http://dx.doi.org/10.5392/JKCA.2012.12.12.489

소셜 컴퓨팅을 위한 연구 · 학습 주제의 계층적 지식기반 구축

Building Hierarchical Knowledge Base of Research Interests and Learning Topics for Social Computing Support

김선호, 김강회, 여운동
한국과학기술정보연구원

Seonho Kim(haebang@kisti.re.kr), Kang-hoe Kim(kimkh@kisti.re.kr),
Woondong Yeo(wdyeo@kisti.re.kr)

요약

본 논문은 연구·학습 주제 지식베이스를 통한 소셜컴퓨팅 지원에 관한 연구로 두 가지 하부 연구로 구성되었다. 첫 번째 연구는 다양한 학문분야에서 전자 도서관 이용자들의 연구 및 학습 주제를 추출하기 위해 분야별로 분류가 잘 되어 있는 NDLTD Union catalog의 석박사 학위 논문 (Electronic Theses and Dissertations : ETDs)을 분석하여 계층적 지식베이스를 구축하는 연구이다. 석박사 학위 논문 이외에 ACM Transactions 저널의 논문과 컴퓨터 분야 국제 학술대회 웹사이트도 추가로 분석하였는데 이는 컴퓨팅 분야의 보다 세분화된 지식베이스를 얻기 위해서이다. 계층적 지식베이스는 개인화 서비스, 추천 시스템, 텍스트 마이닝, 기술기회탐색, 정보 가시화 등의 정보서비스와 소셜컴퓨팅에 유용하게 사용될 수 있다. 본 논문의 두 번째 연구 부분에서는 우리가 만든 계층적 지식기반을 활용하여 4개의 사용자 커뮤니티 마이닝 알고리즘 중에서 우리가 수행중인 소셜 컴퓨팅 연구, 즉 구성원간의 결합도에 기반한 추천시스템,에 최상의 성능을 보이는 그룹핑 알고리즘을 찾는 성능 평가 연구 결과를 제시하였다. 우리는 이 논문을 통해서 우리가 제안하는 연구·학습 주제 데이터베이스를 사용하는 방법이 기존에 사용자 커뮤니티 마이닝을 위해 사용되던 비용이 많이 필요하고, 느리며, 개인정보 침해의 위험이 있는 인터뷰나 설문에 기반한 방법을 자동화되고, 비용이 적게 들고, 빠르고, 개인정보 침해 위험이 없으며, 반복 수행시에도 일관된 결과를 보여주는 방법으로 대체할 수 있음을 보이고자 한다.

■ 중심어 : | 계층적 지식기반 | 관심 연구분야 | 관심 학습분야 | 분류 | 소셜 컴퓨팅 |

Abstract

This paper consists of two parts: In the first part, we describe our work to build hierarchical knowledge base of digital library patron's research interests and learning topics in various scholarly areas through analyzing well classified Electronic Theses and Dissertations (ETDs) of NDLTD Union catalog. Journal articles from ACM Transactions and conference web sites of computing areas also are added in the analysis to specialize computing fields. This hierarchical knowledge base would be a useful tool for many social computing and information service applications, such as personalization, recommender system, text mining, technology opportunity mining, information visualization, and so on. In the second part, we compare four grouping algorithms to select best one for our data mining researches by testing each one with the hierarchical knowledge base we described in the first part. From these two studies, we intent to show traditional verification methods for social community miming researches, based on interviewing and answering questionnaires, which are expensive, slow, and privacy threatening, can be replaced with systematic, consistent, fast, and privacy protecting methods by using our suggested hierarchical knowledge base.

■ keyword : | Hierarchical Knowledge Base | Research Interests | Learning Topics | Classification | Social Computing |

접수번호 : #121022-002
접수일자 : 2012년 10월 22일
심사완료일 : 2012년 11월 29일
교신저자 : 여운동, e-mail : wdyeo@kisti.re.kr

I. Introduction

Research interests and learning topics of patrons in complex information systems, such as digital libraries, are essential information to provide them with advanced personalized services. Although this information usually is provided by the users explicitly, e.g., through filling out registration forms and answering online surveys and questionnaires, some researchers have emphasized the importance of implicit ways of collecting the information, such as collecting information from the data they used. Kelly, in her dissertation[2], tried to figure out users' document preferences based on their implicit feedback tendencies. Nichols[3] and GroupLens[5] showed the great potential of implicit data from users. Human behavior analysis and modeling in information systems, such as by Fidel[12], Belkin[7], Ha[14] and Lin[19], also are techniques to characterize users and communities for personalized service. One of the goals of this paper is to present our effort to build a collection, namely hierarchical knowledge base, of the research interests and learning topics, which is the set of implicit rating data, by analyzing the data they may use, such as ETD, journals, and web pages. Another goal of this paper is to introduce a result of case study of utilizing the collection for social computing support.

