Investigation of Topic Trends in Computer and Information Science by Text Mining Techniques: From the Perspective of Conferences in DBLP*

텍스트 마이닝 기법을 이용한 컴퓨터공학 및 정보학 분야 연구동향 조사: DBLP의 학술회의 데이터를 중심으로

Su Yeon Kim (김수연)**
Sung Jeon Song (송성전)***
Min Song (송민)****

ABSTRACT

The goal of this paper is to explore the field of Computer and Information Science with the aid of text mining techniques by mining Computer and Information Science related conference data available in DBLP (Digital Bibliography & Library Project). Although studies based on bibliometric analysis are most prevalent in investigating dynamics of a research field, we attempt to understand dynamics of the field by utilizing Latent Dirichlet Allocation (LDA)-based multinomial topic modeling. For this study, we collect 236,170 documents from 353 conferences related to Computer and Information Science in DBLP. We aim to include conferences in the field of Computer and Information Science as broad as possible. We analyze topic modeling results along with datasets collected over the period of 2000 to 2011 including top authors per topic and top conferences per topic. We identify the following four different patterns in topic trends in the field of computer and information science during this period: growing (network related topics), shrinking (AI and data mining related topics), continuing (web, text mining information retrieval and database related topics), and fluctuating pattern (HCI, information system and multimedia system related topics).

초 록

이 논문의 연구목적은 컴퓨터공학 및 정보학 관련 연구동향을 분석하는 것이다. 이를 위해 텍스트마이닝 기법을 이용하여 DBLP(Digital Bibliography & Library Project)의 학술회의 데이터를 분석하였다. 대부분의 연구동향 분석 연구가 계량서지학적 연구방법을 사용한 것과 달리 이 논문에서는 LDA(Latent Dirichlet Allocation) 기반 다항분포 토픽 모델링 기법을 이용하였다. 가능하면 컴퓨터공학 및 정보학과 관련된 광범위한 자료를 수집하기 위해서 DBLP에서 컴퓨터공학 및 정보학과 관련된 353개의 학술회의를 수집 대상으로 하였으며 2000년부터 2011년 기간 동안 출판된 236,170개의 문헌을 수집하였다. 토픽모델링 결과와 주제별 문헌 수, 주제별 학술회의 수를 조사하여 2000년부터 2011년 사이의 주제별 상위 저자와 주제별 상위 학술회의를 제시하였다. 주제동향 분석 결과 네트워크 관련 연구 주제 분야는 성장 패턴을 보였으며, 인공지능, 데이터마이닝 관련 연구 분야는 쇠퇴 패턴을 나타냈고, 지속 패턴을 보인 주제는 웹, 텍스트마이닝, 정보검색, 데이터베이스 관련 연구 주제이며, HCI, 정보시스템, 멀티미디어 시스템 관련 연구 주제 분야는 성장과 하락을 지속하는 변동 패턴을 나타냈다.

Keywords: text mining, topic modeling, DMR, topic dynamics, research trend
텍스트 마이닝, 토픽 모델링, 토픽 다이나믹스, 연구동향

---

* This paper is an extended version of 8th International Conference on WIS & 13th COLLNET Meeting presentation.
This work was supported (in part) by the Yonsei University Research Fund of 2014.
** 연세대학교 박사 후 연구원(suyeon@yonsei.ac.kr) (제1저자)
*** 연세대학교 문헌정보학과 대학원생(sj.song@yonsei.ac.kr)
**** 연세대학교 문헌정보학과 부교수(min.song@yonsei.ac.kr) (교신저자)
■ 논문접수일자 : 2015년 2월 24일 ■ 최초심사일자 : 2015년 2월 25일 ■ 게재확정일자 : 2015년 3월 16일
■ 정보관리학회지, 32(1), 135-152, 2015. 〔http://dx.doi.org/10.3743/KOSIM.2015.32.1.135〕

1. Introduction

The field of Information Science has been developed into a cross-disciplinary field. There are several definitions that help us understand the field. One of them is provided by Buckland (2012) who points out that "Information science" has been used to describe various fields such as library and information science, computer science, and information technology. According to Merriam-Webster and American Heritage Dictionary (http://www.merriam-webster.com), Information Science is defined as an "interdisciplinary field that is primarily concerned with analysis, collection, classification, manipulation, storage, retrieval and dissemination of information". Due to the flux of electronic information over the Internet, Information Science has become a pivotal field for managing and processing information.

