cloud)를 구축하는 연구가 활발히 진행되고 있다. 장시간 동안 고성능의 자원을 활용해야 하며 안정적이고 지속적인 자원의 공급을 필요로 하는 대규모 작업 계산 응용이 있다. 이러한 응용의 성공적인 작업 수행과 클라우드 자원을 효율적으로 통합 활용하기 위해 온-디맨드 자원 가상화 특성을 활용한 오토 스케일링 기법이 사용될 수 있다. 오토 스케일링 기법은 효율적이고 통합적으로 클라우드 자원을 제공한다. 그러나 대부분의 오토 스케일링 기법은 단순한 하드웨어의 성능을 기반으로 제공되고 있어 응용의 특성이나 작업의 데드라인 고려가 필요하다. 사용자가 요청한 작업의 데드라인 내에서 작업 수행이 가능하도록 오토 스케일링을 수행하는 알고리즘을 제안한다. 워크플로우와 Bag of Tasks의 응용 패턴에서 효율적인 자원의 활용을 위해 동적 자원 스케일링 기법을 제시하고, 워크플로우 형태 작업의 단백질 주석 처리 응용과 Bag of Tasks 형태의 변광 지수 계산 응용을 하이브리드 클라우드 환경에서 오토 스케일링 기법에 적용하여 결과를 보인다. 두 가지 응용과 응용의 패턴에 따라 효율적으로 자원이 오토 스케일링 되는 결과를 보임으로써 제안하는 오토 스케일링 기법의 우수성을 검증한다.

키워드: 오토-스케일링, 하이브리드 클라우드 컴퓨팅, 워크플로우, 다중 정책

Abstract The appearance of Science Clouds allows scientists to facilitate large-scale scientific computational experiments over cloud environment. Cloud computing enables applications to employ on-demand and scalable resources dynamically. It is necessary for many task computing (MTC) to provide high performance resources in a long phase and certificate stable executions of applications even dramatic changes of vital status of physical resources. Auto-scaling on virtual machines offers efficient and integrated utilization of cloud resources. VM Auto-scaling schemes have been actively studied as effective resource management in order to utilize large-scale data center in a good shape. However, most of the auto-scaling methods just simply support performance metrics such as CPU utilization and data transfer latency but are rarely aware of execution deadline or characteristics of an application. We propose an auto-scaling method, guaranteeing the execution of an application within deadline. It can handle two types of job patterns; Bag-of-Tasks jobs or workflow jobs. As a proof-of-concept, two applications such as variable index computation and protein annotation workflow

· 이 논문은 2013년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. NRF-2013R1A1A3007866)

† 학생회원 : 숙명여자대학교 컴퓨터과학부
ahnysun@sookmyung.ac.kr
jeongsol@sookmyung.ac.kr
†† 종신회원 : 숙명여자대학교 컴퓨터과학부 교수
yulan@sookmyung.ac.kr
(Corresponding author임)

논문접수 : 2013년 11월 18일
(Received 18 November 2013)
논문수정 : 2014년 4월 8일
(Revised 8 April 2014)
심사완료 : 2014년 5월 12일
(Accepted 12 May 2014)

정보과학회논문지: 시스템 및 이론 제41권 제4호(2014.8)

applications were simulated in hybrid cloud environment. The experimental results of the simulation show the method arrange resources reasonably economically by deadline and adaptively on the change of resource status.

Keywords: auto-scaling, hybrid cloud computing, SLA, multi-policies, Bag-of-Tasks

1. 서 론

최근 계산 과학 분야에서 자원을 필요한 만큼 빌려 쓸 수 있는 클라우드 기술을 적용하여 과학 클라우드(Science Cloud)를 구축하는 연구가 세계 여러 곳에서 진행되고 있다. 대규모 작업 계산 응용(MTC: Many Task Computing)은 장시간 동안 고성능의 자원을 활용해야 하며 안정적이고 지속적인 자원의 공급이 필수적이다. 이러한 환경에서 성공적인 작업 수행을 위한 실행관리, 자원 관리 등을 제공하는 계산 문제 풀이 환경(Computational Problem Solving Environment)에 대한 연구가 더욱 필요하다. 이러한 문제 풀이 환경에 필요한 클라우드 자원을 효율적으로 통합 활용하기 위해 온-디맨드 자원 가상화 특성을 활용한 오토 스케일링 기법이 사용될 수 있다. 오토 스케일링 기법은 응용의 수행 도중 효율적인 자원 관리를 위해 가상 자원의 수를 동적으로 변화시킨다.

선행 연구인 [1]에서는 하이브리드 클라우드 컴퓨팅 환경에서 효율적인 자원관리를 위한 오토 스케일링 기법을 제안한다. 하지만 Bag of Tasks (BoT)[2] 형태의 작업뿐만 아니라 워크플로우와 같은 다양한 응용패턴을 지원할 수 있도록 알고리즘을 확대할 필요가 있으며, 다양한 응용의 특성을 고려하여 수행이 가능하도록 하고자 한다. 또한 BoT 형태의 다른 응용에 대한 실험을 추가하여 응용의 특성을 고려한 오토 스케일링 기법의 우수성을 보이고자 한다.

본 논문은 다양한 형태의 응용 패턴과 과학 계산 응용의 요구사항을 반영하여 클라우드 컴퓨팅 인프라를 기반으로 자원들을 효율적으로 활용하기 위해 [1]에서 확장한 동적 가상 자원 오토 스케일링 자동화 방법을 제시한다. 특히 BoT 응용뿐 아니라 워크플로우 응용에 대한 체계적인 자원 활용이 가능한 오토 스케일링 기법을 제시한다.

본 논문에서는 BoT 응용을 위한 알고리즘에 대해 설명하고 워크플로우 응용의 특성에 맞는 오토 스케일링 알고리즘을 제안한다. 제안한 오토 스케일링 알고리즘을 워크플로우 형태 작업의 단백질 주석 처리 응용[3]과 BoT 형태의 변광 지수 계산 응용에 적용하여 성능을 검증한다.

본 논문의 구성은 다음과 같다. 1장의 서론에 이어 2장에서는 관련 연구들을 살펴보고, 3장에서는 오토 스케일링 알고리즘에 대해 설명하며 4장에서 적용 시나리오 및 실험에 대해 설명한다. 마지막으로 5장에서 결론을 맺는다.

2. 관련 연구

클라우드의 가상화 기술을 이용한 자원의 효율적인 관리를 위해 오토 스케일링 기법에 대한 연구가 활발히 진행되고 있다. 자원 제공자 입장에서 사용자가 정의한 규칙을 통해 자원의 규모를 확장하고 축소하는 AWS [4]의 "Auto-scaling"과 Windows Azure[5]에서 사용되는 Paraleap[6], Scalr[7]이 자원 할당 자동화 기법을 사용한다. 규칙 기반의 스케일링 기법은 비교적 단순한 방법으로 동적 자원 할당이 가능하지만 복잡한 수행 패턴을 갖는 작업에 대해서는 실제 필요한 만큼 가상머신의 개수가 동적으로 변화하지 않을 수 있기 때문에 데드라인이나 자원 사용 비용의 문제를 야기 할 수 있다.

이러한 이유로 사용자의 입장을 고려하여 데드라인과 사용 비용에 대한 고려를 한 연구로는 [8-12]이 존재한다. 먼저 [8]은 BoT 응용을 통해 자원 사용 비용의 최소화를 위한 오토 스케일링 기법을 제안했다. 자원의 개수를 변화시키는 수평적(Horizontal) 스케일링과 자원의 하드웨어 사양을 조절하는 수직적(Vertical) 스케일링 기법을 혼합 활용했다. 그러나 이 연구에서도 복잡한 수행 패턴의 작업에 대해 실행 중 자원사용량에 대한 고려가 부족하여 동적인 자원 할당이 가능하다고 하기엔 어려움이 있다.

[9]에서는 워크플로우 응용을 가지고 데드라인과 자원 사용 비용을 고려한 알고리즘을 제안했다. [10]에서는 자원 프로비저닝 프로세스 자동화 및 자원 사용 비용 최소화를 고려했다. 다양한 워크로드 형태와 워크플로우 형태를 고려하였으나, 공용 클라우드 단일 환경에서의 자원 사용만을 고려하였다는 한계가 있다. [11], [12]의 경우 그리드 환경에서 워크플로우 응용을 가지고 오토 스케일링 기법을 수행하여 하이브리드 클라우드 인프라 자원 활용에 참고한다. [11]은 사용자가 원하는 데드라인 내에서 실행 비용을 최소화하기 위한 알고리즘을 제안했다. [12]는 상호의존성이 있는 작업에 대해 효율적인 스케일링 알고리즘을 제안했다. [11]과 [12]의 워크플로우 스케줄링을 참고하여 워크플로우 형태의 작업이 수행 될 수 있는 알고리즘을 제안한다.

본 논문에서는 [1]에서 제안한 오토 스케일링 기법을 확장하여 BoT 형태의 작업뿐 아니라 워크플로우와 같

VM Auto-Scaling for Workflows in Hybrid Cloud Computing

Younsun Ahn
Dept. of Computer Science
Sookmyung Women's University
Seoul, Korea
ahnysun@sookmyung.ac.kr

Yoonhee Kim
Dept. of Computer Science
Sookmyung Women's University
Seoul, Korea
yulan@sookmyung.ac.kr

Abstract— **Appearance of Science Clouds enables scientists to facilitate large-scale scientific computational experiments over cloud environment. Many task computing (MTC) in computational science needs to certificate stable executions of applications even in rapid changes of vital status of physical resources and supports high performance resources in a long period. Auto-scaling approach on virtual machines (VM) increases efficient cloud resources management for the computational problem solving environment. Diverse auto-scaling methods which provide useful resource management presently are being debated and studied. However, most of the auto-scaling methods are just easily considered in performance metrics or execution deadline in specific workloads but not in various patterns of workflow. We propose an auto-scaling method, guaranteeing the execution of various patterns of workflow within deadline time in hybrid cloud environment. The experimental results show the method works dynamically and acceptably on hybrid cloud resources for various workflow patterns having random workload dependency.**

Keywords— auto-scaling; hybrid cloud computing; workflow; workflow dependency;

I. INTRODUCTION

Cloud computing provides on-demand and scalable resources dynamically in order to support application execution. Appearance of Science Clouds enables scientists to facilitate large-scale scientific computational experiments over cloud environment. Many task computing (MTC) needs to certificate stable executions of applications even in rapid changes of vital status of physical resources and support high performance resources in a long time. Therefore, studying computational problem solving environment has been getting more important as it supports the management of task executions or resources in large-scale computation. Auto-scaling approach on virtual machines (VM) increases efficient cloud resources management for the computational problem solving environment. Our previous paper [1] proposed an auto-scaling method to provide efficient resource utilization in a hybrid cloud computing environment. Tasks in Bag-of-Tasks (BoT) [2] can run in parallel while tasks in workflow can be executed in the order of dependency. However, the proposed auto-scaling algorithm limited to specific Bag-of-Tasks in aerodynamics and workflows in protein annotation workflow. We need an auto-scaling method in order to perform applications in a general form of workflows.

This paper proposes an extended version of the auto-scaling method, reflecting in various workflow patterns of tasks based on cloud computing environment. Especially, it dynamically allocates virtual resources depending on tasks in workflow on hybrid cloud environment. We propose an auto-scaling method that can meet a deadline in various workflow patterns. We focus on dynamically allocating VMs in order to maximize resource utilization within a deadline and dealing with task dependency in workflow application. We have evaluated the auto-scaling method with various workflow patterns which have a large number of tasks in hybrid cloud resources. The results of a simulation show the method performs automatically resource allocation satisfying deadline constraints.

The rest of the paper is organized with the sections as follows; we introduce an overview of related works in Section 2. Section 3 explains an auto-scaling algorithm and Section 4 contains contents about a workflow generation. Section 5 explains experiment results. Finally, we conclude the paper and discuss future work in final section.

II. RELATED WORK

Auto-scaling approaches which provide useful resource management presently are being debated and studied. Auto-scaling issues are divided into two sides. First one is rule-based auto-scaling methods such as "Auto-scaling" of AWS [3], Paraleap [4] for Windows Azure [5], and Scalr [6]. This method changes the number of resources by user-defined metrics. However, rule-based auto-scaling methods could lead to execution failure of an individual application in short of consideration on its characteristics such as execution deadline.

In the other side, [7], [8], [9], [10], and [11] are the studies of auto-scaling in consideration of constraints such as a deadline of applications or cost for resource usage. [7] proposes an auto-scaling method minimizing resource usage cost. Horizontal scaling and vertical scaling are used. Horizontal scaling adds or removes the number of VMs and vertical scaling controls the size of a VM. However, this paper only provides resource allocation for Bag-of-Tasks [2] jobs and also dissatisfied resource usage during execution of an application.

This research was supported by Basic Science Research Program through the Natio nal Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT and Future Planning (NRF-2013R1A1A3007866)

978-1-4799-5841-2/14 315.00 © 2014 IEEE
DOI 10.1109/ICCAC.2014.34

However, [9] does not consider various types of workload patterns. [9] lacks considering a mixture of workload patterns. It just performs three special types of workload patterns. It is necessary for our proposed auto-scaling method to consider different workflow patterns to perform automatically. Therefore, we generate various workflow patterns and apply it to our proposed auto-scaling method.

[10] and [11] propose their auto-scaling methods for the execution of workflow applications on Grids. [10] minimizes cost by using sub-deadline and also can reduce execution time for the entire workflow. [11] schedules dependent jobs efficiently. Reference [10]'s and [11]'s workflows are just too simple to evaluate their auto-scaling method. Various patterns of workflows exist in applications. Workflow patterns affect auto-scaling method by the number of tasks and dependency. It may not be possible to schedule an amount of workflow tasks actually needed. We propose an algorithm for workflow referring to [10] and [11].

In this paper, we propose an extended version of an auto-scaling method based on our previous research [1] which can support efficient resource utilization considering the types of jobs in Bag-of-Tasks [2] as well as workflows. Proposed auto-scaling method reflects complex structures of workflow by increasing the number of tasks and dependency. Auto-scaling method can automatically allocate cloud resources by task dependency in a various workflow patterns within deadline.

III. AUTO-SCALING ALGORITHM

We extend [1]'s auto-scaling algorithm which consider only tasks in Bag-of-Task, to support workflow as well. Initial scheduling schedules tasks to prevent waste of VMs within a deadline. Auto-scaling method can perceive delay and deadline violation to comparing actual start time and estimated start time of running tasks during monitoring interval. Algorithm's assumption and notation are referred to [1].

Algorithm 1 – Run-time Scaling
Input – An application,
SLA={a policy *P*, a deadline *D* [, minimum performance requirement *minPM*]}
Output – Scaling decision $S = \{ toStartUp, toShutDown \}$
Scheduling decision $S = \{ tasks \rightarrow VMs \}$

```
1:  If SCALING is TRUE
2:      If Job pattern is bag-of-tasks job
3:          Sort waiting tasks in decreasing order of execution length;
4:          S ← Cost-aware Scheduling(sortedTasks, D, minPM);
5:      If Job pattern is workflow.
6:          Sort waiting tasks in sequential order;
7:          S ← WorkflowScheduling(sortedTasks, D);
8:      each vm where status is running
9:          if no running/waiting tasks on vm then
10:             destroy the vm
11:         send scaling decisions to DRMS
12:         send scheduling decisions to JES
13:     waitForNextInterval();
14:     SCALING ← SLAMonitoring(runningTasks, D);
15:
```

Algorithm 1, Run-time Scaling is extended based on [1]. We newly propose algorithm 2, Workflow Scheduling algorithm to perform tasks in workflow. In the reference [1], Run-time Scaling algorithm can choose an appropriate policy by a pattern of tasks.

Additionally, we extend policies in order to perform workflow as well as Bag-of-Tasks [2] patterns of tasks. Tasks are scheduled with applying one of the two policies such as Cost-aware Scheduling (line 2) and Workflow Scheduling (line 5) of SLA (Service Level Agreements). Cost-aware Scheduling is suitable for a type of tasks in Bag-of-Tasks [2], but Workflow Scheduling is proper to workflow patterns. Cost-aware Scheduling chooses VMs which considered billing time unit to save the cost and user specific minimal performance. Tasks in Bag-of-Tasks [2] are sorted as descending order based on their execution time, while tasks in workflow are performed sequential order.

Algorithm 2 describes our Workflow Scheduling algorithm. The algorithm discovers appropriate a critical path for processing the workflow tasks and schedules the tasks. Proposed Workflow Scheduling algorithm is based on a PCH algorithm [11]. When our auto-scaling method tries to schedule VMs, our method has to adopt a private cloud resource first. When our algorithm tries to allocate public cloud resources for tasks, it considers VMs in the order of running ones, one having the fastest start time, and ones within application deadline. First of all, we find a critical path by using PCH algorithm [11]. Tasks on a critical path are scheduled in a private cloud resource in order to reduce a cost of resource usage (line 3). The total execution time of the critical path is decided by a deadline and additional margin value. It is important to consider task dependency in workflow. Each task could get an EFT (Estimated Finish Time) of a parent task and set an EST (Earliest Start Time) value to EFT of a parent task in order to reflect the order of tasks (line 5). When tasks that are not on a critical path are allocated to VMs, tasks are checked whether their parent tasks have been performed or not. If parent tasks are not allocated in cloud resources, child tasks must wait to be scheduled. The algorithm could calculate an EST of a child task considering EFT of its parent tasks (line 10 (private), line 12 (public)). The algorithm can perform tasks using the appropriate number of VMs. The algorithm can perform tasks using the appropriate number of VMs. Workflow Scheduling algorithm can execute tasks considering task dependency and meeting deadline in a various workflow patterns.

Algorithm 2– Workflow Scheduling
Input –Waiting tasks of the application,
Output –Scheduling decision $S = \{ tasks \rightarrow VMs \}$, VM list *toStartUp*.for the VMs to be newly created

```
1: All tasks are scheduled in sequential order
2: Task in Critical path is allocated VM
3: If There is no running VM, Find a vm on which all task in Critical path can run within the D;
4:     EST_VM is related previous task's EFT_VM
5:     Calculate EFT_VM by adding EST_VM and ET(Execution Time)_VM;
6: Task which has allocated related previous tasks on vm
7: If There is no running private VM, find a private vm
```

Auto-Scaling Method in Hybrid Cloud for Scientific Applications

Younsun Ahn Jieun Choi Sol Jeong Yoonhee Kim
Dept. of Computer Science
Sookmyung Women's University
Seoul, Korea
ahnysun@sookmyung.ac.kr, jechoi1205@gmail.com, {jeongsol, yulan}@sookmyung.ac.kr

Abstract— **Scientists can ease to conduct large-scale scientific computational experiments over cloud environment according to an appearance of Science Clouds. Cloud computing enables applications to apply on-demand and scalable resources dynamically. It is necessary for Many Task Computing (MTC) to offer high performance resources in a long phase and certificate stable executions of applications even dramatic changes of vital status of physical resources. Auto-scaling on virtual machines provides integrated and efficient utilization of cloud resources. VM Auto-scaling schemes have been actively studied as effective resource management in order to utilize large-scale data center in a good shape. However, most of the existing auto-scaling methods just simply support CPU utilization and data transfer latency. It is needed to consider execution deadline or characteristics of an application. We propose an auto-scaling method, guaranteeing the execution of an application within deadline. It can handle two types of job patterns; Bag-of-Tasks jobs or workflow jobs. We simulate a variable index computation application in hybrid cloud environment. The results of the simulation show the method can dynamically allocate resources considering deadline.**

Keywords—auto-scaling; hybrid cloud computing; science cloud; Bag-of-Tasks; workflow

I. INTRODUCTION

Cloud computing enables applications to apply on-demand and scalable resources dynamically. Scientists can ease to conduct large-scale scientific computational experiments over cloud environment according to an appearance of Science Clouds. It is necessary for Many Task Computing (MTC) to offer high performance resources in a long phase and certificate stable executions of applications. Cloud computing offers endless resources by a virtualization technology. Auto-scaling can automatically change the number of Virtual Machines (VM) during execution of an application. Auto-scaling can provide efficient integrative utilization of cloud resources.

Our previous paper [1] proposes an auto-scaling method that dynamically allocates resources satisfying a deadline in a hybrid cloud computing environment. However, proposed auto-scaling algorithm needs to be extended in order to support various patterns of job execution, which are Bag-of-Tasks [2] and workflows.

This paper proposes an extended version of the auto-scaling method, reflected in patterns of jobs and the requirements of an application in hybrid clouds. Chiefly, it dynamically allocates resources depending on the two patterns of jobs (Bag-of-Tasks [2] and workflow) in hybrid clouds. We developed an auto-scaling algorithm and applied it to specialized computation and data analysis application such as variable index computation in hybrid cloud environment as a proof-of-concept.

The rest of the paper is organized with the sections as follows; we introduce an overview of related works in Section 2. Section 3 explains auto-scaling algorithm, and Section 4 contains contents about a scenario of using auto-scaling algorithm and experiment results are discussed. Finally, we conclude the paper and discuss future work in final section.

II. RELATED WORK

Cloud computing provides on-demand and scalable resources by virtualization technology. Auto-scaling topics are currently being discussed and studied as useful resource management. Auto-scaling issues exist in two aspects of consideration. One side use rule-based auto-scaling method. "Auto-scaling" of AWS [3], Paraleap [4] for Windows Azure [5], and Scalr [6] is the example of the rule-based auto-scaling method. Rule-based auto-scaling method elastically increases or decreases resources to meet user-defined metrics, but it is not enough to allocate resources actually needed and it have issues with violation of deadline and cost. On the other hand, [7], [8], [9], [10], [11] are the studies of auto-scaling which consider deadline or cost for resource usage.

[7] proposes an auto-scaling method using Bag-of-Tasks [2] jobs in order to minimize resource usage. It is consider horizontal scaling and vertical scaling. Horizontal scaling adds or removes the number of VM instances and vertical scaling expands or reduces the size of a VM. However, [7] lack of considering resource requirements of dynamic workloads during execution of an application. [8] respects to finish jobs within deadlines at minimum financial cost using their auto scaling method for a workflow and fit deadline on highest performance resource. [9] proposes an algorithm to minimize execution costs while meeting deadline. However, they perform their auto-scaling method in simple environment.

[10], [11] perform their auto-scaling methods using workflow application on Grids. [10] proposes an algorithm that

This research was supported by the MSIP(Ministry of Science, ICT&Future Planning), Korea, under the C-ITRC(Convergence Information Technology Research Center) support program (NIPA-2014-H0401-14-1022) supervised by the NIPA(National IT Industry Promotion Agency)

reduces cost for resource usage within deadline. [11] proposes an economical scheduling algorithm for dependent tasks. We propose a workflow scheduling algorithm referring to Reference [10]'s and [11]'s scheduling algorithms.

In this paper, we propose an extended version of an auto-scaling method based on our previous paper [1] that performs dynamically allocating VM considering the types of jobs from Bag-of-Tasks [2] as well as the workflow and the characteristics of an application.

III. AUTO-SCALING ALGORITHM

A. Assumptions

We use two kind of types of jobs, Bag-of-Tasks [2] and workflow. Jobs in Bag-of-Tasks [2] type can be scheduled on resources separately from each other while jobs in workflow can be performed in order of dependency pattern. In the workflow scheduling algorithm, we use the term 'task' for a job. Auto-scaling method can find delay and deadline violation to comparing actual start time and estimated start time of running jobs. Moreover, monitoring perform every regular term.

B. Algorithms

This paper extends the version of [1]'s algorithm in order to enable the auto-scaling method to schedule jobs in Bag-of-Tasks [2] and workflow. Assumptions and notations for the algorithms are explained in detail at the reference [1]. Algorithm 1, Run-time Scaling is extended based on [1]. We describe algorithm 3, Workflow Scheduling algorithm to deal with tasks in workflow.

In the reference [1], Run-time Scaling Algorithm has two polices and depicts our overall auto-scaling method. Additionally, we extend three policies by adding workflow scheduling. Jobs are scheduled with applying one of the three policies such as Performance-oriented Scheduling (line 5), cost-oriented Scheduling (line 8) and Workflow Scheduling (line 11) of SLA (Service Level Agreements). Jobs in Bag-of-Tasks [2] are sorted as descending order based on their execution time, while tasks in workflow are sorted as sequential order.

Algorithm 1 – **Run-time Scaling**
Input – An application,
SLA={a policy *P*, a deadline *D* [, minimum performance requirement *minPM*]}
Output – Scaling decision *S* = { *toStartUp*, *toShutDown* }
Scheduling decision *S* = { *jobs* → *VMs*}

```
1:  SCALING ← TRUE;
2:  while (true)
3:      if SCALING is TRUE
4:          switch P
5:          case Performance:
6:              Sort waiting jobs in decreasing order of execution length;
7:              S ← PerformanceOrientedScheduling(
                        sortedJobs, D, minPM);
8:          case Cost:
9:              Sort waiting jobs in decreasing order of execution length;
10:             S ← CostOrientedScheduling(sortedJobs, D);
11:         case Workflow:
              Sort waiting jobs in sequential order;
12:             S ← WorkflowScheduling(sortedJobs, D);
13:     each vm where status is running
14:     if no running/waiting jobs on vm then
15:       destroy the vm
16:     send scaling decisions to DRMS
17:     send scheduling decisions to JES
18:     waitForNextInterval();
19:     SCALING ← SLAMonitoring(runningJobs, D);
```

Algorithm2, Performance-oriented Scheduling is suitable for jobs in Bag-of-Tasks [2]. The algorithm basically accompanies cost-effective scheduling and plans jobs to meet deadline using minimum performance requirement. For more information of Performance-oriented Scheduling algorithm, it can be found in the reference [1].

Algorithm 2 – **Performance-oriented Scheduling**
Input –Sorted jobs of the application, deadline *D*, minimum performance requirement *minPM*
Output –Scheduling decision *S* = { *jobs* → *VMs*}, list *toStartUp*.for the VMs to be newly created

```
1: All jobs are scheduled
2: If There is no running private VM, find a private vm
   (PM_vm >= minPM) on which job can start the most quickly within the D;
3: If There is running private VM, schedule job to vm;
4: Continue with the next job;
5: If There is no running public VM, find a public vm
   (PM_vm >= minPM) on which job can start the most quickly within the D;
6: If There is running public VM, Choose Cheaper VM either
   running VM or VM of ITj type
7:     Schedule job to chosen vm;
8: Continue with the next job;
```

Algorithm 3 defines our workflow scheduling mechanism. Proposed Workflow Scheduling algorithm is based on a PCH algorithm [11] and followed the performance-oriented policy. We could get tasks on a critical path and group of tasks using PCH algorithm [11] and then each group can be scheduled. The total execution time of critical path is calculated. Tasks on a critical path are scheduled on a same resource, which can execute all tasks in the critical path in order to reduce communication overhead. Every time we try to schedule tasks, it is a rule to choose private cloud resource prior to public cloud. Each task considers estimated finish time of a parent task and calculates earliest start time. When tasks which are not on a critical path are scheduled to VMs, they check a parent task's estimated finish time.

IV. SCENARIO AND EXPERIMENTS

We simulated variable index computation application in Bag-of-Tasks [2] with CloudSim [12].

We simulated variable index computation application in

Cluster Comput
DOI 10.1007/s10586-014-0371-2

Towards effective science cloud provisioning for a large-scale high-throughput computing

Seoyoung Kim · Jik-Soo Kim · Soonwook Hwang · Yoonhee Kim

Received: 29 September 2013 / Revised: 25 February 2014 / Accepted: 15 March 2014

Abstract The science cloud paradigm has been actively developed and investigated, but still requires a suitable model for science cloud system in order to support increasing scientific computation needs with high performance. This paper presents an effective provisioning model of science cloud, particularly for large-scale high throughput computing applications. In this model, we utilize job traces where a statistical method is applied to pick the most influential features to improve application performance. With these features, a 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 by comparing the proposed model with other policies through experiments and expect noticeable improvements on performance as well as reduction of cost from resource consumption through our model.