As the research interests and learning topics of a person are normally represented by a set of noun phrases, for example, "digital library", "technology opportunity detection", and "web log standardization", the process of building the hierarchical knowledge base is mainly depends on the process of noun phrase extraction. Anick[9] emphasized several reasons that noun phrases are important in information retrieval field, such as the succinctness of describing concepts, the easiness of detecting and extracting, the tightness of expression, the tendency of appearing in actual queries, etc. Therefore, we will identify a digital library patrons' research interests and learning topics by examining the noun phrases extracted from documents she referenced in the ETD, journals, and web pages.

More, in the case study, we use the collection of research interests and learning topics as the standard data set for representing individual users and user communities. As the collection is employed as the standard research interests and learning topics data, it is possible to evaluate the accuracy or the performance of user community finding algorithms in fast and objective manner. Proposed method can substitute the traditional interview based methods of user community findings, which are expensive, time consuming, and potentially invading privacy.

II. Our Approach

The overall goal of this study is to build a collection of research interests and learning topics for various scholarly areas through analyzing well-classified documents, and to make it useful by storing it in a database and providing tools, such as Java Application Programming Interfaces (APIs) and web-based user interfaces. [Figure 1] depicts the overview of all stages to achieve the goal. In stage (1), a large set of documents is collected. This set includes the Networked Digital Library of Theses and Dissertations (NDLTD) ETD Union Catalog[8], which is a collection of ETD metadata provided by 325 member institutions, such as universities and libraries. Although this collection covers all scholarly areas, we added the ACM Digital Library (DL) Transactions[1] and 70 conference web pages in computing field to our data analysis to help guide our focus on computing, because according to our preliminary research[18], many users and information demands in our target digital library, NDLTD, are in the computing field. The NDLTD ETD Union Catalog is collected by using the OAI/ODL Harvester 2.0[4], and the ACM DL Transactions are crawled with SPHINX[13]. The conference web pages are collected manually from the IEEE conference calendar, google searches, and so on. NDLTD ETDs are classified into 76 categories of research interests and learning topics, henceforth referred to as “categories”. ACM Transactions are selected because they are already well-classified into 26 specific topics of computer science and engineering; see the row labeled 8 in [Table 1]. Conference web pages are also classified into 26 in accordance with the ACM Transactions.

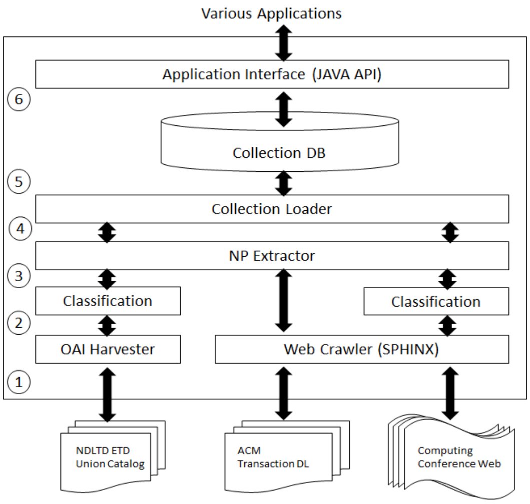


Figure 1. Workflow to build a collection of research interests and learning topics

In stage (2), 102 categories are predefined for classification as shown in [Table 1]. The method we used to classify the documents is specified in the next section.

In stage (3), a noun phrase extractor, developed at the University of Arizona, is employed to extract meaningful noun phrases that represent research interests and learning topics. The noun phrase extractor is tuned to focus on title and subject elements of each ETD's metadata, see [Figure 2], because these parts were entered by its author directly and contain carefully selected phrases that are expected to represent the ETD. The description element, which contains the abstract of the paper, was excluded because it turned out that including the description just adds noise, reducing the accuracy of the whole result. For documents of ACM Transactions, title and keywords fields were analyzed.