Computer science includes a variety of interdisciplinary sciences such as computer system, data mining, and artificial intelligence. Computer science shares the common research interests with Information science. Furthermore, these two disciplines tend to publish new research ideas actively in invisible college venues, i.e., conferences or workshops. Therefore we focus on conference data instead of journal data to identify of knowledge structures of the disciplines and to analyze topic trends over time.

Studying the knowledge structure of a research field is a realm of bibliometrics that conducts the quantitative evaluation of scientific literatures. Using the bibliometric approach, knowledge representation at the level of scientific specialty, intellectual structures, informal and formal networks in both natural and social science were investigated. Janssens et al. (2008) combined bibliometrics with textual content to detect the emerging fields or hot topics in Information Science. The previous studies focus on journals to investigate the knowledge structure of Information Science. In addition, the approaches employed by the previous studies were primarily based on bibliometric analysis rather than content analysis. One of the daunting challenges in bibliometrics is the mapping of the epistemological structure of a given field. In particular, monitoring the dynamics and evolution of the revealed structure of a field has become a focal interest (Glänzel, 2012).

Unlike previous approaches, this paper employs text mining techniques to explore the knowledge structure of the field of computer and information science and topic trends. To this end, we collect 236,170 documents related to computer and information science from 353 conferences in DBLP, which conference data is in form of XML (Extensible Markup Language). The reason for our interests in computer and information science related conferences is because conferences tend to handle the state-of-the-art techniques and trends in a given field. The text mining technique that we have adopted in this paper is Latent Dirichlet Allocation (LDA)-based multinomial topic modeling (a.k.a. Dirichlet Multinomial Regression (DMR)). The main purpose of applying this technique to map the field is to look at the trends and dynamics of computer and information science instead of relying on bibliometric analysis. Analysis based on text mining techniques,

particularly DMR based topic modeling, gives us an insight into understanding the dynamic nature of the field. We identify four different patterns in topic trends in the field between 2000 and 2011: growing, shrinking, continuing, and fluctuating. In addition, we select top authors and conferences for each topic based on content similarity between papers published in conferences and key terms in topics as well as papers written by authors and key terms in topics.

The rest of our paper is organized as follows. Section 2 states the related work. Section 3 describes the proposed approaches and techniques to understanding the field of computer and information science. Section 4 reports the results of the experiments. Section 5 analyzes the similarity between the topics and identifies the topic trend. Section 6 concludes the paper.

2. RELATED WORK

Identifying and predicting the trend or the dynamics in a certain field are getting more spotlights as technology development and the following social changes become faster. Identifying the trend of a field from texts can be done by three different methods: (1) a clustering based method that groups similar documents and looks for the trend by keywords of clusters, (2) a network analysis based method that composes keywords or subject terms and compares those of one year with those of another year, (3) a statistics based method that builds a statistical model to predict the trend. Among these methods, a graph theory based method that constructs a network to analyze the trend was popular. However, a limitation still remains. The limitation is that the network does not represent the features of the topics in the field but only shows the relations between keywords or subject terms simply.

2.1 Clustering based

There are various approaches to studying specialties or dynamics of a particular field. The first approach is based on clustering techniques. This approach, which has been applied for several decades, was used to cluster documents based on the content similarity. Recent studies in this line of research have gone beyond proposing new clustering algorithms and have rather focused on extracting labels and keywords to represent clusters. Researches on labeling clusters adopt various techniques such as unsupervised learning (Cutting et al., 1993; Matsuo & Ishizuka, 2004; Treeratpituk & Callan, 2006), supervised learning (Frank et al., 1999; Hulth, 2004; Kerner et al., 2005; Liu et al., 2008), and graph-based techniques (Brin & Page, 1998; Wan et al., 2007; Liu et al., 2010).

2.2 Network analysis based

The second approach to understanding dynamics of a given research field is based on network analysis. The majority of studies uses citation network for identifying the paradigm shift of researchers, research trends, and the development patterns of a given field.