Keywords Science cloud · High-throughput computing · Job profiling · Cloud provisioning · PCA (Principal components analysis)

S. Kim · J.-S. Kim · S. Hwang
National Institute of Supercomputing and Networking, KISTI, Daejeon 305-806, Korea
e-mail: sssyyy77@kisti.re.kr

J.-S. Kim
e-mail: jiksoo.kim@kisti.re.kr

S. Hwang
e-mail: hwang@kisti.re.kr

Y. Kim (✉)
Department of Computer Science, Sookmyung Women's University, Seoul 140-742, Korea
e-mail: yulan@sookmyung.ac.kr

1 Introduction

Cloud computing nowadays enables numerous scientists to earn advantages by serving on-demand and elastic resources whenever they desire 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, allocating and provisioning virtual machines effectively are still considered as a challenging issue for 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. For these reasons, an appropriate and suitable model for science cloud which can support increasing scientists and computations is required. Fortunately, job profiles which potentially involve 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 Profiling Historical factor for Allocation and Provisioning on Science cloud (**PHAP**) model which is an allocation and provisioning model of science cloud, especially for high throughput computing applications as well as many task computing. In this model, we utilize job traces where a statistical method is applied to pick the most influential features for improving application performance. With the features, 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 achieve-

Published online: 19 April 2014

ments by comparing the proposed model with other policies through experiments. Finally, we expect improvement on performance as well as reduction of cost from resource consumption through our model.

The rest of this paper is structured as follows. Sect. 2 discusses related work and Sect. 3 presents introduction of the proposed system model where application model will be introduced. We discuss allocation and provisioning model in details in Sect. 4, while Sect. 5 discusses experiment and its results. Finally, we conclude this paper and discuss future work in Sect. 6.

2 Related work

With increasing attentions to clouds in several science communities, a lot of efforts have been made to provide optimized and facilitated virtual environments on science clouds. Moreover, almost science applications typically involve long-term computer executions, huge dataset processing and large-scale computations requiring the availability of abundant computing infrastructures. Therefore, it is important to identify requirements of users and their 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.

Wang et al. [1] 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 in order to lease VM. Another work, [2] proposed a controlling system which allocates appropriate resources through monitoring and analysing current workloads of applications. In [2], a virtual server is operated on a group of physical machines and each server is responsible for particular application on the heterogeneous machines. In addition, the system predicts future workload through analysing resource utility and real-time requirements of applications and schedules virtual machines 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, [3] had proposed predicting application runtime and queue wait time. In [3], genetic algorithm had exploited to search similar jobs among job histories. This study predicts two metrics by mining historical workloads corresponding to job execution time and wait time in queue. It firstly searches the most similar job among the histories based on genetic algorithm considering characteristics of both 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 [3] do not lie on resource managements, but only runtime predictions.

Meanwhile, Bhuvan et al. [4] 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 both 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 allows 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 focuses on two kinds of application types; High throughput computing (HTC) and many task computing (MTC) [5] are generally found in most of workloads from various scientific applications. These kinds of workloads which are also referred as 'Bag-of Tasks (BoT)' have several characteristics in common as the followings:

- Massively Paralleled
- Loosely coupled or uncouple in each task
- Adoption of '*throughput*' as a main metric (over a fixed period of time)

Assume that a job(j_i) is submitted by some user and has $\boldsymbol{n}$ number of tasks($t_{x+1}, t_{x+2}, \ldots, t_{x+n}$, here x is arbitrary job id) using 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).

$$throughput(j_i) = \frac{n}{makespan(j_i)} \tag{1}$$

To increase throughput of j_i, we have to reduce its makespan where every n number of tasks should be completed since $\boldsymbol{n}$ is fixed by user when he submits the job. However, it is a

하둡을 이용한 기상요인에 따른 농산물 경락가격 실시간 분석 시스템 설계[1]

(A Design of Real-time Analysis System on the Effects of Weather on the Auction Price of Agricultural Products Using Hadoop)

최지은, 전영란, 김윤희

숙명여자대학교 컴퓨터과학과

jechoi1205@sookmyung.ac.kr, youngr92@hanmail.net, yulan@sookmyung.ac.kr

요 약

기하급수적으로 늘어나는 데이터 속에서 원하는 정보를 찾고 분석하는 기술의 중요성이 높아지고 있다. 기존의 시스템으로는 방대한 데이터를 처리하기에 한계가 분명하여 효율적인 데이터 처리와 관리를 위한 분산 저장 및 병렬 처리 시스템의 연구 및 활용이 중요해지고 있는 상황이다. 한편 농산물은 날씨에 따라 품질과 수확량이 바뀌어 가격이 결정되는 날씨에 민감한 품목이다. 이와 관련하여 농산물이 재배되는 과정에서 날씨가 미치는 영향에 대한 연구는 많지만, 가격이 책정되는 당일의 날씨에 대한 분석은 미흡한 수준이다. 이에 본 연구는 실시간으로 정해지는 농산물 경락가격에 당일 날씨가 미치는 영향을 파악하고자 한다. 본 연구에서는 기상 데이터와 도매시장에서 실시간으로 정해지는 농산물 경락가격의 데이터를 수집하며 비용대비 효율적으로 분석할 수 있는 하둡 프레임워크를 이용하여 분석 시스템을 구축하여, 날씨와 농산물 경락가격의 관계를 빠르게 분석할 수 있음을 보여주었다.

Keywords: 하둡, 농산물, 경락가격, 날씨, 도매시장, 분석 시스템

1. 서론

소셜 미디어의 성장과 스마트폰과 같은 디지털 기기가 삶의 전반에 배치되면서 대규모의 다양한 데이터가 급속히 생성되고 있으며, 기하급수적으로 늘어나는 데이터 속에서 원하는 정보를 찾아내고 분석하는 기술의 중요성이 높아지고 있다. 기존에는 고성능의 값비싼 슈퍼컴퓨터를 이용해 큰 데이터를 처리하기도 했지만 한계가 분명했다. 따라서 데이터를 비용대비 효율적으로 처리할 수 있는 방법론과 함께 데이터 처리와 관리를 위한 분산 저장 및 병렬 처리 시스템의 연구 및 활용이 중요해지고 있는 상황이다[1]. 이와 함께 최근의 대용량 데이터 처리는 대량의 컴퓨팅 자원을 이용한 병렬 처리를 기본으로 하고 있는데, 이는 리눅스나 오픈소스 소프트웨어와 같이 보다 저렴해진 OS와 소프트웨어, 보다 강력한 성능의 서버, 보다

이 논문은 2013 년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF-2013R1A1A3007866)

빨라진 네트워크 기술의 이점에 기인한다. 빅데이터에 대한 병렬 처리는 기존의 단일 머신 상에서 기존의 DBMS로는 어려웠던 대용량 데이터 처리를 보다 저비용으로 실현할 수 있는 방안을 마련한다[1]. 하둡(hadoop)[2]은 간단한 프로그래밍 모델을 사용하는 컴퓨터의 클러스터에 걸쳐 대량의 데이터 세트의 분산 처리를 허용하는 프레임 워크이다. 하둡 프레임워크는 로컬 컴퓨팅 및 스토리지를 제공하고 각각에 하나의 서버에서 최대 확장 할 수 있도록 설계되어있다. 또한 페타바이트 이상의 대규모 데이터를 클러스터 환경에서 저장하기 위한 하둡 분산 파일 시스템(HDFS: Hadoop Distributed File System)과 이를 기반으로 병렬 처리를 지원하기 위한 맵리듀스(MapReduce) 프레임워크로 구성된다. 하둡 프레임워크는 많은 분야에서 대용량 데이터 처리를 위한 핵심 기술로 큰 인기를 누리고 있으며, 다양한 응용에 이용된다. 대용량 데이터를 빠르게 처리하고 다양한 응용을 지원하기 위한 기반 기술로써 빅데이터의 병렬 처리 성능 향상이 필수적이다[1].

한편 농산물은 날씨에 따라 품질과 수확량이 바뀌어 가격이 결정되는 날씨에 매우 민감한 품목이다. 이와 관련하여 농산물이 재배되는 과정에서 날씨가 미치는 영향에 대한 분석은 많이 이루어져있다. 농산물 가격은 유통 과정에서 결정되는 부분이 커서 유통 당일의 날씨에 따라 책정되는데 그 부분에 대한 분석이 미흡하다[3]. 따라서 본 연구에서는 실시간으로 정해지는 농산물 경락가격에 당일 날씨가 미치는 영향을 파악하고자 한다.

다음은 본 논문에서 주요 기여하는 기술 목록을 요약한 내용이다.

1) 본 논문에서는 병렬 처리 성능 향상을 위해 하둡의 프레임워크를 이용한 날씨 데이터와 농산물 경락가격의 상관관계 분석하는 시스템 구현한다.
2) 구현된 하둡 시스템으로 실시간 분석이 가능한 시스템의 전체 알고리즘을 제시한다.
3) 날씨 데이터와 농산물 경락가격의 상관관계 분석을 위한 분석 로직을 제시한다.

본 논문에서는 병렬 처리 성능 향상을 위해 하둡의 프레임워크를 이용하여 날씨 데이터와 농산물 경락가격의 상관관계를 분석하는 시스템을 구현해 보고자 한다. 따라서 본 연구에서는 농산물이 도매시장에서 유통되는 과정에서 정해지는 가격 데이터를 수집하고, 가격이 측정된 해당일의 날씨 데이터를 수집하여 둘의 관계를 분석할 수 있는 시스템을 구현 할 것이다. 본 연구의 목적은 웹 크롤링을 통해 기상 데이터와 농산물 경락가격 데이터를 수집하며, 수집한 빅데이터를 토대로 하둡 시스템을 이용해 기온과 농산물 경락가격의 상관관계를 분석하는 것이다. 본 논문에서는 온도와 농산물 경락가격의 데이터를 수집하고, 기상 날씨에 따른 가격의 평균 가격을 분석하는 시스템에 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

빅데이터가 화두가 되면서, 빅데이터의 응용 및 분석 기법이 중요해지고 있다. 하둡은 이러한 대용량 데이터를 분산 시스템에서 효율적으로 처리할 수 있도록 도와주는 기술로 여러 가지 응용이 활발하게 진행되고 있다. [4]연구에서는 하둡을 이용하여 수집된 데이터를 토대로 광고의 효과를 트위터 상에의 언급도를 측정하는 LiveAD 시스템을 구현하여 광고 효과를 분석할 수 있었다. 그 연구에서는 1대의 하둡 Master server 와 1대의 Slave server를 이용하여 광고 텍스트 데이터 약 4MB와 트위터의 텍스트 데이터 약 6GB를 분석 데이터로 이용하여 시스템을 구축하였다. 또한 하둡의 맵리듀스(MapReduce) 프레임워크를 이용하여 분석된 결과를 그래프를 통해 확인할 수 있었다[4]. [3]연구에서는 배추의 거래량과 거래가격에 영향을 미치는 기상요인들을 밝히고 그 영향 정도를 확인하기 위해 다중 회기 분석을 수행하였다. [7]연구는 기후 변화 속에서 이상기후 현상과 농업의 피해에 대해 알아보고, 기후변화가 농업의 생산과 유통 및 소비에 미친 영향을 분석하였다. [8] 연구는 농산물의 가격 및 유통량과 관련된 데이터가 원시적으로만 존재한다는 문제점과 정보화 시키는 노력이 부족함을 말하고 있다. 또한 막연히 선험적으로만 알고 있던 농산물 유통량과 그 가격의 변화를 통계적으로 검증하였다.

한편 기후변화가 농업과 농산물의 많은 영향을 미치고 있는 만큼 이에 관한 다양한 연구와 정책방안의 필요성이 증가하고 있다. 기존의 국내연구들은 대부분 환경학이나 생물학, 농업학,

A Problem Solving Environment for Distributed Dietary Analysis Using Hadoop

Younsun Ahn[1], Jieun Choi[1], Ik Tae Kim[2], Hyeyoung Jung[3], Yoonhee Kim[1]
[1]Dept. of Computer Science, Sookmyung Women's University, Seoul, Korea
[2]Dept. of Statistics, University of Illinois at Urbana-Champaign, Champaign, Illinois, 61820, USA
[3]Dept. of Food and Nutrition, Sookmyung Women's University, Seoul, Korea
Email: {ahnysun, jechoi1205}@sookmyung.ac.kr, ikim10@illinois.edu, {gracethanks, yulan} @ sookmyung.ac.kr

Abstract: The degree of wellness for a human being is decided by quite a huge amount of measured personal health data, living patterns, and life cycle including dietary or exercise patterns. Especially, as dietary patterns are strongly related to personal health, it is important for personal health to analyze dietary patterns. In the field of food and nutrition, dietary analysis needs to handle large input data and conduct repetitive processes, which is considered as tedious and time-consuming. It is beneficial for scientists to use a proper problem solving environment (PSE) to perform experiments efficiently. We propose a problem solving environment on parallel and distributed computing infrastructure using Hadoop in order to deal with large-scale amount of data and automate repetitive tasks of dietary analysis on food and nutrition data. Our proposed PSE speeds up total analysis dramatically as it automates a series of working tasks composing of sequential and parallel tasks properly in order.

Keywords: Distributed Computing, Wellness, Cloud, Food and Nutrition, Big data, Dietary Analysis;

1. Introduction

Wellness is one of the hottest issues in which people have been recently interested in order to improve and upgrade human being's health and quality of human life. A huge amount of personal health data to indicate the degree of wellness includes diverse information such as physical measurement combining external and internal body status. It does not include only direct measurement from human being but also dietary pattern from food and nutrition, which are recorded daily and continuously from individual's activities. Precisely, as a dietary pattern affects personal health closely, identifying the pattern is necessary for predicting potential health threats, even though no sign of illness exists in current records.

Therefore, the accumulated personal health data (life log, PHR (Physical Health Record), EMR (Electronic Medical Record), etc.) including food and nutrition can be a useful source to find relationship between health conditions and certain disease in order to keep healthy. However, it is difficult to efficiently process, manage, and store tremendously accumulated data for further analysis. It is important to deal with massive data in order to perform the analysis fast, effectively, and objectively.

A problem solving environment (PSE) for various science and engineering domains enables scientists and engineers to perform efficiently and conveniently their experiments, dealing with a large input data and repetitive processes. Especially, a PSE over parallel and distributed computing systems helps to conduct large data in parallel processing and handle repetition of same experiments applied with different data, mostly generated from parametric study problems.

In the field of food and nutrition, a dietary analysis bases on large data and requires tedious and repetitive tasks. Therefore, the statistical analysis in this field has to be based on efficient infrastructure in order to process, manage, and store such a huge amount of data. It is essential for the dietary analysis to be automated with a user-friendly problem solving environment and to be processed quickly by parallel and distributed infrastructure environment. A problem solving environment for dietary analysis on food and nutrition data enables scientists to facilitate analyzing massive data.

We suggest a problem solving environment on parallel and distributed computing infrastructure using Hadoop, which analyzes large dietary data to perform in rapid manner and automates numerous repetitive tasks in parametric study. This speedy analysis identifies dietary patterns and helps to catch potential health threat signals in early stage of health status.

The rest of the paper is organized with the sections as follows; we introduce an overview of related works in Section 2. Section 3 explains a problem solving environment framework for dietary analysis, and Section 4 contains a case study and preliminary results. Finally, we conclude the paper and discuss future work in final section.

2. Related Works

We briefly present some approaches which are representative problem solving environment.

There are many researches on data analysis based on parallel and distributed cloud and grid computing environments processing infrastructure in various domains. CAP (Collaborative Analytics Platform,

2013~2019) [1] project started last year aiming at a developing a standard service platform based on cloud infrastructure in Europe.

Some domain specific PSE for analyzing climate data [2] and numerical data in fluid dynamics [3] are introduced. In [2], a paper suggests a framework for users to analyze large-scale climate data based on cloud. The cloud-based framework adequately uses cloud computing and data storage and harnesses a workflow management and scheduling engine. The framework is designed to be a countermeasure to data analytics challenges, aiding algorithms developers, analysts and end users engaged in data analytics and visualization. In [2], proposed framework to climate studies can perform spatio-temporal data analysis and visualization of a large multi-dimensional climate dataset and reduce execution time. An integrated scientific experiment framework which is suggested in [3] offers experimental tool for users to simulate numerical analysis in the fluid dynamics domain. The system is based on grid computing equipped with UNICORE [4] which is a middleware that offers user interfaces and manages distributed computing resources. In [3], this framework support pre-process for generating input files of numerical analysis and maintain post-process for visualizing analyzed results. In [5], a paper suggests data management and analysis on health care data based on Hadoop which offers distributed computing environment. This proposed distributed computing environment shows that data processing using Hadoop reduces the execution time.

3. A Problem Solving Environment Framework for Dietary Analysis

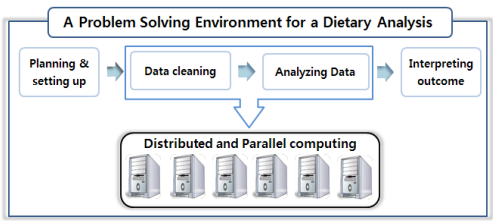


Figure 1. Four steps of a Problem Solving Environment for Dietary Analysis

A problem solving environment for dietary analysis in the field of food and nutrition includes 4 steps: 1. planning and setting up a hypothesis, 2. data cleaning, 3. analyzing data, and 4. interpreting outcomes in figure 1.

First, a user sets up a hypothesis to prove. A User prepares all kinds of raw data set as needed in order to prove the hypothesis. Collecting raw data set measured from various surveys is proceeded. In the data cleaning step, cleaning of raw data set which is provided as surveys is inevitable in order to perform further analysis. Most tasks in this step are tedious and labor-intensive to find appropriate input values from raw data set. Therefore, our system provides an automated and distributed process to generate input data efficiently. After loading the data, input data are cleaned by detecting outliers, eliminating missing values, and setting up certain conditions to remove unnecessary data. Some continuous variables need to be grouped in accordance with a standard value, which is a monotonous work, in order to make the data easy to analyze.

Our framework provides simple processes that look up proper sections rapidly and easy to conduct parametric study by setting diverse parameters at each operation in parallel. After finishing the data cleaning step, raw data are modified into required accuracy and valid data to prove user-defined hypothesis.

In the analyzing step, a user tries to analyze data by adjusting and varying experimental parameters and weights and then select useful analysis strategies which expose their desirable outcome models. All tasks in this step are long-time taken and repetitive. Therefore, each task can be simultaneously processed with using a parallel and distributed processing infrastructure which ensures reduced executing time for an application.

A user has to perform processes of analyzing diverse testing case and confirming result of testing case repeatedly. However, our proposed framework helps a user execute tasks in parallel by using special conditions as needed.

User-defined hypothesis is proved by the suggested statistical models which explain the most of data. In statistical analyzing step, a user needs to repetitively conduct tasks just replacing a kind of statistical models and repeatedly confirm result of the model and then they can get optimal outcome. It is essential for performing an experiment to deal with analyzing step on distributed and parallel computing by using the problem solving environment.

Finally, a user tries to interpret experimental results in the last step where all meaningful outcomes are .

The suggested system offers parallel and distributed processing infrastructure by adapting Hadoop API. So, each task can be simultaneously processed speedy. The suggested system is equipped with Hadoop architecture on top of basic statistics analysis tools like SAS [6] popularly used in food and nutrition domain. Our problem solving environment for dietary analysis is able to experiment efficiently and accurately and also offers a user reduce making mistakes by automating procedure.

For user's convenience, we designed interface to be organized dynamically by controlling event about

개인 맞춤형 건강한 식습관을 위한 모바일 앱 설계

박윤선[1] 유다빈[1] 양다예[1] 최지은[1] 박소현[2] 김윤희[1O]
숙명여자대학교 컴퓨터과학과[1], 숙명여자대학교 식품영양학과[2]
12sun14@gmail.com, udb0408@gmail.com, ekdp1119@gmail.com, jechoi1205@sm.ac.kr,
cjdtjfah3@naver.com, yulan@sm.ac.kr

A Design of a Mobile App for Personalized and Healthy Diet

Yoonsun Park[1] Dabeen Yu[1] Daye Yang[1] Jieun Choi[1] Sohyun Park[2] Yoonhee Kim[1O]
[1]Dept. of Computer Science, Sookmyung Women's University
[2]Dept. of Food and Nutrition, Sookmyung Women's University

요 약

현대인들의 만성질환 예방에 대한 관심이 높아지고 건강관리에 대한 관심이 증가함에 따라 식단관리의 중요성이 대두되고 있다. 이에 따라 효율적인 식단 관리를 제공하는 시스템이 증가하고 있다. 하지만 기존의 수많은 애플리케이션의 경우 식단을 입력하고 간단한 피드백을 주는 수준에 그치고 개인의 특성을 고려한 피드백이 이루어지지 않고 있다. 따라서 개인의 질병 및 신체 조건을 고려한 개인 식습관 관리 시스템이 필요하다. 본 논문에서 제안하는 시스템은 개인의 특성을 고려한 맞춤형 식단 평가 및 추천 기능 기능을 제공하여 사용 효과를 높이고 최종적으로 사용자들의 식생활 습관을 개선할 수 있는 모바일 앱이다. 제안된 모바일 앱은 입력된 사진으로 동료와 식단을 공유하여 사용자의 흥미를 유발해 애플리케이션 이용 지속을 고려하여 설계하였다.

1. 서 론

현대인들의 건강관리에 대한 관심이 증가함에 따라 식단관리의 중요성과 함께 식단관리 모바일 앱을 사용하는 사용자들이 늘고 있다. 하지만 대부분의 경우 대상이 제한되어 있으며 식단 관리에 중점을 두고 있지 않다. 또한 식단에 대한 피드백은 주로 칼로리와 3대 영양소의 섭취량 분석이 전부이다. 따라서 사용자의 특성에 맞춰 구체적인 피드백을 제공하여 좋은 식습관으로의 변화를 유도할 뿐만 아니라 좋지 않은 식습관으로 인한 만성 질병을 예방할 수 있는 시스템의 개발이 필요하다.

본 논문에서는 사용자로부터 식단을 입력받아 사용자의 건강상태와 신체조건을 기반으로 식단 평가 및 추천로직을 설계하여 모바일 앱을 제안한다. 제안된 모바일 앱은 '시콩이'라고 명명하였다. 로직은 기록된 식단에 대한 단순한 분석 결과를 보여주는 것만이 아니라 섭취한 음식의 식품성분이 사용자에게 어떤 영향을 미치는지도 분석한다. 또한 사용자의 식사 패턴을 분석하여 사용자가 좋은 식습관을 실천 할 수 있도록 식단 추천 기능도 제공한다. 다른 사람들의 식단을 공유하는 기능을 통해서 앱의 사용 지속성을 고려하였다. 본 논문에서는 정확하고 효과적인 식단 관리로 건강한 식생활을 유도하는 모바일 앱을 제작하기 위해 식품영양학적 분석[1]을 적용하여 로직을 구성하였다.

본 논문의 구성은 다음과 같다. 1장의 서론에 이어 2장에서는 관련 연구들을 살펴보고, 3장과 4장에서는 시스템의 기능 및 설계에 대해 설명하고 마지막으로 5장에서 결론을 맺는다.

2. 관련연구

대표적인 식단관리 모바일 앱으로 [2], [3]이 있다. [2]는 설문조사를 통한 맞춤형 개별 다이어트 플랜과 식단 관리를 제공하지만 설문조사가 지나치게 많고, 식단을 입력하는 경우에도 직접 입력하는 방식만 제공된다. [3]의 경우 식단을 기록할 때 검색방식이 다양화 되어있지만 사용자로부터 입력받는 데이터가 너무 많다. [2]와 [3] 모두 식단 평가에서는 3대 영양소의 섭취에 대한 평가만을 받고 있으며 식사 패턴 분석이나 사용자의 특성에 맞는 식단 추천을 통한 식단 관리는 어렵다.

한편, [4]는 사진과 동료 피드백을 사용하여 자가 모니터링을 하는 모바일 앱의 사용 지속과 관련된 요인을 분석하였다. [4]에서는 자신의 식단 사진을 입력하는 것과 모바일 앱을 통해 동료와 함께 식단 공유, 동료 피드백 등 소통이 더해진다면 애플리케이션 사용 빈도수가 증가되는 긍정적인 효과가 있다고 한다. 본 논문에서는 식단을 사용자로부터 입력 받아 사용자 맞춤형 관리가 가능하도록 식단의 평가, 추천 기능을 설계하고 이를 모바일 앱으로 구현하였다.

[O]교신저자

1) 이 논문은 2014년도 정부(교육부)의 재원으로 한국과학창의재단(대학생 창의융합형 연구과제 지원사업)의 지원을 받아 수행된 연구임.

3. 기능 설계

시콩이에서 제안하는 시스템의 구조는 그림 1과 같다. 사용자는 시콩이를 통해 식단을 입력한다. 입력된 식단은 식품군, 식품명, 영양정보, 식단 이미지 등의 정보로 식단 데이터베이스에 저장된다. 식단 데이터베이스를 이용하여 각 식단의 영양정보를 분석한다. 영양 평가에서는 저열량 혹은 고열량 식단, 트랜스지방, 콜레스테롤, 나트륨 및 단백질 함량이 높은 식단 등을 기준으로 평가한다. 평가 결과는 사용자 정보 데이터베이스로 부터 가져온 저체중, 비만과 같은 체질량 상태와 당뇨, 고혈압 등의 질병 특성과 결합하여 식단 가이드라인을 생성하는데 이용된다. 시스템은 식단 가이드라인을 가공하여 사용자에게 맞춤형 식단을 추천한다.

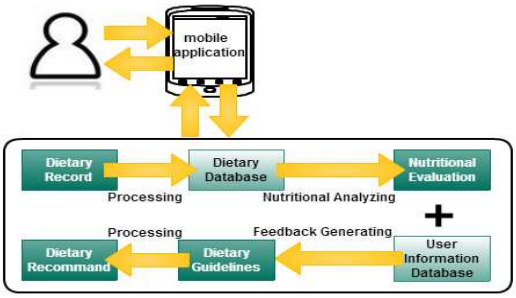


그림 1 시스템 구조도

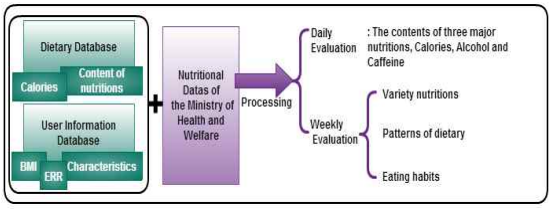


그림 2 평가 로직

시콩이는 크게 식단 평가, 식단 추천 및 검색, 식단 공유의 3가지 기능을 제공한다.

평가 기능은 그림 2와 같은 평가 로직을 갖는다. 먼저 평가할 식단의 칼로리, 영양소 함량 등을 식단 데이터베이스에서 가져오고 체질량 지수, 에너지 필요 추정량 등을 사용자 정보 데이터베이스 내의 나이, 성별, 체중, 신장 등의 값으로 산출한다. 이를 영양학적 정보[5]와 결합하여 평가를 제시한다. 평가는 크게 일별 평가와 주간 평가로 나뉜다. 일별 평가에서는 3대 영양소 섭취, 알코올과 카페인 섭취, 권장 일일 칼로리 섭취를 평가한다. 3대 영양소는 탄수화물 55~70%, 단백질 7~20%, 지방 5~25% 범위 내에서 배분하는 것이 권장됨[5]에 따라 섭취 정도를 부족, 적정, 과다의 상태로 나타낸다. 알코올은 술의 도수별로 200~700ml 이상, 카페인은 녹차와 커피 등 카페인을 함유한 식품을 하루 3잔 이상 마실 경우 경고한다. 권장 일일 칼로리는 총 에너지 소비량을 의미하는데 이는 사용자 정보 데이터베이스에서 사용자의 활동 정도, 나이, 신장, 체중, 성별의 값을 이용하여 산출한다. 그 공식[5]은 다음과 같다.