〈dc oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" dc="http://purl.org/dc/ elements/1.1/" xsi="http://www.w3.org/2001/XMLSchema-instance" schemaLocation ="http://www.openarchives.org/OAI/2.0/oai_dc/ http://www.openarchives.org/OAI/ 2.0/oai_dc.xsd"〉〈title〉Composer-Centered Computer-Aided Soundtrack Composition〈/title〉〈creator〉Vane, Roland Edwin〈/creator〉〈subject〉Computer Science〈/subject〉〈subject〉human computer interaction〈/subject〉〈subject〉 music composition〈/subject〉〈subject〉soundtracks〈/subject〉〈subj ect〉creativity 〈/subject〉〈description〉For as long as computers have been around, people have looked for ways to involve them in music….〈/description〉〈publisher〉University of Waterloo〈/publisher〉〈date〉2006〈/date〉〈type〉Electronic Thesis or Dissertation 〈/type〉〈format〉application/pdf〈/format〉〈identifier〉http://etd .uwaterloo.ca/etd/revane2006.pdf〈/identifier〉〈language〉en 〈/language〉〈rights〉Copyright: 2006

Figure 2. An example of ETD metadata

In stages (4) and (5), the results of the document classifier and noun phrase extractor were classified into research interests and learning topics. They were merged and converted to SQL queries, and entered into the database. [Figure 3] is a logical view of the collection database after stage 5. Research interests and learning topics are stored with their category names along with their frequency in the source documents. Because this database is hierarchical and each node of it consists of research interests and learning topics, it is possible to measure the semantic distances of between two topics using this database.

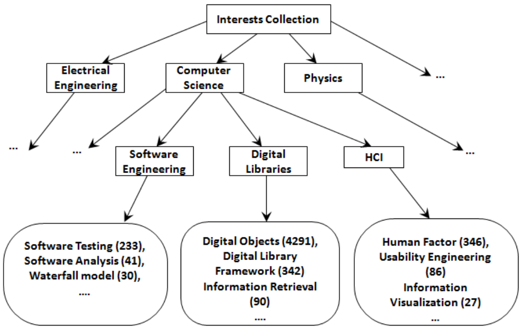


Figure 3. Hierarchical structure of the collection

To make the use of this collection convenient, several Java APIs are implemented for developers and web applications, at stage 6. These APIs include several presentation functions, such as an extraction function dealing with statistical information about the collection, interactive functions implemented with SQL to manipulate the collection, and calculation functions to estimate categories for given user data.

Ⅲ. Document Classification and Noun Phrasing

ACM Transaction documents are already well-classified according to 26 specific topics in the computing fields, but, classifying the NDLTD ETD union catalog is a challenge. Each ETD record is in XML format and contains meta-information describing the document, such as Dublin Core information, title, subjects, description, publisher, date, etc, like the sample given in [Figure 2]. Classification of an ETD was achieved by taking advantage of classification information entered by its author, subject elements, and using a keyword-mapping-based ad-hoc algorithm for the title and subject elements.

The conference web pages in computing field is manually classified according to the classification of ACM Transaction. Classification is targeted to classify the source documents into a predefined set of 102 categories, including categories from the ACM DL, as listed in [Table 1]. The category names and keyword sets for each category are identified by surveying faculty/college systems of universities. This table has an extended categories from our preliminary work[18][1].