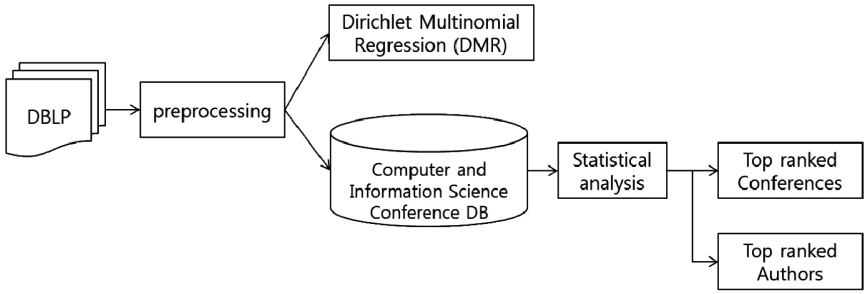
White and McCain (1998) used co-citation analysis to analyze the information science field. Chen and Carr (1999) examined the topic changes of the hypertext field from 1980 to 1998 through co-citation analysis. Xu et al. (2004) formed the criminal network and applied Social Network Analysis (SNS) to the network for identifying crime density, relatedness, etc. over time. Adamic and Adar (2005) analyzed information flow and predicted the friendship among people from their email network.

2.3 Statistical model based

The third approach is based on statistical analysis to understand dynamics of a particular field over time. A well-known technique in this category is Markov assumption. One of the examples is Hidden Markov Model (HMM) and Kalman filter. HMM was also used for ranking algorithms in information retrieval. Kleinberg (2003) used HMM to develop "burst of activity model" that focuses on pattern identification over time. Burst of activity model used outstanding features over time to identify emerging subject fields based on a probabilistic model. In his study, burst is employed to Topic Detection and Tracking (TDT) to understand dynamics of topics by spotting high occurrence of topics in a short period of time. He and Parker (2010) applied improved burst detection to analyze PubMed and MeSH. Recently, there have been rigorous studies applying Latent Dirichlet Allocation (LDA) to understand dynamics of a field LDA, introduced by Blei et el. (2003), a framework of extension of Bayesian network to three tiers, and has been applied to various areas. An understanding of dynamics of a field over time is one of the areas LDA is applied to. Early studies of using LDA arrange documents in chronological order and identify the topical changes of each subset of a document collection after dividing it into subsets (a.k.a. slicing). For example, Griffiths and Steyvers (2004) applied topic modeling to papers published in PNAS proceeding by year and select hot and cold topic based on topical distribution by each year. Wang et al. (2005) proposed and applied the group topic model, a variation of LDA, to U.N. voting records from 1960 to 2000 to examine the topic shift per unit. Blei and Lafferty (2006) proposed dynamic topic models to identify topic changes over time. The dynamic topic model extends the LDA framework to Kalman filter to enable to understand topic changes over time. Wang et al. (2012) proposed temporal topic models that can be applied to a continuous time range which is based on the hyper-parameter of the previous time applied to the following time period. None of studied has applied DMR, which we employ in this paper, to computer and information science field to detect topic trends over time.

3. Proposed Approach

To map the field of computer and information science, several steps are taken as shown in Fig. 1. Step 1 parses DBLP records in XML form (http://www.informatik.uni-trier.de/~ley/db/). The details of this parsing step are provided in the Data



〈Fig. 1〉 Proposed Approach to Mapping Computer and Information Science

Collection section. Step 2 is the step for preprocessing. The fields indexed are title, author, year, and id. The title field is used as input for DMR. Step 3 is the step for topic modeling with DMR with pre-processed data. Step 4 constructs database tables to be served for identifying top conferences and authors in each topic.

3.1 Multinomial Topic Modeling

We used Dirichlet-multinomial regression (DMR) proposed by Mimno and McCallum (2008) for topic modeling. DMR is an extension of Latent Dirichlet Allocation (LDA) proposed by Blei et al. (2003). Topic models are unsupervised techniques for discovering the main themes that pervade a large and otherwise unstructured collection of documents. In our study, we develop our code based on the MALLET topic modeling package written in Java (McCallum, 2002). In contrast to previous methods, topic models based on DMR are able to incorporate arbitrary types of observed continuous, discrete and categorical features with no additional coding, yet inference remains relatively simple.