* 에너지 필요 추정량(ERR)
  = 총 에너지 소비량(TEE) + 생애주기별 부가량
* 총 에너지 소비량(TEE)
= 기초 대사량(BEE)* 활동계수(PA)
* 생애주기별 부가량은 영아, 유아, 아동, 청소년과 성인의 경우 임신 2분기부터 해당함.

주간 평가는 세부 영양소 섭취, 영양소 군별 섭취 횟수, 식습관을 평가한다. 세부 영양소 섭취 평가 시에는 단백질을 제외한 영양소들이 권장 섭취량이 아닌 목표 섭취량을 가진다[5]. 당뇨나 고혈압의 질병이 있는 사용자에 대해서는 질병 별 고위험군 식품을 식품군과 식품명으로 분류하고 이를 key로 하여 사용자가 섭취한 식단의 데이터베이스 내에서 각 key들을 찾아 빈도수를 그래프로 보여준다. 영양소 군별 섭취 횟수 평가는 연령과 권장 섭취 칼로리에 따라 A형 식단(어린이 및 청소년형) 또는 B형 식단(성인용) 권장식사패턴을 적용한다[5]. 식사 패턴은 곡류, 고기·생선·달걀·콩류, 채소류, 과일류, 우유·유제품류, 유지당류의 6가지 영양소 군별로 권장 섭취 횟수를 제시한다.

식습관 평가는 사용자가 입력한 식단의 패턴 분석을 통해 잘못된 식습관을 경고한다. 사용자는 식단을 아침, 점심, 저녁, 간식, 야식으로 분류하여 입력한다. 주 2회 이상 아침 식사를 거르거나 야식을 섭취할 경우 아침 식사를 권장하거나 야식증후군에 대한 경고를 한다. 또한 한 끼의 적정 섭취 칼로리인 500~600kcal를 크게 벗어나거나 부족한 식단에 대해서도 알려주고, 체질량 상태에 따라 주간 섭취 칼로리에 대한 평가를 제공한다. 사용자의 체질량 상태는 아래의 우리나라 기준 체질량 지수(BMI)[5]를 사용하여 판단한다.

* 체질량 지수(BMI) = 체중(kg)/신장(m^2)
* BMI 18.5 미만 = 저체중
  BMI 18.5 ~25 = 정상
  BMI 25 이상 = 과체중

시콩이의 두 번째 기능은 사용자의 특성에 따른 식단 추천 및 검색 기능이다. 당뇨에 걸린 사용자인 경우, 트랜스지방과 콜레스테롤이 높은 식품 등 당뇨에 치명적인 식품들을 제외한 식단을 추천해준다. 비만인 경우 저열량 식단이 추천 기준 내에 적용된다. 식단 검색 시에는 식품군, 식품명 뿐 아니라 100kcal부터 1000kcal 사이의 칼로리 조건 설정 등 세부 기준으로 검색이 가능하다.

세 번째 기능인 공유기능은 식단 입력 시 함께 식사한 사람의 아이디나 이름을 입력하여 식단을 공유하는 기능이다. 이는 식단 사진을 동료와 함께 공유할 수 있게 하여 사용자의 지속적인 사용을 도와 지속적인 식단관리가 가능하게 한다.

식이데이터 분석 자동화 문제풀이환경 설계

최지은 안윤선 송지현 김윤희°
숙명여자대학교 컴퓨터과학과
jechoi1205@sm.ac.kr, ahnysun@sm.ac.kr, jihyun.song0829@gmail.com, yulan@sm.ac.kr

A Design of Problem Solving Environment for Automation of Dietary Data Analysis

Jieun Choi Younsun Ahn Jihyun song Yoonhee Kim°
Dept. of Computer Science, Sookmyung Women's University

요 약

웰니스는 개인 건강 및 삶의 패턴, 운동 또는 식이 패턴에 대한 많은 양의 측정된 데이터에 의해 결정된다. 특히 식이 패턴은 개인 건강과 밀접한 관련이 있어 개인 건강을 분석하기위해 식이 패턴을 분석하는 것은 중요하다. 식품 영양학 분야에서는 식이 분석을 위해 많은 양의 입력 데이터를 처리하는 것이 필요하고 그 단계는 시간이 많이 소모되는 지루하고 반복적인 과정이다. 한편, 과학자들에게 있어서 많은 양의 데이터를 다루고 반복적인 작업이 수행되는 실험의 경우 효율적인 실험과 실험의 편리성을 위해서 문제풀이환경을 구축하는 것이 필요하다. 본 논문에서는 식이데이터 분석 자동화를 위한 문제풀이환경을 제안한다. 제안된 문제풀이환경은 순차적이고 병렬적인 작업들을 적절하게 배치함으로써 전체 실험에 소요되는 시간을 단축할 수 있다.

1. 서 론

웰니스는 사람들이 개인의 건강과 삶의 질을 개선하고 향상시키기 위해 최근 관심을 갖는 가장 인기 있는 분야 중 하나이다. 신체 내외적인 상태를 나타내는 많은 양의 개인 건강 데이터는 물리적으로 측정되는 값뿐만 아니라 식이 패턴과 같이 일상에서 매일 반복적으로 기록되는 개인의 활동도 포함된다. 식이 패턴은 식품 영양학적으로 개인 건강과 밀접한 연관이 있다. 따라서 식품 영양학에서 식이 패턴을 분석하는 것은 개인이 가지고 있을 수 있는 잠재적인 질병을 분석하고 예방하기 위해서 필요하다.

축적된 개인 건강 데이터(life log, PHR (Physical Health Record), EMR(Electronic Medical Record), etc.)는 건강을 유지하기 위해 신체 상태와 특정 질병의 관계를 찾는데 유용하다. 이러한 개인 건강 데이터를 사용하여 통계적 분석을 수행하는 경우 빠르고 효과적으로 많은 양의 데이터를 처리하는 것이 필요하다. 그러나 대량의 축적된 개인 건강 데이터를 효율적으로 처리하고 관리 및 저장하는 것은 매우 어렵다.

본 논문에서는 많은 양의 식이데이터를 처리하고 분석하는 과정에서의 반복적인 작업을 자동화하기 하기 위한 분석 자동화 문제풀이환경을 제안한다. 제안된 문제풀이환경은 병렬 컴퓨팅 환경에서 식이 패턴 분석을 가속화하고 식품 영양학 적으로 식이 데이터 분석을 통한 개인 건강의 잠재적인 위협 요소를 파악한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구를 소개하고 3장에서 설계한 식이데이터 분석 자동화 문제풀이환경을 설명한다. 4장에서는 식품 영양학적 시나리오를 통해 설계한 시스템을 설명하고, 마지막 5장에서 결론을 맺는다.

2. 관련연구

다양한 분야에서 병렬 및 분산 클라우드와 그리드 컴퓨팅 환경에서 데이터 분석하는 것을 돕는 플랫폼에 관한 연구가 진행되고 있다. CAP(Collaborative Analytics Platform, 2013~2019)[1] 프로젝트의 경우 유럽에서 진행되는 클라우드 인프라 기반 표준 서비스 플랫폼 개발을 돕기 위한 프로젝트이다. 한편, 특정 분야를 대상으로 문제풀이환경을 제안하는 관련 연구가 활발하게 진행되고 있다. 기후 데이터 분석을 위한 문제풀이환경에 대한 연구[2]와 유체 역학 실험환경을 위한 문제풀이환경 연구[3]가 진행되었다. [2]에서는 클라우드 기반의 대량의 기후 데이터를 분석하기 위한 프레임워크를 제안한다. 클라우드 기반의 프레임워크는 효과적으로 클라우드 컴퓨팅과 데이터 스토리지, 유용한 워크플로우 관리, 스케줄러 엔진과 같은 컴포넌트를 활용한다. [3]에서 제안된 통합 과학적 실험 프레임워크는 사용자가 유체 역학에 대해 수치 해석적으로 분석 실험하는 것을 돕는다. [4]에서는 하둡을 기반으로 하여 건강 데이터에 관하여 분산 컴퓨팅 환경에서 데이터 관리 및 분석

° "본 연구는 미래창조과학부 및 정보통신산업진흥원의 ICT 융합고급인력과정지원사업의 연구결과로 수행되었음" (NIPA-2014-H0401-14-1022)

하는 것을 제안한다.

3. 식이데이터 분석 자동화 문제풀이환경 설계

식품 영양학에서 식이데이터 분석을 위한 문제풀이환경은 다음과 같은 4가지 과정을 거친다. 1. 가설 설정 및 계획, 2. 데이터 클리닝, 3. 데이터 분석, 4. 결과 해석. 그림1은 식이데이터 분석 자동화 문제풀이환경의 4단계를 나타낸다.

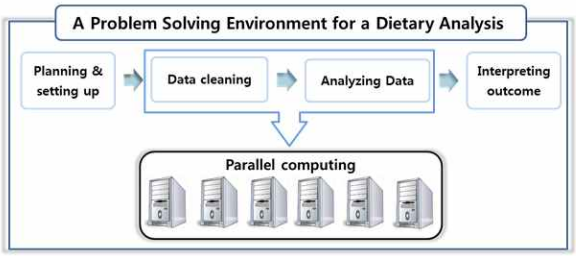


그림1. 식이데이터 분석 자동화 문제풀이환경의 4단계

먼저, 사용자는 데이터를 이용하여 증명하고자 하는 가설을 설정한다. 가설을 증명하기 위해 가공되지 않은 데이터를 준비한다. 가공되지 않은 수많은 독립된 데이터는 하나의 입력 데이터로 구성된다.

데이터 클리닝 단계에서는 데이터 분석을 위해 입력 데이터에 대한 데이터 정제가 필수적이다. 예를 들어 특정 값에 대해 데이터가 누락된 경우와 비정상적인 데이터를 포함한 경우는 해당 값을 정상적인 범위 안에서 수정하거나 제거하는 것이 필요하다. 또는 몇몇 연속적인 값들을 분석에 사용하기 위해 적절한 몇 개의 범주로 나누는 과정이 필요하다. 이 과정에서 사용자는 특정 조건을 사용하여 불필요한 데이터를 제거하거나 반복적인 작업을 통해 연속적인 데이터를 적절한 범주로 나누는 과정을 진행한다. 따라서 데이터 클리닝 단계는 전체 분석 과정에 있어서 시간이 많이 소요되고 사용자의 노동력이 요구된다. 제안된 시스템은 분석을 위한 데이터 클리닝 단계에서의 이러한 반복적인 과정을 자동화하여 사용자가 효율적으로 데이터 분석을 진행하도록 돕는다. 데이터 클리닝 단계가 완료된 후, 사용자는 정의한 가설을 분석하기위한 적절한 데이터를 얻게 된다.

분석 단계에서 사용자는 원하는 가설을 얻기 위해 다양한 데이터 분석 방법을 시도한다. 이 단계에서는 같은 데이터에 대해 여러 통계적 분석을 적용해야 한다. 따라서 많은 양의 데이터를 분석하는데 있어서 오랜 시간이 걸리고 각 분석에는 반복적인 작업이 포함된다. 제안된 시스템은 동일 데이터에 대해 다양한 통계적 분석 방법을 적용하는 것을 병렬 처리하여 보다 빠른 분석을 수행한다.

마지막으로 사용자는 가설을 입증하기 위해 실험 결과를 해석하여 다양한 표와 그래프를 도출해 낸다.

제안된 시스템은 대량의 데이터를 통계적으로 분석하기 위한 자동화된 문제 풀이 환경을 제공한다. 각 반복적인 작업은 분산되어 동시에 수행될 수 있다. 제안된 시스템은 식품 영양학에서 널리 사용되는 통계 분석 시스템 SAS[5]를 활용한다. 병렬 수행을 위해서 따라서 식이 데이터를 분석하는 과정에서 반복적인 작업과 병렬 수행이 가능한 작업이 분산되어 수행되기 때문에 전체 분석과정을 빠르고 효율적으로 수행한다.

4. 사례 연구

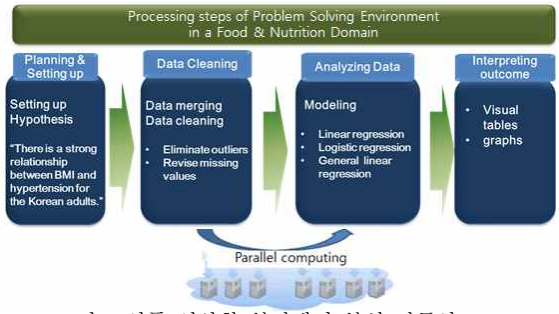


그림2. 식품 영양학 분야에서 분석 자동화 문제풀이환경의 처리 단계

그림 2는 식품 영양학 시나리오에 따라 식이데이터의 분석을 자동화하는 문제풀이환경의 4가지 처리 단계를 보여준다. 사용자는 가설 설정 및 계획 단계에서 "한국 성인의 BMI와 고혈압 사이에 관계가 있다"라는 가설을 세운다.

실험에는 국민 건강 영양 조사[6]의 2007년부터 2012년까지 각 연도별 미 가공된 데이터를 사용한다. 연도별 데이터 크기는 07년도(53.9MB), 08년도(114MB), 09년도(123MB), 10년도(70MB), 11년도(66.6MB), 12년도(63.0MB)이다. 분석에 사용되기 위해 연도별 데이터는 하나의 입력 데이터로 병합된다. 병합된 데이터는 50,405명의 측정 데이터를 갖는다. 입력 데이터는 데이터 클리닝 단계에서 가설 검증을 위해 불필요한 데이터를 제거 하는 단계를 거친다. 예를 들어 암에 대한 데이터는 증명하고자 하는 가설과 관련 없는 데이터이기 때문에 사용자는 특정 조건을 입력하여 불필요한 관련 데이터를 제거한다. 또한 알코올 섭취 및 혈압, 흡연량과 같은 연속적인 변수들에 관해서는 분석을 위해 적절한 그룹으로 나누는 것이 필요하다. 따라서 제안된 프레임워크는 많은 양의 데이터에서 불필요한 값을 처리하는 것과 특정 변수에 대해 파라미터 값을 나누어 절적한 범주를 나누는 것을 자동화 한다. 데이터 클리닝 과정을 마친 후, 가공된 18,220 명의 건강 데이터가 가설 분석에 사용된다.

제안된 프레임워크는 나이, 성별, 연도, BMI, 혈압과 같은 변수를 사용하여 적절한 모델을 찾기 위해, 사용자에게 선형 회귀분석, 로지스틱 회기분석, 일반 선형 모형과 같은 통계학적 분석 방법을 제공한다. 이와 같은 통계학적 모델은 다양한 데이터 타입에 따라 적용되는 것이 나뉜다. 예를 들어 선형 회기분석은 오직 연속적인 변수에 대한 분석에서 적용된다. 통계학적 분석 결과 다음과 같은 선형 방정식이 정의된다. 이 방정식은 선형 회귀 분석을 사용하여 대부분의 데이터를 설명하는데

SNS 환경에서의 광고 오피니언 마이닝 시스템 설계

이수민 유혜진 이아름 방지선 안윤선 김윤희[O]
숙명여자대학교 컴퓨터과학과
2smlee221@gmail.com winnie1219391@gmail.com arsbsp@gmail.com, nin23bix@gmail.com,
ahnysun@sm.ac.kr yulan@sm.ac.kr

A design of Opinion Mining System of Advertisement in SNS Environments

Soomin Lee Hye-Jin Rhu A-Reum Lee Jiseon Bang Younsun Ahn Yoonhee Kim[O]
Dept. of Computer Scinence, Sookmyung Women's University

요 약

SNS(Social Networking System)가 활성화 되면서 많은 사용자들은 실시간적으로 개인적인 의견을 주고받는다. 따라서 SNS 상의 많은 양의 데이터를 분석하여 다양한 새로운 정보들을 생성해 낼 수 있게 되었다. 본 시스템에서는 트위터에 올라오는 제품에 대한 트윗들을 실시간으로 분석하여 해당 제품에 대한 구체적인 광고 반응 평가를 제공하였다. Hadoop을 통하여 많은 양의 데이터들을 실시간으로 분석 하였고 통계적 기법과 오피니언 마이닝을 추가 적용하여 해당 제품에 대한 신뢰도를 분석하였다. 분석 한 신뢰도를 그래프화하여 제공함으로써 광고주들이 실시간적으로 해당 제품의 반응을 파악하기 편리한 시스템을 구축하였다.

1. 서 론

2013년 인터넷이용실태조사[1]에 따르면 인터넷 이용자의 55.1%가 SNS 이용자로 나타났다. 또한 SNS 이용자의 56.2%가 하루에 1회 이상 이용하는 것으로 나타났다. SNS 이용이유로 일상생활에 대한 기록, 개인적 관심사 공유, 전문 정보나, 지식공유가 높은 비중으로 나타난 것으로 보아 SNS 분석을 통해 보다 개인적이고 신속한 정보를 얻을 수 있을 것이다. 따라서 SNS데이터를 이용하여 광고 효과를 분석 할 수 있다고 판단하였다. 기존의 광고효과 측정방법은 체크리스트를 만들어 조사를 하거나 소비자 패널조사 등 사람이 직접 조사하는 방법이 대부분이었다. 이러한 방법은 분석에 소요되는 시간이 오래 걸리고, 높은 정확도를 보장받을 수 없다는 단점이 있다. 따라서, 광고에 대한 효과, 반응을 실시간으로 자동 분석하여 제공되는 프로그램의 필요성이 대두되었다.

본 논문이 웹 시스템은, 이전 논문[2]에서 구현했던 실시간 광고 효과 분석 시스템에, 통계적 기법과 오피니언 마이닝을 추가 적용하여 해당 광고에 대한 평가를 분석하여 광고에 대한 신뢰도를 파악하였다. 이를 통해, 본 웹 시스템은 SNS 상에 광고의 신뢰도 분석과 이에 근거로 단어 카테고리 별 분석을 제공한다. 또한, 동종광고들의 비교 분석 프로그램을 구축하여 광고 반응 및 효과 분석을 사용자가 언제든지 실시간으로 필요한 광고 분석을 할 수 있는 웹 시스템을 구축하였다.

1장의 서론에 이어 2장에서는 관련연구에 대해 서술하였고, 3장과 4장에서는 제안하는 시스템의 기능 설계와 이용방법에 대해 설명하며 마지막으로 5장에서 결론을 맺는다.

2. 관련연구

TV광고의 효과를 간접적으로 계량하기 위한 방법으로 한국방송광고진흥공사(Kobaco)가 프로그램 몰입지수(Program Engagement Index, PEI)를 개발하였다. 이 지표는 TV프로그램의 시청자가 자리를 뜨거나 다른 채널로 돌리지 않았는지, 눈을 떼지 않고 계속 시청했는지 등 TV프로그램의 몰입도를 나타내는 8가지 항목을 측정하여 산출한 것이다. 하지만 이러한 측정 지표는 구체적인 산출 과정에 대한 공개나 검증이 없어 신뢰성에 의문이

[O]교신저자

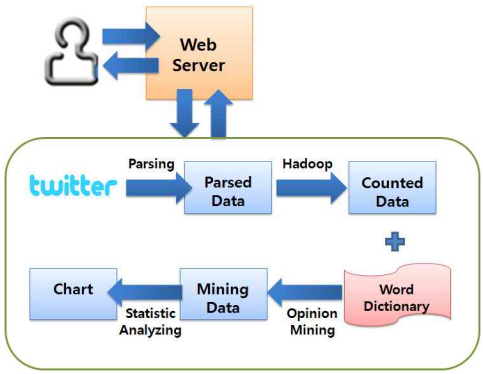
1) 이 논문은 2014년 정부재원(미래창조과학부 여대학(원)생 공학연구팀제 사업)으로 한국연구재단과 한국여성과학기술인지원센터의 지원을 받아 연구되었습니다.

제기되고 있다[3].

현재 SNS 분석 사이트로 소셜메트릭스 인사이트[4] 와 펄스k[5]가 있다. 소셜메트릭스 인사이트는 탐색어 언급도, 긍/부정어 추이 등 다양한 분석을 제공하지만, 검색 시점의 결과만 보여주고 있어 광고효과의 변화 추이는 알 수 없다. 한편 펄스k는 비교분석이나 지역별 분석 등 다양한 분석을 제공하고 있으나 유료이고, 분석 하는데 시간 소요가 있다. 오피니언 마이닝을 통한 텍스트 분석 시스템으로 K-LIWC가 있는데, 글의 구조, 단어 조합의 불균형 등을 통하여 텍스트의 신뢰도를 측정한다. 이를 통해 SNS 데이터 분석과 신뢰도측정이 가능하나, 실시간 분석이 어렵고 체계적인 알고리즘이 부족하다[6].

3. 설 계

본 논문에서 설계한 시스템의 구조는 다음 [그림1]과 같이 하둡을 이용하여 데이터를 수집하여 분산처리 후 오피니언 마이닝과 통계적 기법을 적용하여 웹을 통해 보여주는 형식으로 구성된다. 전체 시스템은 해당 광고의 트윗을 파싱해주는 하둡과 오피니언 마이닝과 통계적 기법을 이용하여 분석해주는 광고 반응 분석기, 사용자에게 분석 결과를 보여주는 웹 인터페이스로 나눌 수 있다. 각 기능에 대한 설명은 다음과 같다.

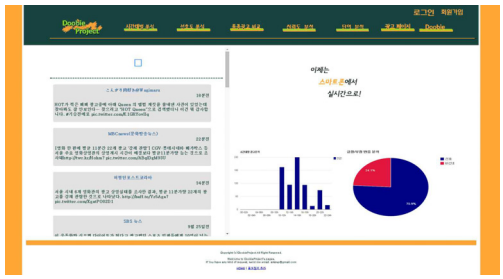


[그림 1] 시스템 구조도

이전 논문[2]에서 구현한 하둡 클러스터에서 파싱해온 데이터에 미리 구축한 단어사전을 토대로 오피니언 마이닝 기법을 추가 적용하여 언급된 문장의 단어 분석을 한다. 본 시스템에서는 긍정어 부정어 유추어 핵심어로 나눈 단어사전을 구축하여 오피니언 마이닝을 적용하여 이를 토대로 각각의 트윗의 단어 구성을 분석한다. 각각의 트윗에서 언급된 단어들은 긍정어 부정어 유추어 핵심어로 마이닝된다. 마이닝된 데이터에 통계적 기법의 식을 적용시켜 광고에 대한 신뢰도를 파악한다. 본 시스템에서는 베이지안 정리를 이용하여 통계적 분석을 한다. 베이지안 정리를 이용하여 단어별 확률을 구한다. 신뢰도 분석 시 마이닝을 통하여 저장해둔 단어 구성에 튜닝을 통해 얻어진 단어별 확률을 적용해 전체적 통계를 구한다. 이를 통해, 본 웹 시스템은 SNS 상에 광고의 신뢰도 분석과 이에 근거로 단어 구성 분석을 도표화 하여 제공한다. 광고주들은 해당제품을 검색하여 웹 서버에서 제공하는 신뢰도 평가와 단어 구성 분석 결과를 얻고 해당 광고에 대한 SNS 사용자들의 인지도 변화와 동종 제품과의 인지도 비교가 가능하다. 본 결과를 통해 구체적이고 신뢰적인 SNS 사용자들의 해당 광고에 대한 반응을 확인할 수 있다.

4. 구 현

먼저, 구현한 웹 서버는 다음 [그림2]과 같다. 실시간으로 파싱된 트윗들은 왼쪽 화면에 띄워져 실시간으로 어떤 내용들이 언급되어지고 있는지 확인이 가능하다. 오른쪽 하단에는 신뢰도와 단어 구성 분포율을 한 눈에 확인 할 수 있는 그래프를 제공한다.



[그림2] 웹 서버 홈페이지

SNS상의 구체적인 반응 분석을 위하여 광고주들은 단어 구성 분석을 참고한다. 단어 구성 분석은 각각의 제품들에 대하여 언급된 트윗들의 긍정어, 부정어, 유추어, 핵심어의 구성 분포율을 보여준다. 또한 동종 제품과의 비교를 할 수 있다. 단어 구성 분석은 다음과 같이 이루어진다. 단어 분석에 앞서 단어 사전을 구축한다. 구축한 단어 사전은 총 1672개의 단어들로, 각각 핵심어, 긍정어, 부정어, 유추어로 나뉘어 데이터베이스에 저장된다. 1시간 별로 crontab을 이용하여 해당 제품에 대해 언급된 트윗들은 하둡으로 파싱되어 모아진다. 모아진 트윗들은 구축한 단어사전을 토대로 오피니언 마이닝을 이용하여 각각 긍정, 부정, 유추, 핵심어로 나누어 단어 구성을 분석한다. 화장품에 대한 단어구성을 검색하면 다음

멀티미디어과학과

메모인더포토: 이미지에 메모 입력을 지원하는 이미지뷰어 구현

남지수, 이현주, 황수진, 이종우
숙명여자대학교 멀티미디어과학과
e-mail : bigrain@sm.ac.kr

Memo In The Photo: An Implementation of Image Viewer Supporting Insertion of Memos

Ji-Soo Nam, Hyeon-Ju Lee, Su-Jin Hwang, Jongwoo Lee
Dept of Multimedia Science, SookMyoung Woman's University

요 약

최근 스마트 기기의 사용이 보편화되고 정보전달의 매체로서 이미지가 많이 사용되고 있다. 따라서 이미지가 나타내는 정보를 사진과 함께 저장하고 해당 이미지를 상세히 안내할 수 있는 어플리케이션의 필요성이 대두되고 있다. 우리는 이에 사용자가 이미지의 원하는 특정 부분에 영역을 지정하고, 영역에 해당하는 텍스트를 입력하여 이를 열람할 수 있게 하려고 한다. 이로써 이미지에 대한 상세 정보전달이 용이한 메모입력 및 열람을 하는 안드로이드 어플리케이션을 제안한다. 사용자가 원하는 글자 색상과 크기로 메모를 저장하고 열람할 수 있다. 메모를 입력한 사진들은 별도의 폴더를 생성하여 저장한다. 메모는 이미지당 최대 2개까지 입력할 수 있으며, 데이터베이스에 저장된 사진과 메모는 '메모보기' 기능을 통해 열람이 가능하며, 메모는 수정과 삭제가 가능하다.

1. 서론

스마트기기의 보급이 보편화되고 기기 내의 카메라 기능이 향상됨에 따라 스마트 기기 내에서 이미지를 다루는 다양한 어플리케이션(이하 앱)이 개발되고 있다. 그 중에서도 이미지와 함께 사용자가 원하는 텍스트를 동시에 표현하여 이미지에 대한 정보를 전달할 수 있도록 텍스트 메모기능을 지원하는 다양한 앱이 주목 받고 있다. 그러나 이미지 자체에 대한 정보를 전달하는 기능을 가진 앱이 대부분이고, 대부분의 앱이 텍스트입력 방식이 아닌 터치입력방식을 통해 메모삽입을 제공하기 때문에 디바이스에 따라 터치펜이 제공되지 않으면 입력이 불편하다. 삼성 갤럭시 시리즈의 포토메모 기능은 사진위에 터치로 메모를 제공하지만, 원본이미지가 아닌 새로 만들어진 불투명한 좌우 대칭된 이미지에 메모를 한다. 따라서 원본 이미지와는 별개의 다른 이미지에 메모를 하게 되어 이미지와 텍스트의 직접적 연관성이 적다. 또한 이미지의 특정부분에 해당하는 메모를 표시할 수 없어 이미지의 특정 영역에 해당하는 구체적인 메모입력이 불가능하다. 본 논문에서는 이미지가 다양한 정보전달의 수단으로 이용되는 오늘날, 이미지와 함께 구체적인 정보를 효과적으로 전달하기 위하여 이미지에 영역별로 텍스트 메모입력이 가능한 앱을 개발하고자 하였다.

2. 메모 삽입 및 보기 구현

전체적인 구조도는 (그림 1)과 같다.