Table 1. 102 categories of research interests and learning topics

	8 high levels categories	102 categories
1	Architecture and Design	Architecture and Construction, Landscape and Architecture
2	Law	Law
3	Medicine, Nursing and Veterinary Medicine	Dentistry, Medicine, Nursing, Pharmacy, Veterinary
4	Arts and Science	Agriculture, Animal and Poultry, Anthropology, Apparel and Housing, Archaeology, Art, Astronomy, Biochemistry, Biology, Botany, Chemistry, Communication, CropSoil and Environment Sciences, Dairy Science, Ecology, Engineering Science, English, Entomology, Family, Food, Foreign Language Literature, Forestry, Geography, Geology, Government International Affair, History, Horticulture, Hospitality Tourism, Human Development, Human Nutrition and Exercise, Informatics, Interdisciplinary, Library Science, Linguistics, Literature, Meteorology, Mathematics, Music Naval, Philosophy, Physics, Plant, Politics, Psychology, Public Administration Policy, Public Affair, Sociology, Statistics, Urban Planning, Wildlife, Wood, Zoology
5	Engineering and Applied Science	Aerospace, Biological Engineering, Chemical, (Computer Science and Engineering)*, Electronics, Environment, Industrial, Materials, Mechanics, Mining and Mineral, Nuclear, Ocean Engineering
6	Business and Commerce	Accounting and Finance, Business, Economics, Management
7	Education	Education

1 This work aimed to reveal the amount of information supply and demand in each scholarly field in NDLTD, by adding the ACM DL collection to specify the computing field. The goal of classification in this study is finding representative documents for each category, rather than classifying all documents precisely. Therefore, ambiguous or inter-category documents were excluded.

8	Computer Science and Engineering	Algorithms, Applied Perception, Architecture and Code Optimization, Asian Language Information Processing, Autonomous and Adaptive Systems, Computational Logic, Computer Systems, Computer Human Interaction, Database, Automation of Electronic Systems, Embedded Computing, Graphics, Information Systems, Information and System Security, Internet Technology, Knowledge Discovery, Mathematical Software, Modeling and Simulation, Multimedia, Programming Language, Sensor Networks, Software Engineering, Speech and Language Processing, Storage, Bioinformatics, Networking

* The Computer Science and Engineering category is expanded in row 8.

The next step is extracting noun phrases from the classified documents. University of Arizona's AZ Noun Phraser[6], is used and tuned to exclude out-of-interest nouns. Along with noun phrases, term frequency, *tf*, is extracted to obtain the frequency in the whole document set, Σtf. Brief statistics of source data, as well as results of classification and noun phrasing, are given in [Table 2].

Table 2. Source data for analysis

Source	Number of Records	Number of Categories	Number of Distinct Noun Phrases Extracted
NDLTD ETD Union Catalog	242,688	76	72,388
ACM DL	3395	26	1,325
Conference Web pages in Computing Field	70	26	352

Ⅳ. A Case Study: A Standard Data Set for Measuring Accuracy of Community Finding Algorithms

A major application of the first part of our study is an evaluation tool for user clustering, or community finding, which is an essential procedure for almost all personalization techniques and recommender systems. A social network mining tools, such as VUDM[17], Visual User model Data Mining tool, provides both an overview and details of users, user communities, and usage trends of digital libraries. [Figure 4] shows how much the result of VUDM depends on various community finding algorithms. Three figures present user model data from 1,100 users, generated by the NDLTD user tracking system[16][17] during the period from September 2005 to June 2006. Users and user communities are represented with icons and spirals, respectively. The distances between user icons, or spirals, represent how similar they are. There are many user clustering methods studied, but, there was no known method to evaluate the performance, and accuracy, of community finding algorithms, except asking the users directly, or

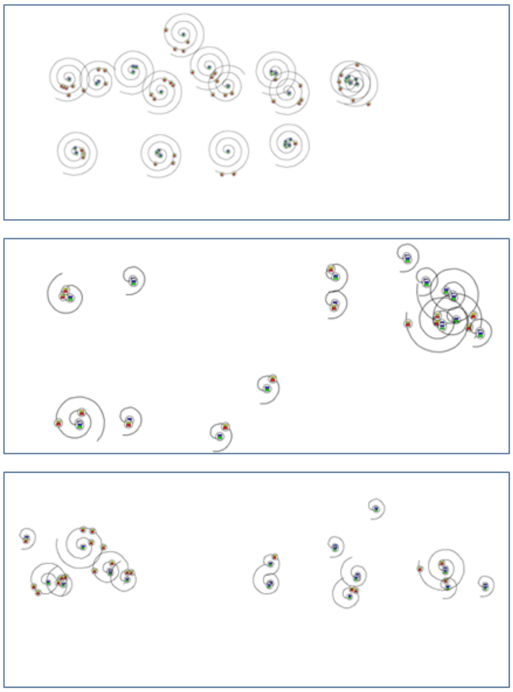


Figure 4. Visualization of user communities in the NDLTD collection. User communities were found by *kNN*(top), FWSMC (middle), and Tanimoto-Single (bottom) algorithm.