With DMR, we first represent a document as bag of words. After that, we apply generative model, a probability distribution, for the document collections. Given a training set of documents, we choose values for the parameters of the distribution that make the training documents have high probability. Then, given a test document, we evaluate its probability according to the model.

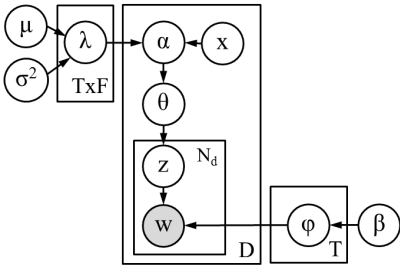
The higher this probability is, the more similar the test document is to the training set. The probability distribution that we use is the multinomial. Mathematically, this distribution is

$$P(\chi;\theta) = \frac{n!}{\Pi_{j=1}^{m} x_j!} \cdot \Pi_{j=1}^{m} \theta_j^{x_j} \quad (1)$$

where x is a vector of non-negative integers and the parameters θ is a real-valued vector. Both vectors have the same length m. Intuitively, θ_j is the probability of word j while x_j is the count of word j. Each time word j appears in the document it contributes an amount θ_j to the total probability, hence the term $\theta_j^{x_j}$. The components of θ are non-negative

and have unit sum: $\sum_{j=1}^{m}\theta_j = 1$.

Each vector x of word counts can be generated by any member of an equivalence class of sequences of words. In Equation (1) the first factor in parentheses is called a multinomial coefficient. It is the size of the equivalence class of x, which is the number of different word sequences that yield the same counts. The second factor in parentheses in Equation (1) is the probability of any individual member of the equivalence class of x. Like any discrete distribution, a multinomial has to sum to one, where the sum is over all possible data points. Here, a data point is a document containing n words. The number of such documents is exponential in their length n: it is n^m. Thus, the probability of any individual document is very small. What is important is the relative probability of different documents. A document that mostly uses words with high probability has higher relative probability. The graphical model for Dirichlet Multinomial Regression is shown in Fig. 2. This is the model originally proposed by Mimno and McCallum (2008). Each node is a random variable and is labeled according to its role in the generative process. The hidden nodes — the topic proportions, assignments, and topics — are un-shaded. The observed nodes — the words of the documents — are shaded. The rectangles are "plate" notation, which denotes replication. The plate denotes the collection words within documents; the D plate denotes the collection of documents within the collection.



〈Fig. 2〉 Graphical Model for DMR

3.2 Data Collection

Candidate conferences for this study were obtained from three sources: 1) Conference search website (http://www.confsearch.org/ confsearch/). From the computer science conference search website, we limit the conference search to the following field: Artificial Intelligence, Databases, Data Mining, HCI, Information Retrieval, Machine Learning, Multimedia, Natural Language, Networks, Theory, Vision, and WWW. 2) International Calendar of Information Science Conferences (http://www.asis.org/Conferences/calendar/), and 3) Information Science Conferences Worldwide (http://www.conferencealerts.com/) that lists upcoming events in information science. From these three sources, we identified 353 conferences that are indexed in DBLP. DBLP is the online bibliography database for Computer and Information Science Bibliography that was developed by the University of Trier. DBLP includes more than 2.3 million bibliographic records as of October 2013. For our study, we download the entire DBLP record in XML format from http://dblp.uni-trier.de/xml/ (the total size of the XML file 1.0G). The sample DBLP record is provided in Fig. 3.