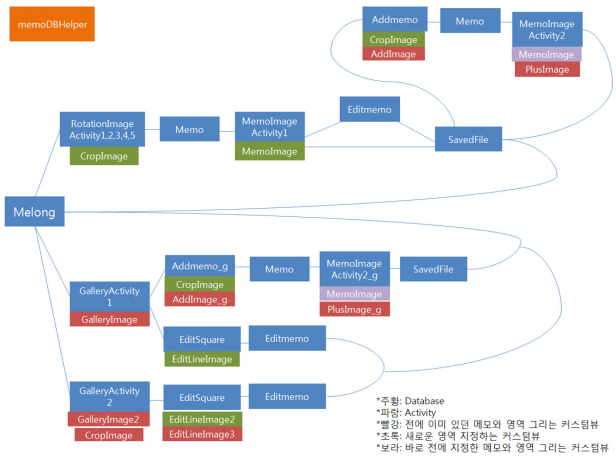


그림 1. 클래스 구조도

먼저 메모를 위해 Melong클래스에서 이미지를 선택하게 되면, sdcard에 Memo in the Photo라는 이름으로 새로운

폴더가 만들어진다. 그 폴더 안에 이미지가 선택된 시간을 구하는 SimpleDateFormat 클래스를 이용해 date를 구한 후[1], imageUrl에 저장한 이미지의 절대경로를 date의 이름으로 저장한다. 이 때, 카메라로 사진을 찍어 메모를 하게 되면 메모리 문제가 생긴다. 핸드폰에 내장되어있는 카메라로 찍은 이미지는 일반 이미지에 비해 용량이 매우 크다. 게다가 정면 카메라보다 후면 카메라의 용량이 약 4배 이상 차이나기 때문에 이미지를 불러올 때 메모리로 인한 문제가 발생한다. 이러한 메모리 누수를 방지하기 위하여 BitmapFactory class의 Options 메소드를 이용한다. 여기서 inSampleSize와 inPurgeable 변수를 사용하는데, 먼저 inSampleSize는 원본 이미지를 필요한 Size의 근접 범위로 줄여 Scale된 이미지를 불러온다. 이를 2로 만들어 주어, 이미지를 1/2 용량으로 줄인다. inPurgeable 변수를 true로 해주게 되면, 필요한 상황일 때 자동으로 메모리 해제를 해준다.

이미지 띄우기 기능은 Melong클래스에서 선택된 이미지에 영역을 지정하기 위해 RotationImageActivity에서 불러올 때, 영역을 지정하고[2] 이미지 위에 메모를 한 후 MemoImageActivity에서 불러올 때, 메모된 이미지를 보기 위해 GalleryActivity에서 불러올 때, 메모 추가 및 수정할 때 사용된다. 메모할 때는 Melong 클래스에서 이미지를 선택할 때 저장된 date를 SharedPreference를 이용해 다음 Activity로 넘겨 이미지를 띄운다. 메모된 사진을 볼 때는 갤러리에서 사진을 선택하게 되면 이미지의 uri이 mImageCaptureUri 변수에 저장되는데, 이를 uri로부터 실제 파일 경로를 가져오는 getRealImagePath 알고리즘을 이용하여 filepath 변수로 변환시킨다[3]. 이 변수를 데이터베이스에 들어있는 값과 비교해 일치하는 레코드의 모든 필드 값들을 불러오고 이 값들을 SharedPreference를 이용해 다음 Activity로 넘겨 이미지를 띄운다.

이미지의 삭제는 GalleryActivity클래스에서 보여 지는 메모를 LongTouch 했을 때 뜨는 Dialog에서 '삭제'버튼을 눌렀을 때 이루어진다. GalleryActivity클래스는 메모가 1개 저장되어 있을 때 메모를 보여주기 위해 나타나는 클래스이기 때문에 메모를 삭제하면 더 이상 그 이미지에 대한 메모가 없어서 이미지의 삭제도 같이 행해져야하기 때문이다. 이미지를 삭제할 때는 따로 정의해준 deleteFile을 이용하는데, 삭제할 이미지의 path를 인자로 받아 그 path에 해당하는 파일이 존재할 경우에는 delete메소드를 통해 파일을 삭제하는 기능을 수행한다[4].

메모가 입력되고 수정, 삭제의 과정을 거칠 때 데이터베이스를 관리하는 savedFile에서 sendBroadcast로 ACTION_MEDIA_MOUNTED를 통해 미디어 스캐너를 실행해 갤러리를 갱신하여 메모와 관련이 있는 이미지의 정보를 최신으로 유지한다.

이미지의 회전은 imageUrl에 저장한 이미지를 RotationImageActivity에서 불러올 때 이루어지는데, '회전'버튼을 누르게 되면 이 때 그 imageUrl에 저장된 bitmap 이미지를 회전 알고리즘을 이용해[5] 90도 회전한 상태로 다시 저장하여 원하는 회전 횟수만큼 다음 RotationImageActivity로 넘어간다.

메모리 정리를 위하여 이미지를 사용하는 모든 Activity에서는 다른 Activity로 넘어갈 때 모든 bitmap을 recycle하여 자원을 반납하고 bitmap과 Imageview를 null로 하여 초기화해준다.

3. 데이터베이스 구현

이미지에 메모를 저장하고 이를 열람하기 위해 memoDBHelper라는 이름의 class를 생성해 sqlite 테이블을 생성했다. memolist라는 이름을 가진 DB테이블은 원본의 URI를 저장하는 'picname', 메모의 내용인 'memo', 메모에 해당하는 영역의 꼭짓점 좌표를 각각 나타내는 'sx', 'ex', 'sy', 'ey', 이미지의 이름을 나타내며 저장된 날짜를 가리키는 'date', 이미지의 URL을 저장하는 'imageUrl', 메모의 색상을 가리키는 'color', 메모의 글자 크기를 나타내는 'size'. 즉, 10개의 필드로 구성된다.

field name	picname	memo	sx	ex	sy	ey	date	imageUrl	color	size
type	STRING	TEXT	FLOAT	FLOAT	FLOAT	FLOAT	TEXT	TEXT	INTEGER	INTEGER

그림 2. memolist 테이블의 필드명과 타입

데이터베이스에 레코드를 삽입하는 insert는 SavedFile 클래스에서 이루어진다. getSharedPreferences를 이용해 SharedPreferences에 저장했던 값들을 불러와 ContentValues형의 변수에 값을 차례로 넣고 앞서 만든 DB에 삽입하는 형태이다.

데이터베이스에 이미 있는 내용을 삭제하는 기능은 1개의 메모가 저장되었을 때 메모를 보여주는 클래스인 GalleryActivity클래스와 메모가 2개 저장되었을 때 메모를 보여주는 GalleryActivity클래스에서 각각 이루어진다. 메모를 LongTouch 했을 때 뜨는 Dialog에서 해당메모에 대한 '수정'과 '삭제'를 선택할 수 있는데[6], '삭제'버튼을 누르게 되면 쿼리를 통해 메모에 대한 레코드를 찾은 후 delete 메소드로 해당 레코드를 삭제한다.

DB 레코드를 수정하는 update는 앞서 말한 Dialog에서 '수정'버튼을 눌렀을 때, EditSquare(EditSquare2)클래스와 Editmemo클래스에서 다시 구한 값들로 행해진다[7]. EditSquare(EditSquare2)클래스에서는 영역에 대한 좌표를, Editmemo클래스에서는 수정할 메모의 내용과 색상, 글자 크기를 다시 설정할 수 있다. 그렇게 해서 재설정된 값은 체크버튼을 누르면 UPDATE를 행하는 쿼리를 통해 기존의 내용대신 레코드에 들어가도록 구현하였다.

4. Color Picker

각각의 색상이 달라 구분하기 쉬운 메모를 저장하기위해 메모의 색상을 선택할 때 안드로이드 ColorPicker를 사용하였다[8]. HoloColorPicker 라이브러리를 추가한 후[9], 채

Journal of Digital Contents Society Vol. 15 No. 2 Apr. 2014(pp. 197-208)
http://dx.doi.org/10.9728/dcs.2014.15.2.197

다중 플랫폼용 실습실 꾸미기 게임 앱 구현

노원빈*, 원문숙**, 이지혜**, 고은별*, 이종우***

요 약

많은 사람들이 다양한 플랫폼의 스마트폰으로 모바일 게임을 즐기고 있다. 그런데 게임개발을 할 때 스마트 단말기 플랫폼 별로 따로 개발을 해야 하는 것이 지금까지의 현실이었다. 이를 해결하기 위해 출시된 것이 게임개발 프레임워크인 Cocos2d-x이다. 기존의 Cocos2d에서 다중 플랫폼으로 발전시킨 것으로 하나의 소스 개발로 여러 모바일 기기 및 웹 브라우저 상에서 사용가능하도록 해준다. 공개 소프트웨어로 누구나 사용이 가능하며 C++과 OpenGL을 기반으로 쉽게 게임을 개발할 수 있다. 본 논문에서는 Cocos2d-x를 이용해 학습을 통한 PC실습실 공간 꾸미기 게임인 아이러브501을 구현하였다. 여기서 501은 저자들 소속 학교의 PC실습실 호수이다. 게임의 순기능 부각을 위해 퀴즈를 풀 수 있는 교육적 기능을 포함하였다. 멀티플랫폼 구현을 위해서 윈도우와 iOS환경에서 동시에 개발하였으며 안드로이드폰, 아이폰, 아이패드, 갤럭시탭 등 다양한 기기에 이식해서 실행 결과를 확인하였다.

키워드 : 다중 플랫폼, Cocos2d-x, 모바일 게임

Implementation of Multiplatform Game Application for Decorating The Lab

Wonbin Rho*, Moonsook Won**, Jihye Lee**, Eunbyul Ko*, Jongwoo Lee***

Abstract

Many people are now enjoying mobile games using various smartphone platforms. However, we have to develop games separately for each smart device platforms so far. Cocos2d-x, a game development framework is released to solve this problem. As a multiplatform version of the existing Cocos2d, Cocos2d-x can make one source code run on various platforms. It is an open software that is able to be used by everyone, and when using it, mobile games can be developed easily based on C++ and OpenGL. In this paper, we implemented a PC laboratory decorating game application, named ILove501, using Cocos2d-x. The 501 is a room number of our PC lab. ILove501 includes an educational feature of solving quizzes in order to highlight positive effects of game. For implementation of a multiplatform game, ILove501 was developed in Windows and iOS environment at the same time, and we verified the results of the execution by porting on a variety of devices such as Android, iPhone, iPad and Galaxy Tab.

Keywords : Multiplatform, Cocos2d-x, Mobile Game

1. 서론

많은 사람들이 다양한 플랫폼의 스마트폰으로 모바일 게임을 즐기고 있다. 이에 따라 모바일 게임 시장이 전체 게임 시장에서 차지하는 위상 역시 나날이 올라가고 있다[1]. 기존에 스마트폰용 게임을 개발하기 위해서는 안드로이드와 아

※ 교신저자(Corresponding Author): Jongwoo Lee
접수일:2014년 02월 05일, 수정일:2014년 02월 27일
완료일:2014년 04월 03일
* 숙명여자대학교 멀티미디어과학과 석사과정
** 숙명여자대학교 멀티미디어과학과 학부연구원
*** 숙명여자대학교 멀티미디어과학과 교수
Tel: +82-2-710-9952 , Fax: +82-2-710-9704
email: bigrain@sm.ac.kr
▣ 본 연구는 숙명여자대학교 2011학년도 교내연구비에 의해 수행되었음(과제번호 1-1103-0517)

이폰용 개발 언어가 달라 두 번 개발해야하는 번거로움이 있었다. 게임 앱은 대개 안드로이드 프로그래밍에서는 JAVA 언어를 기반으로 개발하고, 아이폰에서는 Objective-C 언어를 기반으로 Cocos2d라는 게임엔진을 활용해 개발된다[2]. 그러나 Cocos2d-x라는 스마트 단말기용 게임엔진이 출시되어 이러한 번거로움을 해결해주었다[3]. C++ 기반으로 개발된 하나의 코드로 안드로이드와 아이폰 양측에 이식(porting)할 수 있어 코드의 수정 없이 다중플랫폼(multiplatform) 게임 개발이 가능하기 때문이다.

본 논문에서는 Cocos2d-x를 활용하여 PC 실습실 꾸미기 게임을 구현하고 그 실행결과를 제안한다. 또한 Cocos2d-x를 이용해 하나의 코드를 다중 플랫폼으로 이식하는 과정도 제시한다. 게임의 대략적인 내용은 다음과 같다. 501호라는 낙후된 컴퓨터 실습실을 가정하고, 사용자는 전공퀴즈를 풀어 장학금을 획득한다. 획득한 장학금은 상점에서 아이템을 구매하는데 사용되고 구매한 아이템으로 실습실을 꾸밀 수 있다.

본 논문의 구성은 다음과 같다. 2장에서는 게임 구현에 관련된 기술 및 기존연구에 관해 기술하고, 3장에서는 다중 플랫폼을 위한 구현환경에 관해 기술한다. 4장에서는 구현과정을 소개하며 다중 플랫폼의 이식과정을 구체적으로 보인다, 5장에서는 실행화면을 보인다. 마지막 6장에서는 결론을 제시한다.

2. 관련 기술 및 관련 연구

2.1 기존연구

기존 모바일 2D게임엔진은 플랫폼 별 안드로이드용과 iOS용을 따로 개발해야 하는 실정이었다. iOS는 Cocos2d와 'iTGB for 2D games'[4]라는 게임엔진 등으로 개발했으며, 안드로이드는 ShadingZen 엔진[5], Andengine[6] 등 다양한 게임 엔진으로 개발되었다.

어느 때보다도 속도와 타이밍이 중요한 현 모바일 게임 시장에서 어느 한 쪽 플랫폼의 개발을 완료하고 다른 플랫폼용을 다시 개발하는 방식으로는 이 속도를 따라갈 수가 없다. 그래서 최근 모바일 업계에서 주목받는 기술이 크로스 플랫폼 게임 엔진이며, 그 중 2D 게임 개발은 오픈소스 프로젝트인 Cocos2d-x가 많이 쓰이고 있다. 타 게임엔진들과 비교하면 모바일에만 집중하며 고품질의 빠른 개발을 가능하게 함으로써 경쟁력을 높이고 있다[7]. Cocos2d-x라는 발전된 라이브러리가 나오면서 하나의 소스로 iOS와 안드로이드용 2D게임을 구현할 수 있게 된 것이다. 본 논문에서는 Cocos2d-x의 중요한 역할인 하나의 코드를 멀티 플랫폼용으로 이식하는 과정에 대해 설명한다.

2.1.1 Cocos2d

Cocos2d는 파이썬(Python) 언어로 작성된 게임 엔진을 아이폰 Objective-C 용으로 전환한 것이다[4]. 단일 플랫폼 오픈소스 2d 게임 프레임워크로, OpenGL을 기반으로 한 엔진이지만 엔진을 사용하기 위해 OpenGL을 알아야 할 필요는 없다. 게임 제작 시 공통적으로 사용되는 요소를 추출해 미리 함수나 클래스로 만들어 놓은 그래픽 라이브러리이다[8]. iOS 전용으로 Objective-C 언어를 사용하고, Xcode에서만 개발 가능하다. 이식 방식은 iOS에서 일반 애플리케이션 이식하는 방식과 동일하다. (그림 1)은 Cocos2d로 구현한 게임 사례이다.

(그림 1) Cocos2d로 개발된 Ention Wars

(Figure 1) Ention Wars developed with Cocos2d

2.1.2 AndEngine

AndEngine은 안드로이드 플랫폼용 게임을 쉽게 개발하기 위한 2D게임 엔진이다[6]. 윈도우 기반이나 iOS에서 개발가능하며, 인터넷에서 SDK를 다운받고 설치한다. 그 이후로는 기존의 애플리케이션 개발 형식과 동일한 방법으로 이식할 수 있다.

ISSN 2383-6318(Print) / ISSN 2383-6326(Online)
KIISE Transactions on Computing Practices, Vol. 20, No. 9, pp. 498-506, 2014. 9
http://dx.doi.org/10.5626/KTCP.2014.20.9.498

POI(Point Of Interest) 데이터 검색에서 문자열 유사도 측정 정확도 향상 기법

(Accuracy Improvement Methods for String Similarity Measurement in POI(Point Of Interest) Data Retrieval)

고 은 별 † (EunByul Ko) 이 종 우 †† (JongWoo Lee)

요 약 교통의 발달로 활동범위가 넓은 현대인들은 네비게이션과 지도 앱을 통한 길찾기 검색을 자주 이용한다. 하지만 기존 검색 시스템에서는 부정확한 질의어가 입력되면 원하는 결과를 출력하지 못한다. 이 문제를 해결하기 위해 집합-기반 POI 검색 알고리즘이 등장했고 이어 문자열 유사도 측정 기법, 중복 글자를 고려한 검색 알고리즘이 연구되었다. 본 논문에서는 이전에 연구된 문자열 유사도 측정 알고리즘의 정확도를 향상시킨 기법을 제안한다. 기존 문자열 유사도 측정 기법에서 고려하지 않았던 고유어의 추정단계와 중복 단어를 고려한 블록 및 블록 나열 순서 구하기를 추가하고 측정 기법을 수식화한다. 이를 통해 측정방법을 체계적으로 표현하고 일반화함으로써 POI 검색 결과의 정확도를 향상시킨다. 실험을 통해 본 논문에서 제시하는 기법이 검색 결과 및 검색 순위의 정확도를 향상시킨다는 것을 확인하였다.

키워드: POI 검색, 집합-기반 알고리즘, 문자열 유사도, 네비게이션, 한글 검색

Abstract With the development of smart transportation, people are likely to find their paths by using navigation and map application. However, the existing retrieval system cannot output the correct retrieval result due to the inaccurate query. In order to remedy this problem, set-based POI search algorithm was proposed. Subsequently, additionally a method for measuring POI name similarity and POI search algorithm supporting classifying duplicate characters were proposed. These algorithms tried to compensate the insufficient part of the compensate set-based POI search algorithm. In this paper, accuracy improvement methods for measuring string similarity in POI data retrieval system are proposed. By formulization, similarity measurement scheme is systematized and generalized with the development of transportation. As a result, it improves the accuracy of the retrieval result. From the experimental results, we can observe that our accuracy improvement methods show better performance than the previous algorithms.

Keywords: POI search, Set-based algorithm, string similarity, navigation, the Korean alphabet retrieval

· 이 논문은 2013년도 정부(교육부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임(2013R1A1A2013155)
· 본 연구과제는 숙명여자대학교 교내연구비지원에 의해 수행되었음(과제번호 1-1303-0191)

† 학생회원 : 숙명여자대학교 멀티미디어과학과
iks1614@naver.com
†† 종신회원 : 숙명여자대학교 멀티미디어과학과 교수
(Sookmyung Women's Univ.)
bigrain@sm.ac.kr
(Corresponding author임)

논문접수 : 2014년 5월 7일
(Received 7 May 2014)
논문수정 : 2014년 6월 19일
(Revised 19 June 2014)
심사완료 : 2014년 6월 20일
(Accepted 20 June 2014)

정보과학회 컴퓨팅의 실제 논문지 제20권 제9호(2014. 9)

1. 서 론

현대인들에게 네비게이션과 지도 앱을 통한 길찾기 검색은 매우 흔한 일이고 일상생활에서 많은 편의성을 준다. 대부분 길찾기 검색에서는 지리적 객체를 점으로 표현한 관심 지점이라는 의미의 POI(Point Of Interest)를 활용한다[1,2]. 네비게이션 같은 경우 사용자가 질의어를 입력하면 GPS로 현 위치를 파악하고, 사용자가 입력한 질의어, 즉 사용자의 POI를 파악하여 길 안내를 한다[3]. 하지만 여러 요인으로 인해 목적지 검색에 대한 정확도가 떨어진다. 특히 사용자의 불완전한 기억이나 오타로 인해 오질의어가 입력될 경우 검색 성능이 현저히 떨어진다.

이 같은 문제를 해결하기 위해 고안된 집합-기반 POI 검색 알고리즘은 기존 하드매칭 기법과 비교하여 우수한 성능 향상을 보인다[4]. 또한 집합-기반 POI 검색 알고리즘에 추가적으로 문자열 유사도 측정 기법을 도입해 보다 정확한 검색 결과를 출력하게 되었다[5]. 한편 집합 기반이기 때문에 중복 글자를 구분하지 못하는 문제를 보완한 알고리즘도 연구되었다[6]. 이렇게 보완된 알고리즘은 질의어에 글자의 누락, 추가, 도치가 있어도 정답에 근접한 검색결과를 출력함으로써 현저한 성능 향상을 보였다.

본 논문에서는 기존 측정 기법에서 고려하지 않았던 부분을 추가하여 정확성을 높인다. 또 기존 문자열 유사도 측정 기법에서는 경우를 나누어 점수를 배정했지만 본 논문에서는 문자열 유사도에 영향을 미치는 요소를 질의어, 데이터 길이, 외자 블록, 다자 블록으로 구분하고 각 요소의 중요도에 따라 가중치를 달리 하여 하나의 수식으로 표현한다.

본 논문의 구성은 다음과 같다. 2장에서 기존의 알고리즘에 대해 설명하고, 3장에서 기존 측정 기법에서 고려하지 않은 점을 어떻게 보완했는지 설명한다. 4장에서는 문자열 유사도에 영향을 미치는 요소들에 대해 설명하며, 이들이 통합된 수식을 제시한다. 5장에서는 실험을 통해 본 논문에서 제안하는 측정 기법의 성능을 평가하고, 6장에서 결론을 맺는다.

2. 기존 연구

본 장에서는 선행 연구인 기존 집합-기반 POI 검색 알고리즘과 기존 문자열 유사도 측정 기법의 전반적인 개념을 간략하게 설명한다.

2.1 집합-기반 POI 검색 알고리즘

집합-기반 POI 검색 알고리즘은 집합 개념을 적용하여 POI 데이터를 검색하는 기법으로, 전역적 쿼리 확장[7,8], 지역적 쿼리 확장[7,8], 적합성 피드백[9]처럼 복잡한 계산과 방대한 데이터를 필요로 하지 않아 차량 네비게이션과 같은 독립형 시스템에 적합하다[4].

집합-기반 POI 검색 알고리즘은 n개의 문자로 이루어진 질의어를 m개의 블록으로 균등 분할하여 각 블록에 대해 집합 기반 연산을 수행한다. 또 주어진 레코드에 포함된 질의어 내 글자의 총 개수를 의미하는 '차수'라는 개념을 이용하여 차수가 큰 데이터를 상위에 출력한다. 그래서 모든 글자가 정해진 순서로 정확하게 입력되어야 원하는 결과를 출력하는 하드매칭 기법의 단점을 보완하였다. 그림 1은 집합-기반 POI 검색 알고리즘에서 역 인덱스를 생성하는 과정을, 그림 2는 텍스트 검색 과정을 도식화한 것이다[4].

하지만 기존 집합-기반 POI 검색 알고리즘은 집합 기반이기 때문에 한 레코드 혹은 질의어 내에 같은 글자가 두 번 이상 중복으로 있을 경우 이를 구분하지 못하고 한 개만 포함한 것으로 인식한다. 이 문제를 해결하기 위해 중복된 회수만큼 해당 글자의 아이디를 추가로 부여한 알고리즘이 제안되었다[6,10]. 그림 3은 중복

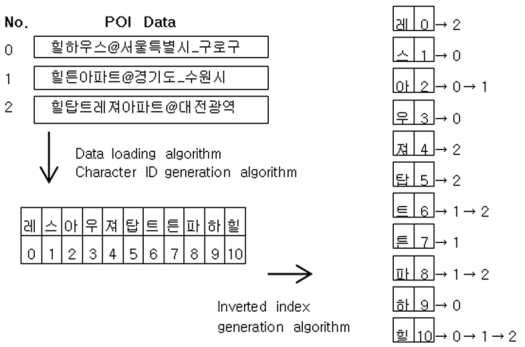


그림 1 역 인덱스 생성 과정

Fig. 1 steps for generating inverted indexes

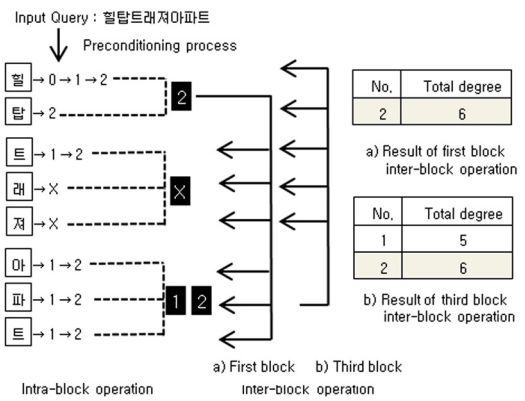


그림 2 텍스트 검색 과정

Fig. 2 steps for searching a text

ISSN 2383-6318(Print) / ISSN 2383-6326(Online)
KIISE Transactions on Computing Practices, Vol. 20, No. 11, pp. 567-572, 2014. 11
http://dx.doi.org/10.5626/KTCP.2014.20.11.567

파형 시퀀스의 공통 특징 추출 기반 모음 'ㅏ' 인식 구현

(Implementation of Korean Vowel 'ㅏ' Recognition based on Common Feature Extraction of Waveform Sequence)

노 원 빈 [†]　　이 종 우 [††]
(Wonbin Roh)　　(Jongwoo Lee)

요 약 최근 네트워크와 컴퓨팅 기술의 발달로 정보기기가 소형화되고 이동성이 중요시되면서 간편하게 제어할 수 있는 음성 인식에 대한 수요가 증가하고 있다. 본 논문은 음성 인식 시스템의 일부로써 한국어 음소 중 모음 'ㅏ' 인식에 대한 연구 결과를 제시한다. 음소는 음성을 구성하고 있는 최소단위로서 음성을 인식하는데 매우 중요한 역할을 한다. 그러나 각각의 음소들을 정확하게 인식하려면 발음의 다양성 등으로 인해 많은 어려움이 존재한다. 본 논문에서는 한국어 음소 중 모음 'ㅏ'를 인식하기 위한 간단하고도 새로운 방식을 제안한다. 제안된 'ㅏ' 인식 휴리스틱은 파형 시퀀스의 공통 특징 추출을 기반으로 이루어졌으며, 이는 기존의 복잡한 방법에 비해 간단하면서도 실험 결과 90% 이상의 성공률로 'ㅏ'를 인식하는 것을 확인하였다.

키워드: 음성인식, 모음, 파형, 시퀀스, 'ㅏ'

Abstract In recent years, computing and networking technologies have been developed, and the communication equipments have become smaller and the mobility has increased. In addition, the demand for easily-operated speech recognition has increased. This paper proposes method of recognizing the Korean phoneme 'ㅏ'. A phoneme is the smallest unit of sound, and it plays a significant role in speech recognition. However, the precise recognition of the phonemes has many obstacles since it has many variations in its pronunciation. This paper proposes a simple and efficient method that can be used to recognize a Korean vowel 'ㅏ'. The proposed method is based on the common features that are extracted from the 'ㅏ'waveform sequences, and this is simpler than when using the previous complex methods. The experimental results indicate that this method has a more than 90 percent accuracy in recognizing 'ㅏ'.