conducting a user survey, whether it is online or offline. Requiring users to enter information explicitly is expensive, and hard to achieve when the communities are large. In this section, we describe an experiment we conducted to see how the collection of research interests and learning topics could be used as an evaluation tool for measuring the accuracy of four community finding algorithms: *kNN*, Fixed Window Size Multi-Classification (FWSMC), Tanimoto - Single, and Cosine - Single

- *kNN*: Similarity is based on the cosine correlation. We find the user community in the nearest k neighbors of a given user. Performance depends on then number of users in the whole system[11].
- FWSMC: Similarity is based on cosine correlation. In contrast to *kNN*, we find the user community within a given threshold of grouping strictness. Performance depends on the given threshold[15].
- Tanimoto - Single: Similarity is based on simple set comparison and single linkage clustering is used. This has a low computational cost[11].
- Cosine - Single: Cosine similarity is based on the popular vector product calculation derived from the Euclidean dot product and single linkage clustering is used[10].

In order to measure the accuracy of user grouping algorithms, we implemented a function, which is based on the cosine coefficient, that estimates categories to which a given user belongs, and their probabilities. Using this function, the accuracy of a user grouping algorithm is measured by calculating the average of the overlap ratio of matched categories among users within a group, see Equation (1). If all members in each group are estimated to belong in the same categories, this value will be 1, otherwise it will approach 0 as the number of matched categories is reduced.

C_n = <Vector> Estimate Categories (User Model n);

$$Accuracy = \frac{G}{\sum_{g \in G} \cup_{n \in g} C_{n,w}} \tag{1}$$

In this equation, G is the set of all interest sharing groups among users, and $|G|$ is the cardinality of set G, which is the total number of elements of set G. g represents an interest sharing group in G, and n represents a user in group g. $C_{n,w}$ is the first w most highly rated research interests and learning topics categories, i.e., “categories”, of user n, where w is the size of the window. The size, or tolerance, of the window is the range of categories considered meaningful, from the estimated list of belonged to categories. Two users are considered to be grouped correctly if both share a category within the tolerant window of size w within their estimated categories. For example, let user a’s categories and user b’s categories be estimated as { c1, c2, c3, c4 }, and { c3, c1, c2, c4, c99 } in descending order of probability, respectively. Then, if w is 1, user a and b will not be considered to be sharing any research interests and learning topics because their first estimated categories c1 and c3 is not the same, and we assume only that the first ranked categories are meaningful. However, if w is 2, they will be considered as sharing some research interests and learning topics because c1 exists in both users’ first two estimated categories. In this paper, we set w to the value 2.

V. Results

[Figure 5] shows the accuracy values for four community finding algorithms as the number of discovered user communities changes. *kNN* performed excellently when it found more than 150 user communities, however, in that case, most groups consist of just one or two users, which is meaningless. Tanimoto – Single showed good performance when the number of communities is small, but the accuracy deteriorates as the number of communities increase. Cosine – Single showed good performance, but the accuracy, when the number of communities is small, was relatively poor than FWSMC. The FWSMC was the most accurate algorithm among the tested algorithms, and shows stable performance regardless of the change in number of communities.

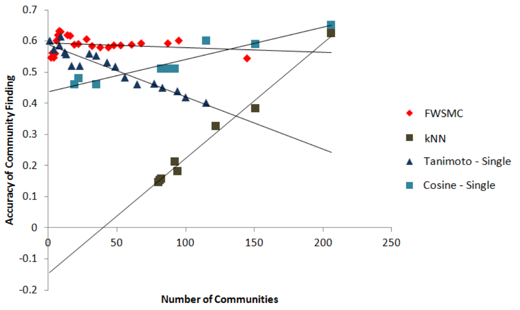


Figure 5. Accuracies of four community finding algorithms.