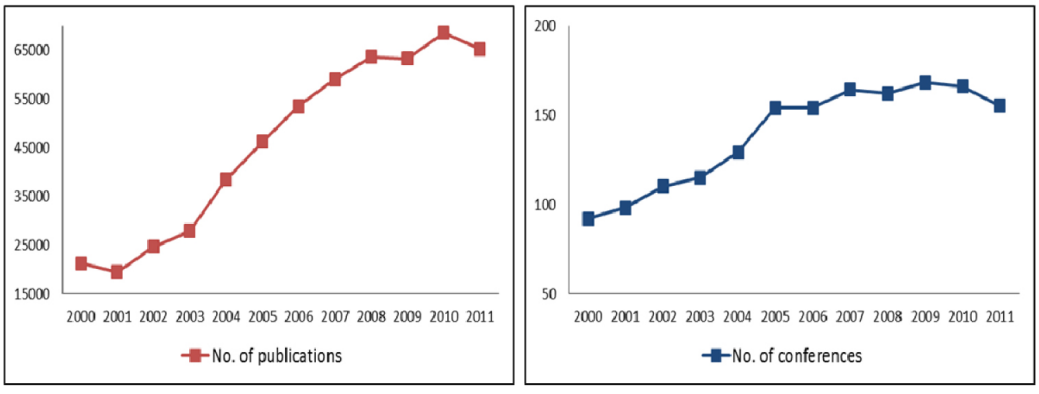
```
<dblp>
  <inproceedings key="conf/ir/FreiQ93" mdate="2002-01-03">
    <author>Hans-Peter Frei</author>
    <author>Yonggang Qiu</author>
    <title>Effectiveness of Weighted Searching in an Operational IR Env.</title>
    <pages>41-54</pages><year>1993</year>
    <booktitle>Information Retrieval</booktitle>
   <url>db/conf/ir/ir93.html#FreiQ93</url>
  </inproceedings>
</dblp>
```

〈Fig. 3〉 A Sample of DBLP Records

4. Result

In this section, we report on the mining results conducted in our study. The total number of records from 353 conferences in DBLP is 236,170. Fig. 4 (a) illustrates the number of publications indexed in DBLP during 2000 to 2011. As indicated in Fig. 4 (a), since year 2003, there is a rapid increase in the publications. This concurs with popularity of the Internet starting from early 2000 where electronic records are produced and shared via the Internet. Fig. 4 (b) shows the number of conferences held in each year. We observe that there is a steady increase in the number of conferences in computer and information science field. An interesting fact is that the growth rate was significantly reduced after year 2005. This may indicate that the community activities of the field of computer and information science has become mature after 2005. The average number of conferences during 2000 to 2011 is 139.

Table 1 shows the ranking of conferences between 2000 and 2011. The ranking was measured by the



〈Fig. 4〉 (a) The number of publications from 2000 to 2011 in DBLP
(b) The number of conferences from 2000 to 2011 in DBLP

〈Table 1〉 Conferences ranking by the number of papers published between 2000-2011

2000	2001	2002	2003	2004	2005
ICPR	ICIP	ICPR	ICIP	ICPR	ICIP
ICIP	CVPR	ICIP	ICEIS	ICIP	ICCSA
ICC	MICCAI	CHI	WWW	ICCSA	CHI
CVPR	ICDAR	ECCV	ICCSA	ICEIS	CVPR
AAAI	ICCV	NIPS	CHI	CHI	AINA
2006	2007	2008	2009	2010	2011
ICPR	GLOBECOM	ICC	ICIP	GLOBECOM	GLOBECOM
GLOBECOM	ICIP	GLOBECOM	GLOBECOM	ICPR	ICC
ICIP	WCNC	ICPR	ICC	ICIP	ICIP
ICCSA	CVPR	ICIP	CHI	ICC	CHI
CHI	AINA	WCNC	WCNC	CHI	AAAI

number of papers published by a conference. During 2000 to 2011, ICPR and ICIP are two conferences that published most papers. ICPR stands for International Conference on Pattern Recognition and ICIP stands for International Conference on Image Processing. Those two conferences are well-known in image and video processing. CHI, which stands for Human Factors in Computing Systems that is the most prestigious conference in Human Factor, is also the conference that publishes a large volume of papers. GLOBECOM, Global Communications Conference which is a leading conference in telecommunication, makes the first appearance in 2006, and since then, GLOBECOM was ranked top5.

Table 2 shows the results of DMR. In order to find the proper number of topics, we experiment in iterated process. We set the parameter of LDA with 20 topics, 1000 iterations, 1 alpha value, and 0.01 beta value. Alpha and beta values are set by the default value. In results, we found that topic 1 and 16 are about human computer interaction. Topic 4 and 20 have something to do with the research on image/video. Topic 5, 7, 9, 11, 13, and 14 are more or less related to wireless/distributed network. Topic 6 is about Web. Topic 8 and 12 are associated with data mining. Topic 18 is related to text mining. Topic 19 is about information retrieval.