Keywords: speech recognition, vowel, wave, sequence, 'ㅏ'

· 이 논문은 2014년도 정부(교육부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임(NRF-2013R1A1A2013155)
· 본 연구과제는 숙명여자대학교 교내연구비지원에 의해 수행되었음 (과제번호1-1303-0191)

† 학생회원 : 숙명여자대학교 멀티미디어과학과
wonbin3371@naver.com
†† 종신회원 : 숙명여자대학교 멀티미디어과학과 교수
(Sookmyung Women's Univ.)
bigrain@sm.ac.kr
(Corresponding author임)

논문접수 : 2014년 2월 7일
(Received 7 February 2014)
논문수정 : 2014년 9월 25일
(Revised 25 September 2014)
심사완료 : 2014년 9월 26일
(Accepted 26 September 2014)

정보과학회 컴퓨팅의 실제 논문지 제20권 제11호(2014. 11)

1. 서 론

음성 인식 기술은 사람이 일상생활 속에서 마우스나 키보드 등을 사용하지 않고 목소리를 통해 원하는 기기 및 정보 서비스의 이용을 제어할 수 있는 기술이다[1]. 1950년대부터 시작된 음성 인식 연구는 지난 60년 이상의 연구 개발 역사를 가지고 있으며, 꾸준한 개발로 지속적인 발전을 이루어왔다[2,3]. 그러나 2000년대 중반까지 낮은 음성 인식률로 대중화가 되지 못했다[1]. 아직 인간의 모든 언어 표현을 이해하는 음성인식 기술은 개발되지 못했지만, 비교적 정형화된 문장이나 일정 범위의 어휘로 한정될 경우 현존 기술로도 높은 정확도를 담보할 수 있다는 측면에서 상당한 수준의 기술적 진보를 달성하였다. 네트워크와 컴퓨팅 기술의 발달로 음성 인식률이 개선되었을 뿐만 아니라 정보기기가 소형화되고 이동성이 중요시되면서 음성으로 간편하게 제어할 수 있는 음성인식에 대한 수요는 더욱 증가할 것으로 전망되고 있다[1].

음성 인식 시스템에서 분석의 기본 단위로는 단어, 어절, 음소 등이 있으나, 음소 단위는 단어 및 음절 단위보다 그 종류가 작고 음향적인 특성을 인식기에 고르게 반영할 수 있기 때문에 많이 사용된다[4]. 그러나 발음의 변화 등 여러 가지 요인으로 인해 각각의 음소들을 정확하게 인식하는 것은 무척 어렵다[5]. 기존 연구에서는 단어, 어절, 음소 등으로 나누어 시스템을 개발하는 데에 복잡한 알고리즘과 신경망을 이용했다. 그 과정으로 얻은 학습 효과를 통해 음성 인식 연구들이 진행되었다. 또한, 현재 각 모바일, 태블릿 및 웹에서 사용되고 있는 서버 기반의 음성인식은 데이터베이스 센터에 저장되어 있는 방대한 디비를 사용해야 하므로, 인식할 때마다 인터넷 망에 연결되어 있어야 하는 번거로움이 있다.

따라서 본 논문에서는 음소를 기본 단위로 하는 음성 인식 시스템의 일부로 한국어 음소 중 모음 'ㅏ'를 인식하고자 한다. 모음 'ㅏ'의 여러 파형 시퀀스들을 관찰한 결과 특정한 패턴을 가지고 있음을 발견하였고, 이를 바탕으로 모음 'ㅏ'의 파형 패턴을 추출하는 연구를 수행하였다.

본 논문의 구성은 다음과 같다. 2장에서는 본 연구와 관련된 기존 연구에 대해 소개한다. 3장에서는 모음 'ㅏ' 파형의 특징들을 자세히 분석하고, 이를 인식하기 위한 휴리스틱을 제안한다. 4장에서는 본 논문에서 제안하는 모음 'ㅏ' 인식 휴리스틱의 성능을 실험하고 평가하며, 마지막으로 5장에서 결론을 맺는다.

2. 기존 연구

음성의 경우 동일한 단어라도 발성자의 습관이나 성별, 지속 시간의 차이, 그리고 조음 환경에 따른 조음 결합 현상 등의 영향으로 그 특성이 다르다. 이러한 특성 때문에 컴퓨터가 음성을 정확히 인식하는 것은 매우 어렵다. 현재까지 개발된 음성인식 시스템은 부분적으로 음성을 어느 정도 인식하는 수준에 머물러 있으며, 인식된 음성을 분석하는 영역만이라도 높은 성능을 나타내도록 하는 연구가 활발히 이루어지고 있다[6,7].

본 논문에서는 음성을 정확히 인식하는 것에 초점을 맞추었다. 즉, 음소 단위 기반의 고립 단어 음성을 인식 대상으로 한정시켜 인식의 정확도를 높이고자 한 것이다. 이를 위한 단계별 연구로써 고립단어를 대상으로 모음 'ㅏ' 인식을 하게 되었다. 모음 'ㅏ'의 공통 특징을 먼저 실험해봄으로써, 간단한 휴리스틱이 기존의 복잡한 알고리즘과 대비할 수 있는 지 알아보고자 하였다. 따라서 휴리스틱의 실험적 구현을 위해 한글 모음 중 가장 많이 사용되는 'ㅏ'만을 우선 실험 대상으로 하였다.

기존의 한국어 음성 인식 시스템에 대해 간략하게 설명하면 다음과 같다. 1980년대 이후 한국어 음성 인식을 위한 신경망 연구는 단어, 음절, 음소 등 각각의 특성에 맞게 활발하게 진행되었다. 기존 음성 인식 연구에는 HMM(hidden markov model)[8]과 TDNN(Time-Delay Neural Network)[9] 등이 있다. HMM은 음운, 단어와 같은 인식하고자 하는 음성의 단위를 통계적으로 모델화한 것으로, 1980년대 후반부터 성행한 음성 인식 기술이다. HMM은 기준 패턴이 되는 단어 중에서 테스트 패턴과의 유사도를 계산하여 그 중 가장 비슷한 단어로 인식하는 기본 철학으로 시작한다[10]. HMM으로 음소를 인식하는 과정은 다음과 같다. 먼저 미지의 입력 음성으로부터 전처리를 한 후, 입력 음성의 특징 파라미터를 추출한다. Viterbi 알고리즘을 이용하여 각 기준음소 HMM에서 입력 음성의 관측 확률을 계산하여, 가장 높은 관측확률을 나타내는 HMM의 음소로 인식하게 된다[11]. HMM은 정확도가 높지만, 미리 방대한 양의 학습을 필요로 하고, 확률 모델을 사용하여 인식 과정이 복잡하다.

TDNN은 Waibel에 의해 제안된 음운 식별용 신경망으로 화자종속 음운 인식에 있어서 매우 높은 인식률을 나타내는 시스템이다[12]. TDNN 구조를 이용한 한국어 음소 인식 시스템은 자음/모음 인식기, 자음그룹 인식기와 모음그룹 인식기, 각 음소 그룹에서의 음소인식기로 나누어져 있으며, 각각의 인식기는 TDNN으로 구성되어 있다. 연속 단어 화자 독립이나 실시간처리가 아니며 모음의 인식률이 80%정도로 낮은 편이다[10].

또한, 음소를 기본 단위로 하는 음성 인식[13,14]에는 음소 'ㅅ', 'ㅈ', 'ㅊ' 인식에 대한 연구가 있다. 이는 부호분포 변동성이라는 지표를 활용하여, 기존의 영교차율

ISSN 2383-6318(Print) / ISSN 2383-6326(Online)
KIISE Transactions on Computing Practices, Vol. 20, No. 12, pp. 711-716, 2014. 12
http://dx.doi.org/10.5626/KTCP.2014.20.12.711

집합 기반 POI 검색을 이용한 문장 유사도 측정 기법

(Sentence Similarity Measurement Method Using a Set-based POI Data Search)

고 은 별 † 이 종 우 ††
(EunByul Ko) (JongWoo Lee)

요 약 최근 논문 표절 논란과 지능형 텍스트 검색서비스에 대한 관심이 증가하면서 문장 유사도 측정의 필요성이 증가하고 있다. n-gram, 편집거리, LSA 등 기존의 다양한 방향으로 선행 연구가 있었지만 각 기법마다 장단점이 존재한다. 본 논문에서는 집합 기반 POI 검색 기법을 이용한 새로운 방향의 문장 유사도 측정 기법을 제안한다. 집합 기반 POI 검색 기법은 하드매칭에 비해 단어의 도치, 누락, 삽입, 변경에 현저한 성능 향상을 보인다. 이 기법을 이용하면 보다 정확하고 빠른 문장 유사도 측정이 가능하다. 제안하는 기법은 기존 집합 기반 POI 검색 기법의 데이터 로딩 알고리즘과 텍스트 검색 알고리즘을 변형하고 어절 연산 알고리즘을 추가하여 두 문장의 유사도를 백분율로 표현한다. 실험을 통해 본 논문에서 제시하는 기법이 정확도와 속도에서 n-gram과 기존 집합 기반 POI 검색 기법에 비해 우수함을 확인하였다.

키워드: POI 검색, 집합-기반 검색 알고리즘, 문장 유사도, 표절 검사, 텍스트 검색

Abstract With the gradual increase of interest in plagiarism and intelligent file content search, the demand for similarity measuring between two sentences is increasing. There is a lot of researches for sentence similarity measurement methods in various directions such as n-gram, edit-distance and LSA. However, these methods have their own advantages and disadvantages. In this paper, we propose a new sentence similarity measurement method approaching from another direction. The proposed method uses the set-based POI data search that improves search performance compared to the existing hard matching method when data includes the inverse, omission, insertion and revision of characters. Using this method, we are able to measure the similarity between two sentences more accurately and more quickly. We modified the data loading and text search algorithm of the set-based POI data search. We also added a word operation algorithm and a similarity measure between two sentences expressed as a percentage. From the experimental results, we observe that our sentence similarity measurement method shows better performance than n-gram and the set-based POI data search.

Keywords: POI search, Set-based algorithm, sentence similarity, piracy test, text search

· 이 논문은 2013년도 정부(교육부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임(2013R1A1A2013155)

† 학생회원 : 숙명여자대학교 멀티미디어과학과
iks1614@naver.com
†† 종신회원 : 숙명여자대학교 멀티미디어과학과 교수
(Sookmyung Womens Univ.)
bigrain@sm.ac.kr
(Corresponding author임)

논문접수 : 2014년 9월 30일
(Received 7 May 2014)
심사완료 : 2014년 10월 23일
(Accepted 20 June 2014)

정보과학회 컴퓨팅의 실제 논문지 제20권 제12호(2014. 12)

1. 서 론

최근 지적재산권에 대한 관심이 증가하면서 특허전쟁 및 논문 표절 논란이 중요한 이슈가 되고 있다[1]. 또한 운영체제 내의 지능형 파일 내용 검색에 대한 관심도 증가하고 있다[2]. 이런 사용자들의 요구를 충족하기 위해서는 기초적인 문장 유사도 측정 알고리즘 연구가 선행되어야 한다.

문장 유사도 측정 알고리즘은 표절 검사에 가장 많이 응용된다. 영어에 대한 연구는 많이 진행되었지만 한국어에 대한 연구는 미미한 실정이며 영어에 비해 어순이 자유로운 한글은 기존에 많이 연구되었던 알고리즘을 적용하는 데에 한계가 있다.

이에 한글에 맞는 문장 유사도 측정 알고리즘에 대한 연구가 진행되었다. 편집거리와 n-gram 기반 문장 유사도 측정법[3,4], LSA와 n-gram 기반 문장 유사도 측정법[5] 등의 다양한 연구가 그것이다. 하지만 여전히 기존 측정방법에는 단점이 존재한다.

본 논문에서는 기존의 접근 방향과는 다른 방향의 측정 방법을 제시한다. 하드매칭 기법에 비해 단어의 도치, 누락, 오타에 뛰어난 성능을 보였던 집합 기반 POI 검색 알고리즘[6]을 이용한 접근법이다. 집합 기반 검색 알고리즘을 변형하여 문장 유사도 측정에 응용하고 그 성능을 검증한다.

본 논문의 구성은 다음과 같다. 2장에서는 문장 유사도 측정 관련 연구와 집합 기반 POI 검색 알고리즘을 소개하고, 3장에서는 집합 기반 문장 유사도 측정 알고리즘을 설명한다. 4장에서는 제안한 알고리즘의 성능을 검증하고, 5장에서 결론을 맺는다.

2. 관련 연구

본 장에서는 기존에 연구되었던 문장 유사도 측정 방법과 집합 기반 POI 검색 기법에 대해서 간략히 설명한다.

2.1 편집거리와 n-gram 기반 유사도 측정

편집거리란 한 문자열이 다른 문자열로 변경되기 위해 필요한 삭제, 삽입, 치환 연산의 수를 말한다[3,4]. 예를 들어, ‘string’이라는 문자열이 ‘strong’이 되기 위해서 ‘i’가 ‘o’로 치환되어야 하므로 편집거리는 1이 된다. n-gram은 한 문자열을 n개 단위로 잘라 추출한 집합과 다른 문자열에서 추출한 집합의 교집합을 이용하여 유사도를 측정하는 방식이다. 예를 들어, ‘string’의 경우 tri-gram은 ‘str’, ‘tri’, ‘rin’, ‘ing’이며, ‘strong’은 ‘str’, ‘tro’, ‘ron’, ‘ong’ 이고, 이 둘의 교집합은 ‘str’이다.

편집거리와 n-gram을 이용한 방법은 문자열의 양이 많아질수록 많은 양의 저장공간과 연산이 필요하기 때문에 효율이 떨어지는 문제가 있다. 또 문맥을 전혀 고려하지 않기 때문에 다른 문장임에도 불구하고 같은 문장이라고 오인하기 쉽다.

이런 문제를 해결하기 위해 접두사 원소 선별을 이용하여 연산의 양을 줄이는 연구가 제안되었다[3]. 하지만 연산 속도는 현저히 향상되었지만 정확성은 입증되지 않았다는 문제점이 있다. 또 문맥가중치를 반영하여 조사, 어미같은 기능어를 제거하는 연구가 있었다. 문맥가중치를 반영한 측정법은 매개변수에 의해 검색 성능이 달라진다[4].

2.2 LSA와 n-gram 기반 유사도 측정

LSA(Latent Semantic Analysis)는 벡터 공간 모델의 단점을 극복하기 위해 제안된 모델이다[5]. 벡터 공간 모델 방식은 정보 검색 모델의 한 종류로써, 각 문장의 색인어를 벡터로 표현하여 문장간의 유사성을 측정하는 방법이다. 벡터 공간 모델 방식은 색인어로 추출되는 키워드가 정확히 일치해야한다는 문제가 있는데 이런 문제를 해결하기 위해 제안된 모델이 LSA이다.

LSA 모델은 의미적 유사성을 탐지하는데 유용하고, n-gram은 어순같은 형태적 변경을 탐지하는데 유용하다. 이 두 가지 모델을 조합하여 보다 성능이 뛰어난 모델이 제안되었다. 이 모델은 원문복사, 단어치환, 어순변경, 문장요약 4가지 유형의 표절에 잘 작동한다.

2.3 집합 기반 POI 검색 기법

집합 기반 POI 검색 기법[6]이란 집합 개념을 적용하여 POI 데이터를 검색하는 기법으로, 계산이 복잡하고 방대한 양의 데이터를 필요로 하는 기법을 사용하기 힘든 차량 네비게이션과 같은 독립형 시스템을 위한 알고리즘이다. 부정확한 POI 질의어 입력으로 인한 POI 검색 서비스 성능 저하 문제를 시스템 자체 내에서 해결하기 위해 제시한 새로운 알고리즘이다.

집합 기반 POI 검색 기법은 데이터 로딩 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.

그림 1은 세 개의 레코드에 대해 기존 집합 기반 POI 검색 기법에서 데이터 로딩부터 글자 아이디와 역 인덱스를 생성하는 과정 예를 표현한 것이다. 역 인덱스란 해당 글자가 어느 위치에 있는지 나타내는 자료구조를 의미하며, 그 글자를 포함하고 있는 데이터의 줄 번호를 저장한다.

이렇게 생성된 역 인덱스는 텍스트 검색 알고리즘에서 이용된다. 텍스트 검색을 할 때, 입력 받은 n개의 문자로 이루어진 질의어를 m개의 블록으로 균등 분할하여 각 블록 내 집합 연산을 수행하고 이 결과를 이용하여 다시 블록 간 연산을 수행한다. 이때 ‘차수’라는 개념을 사용하는데, 차수란 주어진 레코드에 포함된 질의어

Journal of Korea Multimedia Society Vol. 17, No. 3, March 2014(pp. 385-392)
http://dx.doi.org/10.9717/kmms.2014.17.3.385

글자가 움직이는 동영상 자막 편집 어플리케이션 개발

하예영[†], 김소연[††], 박인선[†††], 임순범[††††]

요 약

본 논문에서는 사용자가 스마트폰을 이용하여 직접 찍은 동영상에 움직이는 자막을 즉석에서 간편하게 편집할 수 있는 안드로이드 어플리케이션인 VIVID를 개발하였다. VIVID를 통해 휴대폰에 저장된 동영상에 원하는 자막의 시간, 텍스트, 위치, 모션(자막의 움직임)을 간편하게 설정할 수 있으며, 결과물을 html 형식으로 웹 서버에 업로드하여 다른 사람들과 함께 공유할 수 있게 하였다.

Development of Video Caption Editor with Kinetic Typography

Yea-Young Ha[†], So-Yeon Kim[††], In-Sun Park[†††], Soon-Bum Lim[††††]

ABSTRACT

In this paper, we developed an Android application named VIVID where users can edit the moving captions easily on smartphone videos. This makes it convenient to set the time range, text, location and motion of caption text on the video. The editing result is uploaded to web server in html and can be shared with other users.

Key words: Android Application(안드로이드 어플리케이션), Video Caption Editor(동영상 자막 편집), Kinetic Typography(키네틱 타이포그래피)

1. 서 론

요즘 스마트폰을 사용하다보면 SNS나 메신저 어플리케이션을 통해서 실시간으로 촬영한 재미있는 동영상을 공유하는 모습을 쉽게 볼 수 있다. 이처럼 스마트폰의 고성능화로 인하여 스마트폰 카메라로 촬영하는 비중이 높게 나타남에 따라 다양한 유형의 모바일 동영상 편집 어플리케이션들이 개발되어 활용되고 있다[1,2]. 기존의 동영상 편집 어플리케이션들은 주로 구간 자르기, 화면 전환 효과, 타이틀 자막 삽입 기능을 제공하고 있다. 그러나 모바일 환경에서 여러 자막을 편집하고 그 자막에 다양한 움직임을 줄 수 있는 키네틱 타이포그래피를 활용한 동영상 편집 어플리케이션은 찾아보기 힘들다.

잘 만들어진 동영상들을 보면 알 수 있듯이 화면에 어울리는 자막은 그 전달 효과가 매우 크다. 최근 예능 프로그램에서 종종 등장하고 있는 움직이는 자막들은 재미를 극대화하는 효과를 발휘하고 있다[3]. 동영상에 자막을 편집하려면 PC환경에서 정교한 자막 편집이 가능한 전문 동영상 편집 혹은 애니메이션

※ 교신저자(Corresponding Author) : 임순범, 주소 : 서울시 용산구 청파동 53-12 숙명여자대학교(140-742), 전화 : 010-5329-9919, FAX : , E-mail : sblim@sm.ac.kr
접수일 : 2013년 9월 12일, 수정일 : 2014년 1월 20일
완료일 : 2014년 2월 4일
[†] 준회원, 숙명여자대학교 멀티미디어과학과
(E-mail : yyh0620@gmail.com)
[††] 준회원, 숙명여자대학교 멀티미디어과학과
(E-mail : ksy1224@naver.com)
[†††] 준회원, 숙명여자대학교 멀티미디어과학과
(E-mail : bestellie@hanmail.net)
[††††] 종신회원, 숙명여자대학교 멀티미디어과학과
※ 본 논문은 지식경제부 산업원천기술개발사업의 과제번호 10041794 과제의 연구결과로 수행되었음.
※ 본 논문은 숙명여자대학교의 2013년도 교내 연구비 지원에 의해 수행되었음.

편집 프로그램을 사용하고 있지만 모바일 환경에서는 바로 찍은 동영상에 여러 자막을 간편하고 짧은 시간 안에 편집해야 한다. 더구나 움직이는 자막을 효과적으로 편집할 수 있도록 키네틱 타이포그래피 기술을 적용하는 모바일 어플리케이션은 아직 시작 단계이다[4].

따라서 본 논문에서는 프리미어나 애프터 이펙트 같은 전문 PC용 프로그램을 이용하지 않고 스마트폰에서 즉석으로 동영상에 어울리는 움직이는 자막을 편집할 수 있는 어플리케이션을 구현하였다.

2. 관련 연구

2.1 키네틱 타이포그래피 기술동향

키네틱 타이포그래피의 기술연구는 MIT와 CMU 등에서 활발하게 진행되고 있다. 키네틱 타이포그래피의 실행을 위한 엔진 구현과 기반으로 한 타이포그래피의 활용을 위한 제작 기술 등에 대해 주로 연구하고 있다. 이러한 연구는 키네틱 타이포그래피가 다양한 응용 서비스 시스템에서 인간의 표현력이 보다 풍부해질 수 있음을 보여주고 있다[5].

본격적으로 키네틱 타이포그래피를 제작하려면 Adobe사의 Flash와 After Effect, Apple사의 Motion 등 전문적인 소프트웨어 도구를 사용하고 있다. 그러나 이러한 도구들은 데스크탑 PC에서 편리한 입력장치를 요구하며, 일반 사용자들에게는 사용 방법이 다소 어렵고 전문적이라는 단점이 있다[4].

2.2 동영상 편집 및 자막 편집 어플리케이션 분석

'AndroVid'[6]와 'Magisto Video Editor & Maker'[7]는 대표적인 동영상 편집 어플리케이션으로 기능적인 측면에서 서로 다른 성격을 갖고 있다. 'AndroVid'는 다양한 포맷의 동영상에 대하여 자르기, 회전, 변환, 효과, 이름 바꾸기, 공유 등의 다양하고 세부적인 편집 기능을 제공하는 반면, 'Magisto Video Editor & Maker'는 편집할 영상, 테마, 배경음악, 영상 제목을 정해주면 어플리케이션이 알아서 자동으로 편집한 뒤 사용자에게 결과물을 만들어준다.

'헤헤케케 패러디'[8]와 '이피디'[9]라는 자막 편집 어플리케이션의 경우, 직접 찍은 사진을 각종 예능 방송의 자막 테마로 꾸밀 수 있는 기능을 제공하고 있다. '헤헤케케 패러디'는 정해진 틀 안에서 텍스트를 입력하면 자동으로 자막을 합성해주는 방식이고, '이피디'는 원하는 자막 이미지를 사진 위로 직접 위치와 크기를 지정해줄 수 있도록 되어 있다. 널리 쓰이고 있는 사진 편집 어플리케이션 중 '라인 카메라'[10]는 다양한 스타일의 스탬프와 프레임, 세련되고 귀여운 필터를 제공한다. 사진에 메시지를 넣고 싶을 때에는 텍스트 내용을 입력하고 글꼴, 글자색을 설정한 뒤 텍스트가 화면에 나타나면 드래깅으로 원하는 위치에 움직일 수 있으며 크기를 자유자재로 조절할 수 있다.

2.3 html5의 〈canvas〉와 〈video〉요소

html5는 html의 차기 제안 버전이다. <audio>, <video>, <canvas> 등의 요소를 추가하여 플러그인 없이 최신 멀티미디어 콘텐츠를 브라우저에서 쉽고 용이하게 볼 수 있게 한다[11].

<canvas> 요소는 JavaScript를 이용하여 그래픽을 그릴 때 사용된다. 캔버스는 html 페이지 위에 직사각형의 프레임을 제공하며 JavaScript를 이용하여 실제 그래픽을 그린다. 캔버스는 도형, 문자를 그리거나 이미지를 삽입하는데 필요한 함수들을 API로 제공하고 있으며 기존의 플래시가 갖고 있던 강점인 화려한 애니메이션과 인터랙티브한 효과를 html 문서 내에서 표현할 수 있다.

<video> 태그를 사용하면 html 페이지에서 동영상을 임베드(embed)할 수 있다. html5에서는 <video> 요소에서 다양한 함수, 이벤트 및 프로퍼티 등을 제공하며 자바 스크립트로 동영상을 컨트롤할 수 있다.

2.4 KineticJS

KineticJS는 데스크탑이나 모바일 어플리케이션에서 고성능의 애니메이션, 필터링, 이벤트 처리 등을 편리하게 할 수 있도록 하는 html5 Canvas JavaScript framework이다[121]. 캔버스 API를 확장한 오픈소스 라이브러리로서 그래픽과 상호작용하는데에 미흡한 점이 있는 html5 캔버스를 보완한다. stage 위에 그래픽을 그릴 수 있으며, 이벤트 리스너를 추가하여 각 그래픽의 이벤트를 독립적으로 수행할 수 있다. stage 위에 추가된 각 layer는 shape 또는 shape 그룹을 포함하고 있으며 stage, layer, group,

문자 계층 구조를 활용한 키네틱 타이포그래피 모션 전달 기법

(Kinetic Typography Motion Propagation Technique based on Text Hierarchy)

김 지 성 † 김 민 우 † 임 순 범 †† 최 윤 철 †††
(Jisung Kim) (Minwoo Kim) (Soon-bum Lim) (Yoon-chul Choy)

요 약 본 논문에서는 문자의 계층구조를 활용한 키네틱 타이포그래피 모션 전달 기법을 제안한다. 제안하는 기법을 활용하면 사용자는 한 번의 모션 입력으로 동일한 계층에 있고 선택한 문자 뒤에 위치한 모든 문자에 사용자가 입력한 모션을 전달할 수 있다. 제안하는 기법과 사용자 인터페이스는 터치 인터페이스를 기반으로 하는데, 이것은 기존에 키네틱 타이포그래피 저작 과정에서 일정한 시간 간격으로 움직이되 동일한 키네틱 타이포그래피 모션을 갖는 여러 문자에 모션을 입력할 때 발생하는 비효율적인 반복 작업을 피할 수 있게 한다. 또한, 논문에서 제안하는 문자 선택 기법은 사용자가 문자를 조작하기 전에 작업 공간에 존재하는 문자를 직관적으로 선택할 수 있도록 한다. 이 기법을 활용하면 사용자는 간단한 조작만으로 조작하고자 하는 문자를 계층이나 그 수와 관계없이 선택할 수 있다.

키워드: 키네틱 타이포그래피, 모션 전달, 문자 선택, 터치 기반 인터페이스

Abstract In this paper, we propose a technique that utilizes text hierarchy for propagating kinetic typography motions to the following texts with the same class in the hierarchy as the preceding text. The proposed technique and interface is based on a touch interface and allows users to avoid repetitive work that might happen when applying identical kinetic typography motions to the following texts with a certain delay. In addition, we also present a intuitive text selecting technique to facilitate users to easily choose the desired unit(passage, word and character) for controlling the text objects regardless of text hierarchy or the number of them.

Keywords: kinetic typography, motion propagation, text selection, touch-based interface

· 본 연구는 산업통상자원부 및 한국산업기술평가관리원의 산업융합원천기술개발사업(정보통신)의 일환으로 수행하였음. [10041794, 소셜미디어 환경에서 감성전달을 위한 키네틱 타이포그래피 기술 개발]

† 학생회원 : 연세대학교 컴퓨터과학과
kimjs3318@gmail.com
kimmw2010@naver.com
†† 종신회원 : 숙명여자대학교 멀티미디어학과 교수
sblim@sookmyung.ac.kr
††† 종신회원 : 연세대학교 컴퓨터과학과 교수
ycchoy@rainbow.yonsei.ac.kr
(Corresponding author임)

논문접수 : 2014년 4월 17일
(Received 17 April 2014)
논문수정 : 2014년 5월 7일
(Revised 7 May 2014)
심사완료 : 2014년 5월 8일
(Accepted 8 May 2014)

정보과학회논문지: 컴퓨팅의 실제 및 레터 제20권 제8호(2014.8)

1. 서 론

최근 스마트 폰, 태블릿 PC 등 다양한 모바일 기기들이 보편화됨에 따라 터치 디바이스와 그 인터페이스는 우리에게 친숙한 환경으로 자리 잡았다. 이와 함께 소프트웨어의 형태도 모바일 환경의 특징인 실시간성을 활용한 다양한 어플리케이션이 출시되는 등 많은 변화가 일어나고 있다. 특히 단순한 소통 수단인 기존의 SMS를 넘어 때와 장소를 가리지 않고 개인의 상태와 감정을 나타내는 소셜 네트워크 서비스 그리고 사진, 문자, 동영상 등 다양한 컨텐츠를 주고받을 수 있는 모바일 메신저에 관련된 소프트웨어가 특히 주목을 받으며 성장하고 있다.