The scatter plot in [Figure 6] shows the change in the accuracy of the FWSMC and Tanimoto – Single algorithms relative to changes of the value “strictness of community finding”, which is a threshold used to decide whether two users should be grouped or not. This value is entered through a slide bar in the VUDM user interface for interactive filtering. *kNN* is not considered because it instead uses the k-value for the threshold. In this figure, the accuracy of Tanimoto – Single increases as the value “strictness of community finding” increases, while the accuracy of FWSMC is high and doesn’t exhibit any significant change. From this result, we see that both algorithms are selectable for two different intentions of filtering, one to control the number of communities by changing the strictness of community finding, and the other to control the number of communities while maintaining a high level of overall accuracy.

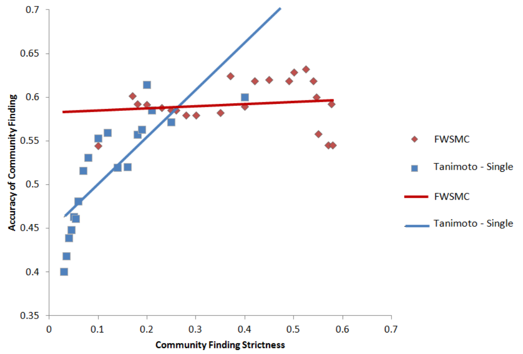


Figure 6. A comparison of accuracies of two community finding algorithms with regard to the change of the strictness of grouping.

Ⅵ. Conclusions and Future Works

We built a collection of research interests and learning topics of various scholarly areas, for social network research support, by classifying and by identifying noun phrases from a large well-classified scientific collection, such as NDLTD Union catalog, journal articles, and conference web pages. For journal articles, the ACM Transactions are selected to put our focus on the computing fields. For classification, 102 categories, includnig76 categories from non-computing fields and 26 categories from computing fields, have been identified and used. Collected noun phrases are entered into a database and Java APIs are developed to facilitate the use of the collection. We also described a case study, an experiment we conducted, on utilizing the output of the study, the collection of research interests and learning topics, for social network research. In the experiment, the collection is used as an evaluation tool, a standard data set, to compare the accuracy and related characteristics of four user community finding algorithms. This experiment shows the traditional expensive, slow, and privacy threating methods of social computing research, which are mainly based on interviewing and answering questionnaires, can be replaced by systematic, fast, cheap, and consistent method by using the collection of research interests and learning topics as the evaluation tool.

In our planned future study, we will expand the categories in detail for all scholarly areas. The experiment also needs to be extended with a larger amount of user data to verify that our result is independent of the size of data. Additional variables should be tested, such as: number of users, strictness of community finding, k-value of *kNN*, prefixed preferred group size, tolerant window size, length of noun phrases extracted, etc. More, the potential application of our collection, such as recommender system, mining technology opportunity for SME(small and medium enterprises), detecting concept drifts of information service patrons, visualizing document topics, or visualizing communities based on interests and topics, will be studied.

Ⅶ. Acknowledgements

We thank the: people and organizations working on NDLTD, developers of the AZ Noun Phraser, and the NDLTD users who answered our special survey.

참 고 문 헌

[1] ACM Transactions, ACM DL: Transactions, http://portal.acm.org/browse_dl.cfm?linked=1&part=transaction&coll=ACM&dl=ACM&CFID=19463219&CFTOKEN=17951970, 2007.

[2] D. Kelly, "Understanding Implicit Feedback and Document Preference: A Naturalistic User Study," Doctoral dissertation, Rutgers University, 2004.

[3] D. Nichols, "Implicit Rating and Filtering," Proceedings of the 5th DELOS Workshop on Filtering and Collaborative Filtering, Budapest, Hungary, pp.31-36, 1998.

[4] H. Suleman, "Introduction to the Open Archives Initiative protocol for metadata harvesting," Proceedings of the ACM/IEEE 2nd Joint Conference on Digital Libraries (JCDL 2002), Portland, OR, Software available at http://oai.dlib.vt.edu/odl/software/harvest/, 2002.

[5] J. Konstan, B. Miller, D. Maltz, J. Herlocker, L. Gordon, and J. Riedl, "GroupLens: Applying Collaborative Filtering to Usenet News," In Communications of the ACM, Vol.40, No.3, pp.77-87, 1997.

[6] K. Tolle and H. Chen, "Comparing noun phrasing techniques for use with medical digital library tools," Journal of American Society for Information Science, Vol.51, No.4, pp.352-370, 2000.