Table 3 shows the list of top conferences per topic. The decision on top conferences was made by the number of matched terms in the title published in a conference and key terms in a topic. The reason for that is to find the leading conferences in a particular topic. Leading conferences in topic 1 include CHI, HRI, IUI, and INTERACT. There conferences are related to human computer interaction and social computing. Top conferences in topic 2 are ICDAR, ACL, ICPR, NAACL, and CICLING, and these conferences are associated with computational linguistics. Topic 8 includes conferences like ICDE, SIGMOD, EDBT which are pertinent to database and data mining. As the similar conferences, topic 12 include leading conferences such as ICDM, KDD, PAKDD, and PKDD, and these conferences

〈Table 2〉 Topic Models by DMR

Topic 1 HCI	Topic 2 Computational linguistics	Topic 3 Information system	Topic 4 Image processing	Topic 5 Neural network system
human	word	information	image	network
computer	speech	system	video	neural
interaction	recognition	business	compression	distributed
user	language	knowledge	segmentation	real-time
interface	translation	management	algorithm	system
Topic 6 Web	Topic 7 Network	Topic 8 Database	Topic 9 Mobile service	Topic 10 Graph algorithm
music	MIMO	database	mobile	graph
social	OFDM	efficient	network	algorithm
retrieval	channel	data	service	planar
classification	frequency	xml	security	paths
web	codes	processing	key	approximation
Topic 11 Network	Topic 12 Data mining	Topic 13 Wireless network	Topic 14 Wireless network	Topic 15 Artificial intelligence
network	data	wireless	wireless	automata
traffic	mining	network	network	logic
routing	knowledge	sensor	channel	model
packet	discovery	routing	scheduling	theory
streaming	algorithm	protocol	allocation	learning
Topic 16 Artificial intelligence	Topic 17 Game	Topic 18 Text mining	Topic 19 Information retrieval	Topic 20 Multimedia system
proceedings	algorithm	classification	information	image
conference	online	clustering	image	video
workshop	game	semi	retrieval	tracking
artificial	genetic	supervised	ranking	detection
intelligence	optimization	learning	evaluation	recognition

are related to data mining and text mining. Topic 15 includes artificial intelligence related conferences such as AAAI, ICALP, MFCS, ECAI, and UAI. Top conferences in topic 16 also are related to artificial intelligence. Topic 19 includes conference such as SIGIR, CIKM, ACL, ECIR, and ICIP, and these conferences are related to information retrieval.

Table 4 shows the topic similarity based on publications of top authors. Since a researcher tends to continuously carry out studies in a chosen research field, the papers published by the researcher thematically similar. This characteristic enables us to merge similar topics or identify relatedness among topics. From Table 4, we select papers published by top 5 researchers in each topic and compute the similarity among topics by title term match to understand the thematic relatedness among topics. Among all top authors, we select some sample authors to describe their research interests and activities. Top authors in topic 1 are Steve Benford, Hiroshi Ishiguro, Stephen Brewster, and Laurence Nigay. Steve Benford is Professor of Collaborative Computing in

〈Table 3〉 Conferences that key terms appears frequently per topic

Topic1	Topic2	Topic3	Topic4	Topic5
CHI	icdar	ECIS	ICIP	DSN
HRI	acL	ICEIS	ICPR	ICDCS
ICML	icpr	PACIS	PCM	NIPS
IUI	NAACL	BIS	VCIP	ICCSA
INTERACT	CICLING	CHI	DCC	ICEIS
Topic6	Topic7	Topic8	Topic9	Topic10
ISMIR	GLOBECOM	ICDE	CCS	SODA
SIGIR	ICC	SIGMOD	ICCSA	FOCS
MMM	WCNC	CIKM	AINA	GD
RECSYS	ICIP	EDBT	PERCOM	WG
IUI	WIMOB	ICEIS	GLOBECOM	ISAAC
Topic11	Topic12	Topic13	Topic14	Topic15
INFOCOM	ICDM	GLOBECOM	GLOBECOM	AAAI
GLOBECOM	KDD	ICC	ICC	ICALP
ICC	PAKDD	INFOCOM	WCNC	MFCS
LCN	PKDD	WCNC	INFOCOM	ECAI
NETWORKING	DIS	MASS	ICCCN	UAI
Topic16	Topic17	Topic18	Topic19	Topic20
IHM	AAAI	NIPS	SIGIR	ICIP
AAAI	WINE	ICML	CIKM	CVPR
ECAI	SODA	ICIP	ACL	ICPR
PERCOM	ICALP	ICPR	ECIR	ICCV
CIKM	NIPS	ICDM	ICIP	ECCV