이러한 소프트웨어에서 사용자들은 기본적으로 문자로 의사소통하며 자신의 감정을 나타내지만, 정적인 문자로 감정을 나타내는 데는 한계가 있기 때문에 감정을 표현하는 보조 수단으로 이모티콘 등이 활용되고 있다. 움직이는 문자를 통해 감정이나 의사를 전달하는 방법으로 키네틱 타이포그래피가 있지만, 키네틱 타이포그래피 컨텐츠를 초보자들이 스스로 제작하기는 쉽지 않다. 기존의 범용 저작 도구인 Adobe After Effect와 Flash[1]를 이용해서 제작하면 다양한 효과와 함께 정교한 결과물을 만들어낼 수 있지만, 기존 도구의 복잡한 인터페이스는 초보자들이 일상적으로 쉽게 사용할 수 있는 간단한 방식이 아니며 오히려 인터페이스를 배우는데 더 많은 시간이 걸리기 때문에 복잡한 효과를 필요로 하지 않는 결과물을 저작하는 데에 있어서는 적합한 시스템이라 할 수 없다. 초보자가 얻고자 하는 결과물은 대부분 간단하거나 같은 움직임이 반복되는 모션이다. 예를 들어 여러 문자가 같은 동작으로 움직이고자 하는 경우, 기존의 방법으로는 문자를 선택하고 모션을 입력하는 동일한 작업을 여러 번 반복해야 하며 시간에 따라 모션이 나타나거나 사라지게 하기 위해 개별적으로 타이밍을 조절해야 한다. 이것은 간단한 작업임에도 같은 작업을 반복해야 하는 비효율적인 면과 문자가 많아질수록 더 많은 시간이 요구된다는 문제가 있다.

본 논문에서는 기존에 선행된 연구를 토대로 문자에 반복적으로 모션을 입력하는 기존 방식의 비효율성을 개선하기 위한 모션 전달 기법을 제안한다. 사용자는 이 기법을 통해 한 번의 모션 입력을 통해 나머지 문자에 동일한 모션을 입력할 수 있으며 타임라인을 활용하여 모션이 실행되는 시점을 자유롭게 조작할 수 있다. 또한, 문자 선택을 위해 제안한 인터페이스는 터치 디스플레이가 가지는 시각적 표현의 한계를 해결하여 보다 직관적이고 편리한 사용자 환경을 지원할 수 있는 특징을 가지고 있다.

본 논문의 2장에서는 키네틱 타이포그래피에 관련된 기존의 연구를 살펴본다. 이어 3장에서는 본 논문에서 활용하고 있는 문자의 계층구조에 대해 설명한다. 그리고 4장에서는 모션 입력을 위해 문자의 조작 단위를 변경하고 문자를 선택하는 제스처에 대해서 서술하고, 5장에서는 모션 전달 기법을 지원하는 키네틱 엔진과 동작 과정, 그리고 타임라인 조작 인터페이스에 대해 설명한다. 이어 6장에서 사용자 평가에 따른 결과를 나타낸 후 7장에서 결론 및 향후연구를 언급하며 끝을 맺는다.

2. 관련 연구

2.1 키네틱 타이포그래피 엔진

J. C. Lee[2]가 제안한 키네틱 타이포그래피 엔진은 일반적으로 애니메이션 제작에 사용되는 기본적인 모션인 이동, 크기, 회전 효과 등을 지원하고 텍스트의 색상, 밝기 등 다양한 속성 변경 기능을 제공한다. 하지만 인터페이스가 제공되지 않고 엔진만 존재하기 때문에 실제 제작 도구로 활용하기 위해서는 인터페이스를 직접 개발해야 한다는 문제가 있다.

Active Text[3]에서는 기존의 1차원 배열 방식으로 문자를 저장하고 나타낼 때 발생하는 문제를 문자의 계층 구조를 활용하여 해결하고자 했다. 기존에 문자를 스트림 형태로 관리하는 방법은 정적인 표현에서는 효과적이지만, 여러 문자를 실시간으로 처리할 때 한계가 있었기 때문에 실시간으로 움직이는 인터랙티브한 문자를 생성하고 처리하기 위해 계층구조를 통해 의미 있는 단위로 문자를 나누어 관리하는 방법을 제안했다. 하지만 이 시스템 역시 키네틱 타이포그래피 제작을 위해 필요한 간단한 기능만 제공했기 때문에 실제로 사용하는 데는 한계가 있다.

2.2 키네틱 타이포그래피 제작 인터페이스

키네틱 타이포그래피 제작에 특화된 처리 엔진에 관련된 연구과 함께 키네틱 타이포그래피 제작을 위한 인터페이스 역시 함께 연구되었다. 먼저 The Kinedit System[4]은 GUI 방식을 적용한 키네틱 타이포그래픽 저작 도구로 그림 1(a)와 같이 에디터 형식의 인터페이스를 제공한다. 사용자는 하단의 입력창에 텍스트를 입력하고 원하는 문자를 선택한 후 키네틱 효과를 지정할 수 있다. 이 방법은 사용자에게 친숙한 방식이지만 미리 정의된 효과만 적용할 수 있다는 점에서 사용자가 원하는 컨텐츠를 제작하는 데는 제약이 있다.

Zhinquan Yeo[5]는 [2]의 엔진을 활용하여 인스턴스 메시지에 정의된 감정을 적용하여 움직이는 문자를 표현할 수 있는 인터페이스 환경을 제공하였다. 하지만 정의된 9개의 감정을 나타내는 모션만을 적용할 수 있기 때문에 사용자가 원하는 모션을 직접 입력할 방법이 없다는 것이 단점이다.

Journal of Korea Multimedia Society Vol. 17, No. 8, August 2014(pp. 1025-1032)
http://dx.doi.org/10.9717/kmms.2014.17.8.1025

디지털 음성 도서에서 MathML 수식의 수준별 독음 변환 기법

황정수[†], 임순범[††]

A Study on Phased Reading Techniques of Mathematical Expression in the Digital Talking Book

Jungsoo Hwang[†], Soon-Bum Lim[††]

ABSTRACT

Until now, there were few supports on reading the mathematical expressions except text based expressions, so it is important to provide the reading of the mathematical expressions. Also, there are various of obstacles for people who are not visually impaired when reading the mathematical expressions such as the situation of presbyopia, reading the mathematical expressions in the vehicles, and so on. Therefore, supports for people to read mathematical expressions in various situations are needed. In the previous research, the main goal was to transform the mathematical expressions into Korean text based on Content MathML. In this paper, we expanded the range of the research from a reading disabilities to people who are not reading disabilities. We tested appropriacy of the rules we made to convert the MathML based expressions into speech and defined 3 math-to-speech rules in korean based on levels. We implemented the mathematical expressions by using 3 math-to-speech rules. We took comprehension test to find out whether our math to speech rules are well-defined or not.

Key words: Content MathML, Math to Speech

1. 서 론

최근 독서 환경 개선을 위한 다양한 독서 장애인용 디지털 음성 도서 기술이 개발되면서 장애인용 전자책 시장이 활성화되어가고 있다. 특히 DAISY 형식으로 작성된 XML 기반의 전자책은 많은 독서 장애인들이 사용하고 있다.

현재까지는 텍스트를 제외한 특수 콘텐츠, 특히 수학 기호의 독음은 아직까지 미흡하여 수식 표현을 제공하는 것이 필요하다. 또한 시각 장애인뿐만 아니라 일반인이 여러 단말기를 통하여 웹을 접근할 때 나타날 수 있는 여러 장애가 존재한다. 예를 들어, 한 손으로 모바일 기기를 사용하는 상황, 버스에 앉아서 기기를 사용하는 상황, 노안으로 인해 기기를 정상적으로 볼 수 없는 경우 등이 있다. 그러므로 일반인이 사용하기에도 적합한 수식 독음을 위해 연구의 대상을 확장하는 것이 필요하다. 그러므로 다양한 환경 및 상황에 따른 대상자 확장을 통해 전자책 내

※ Corresponding Author : Soon-Bum Lim, Address: (140-742) Sookmyung Women's Univ. Cheongpa-dong 2-ga, Yongsan-gu, Seoul, Korea, TEL : +82-2-710-9424, E-mail : sblim@sookmyung.ac.kr
Receipt date : Jun. 11, 2014, Revision date : Jul. 2, 2014
Approval date : Jul. 16, 2014
[†] Dept. of Multimedia Science, Sookmyung Women's University (E-mail : hjs6344@naver.com)
[††] Dept. of Multimedia Science, Sookmyung Women's University
※ This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(20100024154).
※ This research was supported by the Sookmyung Women's University Research Grants(1-1303-0209).

에 표현된 MathML 수식 콘텐츠의 독음 연구가 필요하다.

본 연구에서는 독서 장애인 이외에도 일반인, 학생 등으로 연구 대상을 확대하고자 한다. 중학교 및 공통 수학 교과 과정에 포함된 수식을 대상으로 Content MathML로 표현된 수준별 독음 규칙을 정리하고, 이에 따른 독음 변환 기법을 연구한다.

2. 관련 연구

2.1 DAISY

DAISY(Digital Accessible Information SYstem)는 독서장애인을 위한 디지털 국제 표준(ANSI/NISO Z.39.86-2005) 문서 형식으로 현재 국내 여러 기관을 통해서 제작되고 있다[1]. 초기 디지털 음성도서는 시각 장애인과 독서장애인을 위해 낭독한 음성파일을 디지털화하여 CD 등에 저장한 도서 등이 주류였다. 하지만 단순히 음성을 디지털화하는 것뿐만 아니라 디지털화한 음성의 녹음과 함께 추가적인 모든 정보를 제공하는 디지털 음성정보 시스템(Digital Audio-based Information System, DAISY)으로 즉, 보존, 배포, 이용에 이르기까지 하드웨어와 소프트웨어가 한 세트로 된 것을 디지털 음성도서라 확대 정의하였다. 또한 정보 접근의 중요성에 입각하여 Digital Access Information System(DAISY)으로 추후 개정되었다. 따라서 DAISY는 '국제 표준의 독서장애인용 멀티미디어 콘텐츠(파일) 포맷'이다[2].

DAISY는 텍스트북과 오디오북의 혼합 형태이며, DAISY 표준의 목적은 독서 장애와 상관없이 정보와 지식에 동등하게 접근하도록 하는 것이다. 1998년 DAISY 2.0이 발표되었고, 2005년에 ANSI/NISO Z.39.86-2005 표준으로 DAISY 3이 발표되었다[3].

2.2 수식 마크업 MathML

HTML과 이미지 파일을 이용한 기존 웹 상에서의 수식은 표현과 인코딩 면에서 문제점이 있다. 표현의 문제는 텍스트의 크기가 커지거나 작아지면 수식을 표현하는 이미지와 텍스트 간의 균형이 깨어진다는 것이 하나의 예가 될 수 있다. 인코딩의 문제는 수식 표현을 검색하고자 할 때 사용할 수 있는 유용한 도구가 없고, 특정 표현을 잘라 붙이는 것과 같은 과정이 매우 복잡하고 어렵다는 것들이다.

MathML은 이러한 웹에서의 수학적 표현의 한계를 극복하기 위해 만들어진 XML의 응용분야이다. 이는 컴퓨터 통신에 있어 수학적인 내용이나 표기를 다루는 방법에 대한 저수준 포맷으로 웹에서 수식을 표현하고, 처리, 공유하는 데에 필요한 많은 해결책을 제공하는 마크업 언어이다[4]. W3C에서는 2001년 2월 MathML 2.0의 표준안을 발표했고, 2010년 10월에 표준안으로 발표된 MathML 3.0은 MathML 2.0의 개정판이다. W3C는 2014년 4월에 MathML3 2nd Edition의 표준안을 발표하였다.

수식을 전달하기 위해서 사용할 수 있는 방법은 수식의 모양을 보이는 대로 설명하는 방법과 수식의 의미를 전달하는 두 가지 방법이 있을 수 있다. MathML에서는 두 가지 방법을 모두 사용하고 있다. 전자의 방식을 채택하여 수식의 표현에 중점을 둔 표현 MathML과 후자의 방식을 채택하여 수식의 의미 전달에 초점을 둔 내용 MathML이 있다. 사용자는 상황이나 목적에 맞게 두 가지 방법 중 하나를 선택하여 사용하거나 두 가지 방법을 적절히 섞어 사용할 수 있다.

2.1.1 표현 MathML

표현 MathML은 수학의 레이아웃 표현에 중점을 둔 마크업이다[5]. 표현 MathML은 레이아웃 박스 개념을 기초로 만들어졌고, 약 30개 정도의 요소와 50개 정도의 속성을 가지고 있다. 요소들은 수식에서 레이아웃 박스들을 기술하기 위한 것들로 구성되어 있고, 각 요소들은 하위 요소들이 어떻게 논리적으로 연결되는지를 설명해주는 레이아웃 스키마를 가진다.

수식은 토큰 요소들을 기본으로 구성되며 이들은 레이아웃 스키마들과 조합하여 하나의 수식을 완성하게 된다. 토큰 요소들은 직접적으로 문자 데이터를 포함할 수 있으며 식별자, 연산자, 그리고 숫자들이 각각의 요소들에 둘러싸여 나타나야 한다.

2.1.2 내용 MathML

내용 MathML의 목적은 식의 내부 수학적 의미에 중점을 두고 명확한 인코딩이 가능하게 하는 마크업이다[6]. 따라서 내용 MathML로 표현된 수식은 표현 MathML로 표현된 수식보다 모호성이 적고, 기계

Spatial Hypertext Modeling for Dynamic Contents Authoring System based on Transclusion

Ja-Ryoung Choi
Dept. of Multimedia Science
Sookmyung Women's University
Seoul, Korea
j2arlove@gmail.com

Sungeun An
Dept. of Multimedia Science
Sookmyung Women's University
Seoul, Korea
imsung110@gmail.com

Soon-Bum Lim
Dept. of Multimedia Science
Sookmyung Women's University
Seoul, Korea
sblim@sookmyung.ac.kr

ABSTRACT
This paper proposed a web content collecting model to reuse a variety of web contents based on Transclusion. Transclusion is a model for collecting existing web contents and including them into a new document. However, Transclusion lacks consideration of copyright issues and dynamic changes. Therefore, we classified Transclusions into three different types based on copyright restrictions: Trans-quotation, Trans-reference and Trans-annotation. Then we represented Transclusions in each different type of spatial hypertext model. Also, we designed RVS(ReVerse Syndication) model in order to trace the dynamic changes.

Categories and Subject Descriptors
H.5.4 [**Information Interfaces and Presentation**]: Hypertext/Hypermedia – *Navigation, User issues*

General Terms
Design; Human Factors; Theory.

Keywords
Transclusion; Spatial hypertext; Dynamic contents.

1. INTRODUCTION
Web technologies have evolved to give users the ability to produce and consume web content. For example, when making textbook or lecture notes, authors access to various web contents and collect the pieces they want to reuse and paste it in their documents. But as web is open to everyone, illegal content reproduction issues arise. Moreover, web contents are subject to change, and they cause dynamic changes such as content loss, update and modification made by various authors. In order to solve such issues, we proposed a web content collecting model. In this paper, to reuse various web contents, we classified Transclusions into three different types based on copyright restrictions. All of these are effectively managed by strong connection with the original source, so as to adapt to dynamic changes of web contents.

HT'14, September 1–4, 2014, Santiago, Chile.
ACM 978-1-4503-2954-5/14/09.
http://dx.doi.org/10.1145/2631775.2631780

2. RELATED WORK
Project Xanadu is an existing technique for collecting and connecting documents and files on web[1]. Based on unbreakable links for connecting to original sources, Project Xanadu is designed to provide a view of two documents on one screen, with full support on versioning. Transclusion proposed by Project Xanadu is a technique for the inclusion of existing content into a new document. Contents authors actually do not make a copy of an existing document, but make a reference to the virtually stored original content.

TED(Transclusions in Enterprice Documents) system is one of such examples. This system proposed multiple classes of XML-based transclusions[2,3]. As the original content changes, the referenced content on target document is revised accordingly.

WARP system is another study of Transclusions. WARP is a web-based dynamic spatial hypertext that runs in a web browser. WARP has the ability to transclude other spatial hypertexts as collections[4]. Project Breadcrumbs is another study similar to WARP[5].

3. SPATIAL HYPERTEXT MODEL FOR TRANSCLUSION
In this system, we classified Transclusions into three different types based on copyright restrictions: Trans-quotation, Trans-reference and Trans-annotation. These restrictions are concerned about the rights on source reproduction and distribution (see Figure 1).

- **Trans-quotation**

  In cases where both source reproduction and distribution are allowed, Trans-quotation includes a direct quote from its original source.

- **Trans-reference**

  In cases where source distribution is allowed but not source reproduction, Trans-reference includes a quote paraphrased by user from its original source.

- **Trans-annotation**

  In cases where neither source reproduction nor distribution is allowed, Trans-annotation separates content from the original source. Since direct quotation is not allowed for this type, users can access the original contents through the link address instead.

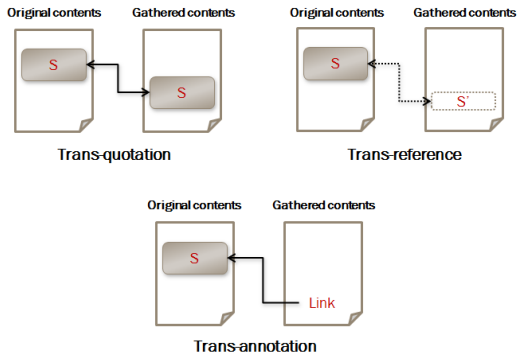


Figure 1. Classification of Transclusions

Each Transclusion is expressed in a different form of spatial hypertext model. Trans-quotation is displayed as it is because directly quoted. If users want to see the original content web page, they can click on the transcluded region. For Trans-reference, when users place the cursor over the transcluded region, the original web document pops up; so as to see where the quote came from and what it originally states. Since copyright restrictions are quite strict for Trans-annotation, its transcluded region contains only URL address that takes users to the original source web page.

4. DYNAMIC CONTENTS SYNDICATION MODEL

The web content collecting model is based on Transclusion to reuse web contents. Metadata for tracking information is designed to manage this system. Tracking information contains necessary information to maintain connection to its original content, and the concept of this metadata is opposite concept of RSS.

As Figure 2 shows, existing RSS is created by the content provider, whereas RVS(Reverse Syndication) is created by the content collector. Using this RVS model, the content collector is able to trace dynamic changes such as content loss or update.

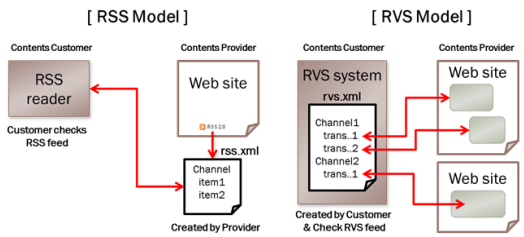


Figure 2. RSS Model vs. RVS Model

Figure 3 describes the RVS metadata structure in RELAX NG format. Since RVS is an XML document, we described the structure of RVS document using RELAX NG; which is a schema language for XML. It has <rvs> as a root node, <document> and <channel> as child nodes. <channel> node records information of web contents collected by users, and has child nodes as follows: <title>, <link>, <description>, <copyright>, <pubDate>, <category>, <managingEditor>, <language>, <ttl>, <skipDays> and <transclusion>. In order to trace the original source, <transclusion> node records portions of the content that have been referenced by users. Such RVS file enables continuous connection with the original source, enabling us to check for updates.

```
element rvs {
  element document {
    element title {text}
    element link {text}
    element description {text}
  }
  element channel {
    element title {text}
    element link {text}
    element description {text}
    element copyright {text}
    element pubDate {text}
    element category {text}*
    element managingEditor {text}*
    element language {text}
    element ttl {text}
    element skipDays {text}

    element transclusion {
      element title {text}
      element link {text}
      element description {text}
      element pubDate {text}
      element right {text}
      element author {text}*
      element category {text}*
      element extraction {
        attribute start {text}
        attribute end {text}
      }
      element relation {
        attribute type {
          "quotation" | "reference" | "annotation"}+
        }
      }
    }+ // the end of transclusion
  }+ // the end of channel
} // the end of rvs
```

Figure 3. Metadata structure for Dynamic Contents Syndication

5. CONCLUSION

In this paper, we designed a system that collects various web contents and displays them as spatial hypertext. We analyzed patterns of web content, and classified Transclusions into Trans-quotation, Trans-reference and Trans-annotation. These Transclusions are displayed as transcluded regions. Also, we defined RVS model for metadata to adapt to dynamic changes. This model can show whether there is any change in the original document.

The system will be evaluated in terms of the feasibility of e-Book. Concluding from this evaluation, the paper will explore the future work necessary to enhance the model that automatically handles problem with dynamic changes.

6. ACKNOWLEDGMENTS

This research is supported by Ministry of Culture, Sports and Tourism(MCST) and Korea Creative Content Agency(KOCCA) in the Culture Technology(CT) Research & Development Program 2014.

7. REFERENCES

[1] T.H. Nelson. The new XANADU structure for the web[Online], http://xanadu.com/nxu, 2001.

[2] A.D. Iorio, J. Lumley. From XML inclusions to XML transclusions. In Proceedings of the 20th ACM conference on Hypertext and hypermedia (pp. 147-156), 2009.

[3] A.D. Iorio, et al. Towards XML Transclusions. In Proceedings of the First Workshop on New Forms of Xanalogical Storage and Functions, ACM Hypertext Conference (pp. 23-28), 2009.

[4] L.Francisco-Revilla, F. M.Shipman, WARP: a web-based dynamic spatial hypertext. In Proceedings of the 15th ACM conference on Hypertext and hypermedia (pp. 235-236), 2004.

[5] L.Francisco-Revilla, A.Figueira, Adaptive spatial hypermedia in computational journalism. In Proceedings of the 23rd ACM conference on Hypertext and social media (pp. 313-314), 2012.

A study on 3D Representation of Declarative Content for Web Platform based Hybrid TV

Hyun-Jeong Yim[1], Hee-Jin Lee[2], Soon-Bum Lim[2], *Member, IEEE,* Byungjun Bae[1], Heung Mook Kim[1], *Member, IEEE*, and Namho Hur[1], *Member, IEEE*

[1]Broadcasting System Research Group, ETRI, Republic of Korea
[2]Dept. of Multimedia, Sookmyung Women's University, Republic of Korea

***Abstract*— This paper proposes a 3D representation method that can be used in web platform based hybrid 3DTV. In order to provide declarative content for stereoscopic TV service, we studied content separation and 3D representation techniques, including 3D positioning. As the result of this study, we have designed and implemented a 3D view volume for three dimensional spatial layout and stereoscopic rendering of each element by extending CSS. To verify the suggested method, we generated sample content and it demonstrated that the proposed method enables 2D-based content to be reused in 3D content. Moreover, this approach can be easily adopted for web platform-based content creation.**

***Index Terms*— Declarative Content, 3D representation, CSS, Stereoscopic, Web platform**

I. INTRODUCTION

In recent years, broadcasters around the world have developed commercial hybrid broadcasting services, as well as and related standardizations are in process for providing rich content. In hybrid broadcasting, the role of data content, whether bounded or unbounded with linear broadcasting audio/video stream, is gaining importance, as it can provide additional information and interactivities to viewers.

To provide data content in hybrid broadcasting, declarative content is now used for data representation. The Hybrid Broadcast Broadband TV (HbbTV)[1] provides applications made with declarative content, and the Advanced Television System Committee(ATSC) 2.0[2] supports declarative object for its interactive service. Japan's national public broadcaster, NHK, has been developing hybrid broadcasting service platform based on HTML5 and has launched related services. The Web and TV Interesting Group in the W3C is currently researching for applying HTML5-based web technologies to TV.

Most declarative content consists of multimedia data content and markup languages such as eXtensible Markup Language (XML), Hyper Text Markup Language (HTML), and Cascading Style Sheet (CSS)[3]. The CSS is a style sheet language used for describing the presentation semantics of declarative content, and it has several profiles for supporting various devices.

The CSS TV profile[4] is for TV devices that specify appropriate data representation and its display on a TV screen. This profile identifies a minimum set of properties, values, selectors, and cascading rules. The CSS TV profile is a superset of the CSS Mobile profile, employing the 'TV' media type.

However, CSS including CSS TV profile did not include 3DTV service guidelines when it was standardized, and does not yet specify how to provide stereoscopic TV service to viewers based on declarative content. Therefore, the development of a method for providing stereoscopic content using declarative content is required.

To resolve such issues, this paper introduces a content representation method that can support declarative 3D content using the CSS. Since most hybrid broadcasting usage scenarios include a means by which viewers can access web pages via the Internet, we also considered the reuse of 2D-based web pages or declarative data content as 3DTV content.

II. THE PROPOSED 3D REPRESENTATION METHOD FOR WEB PLATFORM BASED HYBRID TV CONTENT

A. Media type extension

We have defined a media type extension to support 3D-capable devices. CSS currently supports media-dependent style sheets tailored for different media types. *Media queries*[5] extends the functionality of media types by allowing more precise labeling of style sheets. However, the current version of the CSS media query does not include 3D media type. Thus, in this paper, the CSS media query extension has been defined in order to adaptively change of stereoscopic content according to the receiving devices.

This research was supported by the Ministry of Science, ICT and Future Planning, Rep. of Korea, under the support program supervised by the Korea Communications Agency (KCA)

H. J. Yim, B. J. Bae, H. M. Kim, and N. Hur are with the Broadcasting System Research Department, Electronics and Telecommunications Research Institute, Daejeon, 305-700, Korea (e-mail: {hjyim7, 1080i, hmkim, namho}@etri.re.kr).

H. J. Lee and S. B. Lim are with the Dept. of Multimedia, Sookmyung Women's University, Cheongpa-ro, 47-gil 100, Youngsan-gu, Seoul, 140-742 Korea (e-mail: {wmfvv, sblim}@sm.ac.kr).

B. *Content fragmentation*

In order to reconstruct the spatial layout of declarative content from 2D to 3D, declarative content consisting of multimedia elements are separated. For content separation, we used various web page separation methods that utilize the CSS3 module for the content provider to reconstruct 2D webpages in the 3D space effectively. We considered the media type-based separation approach, semantic based separation approach, and arbitrary separation approach, each of which can be used to realize the user's intent.

C. *3D view volume generation*

Next, we generated a 3D view volume using CSS3 modules. To transform 2D-based content into 3D content, 3D space, which is called 3D view volume, needs to be settled. The 3D view volume is a space surrounded by a near plane, far plane and side planes; using this view volume, it is possible to display only the objects located in the interior, and these are projected onto the windows of the view plane.

To obtain a stereoscopic viewing effect, the 3D view volume requires two virtual cameras, one for the left-eyed view volume and one for the right-eyed view volume; the objects must be located in the crossed-area of each view volume. Accordingly, we analyzed the parameters of the 3D view volume setting and then investigated related properties in the CSS3 module.

D. *Defining 3D spatial layout*

After setting the 3D view volume, content positioning was conducted. Through this process, separated multimedia elements - consisting of web pages or declarative data content- obtain their z-index within 3D view volume. Without this process, all of the web elements would be located on the same plane, called the convergence plane, which has zero coordinates in 3D space. The z-index represents the relative z coordinate for content positioning. Assigning positive or negative values for the z-index will result in different parallaxes related to the content elements. A brief diagram outlining the suggested 3D representation technique is shown in Fig. 1.