[7] N. Belkin, "Interaction with Texts: Information Retrieval as Information-Seeking Behavior," In Information Retrieval, pp.55-66, 1993.

[8] NDLTD, Networked Digital Library of Theses and Dissertations, http://www.ndltd. org, 2007.

[9] P. Anick and S. Vaithyanathan, "Exploiting Clustering and Phrases for Context-based

Information Retrieval," Proceedings of the 20th annual international ACM SIGIR conference on research and development, Philadelphia, PA, pp.314-323, 1997.

[10] R. Baeza-Yates, *Modern Information Retrieval*, Addison Wesley, Santiago, Chile, 1999.

[11] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, A Wiley-Interscience Publication, 2000.

[12] R. Fidel and A. Pejtersen, "From Information Behaviour Research to the Design of Information System: the Cognitive Work Analysis Framework," *Information Research*, Vol.10, No.1, 2004.

[13] R. Miller and K. Bharat, "SPHINX: a framework for creating personal, site-specific web crawlers," Proceedings of the 7th World-Wide Web Conference, 1998.

[14] S. Ha, K. Im and H. Pae, "Comparison of Online Game User Communities by using Social Network Analysis," The Journal of the Korea Contents Association, Vol.9, No.8, pp.178-189, 2009.

[15] S. Kim, E. Fox, W. Fan, C. North, D. Tatar, and R. Torres, "Design and Evaluation of Techniques to Utilize Implicit Rating Data in Complex Information Systems," Computer Science Technical Report, Blacksburg, VA, Virginia Tech, http://eprints.cs.vt.edu/archive/00000980/, TR-07-20, 2007.

[16] S. Kim, NDLTD, "Search Interface Embedded User Tracking System," available at http://boris.dlib.vt.edu:8080/controller/ index.jsp, 2006.

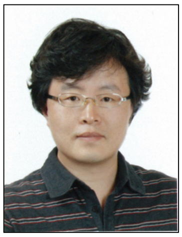
[17] S. Kim, S. Lele, S. Ramalingam, and E. Fox, "Visualizing User Communities and Usage Trends of Digital Libraries based on User Tracking Information," Proceedings of the 9th International Conference on Asian Digital Libraries (ICADL '06), Kyoto, Japan, Lecture Notes in Computer Science 4312, Springer-Verlag, Berlin Heidelberg New York, pp.111-120, 2006.

[18] S. Kim, S. Yang, and E. Fox, "Supply and Demand Analysis in NDLTD Based on Patron Specialty Survey and Contents Statistics," Proceedings of the 9th International Symposium on Electronic Theses and Dissertations (ETD '06), http://www6.bibl.ulaval.ca:8080/etd2006/pages/prog.jsf, 2006.

[19] S. Lin and N. Belkin, "Validation of a model of information seeking over multiple search sessions," Journal of American Society for Information Science & Tech.(JASIST), Vol.56, No.4, pp.393-415, 2005.

저 자 소 개

김 선 호(Seonho Kim) 정회원



- 1995년 2월 : 연세대학교 전산학과(공학사)
- 1997년 2월 : 포항공과대학교 정보통신학과(공학석사)
- 1997년 3월 ~ 현재 : 한국과학기술정보연구원(선임연구원)
- 2008년 5월 : 미국 버지니아공과대학교 전산학과(공학박사)

<관심분야> : 데이터마이닝, 디지털콘텐츠, 미래유망기술탐색, 자연언어처리, 인공지능

김 강 회(Kang-hoe Kim) 정회원

- 2002년 8월 : 한국외국어대학교 무역학과(석사)
- 2006년 8월 : 한국외국어대학교 무역학과(박사수료)
- 2001년 1월 ~ 현재 : 한국과학기술정보연구원 책임연구원

<관심분야> : 국제공동연구, 정보분석기법, 비즈니스 인텔리전스 연구

여 운 동(Woondong Yeo) 정회원

- 2002년 2월 : 경북대학교 전자공학과(공학석사)
- 2009년 8월 : 고려대학교 컴퓨터학과(박사수료)
- 2002년 4월 ~ 현재 : 한국과학기술정보연구원 선임연구원

<관심분야> : 데이터마이닝, 과학기술 계량정보분석