the Mixed Reality Laboratory at Nottingham. He has received best paper awards at CHI 2005, CHI 2009, CHI 2011, CHI 2012. He was elected to the CHI Academy in 2012. Hiroshi Ishiguro is director of the Intelligent Robotics Laboratory in the Department of Systems Innovation at Osaka University in Japan. He received the best paper award at HRI in 2009 and the best paper and poster awards at HRI in 2007. Stephen Brewster is Professor of Human-Computer Interaction in the Department of Computing Science at the University of Glasgow in UK. Laurence Nigay is also ranked high in topic 16. He is a Professor at l'Université Joseph Fourier, and his research interests are engineering for HCI. Based on top authors' research expertise, we confirm that topic 1 is about HCI. Three authors in topic 11, a professor Donald F. Towsley, a professor James F. Kuros, and Zhi-Li Zhang, are affiliated with the Department of Computer Science at University of Massachusetts Amherst. Donald F. Towsley is a Distinguished University Professor in the Department of Computer Science at the University of Massachusetts Amherst. His research interests include network measurement, modeling, and analysis. James F. Kurose is a computer science professor at University of Massachusetts Amherst. He is a

〈Table 4〉 Top Authors whose papers appear most frequently in a topic

Topic1	Topic2	Topic3	Topic4	Topic5
Steve Benford	Cheng-Lin Liu	Shazia Wasim Sadiq	Feng Wu	Miguel Correia
Hiroshi Ishiguro	Masaki Nakagawa	Michael Rosemann	Jizheng Xu	Azzedine Boukerche
Stephen Brewster	Horst Bunke	Joseph Barjis	Min-Jen Tsai	Hongyi Wu
Takeo Igarashi	Eiichiro Sumita	Wasana Bandara	David Taubman	Jie Wu
Laurence Nigay	Xiaoqing Ding	Brian J. Corbitt	Akio Miyazaki	Yu Wang
Topic6	Topic7	Topic8	Topic9	Topic10
Gerhard Widmer	Norman Beaulieu	Gonzalo Navarro	Dongho Won	Takao Nishizeki
Douglas Eck	Robert Schober	Divesh Srivastava	Xuemin Shen	Stephen G. Kobourov
Meinard M	Chintha Tellambura	Philip S. Yu	Ghita Most	Xiao Zhou
Qing Li	Khaled Ben Letaief	Wing-Kai Hon	Rongxing Lu	Hans L. Bodlaender
Barry Smyth	Inkyu Lee	Clement T. Yu	Pin-Han Ho	Dimitrios M. Thilikos
Topic11	Topic12	Topic13	Topic14	Topic15
Donald F. Towsley	Jiawei Han	Azzedine Boukerche	Ying-Chang Liang	Stephen Muggleton
James F. Kurose	Philip S. Yu	Garcia-Luna-Aceves	Xinbing Wang	Luc De Raedt
Christophe Diot	Shusaku Tsumoto	Jie Wu	Husheng Li	Oscar H. Ibarra
Jiangchuan Liu	Heikki Mannila	Xuemin Shen	Zhu Han	Juraj Hromkovic
Zhi-Li Zhang	Nick Cercone	Xiang-Yang Li	Khaled Letaief	Manuel Ojeda Aciego
Topic16	Topic17	Topic18	Topic19	Topic20
Forest Fisher	Paul G. Spirakis	Dit-Yan Yeung	Jimmy J. Lin	Larry S. Davis
Jean-Yves Marion	Tuomas Sandholm	Changshui Zhang	W. Bruce Croft	Luc J. Van Gool
Laurence Nigay	Kevin Leyton-Brown	Michael I. Jordan	Eric Nyberg	Rama Chellappa
Steve A. Chien	Thomas Henzinger	Rong Jin	Eugene Agichtein	Stan Z. Li
Susanne Albers	Christos Papadimitriou	Eric P. Xing	Sanda Harabagiu	Takeo Kanade

coauthor of the well-known textbook Computer Networking: A Top-Down Approach. Zhi-Li Zhang is currently a professor in the Department of Computer Science and Engineering at University of Minnesota, and he was a student of Donald F. Towsley and James F. Kurose. Jiangchuan Liu is an Associate Professor in the School of Computing Science at Simon Fraser University in Canada. His research interests are in networking and he received the best paper award in GLOBECOM in 2011 and ACM Multimedia in 2012. Top authors in topic 11 also confirm that topic 11 is about network.