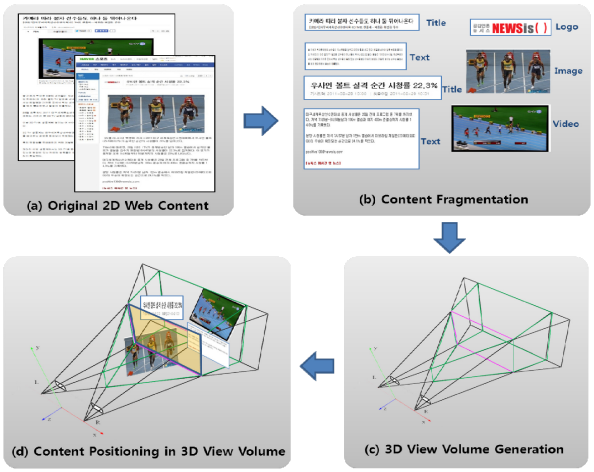


Fig. 1. Suggested method for 3D representation from 2D web contents

III. Experimental Results

To verify the proposed method for content creation, we developed a rendering engine emulator that produces stereoscopic images from virtual left-eyed and right-eyed cameras. The proposed method was implemented using JavaScript API, HTML5, CSS, and related multimedia data for sample content creation. We checked the sample content generated by the emulator on the 3D stereoscopic monitor, and the result on the 3DTV is shown in Fig. 2.

Fig. 2. Results of proposed 3D representation method on 3DTV

IV. Conclusions

A 3DTV service currently requires a considerable financial investment in new equipment and in creating the 3D content; hence, an ordinary 3D video service is limited in its ability to generate enough revenue for broadcasters. Therefore, to be able to reuse existing web content as 3DTV content, this paper suggested and implemented web standard-based content description method for 3D representation on TV.

The proposed 3D representation technique for declarative content can be applied to various business models, such as 3D home shopping, news, and educational content. It is therefore expected that the proposed service will revitalize the overall 3DTV environment, including content generation and available services.

References

[1] ETSI, Hybrid Broadcast Broadband TV (for HbbTV-1.5), Etsi ts 102 796 V1.2.1, Nov 2012.
[2] ATSC, A/105: ATSC Candidate Standard – Interactive Services Standard, Oct 2013.
[3] W3C, Cascading Style Sheets, level 1 W3C Recommendation, Dec 1996. Available: http://www.w3.org/TR/CSS1/
[4] W3C, CSS TV Profile 1.0, May 2003. Available: http://www.w3.org/TR/css-tv
[5] W3C, Media Queries W3C recommendation, Jun 2012. Available: http://www.w3.org/TR/css3-mediaqueries

International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 9, Number 23 (2014) pp. 18527-18534

http://www.ripublication.com

Development of a Mobile Voice Book Reader Enabling Interval Listening Basd on HTML5

Joo Hyun Park

Dept. of Multimedia Science, Sookmyung Women's University
Cheongpa-ro 47-gill 100, Yongsan-gu, Seoul, 140-742, Korea

Ja Rang Seo

Dept. of Multimedia Science, Sookmyung Women's University
Cheongpa-ro 47-gill 100, Yongsan-gu, Seoul, 140-742, Korea

Joo Hye Hong

Dept. of Multimedia Science, Sookmyung Women's University
Cheongpa-ro 47-gill 100, Yongsan-gu, Seoul, 140-742, Korea

Soon-Bum Lim

Dept. of Multimedia Science, Sookmyung Women's University
Cheongpa-ro 47-gill 100, Yongsan-gu, Seoul, 140-742, Korea

Abstract

Recently, various eBook applications are provided as reading activities through mobile devices and audio books are becoming popular. The existing eBook viewer applications have limitations when listening to books about economy, natural science and engineering. In this research, we developed a mobile android system which enables the user to repeatedly listen to specific passages, and applies mathematical formula reading techniques using document structuralizing elements in HTML5. In particular, this system can be discriminated from other systems in the way that it was

designed to enable not only screen display but also vocal output, considering situations in which the user can be with environmental or physical limits. Also, this system can be a solution to the existing audiobook's expensive producing budget and excessive memory problem by choosing the voice book method, which converts text data into voice using the android Text-To-Speech(TTS) engine instead of playing recorded audio files.

Keywords: Structured Tag Information, Section Selection, HTML5 Tag, EPUB3, Mobile Application, Voice Book Reader

1. Introduction

With the growth of multimedia technology, the emergence of EPUB[1], which is the standard format for eBook, is becoming spotlighted as it digitalizes paper books. EPUB3, the latest version of EPUB, is leading the advance of eBook into various multimedia formats by enabling the eBook to change from black-and white to that with color, sound and image, or to provide moving pictures and sounds as a result of pushing the screen. As types of eBooks which are growing with EPUB, there are 'Audiobook' and 'Voicebook'[2].

Audiobook, which plays pre-recorded voice file with text, is not showing a big ripple effect due to practical problems such as the increase of production fee in the process of pre-recording, or bigger application size as a result of including voice files. Voicebook, which applies the Android OS-based TTS(Text-to-Speech) engine, can be solution to the problems of Audiobook. Once TTS engine is applied, the process of producing record files will no longer be necessary since the device directly reads the text data which is included in eBook. Therefore, by reducing the size of program and simplifying the production process, it will supplement the weak points of Audiobook. However, existing Voicebook system also only provides the function of simply reading the text. Hence it still has weak points in the way that it has limits in the fields of science and technology which include formula and symbols, and that it can only be applied in literature such as novels[3]. Also as a mobile service, it must enable reading anytime at anyplace, but it is currently inconvenient since reading can only be done in limited places[4].

Therefore in this paper, we proposed and implemented 'The Reader System', which enables selecting desired passages, repeatedly listening to them and reading various formulas and symbols, using structuralizing elements of HTML5. In particular, this system can be discriminated from other systems in the way that it was designed to consider situations in which the user can be with environmental or physical limits, so that it provides not only screen display but also vocal output.

2. Related Work

As for standard eBook format for converting an analogue voicebook into digital file, DAISY(Digital Accessible Information System)[5] is the representative. DAISY is an international standard digital document form for the blind who cannot

International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 9, Number 23 (2014) pp. 18535-18542

http://www.ripublication.com

Resource Tracing Scrap Modeling for E-Textbook Authoring System based on Transclusion

Ja-Ryoung Choi

Dept. Of Multimedia Science, Sookmyung Women's University
Cheongpa-ro 47-gil 100, Yongsan-gu, Seoul, 140-742, Korea

Soon-Bum Lim

Dept. Of Multimedia Science, Sookmyung Women's University
Cheongpa-ro 47-gil 100, Yongsan-gu, Seoul, 140-742, Korea

Abstract

In this paper, we suggest an E-textbook Authoring System which creates original-traceable scrap notes in order to enable educators to collect various web contents and make use of them in an E-learning environment. In particular, we enabled the system to scrap contents according to permissible copyright range as the contents scrapped from web must not infringe the copyrights. In addition, using the dynamism of web, we designed a model called Resource Tracing Scrap Model to adjust to situations such as modifying or deleting contents. Educators can access original contents by recording information for tracing while in the process of creating scrap notes, and can maintain scrap notes continuously. Through such process, we expect to produce E-textbook applying various web contents.

Keywords: Transclusion, Web Contents, Dynamic changes, Web Scrap, e-Textbook, e-Learning

1. Introduction

Recently, there were many changes in education environment due to the growth of web technologies and digital media devices. As paper-based lectures are being

digitalized, it is becoming common to utilize resource from web to make teaching materials and use them in lectures. Due to such change of education environment, a system which enables educators to reuse contents from web and make class materials is necessary, and accordingly, comprehension about web contents should be done. In particular, web 2.0 environment is an open platform based on user participation, which encourages multitudinous production and share of contents. Moreover, beyond sharing, users' demand to collect and reuse contents that are made by others is expanding.

However, anyone can use the web as it is an open platform, and since contents from web are digital, problems about illegal reproduction exist. Also, web has the characteristics of dynamism, as contents or structures can change with flow of time. Therefore, in order to resolve these matters, an E-textbook Authoring System applied with new technique based on web contents collection is necessary.

In this paper, we suggest a Resource Tracing Scrap Model which creates scrap notes of original contents according to permissible copyright range, records tracing information and adjusts to dynamic changes, when an educator makes teaching materials by reusing resources collected from web.

2. Related Works

Since web-based education can provide service regardless of time or space, study on E-learning is actively progressed. Web technology is becoming a learning space for providing new ways of training such as distance training courses[1], U-Learning[2] and mobile Learning[3]. Learning Management System is a system which established an overall E-learning environment. It has strengths in the way that the educator can utilize study planning, lecture progress, evaluation and more as study tools, and student can get customized systematic education service. However, LMS is a system to manage overall lecture process rather than to produce teaching material, which makes it off the point of this study.

As study for providing customized educational contents, methods are being progressed to construct the contents dynamically depending on the educator or learner. In particular, since it is a recent trend to use web resources or to educate online, there is a system which enables authoring reusable teaching materials, as study to construct education contents from an educator's perspective.

ReLOAMS(Reusable Learning Object Authoring and Management System)[4] and TMCD(Teaching-Material Design Center)[5] are representative studies. They are systems which divide the contents into small modules, therefore enabling the educators to compose teaching materials according to their own lecture.

To collect original contents from web, there are ways of saving the address of webpage using bookmark or copy-and-pasting resources. These methods have limits when managing scrapped resources and lack of consideration about the characteristics of web.

Transclusion method is a study to utilize web contents reflecting the characteristics of web. Transclusion was suggested in Xanadu Project, which was initiated by Ted Nelson. It is a method to insert other documents partially or entirely, maintaining the

A Study on 3D Stereoscopic Representation of Web Content using CSS3

Hee-Jin Lee
Dept. of Multimedia science
Sookmyung Women's Univ.
Seoul, Republic of Korea
leeheejin881110@gmail.com

Hyun-Jeong Yim
Dept. of Broadcasting System Research
ETRI
Daejeon, Republic of Korea
hjyim@etri.re.kr

Soon-Bum Lim
Dept. of Multimedia science
Sookmyung Women's Univ.
Seoul, Republic of Korea
sblim@sm.ac.kr

***Abstract*—In this paper, we represented the HTML5 webpage in 3D space for 3D stereoscopic display by using CSS3 stylesheet. At first, we suggested various webpage separation methods in order to reconstruct webpages in 3D space effectively. Next, we suggested 3D view volume setting method using CSS3 modules. For the resulting stereoscopic images, we developed a rendering engine emulator which produces dual images from virtual left and right-eyed cameras. And we have checked the sample contents displayed on the 3D stereoscopic monitor.**

Keywords—3D, Web contents, Stereoscopic, HTML5, CSS3

I. INTRODUCTION

Recently in 3DTV, content can be represented independently of the platform using web standards such as XML, HTML(Hypertext Markup Language), CSS(Cascading Stylesheet) regardless of the type of devices. In particular, browser-based content of TV can be expressed in CSS stylesheet using CSS TV Profile 1.0[1], which is defined considering requirements and constraints of TV device. However, the current standard of CSS TV Profile, which was defined before the appearance of CSS3[2], does not contain specification of methods to place content in the 3D space or represent it stereoscopically. This paper presents the method of converting 2D web content written in HTML5 into stereoscopic display in 3D space by using CSS3.

II. RELATED TECHNOLOGY

According to the increase of 3D data application in webpage, 3D representation-related study on the webpage is being progressed[3]. Now, 3D modeling representation is available on WebGL and Web3D, but it focus on entirely 3D objects in 3D space. Also, related studies on representation for the stereo 3D web are being progressed in W3C. It extends CSS specification to represent stereoscopic 3D on the web as shown in <Table 1>.

<Table 1> CSS specification extension for representing stereoscopic 3D[4]

Composition	CSS property
Extended CSS 3D property	perspective-baseline
Stereo properties to specify stereo 3D contents	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 (Device API for stereo rendering)	non-normative

In particular, browser-based content of TV can be expressed using CSS TV Profile 1.0, which is defined considering requirements and constraints of TV device is shown in <Table 2>. Next, CSS3 which define webpage layout and style extends properties of CSS2 is organized with various modules; e.g. CSS3 Fragmentation[5], CSS3 Positioned Layout[6], CSS 3D Transform, etc. CSS3 3D Transform Module[7] is only concerned in 3D transformation (T, R, S) on the objects in 2D space, but there is no attribute to layout objects in 3D space.

<Table 2> CSS TV Profile 1.0 specification

Related modules	Excluded specification from current CSS
Backgrounds and Borders module Level 3	Background-attachment, border-collapse, border-spacing
Positioned Layout module Level 3	none
Font module Level 3	font-size-adjust, font-stretch

ISSN 1975-4736 ©MITA 2014

Text Level 3	text-shadow, word-spacing, white-space
Lists and Counters Module Level 3	none
Basic UI module Level 3	cursor

III. 3D VIEW VOLUME SETTING USING CSS STYLESHEET

In order to place multimedia objects comprising the web content into 3D space, 3D view volume must be set. Camera, lens, clipping and other setting information are required by default. <Table 3> summarizes default information required to set the 3D view volume and the properties of view volume setting in existing CSS.

<Table 3> 3D view volume setting information and related property of CSS[8]

category	3D view-volume set-up information	Related attribute of CSS
viewpoint Setup	viewpoint(x, y, z)	perspective-origin: x, y
	up-vector(x, y, z)	fixed(0, 1, 0)
	view direction(x, y, z)	fixed(0, 0, 1)
view space Setup	view angle($\propto_w$, $\propto_h$)	fixed(Screen size)
	near, far clipping distance(d_n, d_f)	none
	convergence distance(d_c)	perspective
projection setup	projection mode setup	transform-style : 'flat | preserve-3d'
placement of web content	none	z-index: <element>

Currently, there is no way to set all the information to set the rich view volume using CSS. We can set the view volume using following three properties; perspective, perspective-origin, transform-style of the CSS 3D Transform module[7]. The 'perspective' property allows the user to make elements with three-dimensional transforms appear as if they positioned in a common 3D space. It means distance from viewpoint(x, y, 0) to convergence plane. The 'perspective-origin' property provides control over the origin at which perspective is applied, effectively changing the location of the vanishing point. It specifies a reference point of x, y coordinates at viewpoint. And 'transform-style' property determines if the element is positioned in three-dimensional or flattened space when it applied to an element. It is used in order to setup graphics projection, for example, plane or perspective projection.

In order to place the separated multimedia objects in 3D space, three-dimensional coordinates must be allocated to the separated objects. However, because 'z-index' property of the current CSS3 Positioned Layout module[6] only determines the order of elements to be placed in the same plane, it cannot set the depth information between elements. Therefore, 'translateZ (<length>)', 'transform3d (<translate-value>, <translate- value>, <length>)' properties of the existing CSS 3D Transform module were used to move the object forward and backward according to the Z value of the current object location, and this enables the input of Z value into each objects and expression of relative depth information. In order to confirm the result of placing 2D content in 3D space in document, the result was rotated by a predetermined angle as shown in [Fig. 1].

[Fig. 1] Placement of web content in view volume

IV. RENDERING ENGINE EMULATOR AND RESULTS

The user set the mono view volume beforehand using view volume setting properties in CSS, and placed the objects in a 3D space. The emulator automatically sets the stereo view volume which has a standard binocular distance of 65mm based on user location in mono view volume.

Thus, the view volume of the left and right eyes are set in 'perspective-origin' location (<user-set x-coordinate of the view volume ± 32.5mm>, <user-set y-coordinate of the view volume>), and objects are placed in stereo, finally obtaining a left and right images which had binocular parallax. Result of expressing the final output stereo image 3D-stereoscopically in a 3DTV display is shown in [Fig. 2].

Primitive Motion API for Kinetic Typography Rendering Engine

YoonAh Cho
Dept. Multimedia Science
Sookmyung Women's University, Korea
fanwith@naver.com

JooHyun Park
Dept. Multimedia Science
Sookmyung Women's University, Korea
park.Joohyun5@gmail.com

Hyo-Yong Kim
Dept. Digital Cultural Technology and Contents,
Hansung University, Korea
hykim@hansung.ac.kr

Soon-Bum Lim
Dept. Multimedia Science
Sookmyung Women's University, Korea
sblim@sookmyung.ac.kr

***Abstract*— Kinetic typography articulates movement into static text; it covers time, space and speed to convey emotions. This paper examines primitive motion API, the way to express kinetic typography. The main purpose of primitive motion API is to replace existing software and expert coding with it. To achieve the purpose, the thesis classifies the movement systematically and develops Primitive Motion API which renders Kinetic typography.**

Keywords—Kinetic Typography; Primitive Motion; Motion API; Motion Component; Motion Library; Systematic Primitive Motion API; Rendering Engine

I. INTRODUCTION

A considerable amount of literature has been published on kinetic typography which established itself as a tool of communication along with the rapid development of visual culture and digital technology. Kinetic typography is a medium to present movement in text, adding the variety of effects such as time, space, speed and weight to static text, and showing both temporal and spatial representation of the text. The first serious discussions and analyses about kinetic typography emerged in 1995 from temporal typography. In addition, a technical research is progressed in MIT and CMU; implementation of the engine for kinetic typography [1] and production of time-based technology [2] are typical researches. It shows that kinetic typography can be used in various application systems and expresses texts. Despite its influence and expressiveness, kinetic typography undergoes several major drawbacks. Most studies in the field of kinetic typography have focused on artistic significance and creative presentation. We have to use professional animation software such as Flash [3] and After Effect [4], or other inconvenient programming to create kinetic typography; they are time-consuming tasks and requiring professional skills and not in any condition to represent kinetic typography in social media. This paper will examine Primitive Motion API, the medium to compose kinetic typography. There are two primary objectives of the study: 1)to analyze the representative motion systematically and 2) to propose Primitive Motion API to replace existing software and expert coding.

II. RELATED WORKS

2.1 Engine for Kinetic typography

Studies on technical method of kinetic typography have been conducted since the late 1990s. Kinetic typography engine [5] is a foundational framework for generating kinetic typography; it provides basic maneuvers such as moving, rotating and twisting of texts. Recently, KTE2[6], has been published with its additional features. However, the engine is inappropriate for authoring tools in Korean kinetic typography and unequipped with the interface for it.

2.2 Preceding research

Mobile Instant Messenger Using Kinetic Typography [7] was designed to create kinetic typography quickly and easily in any users' skill and proficiency. In the process, the Typography analyzed, organized and divided the motion into four groups of motion libraries. Primitive Motion API re-analyzed the motion with the systematic classification based on the precedent research.

III. CLASSIFICATION THE MOTION

The construction of kinetic typography is distinguished from the production with existing PC-based animation manufacture tools in its implication and environment. Existing PC-based animation programs are often utilized for art or media art, requiring professional and sophisticated skills. Therefore, it is relatively unbound by restrictions as screen size and time.

However, it is bound in mobile environment; for instance, the range of motion and method to select and set up the menu are restricted in mobile system. Thus, problems on expressing motions will occur with existing PC-based animation manufacture tools in social-media background. In consequence, the paper proposes *Primitive Motion API* to be utilized efficiently in social-media environment.

ISSN 1975-4736 ©MITA 2014

3.1 Classification the Motion

In this paper, it analyzed PC-based kinetic typography manufacturing tools of Adobe's Flash and AfterEffect, and animation effects on Microsoft's PowerPoint, and Apple's KeyNote. As a result, the research classified the components of kinetic typography into two major parts: general factors and motion factors, which are further categorized into 8 general factors and 7 motion factors each. General factors are composed of font, size, color, underline, italics, and space; motion factors are comprised of opacity, speed, direction, location, separation, size, and repetition.

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar.

3.2 Grouping the Motion

According to the way for the typography to move, motion of kinetic typography is classified into three basic motions. each motion is grouped in sub-motions; basic motions are classified into static motion, straight motion, curved motion, and free motion, and are defined with the attribute values speed, separation, rotation, size of motion, size of letter, and opacity which influence the motion.

Static motion lacking direction, does not have sub-motion; straight motion has sub-motions distinguished by right directions of up, down, right and left and oblique directions of the angle of divergence; curved motion has sub-motions differentiated by shape of motion (basically into circular and Bezier motions); free motion, lacking regularity does not have sub-motion. The classification system referenced SMIL Transition to categorize kinetic typography motions.

3.3 Designing the Primitive Motion API

Primitive Motion API is generally classified into static motion, straight motion, and curved motion and subdivided into sub-motion APIs like [TABLE Ⅰ]. Also, each API has its attribute value to express the motion.

TABLE I. DESIGN THE PRIMITIVE MOTION API

```
Primitive Motion API =
{ <Static Motion>, <Line Motion>, <Curved Motion> }

Sub-motion API of Static Motion =
{ <Static Rotation>, <Static Size Change>,
{ <Static Opacity Change>, <Static Shaking> }

Sub-motion API of Line Motion =
{ <Line Movement> }

Sub-Motion API of Curved Motion =
{ < Circle Movement>, <Bezier Movement>  }
```

[TABLE Ⅱ] arranged the attribute values of each sub-motion API; the arrangement is to enable users to produce kinetic typography simply by selecting and changing the motion attribute values.

TABLE II. PRIMITIVE MOTION API AND PARAMETERS

Primitive Motion	Sub-Motion	Parameter	Usage Example
Static Motion	Rotation	{ object, angle, duration time }	
	Scale	{ object, size, duration time }	
	Opacity	{ object, opacity value, duration time }	
	Shaking	{ object, size, duration time }	
Line Motion	Line	{ object, direction, length, duration time }	
Curve Motion	Circle	{ object, angle, radius, duration time }	
	Bezier	{ object, (x , y) * , duration time }	

4 IMPLEMENT OF PRIMITIVE MOTION API

Primitive Motion API is designed by extracting the properties of each group and building templates for all groups. It can display movement which will be applied in application.

4.1 Structure of Primitive Motion API

Primitive Motion API controls geometrical movements including Font Rendering Module and Movement Controlling Module. It is designed by forming the text object, setting up the attribute values of the texts and creating it as a function. Primitive Motion API can be used by calling.

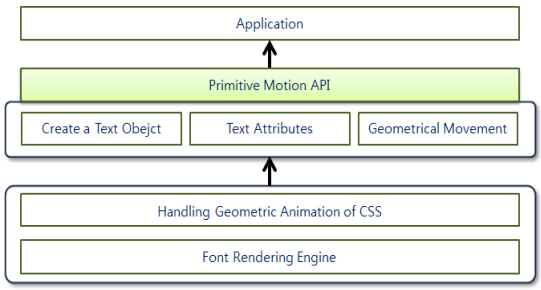


FIGURE I. STRUCTURE OF PRIMITIVE MOTION API

4.2 Environment of implementation

Primitive Motion API is developed in JavaScript Library which is the basis for standard web environment. User should import into the web and mobile application, and call the API to element of HTML file made by user. Its aim is providing method to generate kinetic typography for not the experts but ordinary people. They articulate movement into texts using Primitive Motion API quickly and easily.

의약과학과

Regulatory Toxicology and Pharmacology 69 (2014) 207–216

Contents lists available at ScienceDirect

Regulatory Toxicology and Pharmacology

journal homepage: www.elsevier.com/locate/yrtph

Carcinogenicity study of CKD-501, a novel dual peroxisome proliferator-activated receptors α and γ agonist, following oral administration to Sprague Dawley rats for 94–101 weeks

Hee Su Lee [a,1], Minsun Chang [b,1], Ji-Eun Lee [a,c], Woojin Kim [a], In-Chang Hwang [d], Dal-Hyun Kim [d], Hyun-Kyu Park [a], Hyun-Ji Choi [a], Woori Jo [a], Shin-Woo Cha [e], Woo-Chan Son [a,*]

[a] Department of Pathology, Asan Institute for Life Sciences, University of Ulsan College of Medicine, Asan Medical Center, Seoul, Republic of Korea
[b] Department of Medical and Pharmaceutical Science, Sookmyung Women's University, Seoul, Republic of Korea
[c] College of Veterinary Medicine, Chonnam National University, Gwangju, Republic of Korea
[d] Chong Kun Dang Research Institute, CKD Pharmaceuticals Inc., Kyunggi-do, Republic of Korea
[e] Division of Non-clinical Studies, Korea Institute of Toxicology (KIT), Daejeon, Republic of Korea

ARTICLE INFO

Article history:
Received 27 November 2013
Available online 18 April 2014

Keywords:
CKD-501
Carcinogenicity
Dual PPARα/γ agonist

ABSTRACT

CKD-501 is a peroxisome proliferator-activated receptor (PPAR) agonist. The current study was conducted in Sprague Dawley (SD) rats for 94–101 weeks to investigate the carcinogenic potential of CKD-501. 60 males received 0, 0.03, 0.12, or 1.0 mg/kg/day, which was changed after 66 weeks to 0.24 mg/kg/day due to increased mortality, while 60 females received 0, 0.03, 0.06, or 0.12 mg/kg/day throughout the study period. After switching the dosage, no significant changes in the survival rates were observed. Non-neoplastic lesions such as bladder transitional cell hyperplasia and a diminished corpus luteum were observed in females administered 0.12 mg/kg/day and the right chamber dilation and left ventricular hypertrophy were increased dose dependently in both males and females. Non-neoplastic lesions such as bone marrow hypoplasia and fat cell proliferation and neoplastic lesions such as lipomas and liposarcomas observed in males and/or females were considered expected pharmacological effects for this compound. Compared to rosiglitazone, CKD-501 had a 4.4-fold higher margin of safety for tumor induction and did not cause bladder carcinoma as was observed with pioglitazone.

1. Introduction

CKD-501 is a dual agonist of peroxisome proliferator-activated receptors (PPARs) α and γ. PPARs normally function as transcriptional factors regulating metabolic homeostasis. They play an important role in the control of lipid and glucose metabolism (Gurnell, 2005; Gross and Staels, 2007). The activation of PPARs by selective ligands or agonists first initiates heterodimerization with the 9-cis-retinoic acid receptor leading to binding of peroxisome proliferator response elements. Dual agonists activating both PPARα and γ are efficacious agents for the treatment and management of a number of metabolic diseases including type 2 diabetes (Varley and Southgate, 2008).

Despite these promising actions, safety concerns regarding PPAR agonists arose following the controversial results of numerous toxicology studies in animals as well as in clinical trials (Rubenstrunk et al., 2007). Clinical trials revealed additional adverse effects of PPARγ agonists such as tesaglitazar and muraglitazar, including augmentation of serum creatinine levels and an increased risk of major cardiovascular events (muraglitazar) as well as a reduced rate of glomerular filtration (tesaglitazar) Rubenstrunk et al., 2007. Many investigations have also reported the carcinogenicity of this class of drugs *in vivo*. For example, troglitazone, a PPARγ agonist, induced hemangiosarcomas in B6C3F1 mice in a 2-year bioassay (Duddy et al., 1999; Herman et al., 2002). Moreover, PPAR agonists have been linked to carcinogenesis in the epithelial lining of the urinary tract in humans (Varley and Southgate, 2008). The recent studies, however, suggest that PPARγ agonists may actually inhibit the growth of colon cancer cells (Ban et al., 2010; Shen et al., 2007). As such, the current study was conducted to evaluate the carcinogenic potential of CKD-501 following 94–101-week oral administration to Sprague Dawley (SD) rats. The study was carried out in accordance with the standards set forth by the Organization for Economic Co-operation and Development (OECD) Test Guideline 451 (Carcinogenicity Studies) and International Conference on Harmonization (ICH) Harmonized Tripartite

* Corresponding author.
E-mail address: wcson@amc.seoul.kr (W.-C. Son).
[1] Two authors equally contributed to this work.

http://dx.doi.org/10.1016/j.yrtph.2014.04.003
0273-2300/

Guidelines (OECD guidelines for the testing of chemicals/Section 4: Health effects, 1981; International Conference on Harmonisation, 1995; International Conference on Harmonization, 1997; International Conference on Harmonisation, 2008).

2. Materials and methods

2.1. Test materials and formulation analysis

CKD-501 was produced at 98.73% purity by Chong Kun Dang (Seoul, Korea). Prior to use, the structure and purity of the drug were verified and its stability was validated at all concentrations employed by referring to the results of toxicology tests (tests #G05130 and #G05136). In order to obtain fresh-dose formulations, CKD-501 was dissolved in 5% gum Arabic (a water-soluble gum derived from the acacia tree) every day and stored in a refrigerator. Gum Arabic, purchased from Sigma Aldrich (St. Louis, MO, USA), was diluted to 5% with distilled water (JW Pharmaceutical, Seoul, Korea). The amount of test substance for the preparation of formulations and the amount actually used were decided on each occasion based on the most recent body weight (measured every week for the first 14 weeks, and every 4 weeks thereafter). Before the formulations were used, any discrepancies between these amounts were examined. The dose formulations were periodically analyzed for concentration and homogeneity and were found to be within ±10% of the nominal concentration.

2.2. Animal husbandry and maintenance

Four-week-old male and female Crl:CD(SD) IGS rats from a specific pathogen free colony were collected from Orient Bio (Seoul, Korea) and were quarantined for 7 days prior to the study. They were approximately 5 weeks old on the first day of the study and were housed 2/cage after dosing administration commenced. Rats were given water *ad libitum* and an autoclaved pellet diet (Lab Diet® #5002, PMI Nutrition International, St. Louis, MO, USA) *ad libitum*, except for one night of fasting before their planned sacrifice. The environmental conditions were kept within the following ranges (temperature, 23 ± 3 °C; relative humidity, 50 ± 10%; 12-h light/dark cycle, illumination, 150–300 Lux; ventilation, 10–20 times/h) and monitored every hour throughout the study. The animals were managed at the Korea Institute of Toxicology (KIT), an accredited animal facility, by following the AAALAC International Animal Care Policies. The Animal Care and Use Committee of the KIT reviewed and approved all the study protocols.

2.3. Experimental design and rationale for dosage selection

Oral administration was selected since it is the intended administration route of CKD-501 in humans. Groups of 60 male rats were administered with CKD-501 at dosages of 0, 0.03, 0.12, or 1.0 mg/kg/day for 94–101 weeks and groups of 60 female rats were treated with CKD-501 at dosages of 0, 0.03, 0.06, or 0.12 mg/kg/day via oral gavage (Table 1). A preliminary 13-week toxicity study was conducted to identify the appropriate target organs and dosages for the rat carcinogenicity study (dosages 0.1, 1.0, and 10.0 mg/kg/day: unpublished data). None of the male rats in any of the groups died during the course of the preliminary study. However, all of the female rats administered 10.0 mg/kg/day and one female administered 1.0 mg/kg/day died. Histopathological examination revealed adipose tissue hyperplasia in a dose dependent manner in all of the CKD-501-treated rats, males and females. In addition, hypertrophy of the myocardium was observed for both genders administered 1.0 mg/kg/day. The maximum tolerated dosage was estimated to be 1.0 mg/kg/day in males and 0.1 mg/kg/day in females according to the test results of the 13-week preliminary study. The high dosage level for males (1.0 mg/kg/day) was changed to 0.24 mg/kg/day after 66 weeks due to increased mortality (26.7%) at 65 weeks. The low-dose level was set at 0.03 mg/kg/day for both males and females with no adverse effects expected and the mid-dose level was the geometric mean between the low and high dose levels, or 0.12 mg/kg/day for males and 0.06 mg/kg/day for females. Rats assigned to the male and female control groups were administrated 5% Gum Arabic instead of CKD-501. The daily dose volume was calculated on the basis of 10.0 mg/kg.

2.4. In-life phase examination

Visual inspections and detailed physical examinations were performed on each animal every week to observe any superficial palpable nodules. The rats were sacrificed after blood samples for hematology and clinical biochemistry were collected. Morbidity, mortality, location, size, consistency, and time of the first observation of any external swelling were carefully examined and documented using the Path/Tox System (Version 4.2.2) (Xybion Medical Systems, Cedar Knolls, NJ, USA). The following measurements were recorded weekly for the first 14 weeks and every 4 weeks thereafter: body weights and weights of stool output, water and food provided to each cage. These data were used to calculate the mean weekly consumption per animal (mg/rat/week). At the end of the study, all rats were weighed after being anesthetized by inhaling isoflurane. For hematology and blood biochemistry parameters, blood samples were obtained from the caudal vein and the rats were exsanguinated by the abdominal aorta. Due to the increased mortality rate, male rats in the control, 0.03, 0.12, and 1.0/0.24 mg/kg/day groups were sacrificed at 101, 101, 100, and 94 weeks, respectively, while female rats in the control, 0.03, 0.06, and 0.12 mg/kg/day groups were sacrificed at 101, 101, 97, and 100 weeks, respectively.

2.5. Hematology and blood chemistry

Blood for hematology determinations was added in tubes with an anticoagulant, EDTA-2K. Blood samples were collected as

Table 1
Composition of test groups.

Group	Gender	Number of animals	Animal number	Dose volume (ml/kg/day)	Dosage (mg/kg/day)
Control (gum Arabic)	Male	60	1–60	0	0
	Female	60	241–300		
T1	Male	60	61–120	0.03	0.2
	Female	60	301–360		
T2	Male	60	121–180	0.12	1.0
	Female	60	361–420	0.06	
T3	Male	60	181–240	1.0 (Week 1–Week 65) 0.24 (Week 66–termination)	6.0
	Female	60	421–480	0.12	

BIOMOLECULES & THERAPEUTICS

Original Article
Biomol Ther 22(4), 347-354 (2014)

Selective Estrogen Receptor Modulation by *Larrea nitida* on MCF-7 Cell Proliferation and Immature Rat Uterus

Hye-Na Ahn[1], Si-Yeon Jeong[1], Gyu-Un Bae[1], Minsun Chang[2], Dongwei Zhang[2], Xiyuan Liu[2], Yihua Pei[5], Young-Won Chin[5], Joongku Lee[3], Sei-Ryang Oh[4] and Yun Seon Song[1,*]

[1]College of Pharmacy, [2]Department of Medical and Pharmaceutical Science, College of Science, Sookmyung Women's University, Seoul 140-742, [3]College of Pharmacy, Dongguk University-Seoul, Goyang 410-820, [4]International Biological Material Research Center, Korea Research Institute of Bioscience & Biotechnology, Daejeon 305-806, [5]Natural Medicine Research Center, Korea Research Institute of Bioscience & Biotechnology, Cheongwon 363-883, Republic of Korea

Abstract

Larrea nitida is a plant that belongs to the Zygophyllaceae family and is widely used in South America to treat inflammatory diseases, tumors and menstrual pain. However, its pharmacological activity remains unclear. In this study we evaluated the property of selective estrogen receptor modulator (SERM) of *Larrea nitida* extracts (LNE) as a phytoestrogen that can mimic, modulate or disrupt the actions of endogenous estrogens, depending on the tissue and relative amount of other SERMs. To investigate the property of SERM of LNE, we performed MCF-7 cell proliferation assays, estrogen response element (ERE)-luciferase reporter gene assay, human estrogen receptor (hER) binding assays and *in vivo* uterotrophic assay. To gain insight into the active principles, we performed a bioassay-guided analysis of LNE employing solvents of various polarities and using classical column chromatography, which yielded 16 fractions (LNs). LNE showed high binding affinities for hERα and hERβ with IC_{50} values of 1.20×10^{-7} g/ml and 1.00×10^{-7} g/ml, respectively. LNE induced 17β-estradiol (E2)-induced MCF-7 cell proliferation, however, it reduced the proliferation in the presence of E2. Furthermore, LNE had an atrophic effect in the uterus of immature rats through reducing the expression level of progesterone receptor (PR) proteins. LN08 and LN10 had more potent affinities for binding on hER α and β than other fractions. Our results indicate that LNE had higher binding affinities for hERβ than hERα, and showed SERM properties in MCF-7 breast cancer cells and the rat uterus. LNE may be useful for the treatment of estrogen-related conditions, such as female cancers and menopause.

Key Words: *Larrea nitida*, Lignan, Phytoestrogen, Estrogen receptor, MCF-7 cell, Uterus

INTRODUCTION

The genus *Larrea* (Zygophyllaceae) has a long history of ethnobotanical use by native people of south America. The genus comprises six species: *Larrea ameghinoi, Larrea cuneifolia, Larrea divaricata, Larrea mexicana, Larrea nitida* and *Larrea tridentate*. These plants have been used to treat a variety of illnesses, including cancer, inflammation and menstrual pain (Brent, 1999). The extract of *Larrea divaricata*, a species in the *Larrea* genus, has a biphasic effect on cell proliferation, exerting a stimulatory effect at low concentrations but inhibiting the proliferation of BW 5147 T lymphoma cells at high concentrations (Anesini *et al*., 1999). Methanolic extract of *Larrea divaricata* exhibited anti-tumor activity in the MCF-7 human breast adenocarcinoma cell line (Bongiovanni *et al*., 2008). As a folk remedy, the genus *Larrea* was not recommended for pregnant and lactating females, suggesting that the *Larrea* genus regulates hormonal functions in females. Estrogen, a major female hormone, regulates various physiological responses in many target tissues and plays important roles in the development and progression of breast and endometrial cancers (Hayashi *et al*., 2003; Bryant *et al*., 2005).

Estrogen exerts its biological effects by binding to estrogen receptors (ERs), which are present mainly in the nucleus and comprise ERα and ERβ (Hayashi and Yamaguchi, 2008). Selective ER modulators (SERMs) bind to ERs and exert agonistic actions in several target tissues, while acting as ER antagonists in others (Shang and Brown, 2002). Progesterone (P4)

Open Access **http://dx.doi.org/10.4062/biomolther.2014.050**

Received May 7, 2014 **Revised** Jun 14, 2014 **Accepted** Jun 19, 2014
Published online Jul 31, 2014

***Corresponding Author**
E-mail: yssong@sookmyung.ac.kr
Tel: +82-2-2077-7231, Fax: +82-2-710-9871

bound to the progesterone receptor (PR) is important in mammary gland and reproductive tract development, especially in the uterus and ovaries (Graham and Clarke, 1997; Mulac-Jericevic *et al*., 2003). PR has two predominant isoforms, PR-A and PR-B. PR-B mediates the proliferative effect of P4 in the mammary gland and PR-A is essential for the functional response of P4 in the uterus and ovaries (Mulac-Jericevic *et al*., 2003). Therefore, Kuhl and Stevenson suggested that the regulators of PR or ER are used for the prevention and treatment of breast cancer and postmenopausal osteoporosis (Kuhl and Stevenson, 2006). Phytoestrogens are compounds naturally found in plants that possess estrogenic properties associated with the prevention of cancer, heart disease, menopausal symptoms and osteoporosis (Bedell *et al*., 2014). The main classes of phytoestrogens in terms of chemical structure are isoflavones, prenylated flavonoids, coumestans, and lignans (Ward and Kuhnle, 2010).

Phytochemical studies have indicated that the genus *Larrea* contains lignans, flavonoids, condensed tannins, triterpene saponins and naphthoquinones with lignans being the major component (Gisvold and Thaker, 1974; Konno *et al*., 1990). Although various species of the genus *Larrea* have been studied to have therapeutic effects, pharmacological properties of *Larrea nitida* associated female hormones and their interaction with hormone receptors have not reported.

In this study, we evaluated the SERM activity of *Larrea nitida* to determine its pharmacological importance as a source of phytoestrogens. Extract of *Larrea nitida* (LNE) was investigated in terms of its ability to bind to purified hERα and hERβ, and to modulate ER-mediated gene transcription, proliferation in MCF-7 breast cancer cells, and its uterotrophic effect in rats.

MATERIALS AND METHODS

Cell culture and materials

The MCF-7 cell line is a steroid-sensitive subline developed from MCF-7 human breast cancer cells (kindly provided by Dr. Anna Soto, Tufts University, Boston, MA, USA). MCF-7 cells were cultured in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% fetal bovine serum (FBS) and 1% antibiotic-antimycotic solution. Cells were maintained in humidified 5% CO_2 atmosphere at 37°C. All cell culture reagents were purchased from Gibco (Carlsbad, CA, USA). [2,4,6,7-^{3}H]-17β-Estradiol ([^{3}H]-E2, 88.0 Ci/mmol) was purchased from PerkinElmer Inc. (Boston, MA, USA). Reagents and chemicals were purchased from Sigma-Aldrich (St. Louis, MO, USA) and unless stated otherwise, were of research grade suitable for cell culture or of the highest grade available. Each chemical was dissolved in either ethanol or DMSO before being tested in the various assay systems. Stock solutions were stored at -20°C. The final solvent concentration in the culture medium did not exceed 0.5%.

Plant materials

The stems and leaves of *Larrea nitida* were collected in Jarilla, Chile in 2007 and identified by Dr. Joongku Lee, Korea Research Institute of Bioscience and Biotechnology (KRIBB). A voucher specimen (access number FBM026-052) was deposited in the herbarium of the KRIBB.

Preparation of LNE and its chromatography profile

Air-dried stems and leaves of *Larrea nitida* (52 g) were pulverized and extracted with methanol (MeOH) to yield the crude extract of *Larrea nitida* (LNE) (14 g). A portion of this LNE (12 g) was subjected to a reversed-phase silica column chromatography eluting with a gradient mixture of H_2O-MeOH (90:10-0:100) to give 16 sub-fractions (LN01-16). Separation of LN 08 by aforementioned HPLC method yielded nordihydroguaiaretic acid (NDGA) (6.8 mg, t_R = 38.04 min). This compound exhibited comparable spectroscopic data (^{1}H and ^{13}C NMR) and HRESIMS *m/z*: 303.1591 $[M+H]^+$ (calculated for $C_{18}H_{23}O_4$, 303.1596) to published values (Son *et al*., 2005).

Cell proliferation assay

MCF-7 cells were plated in 12-well plates at an initial density of 6×10^4 per well (Soto and Sonnenschein, 1984; Soto and Sonnenschein, 1985). Cells were allowed to attach for 24 h and the medium was then replaced with experimental medium (phenol red-free DMEM containing 5% charcoal dextran-treated FBS (CD-FBS)). LNE, E2, ICI 182,780 (ICI; pure ER antagonist), or a combination of these compounds was added to the medium in a range of concentrations and the cells were incubated for 6 days (late exponential phase). Cell growth was determined using the 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyl tetrazolium bromide (MTT) assay.

ERE-luciferase reporter gene assay

MCF-7 cells were seeded in 12-well plates (6×10^5 cells/well) and cultured in phenol red-free DMEM containing 5% CD-FBS in triplicate. Cells were then transiently transfected with pERE-Luc plasmid (0.5 μg/well) and 0.2 μg of an inactive control plasmid encoding a β-galactosidase gene using Lipofectamine 2000 Reagent (Invitrogen, Carlsbad, CA, USA). The pERE-Luc plasmid contains three copies of the *Xenopus laevis* vitellogenin A2 ERE upstream of firefly luciferase (a gift from Dr. V. C. Jordan, Lombardi Comprehensive Cancer Center, Georgetown University) (Catherino and Jordan, 1995). One day after transfection, various concentrations of LNE were added in the presence or absence of E2. Cells were lysed for luciferase activity analysis after incubation for 24 h. Luciferase activity was measured using a luciferase reporter assay system (Promega, Madison, WI, USA). Luminescence was detected with a TD-20/20 luminometer (Turner Design, Sunnyvale, CA, USA). Finally, luciferase activities were normalized to β-galactosidase activity.

ER ligand-binding assay

The binding ability of the test compound to control or treated recombinant full-length human estrogen receptor α (hERα) was assessed as described previously (Obourn *et al*., 1993). hERα (2088 pmol/mg estrogen binding activity, Invitrogen Corp.) was diluted to a concentration of 5 nM with a binding buffer containing 10 mM Tris (pH 7.5), 10% glycerol, 1 mM DTT and 1 mg/ml bovine serum albumin (BSA). The hERα was labeled with [^{3}H]-E2 (3 nM) in the presence or absence of 10^{-6} M unlabeled E2, various concentration of LNE or LNs fractions. For the hERβ binding assay, hERβ (4500 pmol/mg estrogen binding activity, Invitrogen Corp.) was diluted to a concentration of 10 nM with the binding buffer and labeled with [^{3}H]-E2 (10 nM) in the presence or absence of 10^{-6} M unlabeled E2, LNE or LNs. After incubation for 2 h at 27°C, the reactions were terminated by rapid filtration through glass-fiber filters (Packard

Biochemical and Biophysical Research Communications 452 (2014) 840–844

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

journal homepage: www.elsevier.com/locate/ybbrc

Human glutathione S-transferase P1-1 functions as an estrogen receptor α signaling modulator

Xiyuan Liu [a], Byoung Ha An [b], Min Jung Kim [a], Jong Hoon Park [a], Young Sook Kang [c], Minsun Chang [d,*]

[a] Department of Biological Science, Sookmyung Women's University, Seoul, Republic of Korea
[b] Department of Food and Nutrition, College of Life Science, Sookmyung Women's University, Seoul, Republic of Korea
[c] Department of Pharmacy, College of Pharmacy, Sookmyung Women's University, Seoul, Republic of Korea
[d] Department of Medical and Pharmaceutical Science, College of Science, Sookmyung Women's University, Seoul, Republic of Korea

ARTICLE INFO

Article history:
Received 24 August 2014
Available online 16 September 2014

Keywords:
Estrogen receptor α
Glutathione S-transferase P1-1
Estrogen response element
Regulatory interacting protein 140
Breast cancer

ABSTRACT

Estrogen receptor α (ERα) plays a crucial role in estrogen-mediated signaling pathways and exerts its action as a nuclear transcription factor. Binding of the ligand-activated ERα to the estrogen response element (ERE) is a central part of ERα-associated signal transduction pathways and its aberrant modulation is associated with many disease conditions. Human glutathione S-transferase P1-1 (GSTP) functions as an enzyme in conjugation reactions in drug metabolism and as a regulator of kinase signaling pathways. It is overexpressed in tumors following chemotherapy and has been associated with a poor prognosis in breast cancer. In this study, a novel regulatory function of GSTP has been proposed in which GSTP modulates ERE-mediated ERα signaling events. Ectopic expression of GSTP was able to induce the ERα and ERE-mediated transcriptional activities in ERα-positive but GSTP-negative MCF7 human breast cancer cells. This inductive effect of GSTP on the ERE-transcription activity was diminished when the cells express a mutated form of the enzyme or are treated with a GSTP-specific chemical inhibitor. It was found that GSTP inhibited the expression of the receptor interacting protein 140 (RIP140), a negative regulator of ERα transcription, at both mRNA and protein levels. Our study suggests a novel non-enzymatic role of GSTP which plays a significant role in regulating the classical ERα signaling pathways via modification of transcription cofactors such as RIP140.

1. Introduction

Estrogen receptor α (ERα) is a transcription factor crucial for estrogen-responsive gene transcription. The dimers of ligand-bound ERα interact with the specific DNA sequences such as the estrogen response element (ERE; GGTCAcagTGACC), activating protein (AP)-1 binding sequences (ATGAGTCAT), or GC rich specificity protein (Sp) binding sequences ((G/T)GGGCGG(G/A)(G/A)(C/T))-located in the regulatory region of target gene promoters [1]. Interaction of ERα with the specific responsive elements leads to the induction of target gene transcription depending on the cell type and promoter context by recruiting a variety of coregulatory proteins that stabilize the transcription initiation complex or alter chromatin structure.

ERα target genes including cyclin D1, progesterone receptor, and trefoil factor 2 are tightly regulated since these genes are closely implicated in cell proliferation, survival and/or death, thus playing vital roles in development, growth, and tumorigenesis [2]. The classical mechanism of ER action involves binding of ERα to the ERE and is attributed to approximately two thirds of estrogen activated ERα action [3]. Thus, regulation of the classical ERα action is a critical part of in the management of receptor-mediated disease conditions.

Human glutathione S-transferase P1-1 (GSTP; EC 2.5.1.18) catalyzes thioether conjugation of glutathione (GSH) with potentially toxic electrophile reactive intermediates, acting as the detoxifying enzyme [4]. In addition, GSTP has a diversity of regulatory functions such as JNK signaling pathways and Cdk5 kinase activity, and it is therefore involved in cell death and/or survival mechanisms [5,6]. Elevated GSTP expression of has been demonstrated in tumors and believed to play a role in drug resistance. For

Abbreviations: AF, activating function; AP, activating protein; ANOVA, analysis of variance; DMSO, dimethylsulfoxide; ERα, estrogen receptor α; ERE, estrogen response element; GSTP, human glutathione S-transferase P1-1; 4-OHT, 4-hydroxytamoxifen; IgG, immunoglobulin G; PBS, phosphate buffered saline; PVDF, poly vinylidene fluoride; RIP, regulatory interacting protein; SERM, selective estrogen receptor modulators; Sp, specificity protein; TBST, tris-buffered saline with 0.5% Tween-20.

* Corresponding author.
E-mail address: minsunchang@sm.ac.kr (M. Chang).

http://dx.doi.org/10.1016/j.bbrc.2014.09.017
0006-291X/

example, a study showed that GSTP expression indicates chemotherapy resistance and has been associated with a poor prognosis in breast cancers [7].

The recent studies show that the presence of GSTP predicts poor pathological complete response in ERα-negative breast cancer during neoadjuvant chemotherapy [8]. Various studies also indicate that GSTP expression accompanying loss of ERα at the either mRNA or protein level may contribute to poor prognosis in breast cancer and the mechanisms of drug resistance [9,10]. Although the regulation of GSTP at either the gene or protein level via ERα and its role in the development of endocrine therapy resistance remain unclear, GSTP expression is increased with decreased expression of ERα and has been associated with an altered response of selective estrogen receptor modulators (SERM) to ERα [11,12]. It was shown that GSTP protein expression occurs in the presence of ERα in drug-resistant cancer cell lines, implying that GSTP may play a role in the activation of ERα signaling leading to an abnormal response to chemotherapeutic agent such as tamoxifen.

In the present study, we demonstrate that the ectopic expression of functionally complete GSTP enhances the classical ERα signaling activities in ERα-positive but GSTP protein-negative human mammary epithelial cancer MCF7 cells. ERE-mediated ERα transcriptional activities were increased in the presence of GSTP and GSTP enzyme inhibitor decreased the ERE-mediated ERα transcriptional activities. One of molecules involved in GSTP-induced modulation of ERα signaling has been identified as receptor interacting protein 140 (RIP140). Our study proposes a novel non-enzymatic role of GSTP in regulating ERα classical signaling events, which supports the idea that GSTP has additional nonenzymatic functions other than regulation of kinase signaling pathways.

2. Materials and methods

2.1. Chemical and reagents

All chemicals and reagents were purchased from Sigma (St. Louis, MO) unless stated otherwise. Ezatiostat hydrochloride was purchased from APExBio (Boston, MA). The compounds were dissolved in DMSO and stored at −20 °C.

2.2. Cell culture

All cell culture reagents were purchased from Invitrogen (Grand Island, NY) unless stated otherwise. The MCF7 human breast cancer cell line (ATCC® HTB-22™) and HeLa human cervical cancer cell line (ATCC® CCL-2™) were obtained from American Type Culture Collection (Manassas, VA). Cells were routinely maintained in DMEM (HyClone, GE Healthcare Life Sciences, Logan, UT) containing 10% fetal bovine serum (FBS), 1% nonessential amino acids at 37 °C in 5% CO_2. Estrogen-free media were prepared by supplementing 3 × dextran-coated charcoal-treated FBS to phenol-red free DMEM (HyClone) while other components remained the same.

2.3. Plasmids and transfection

The sequences for the primers and siRNA oligonucleotides are listed in Table S1. The expression vector coding for GSTP (GenBank Accession No. NM_000852.3) was generated by inserting the 0.6 kbp long full-length GSTP into the EcoRI/BamHI sites of pIRES-neo vector (Clontech, Mountain View, CA) to produce the pIRES-neo/GSTP plasmid. The ERE-luciferase plasmid contains three copies of the *Xenopus laevis vitellogenin* A2 ERE upstream of the fire fly luciferase [13] and is a gift from Dr. V.C. Jordan (Lombardi Comprehensive Cancer Center, Georgetown University Medical Center, Washington, DC). The pSG5 plasmid containing the wild-type human ERα was originally engineered by Professor P. Chambon (Institut National de la Sante et de la Recherche Medicale, Strasbourg, France). Knockdown of the target gene was performed through the transfection of siRNA duplexes to cells. siRNA oligonucleotides with TT overhang to the end of 3′ position were purchased from Bioneer (Daejeon, South Korea). Scrambled inhibitory RNA (siControl) was derived from a message transcribed from the chloroplast genome of *Euglena gracilis* (GenBank Accession No. X70810; position 24750–24768). HeLa cells were incubated in estrogen-free media for 24 h before seeding in the 24-well plates. siRNA duplexes (30 nM) were transfected using Lipofectamine RNAiMAX reagent (Invitrogen) according to manufacturer's instruction. Six hours after transfection, the cells were put into fresh estrogen-free media and incubated for another 18 hr. ERE-luciferase plasmids (0.25 μg/well) were transfected using Lipofectamine 2000. MCF-7 or HeLa cells were cultured in estrogen-free DMEM media for 4 days before transfection and plated (1.5×10^5 cells/well) in triplicate in a 24-well plate. Cells were transiently transfected with pIRESneo/GSTP (0.5 μg/well) and ERE-luciferase plasmid (0.25 μg/well) and with an internal control plasmid pRL-Tk (0.1 μg/well) using Lipofectamine 2000 Reagent (Invitrogen).

2.4. Luciferase reporter assay

The cells were harvested with Passive Lysis buffer (Promega, Madison, WI). Luciferase activity present in the cell lysates was measured using the Dual Luciferase Assay kit (Promega) with a VICTOR3™ (PerkinElmer, Waltham, MA) in a 96 well plate format. Data are reported as relative luciferase activity (firefly luciferase reading divided by the Renilla luciferase reading).

2.5. Total RNA isolation and PCR analysis

The MCF7 cells grown in estrogen-free media were treated with appropriate test compounds or vehicle control. Total RNA was extracted using RNeasy kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. Total RNA (5 μg) was reverse transcribed with iScript cDNA synthesis kit (Bio-Rad, Hercules, CA). Real-time PCR was performed at 95 °C for 30 s, 60 °C for 30 s, and 72 °C for 30 s for 30–35 cycles (Palm-Cycler™, Corbett Life Science, Sydney, Australia) in a reaction mixture with a total volume of 10 μl using the real-time PCR SYBR green kit (QIAGEN). Rotor-Gene 6.1 software was used for the estimation of the CT parameter. The $2^{-\Delta\Delta}$CT method was applied for the quantification of the gene of interest. Data normalization was performed by dividing the expression level of a *RIP140* gene by that of *18S rRNA*. Finally, results were expressed as fold-induction where the gene expression level in vehicle control-treated cells was set as 1.

2.6. Western blot

Cells were trypsinized, pelleted, washed in PBS, resuspended in lysis buffer (50 mM HEPES, 150 mM NaCl, 1 mM EDTA, 2.5 mM EGTA, 10 mM β-glycerophosphate, 10% glycerol, and 0.5% NP-40, pH 8.0) containing protease inhibitors (Complete protease inhibitor cocktail tablets, Roche, Indianapolis, IN), mixed, and centrifuged at 12,000×*g* for 10 min. Protein concentrations were measured by Bradford assay (Bio-Rad). Equal amounts of protein were electrophoresed on a 10% polyacrylamide gel and transferred to a PVDF membrane using Trans-Blot Turbo transfer system (Bio-Rad). The membrane was blocked in TBST containing 10% nonfat dry milk at 4 °C overnight and then incubated for 2 h with anti-human GSTP (3F2) mouse antibody (dilution 1:1000, Cell Signaling Technology, Danvers, MA), rabbit anti-RIP140 antibody (dilution 1:200, Santa Cruz Biotechnology, Dallas, TX) or rabbit anti-ERα

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 과한 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 소정의 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 총장 의 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.