5. DISCUSSION

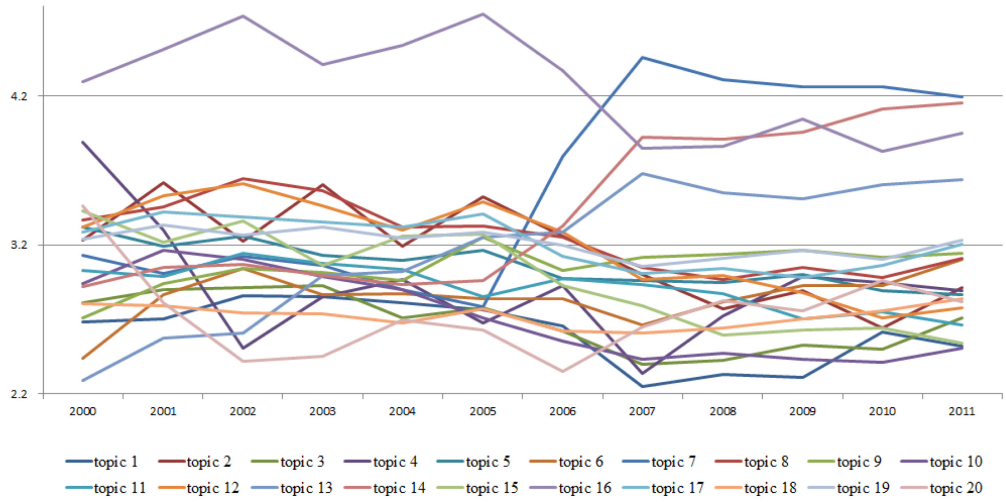
In this section, we analyze topic similarity between topic pairs and topic trends during the time span of 2000 through 2011. Fig. 5 shows the matrix of topic similarity among topics. Columns show the number of papers written by top 5 authors in at most three topics. If more than two terms in a title of a paper are matched with topic terms in each topic, we treat the paper as belonging to a particular topic. Fig. 5 shows the result of the topic similarity based on publications of top authors. One interesting

observation is that top authors in topic 5 published more papers in topic 13 than in topic 5. On the contrary, top authors in topic 13 published only 10 papers in topic 5. Topic 5 is related to neural network system and topic 13 is related to wireless network. It may imply that some of authors in topic 5 move to topic 13. According to Fig. 5, topic 5 is likely to be a mature field and is slightly shrunken after 2006. The similar pattern is observed in topic 9. Top authors in topic 9 published 33 papers in topic 9 and 32 papers in topic 13. However, top authors in topic 13 published only 11 papers in topic 9. This indicates that top authors in topic 9 conduct researches in both topics but top authors in topic 13 carry out researches mainly in topic 13. Research exchanges between topics are observed in topic 8-topic 12, topic 7-topic 14 based on our approach. Other than these topics, the rest of the topics show that there is no significant topical similarity.

Fig. 6 shows topic trends of 20 topics drawn from computer and information science -related conferences in DBLP between 2000 and 2011. By and large, topic trends shown in Fig. 6 can be classified into 4 different categories: growing, shrinking, continuing, and fluctuating. The growing category is that there is a steady increase in the topic probability in a topic over time. Topic 7, 13, and 14 belong to this category. Topics in the shrinking category have a trend of decrease in the topic probability in a topic. Topic 2, 4, 10, 12, 15, and 16 belong to this category. Topics in the continuing category tend to show a steady topic probability, and topic 5, 6, 8, 9, 11, 17, 18, and 19 belong to this category. The last category is fluctuating which denotes that there is not a consistent pattern in the topic trend. Topic 1, 3, and 20 belong to this category.

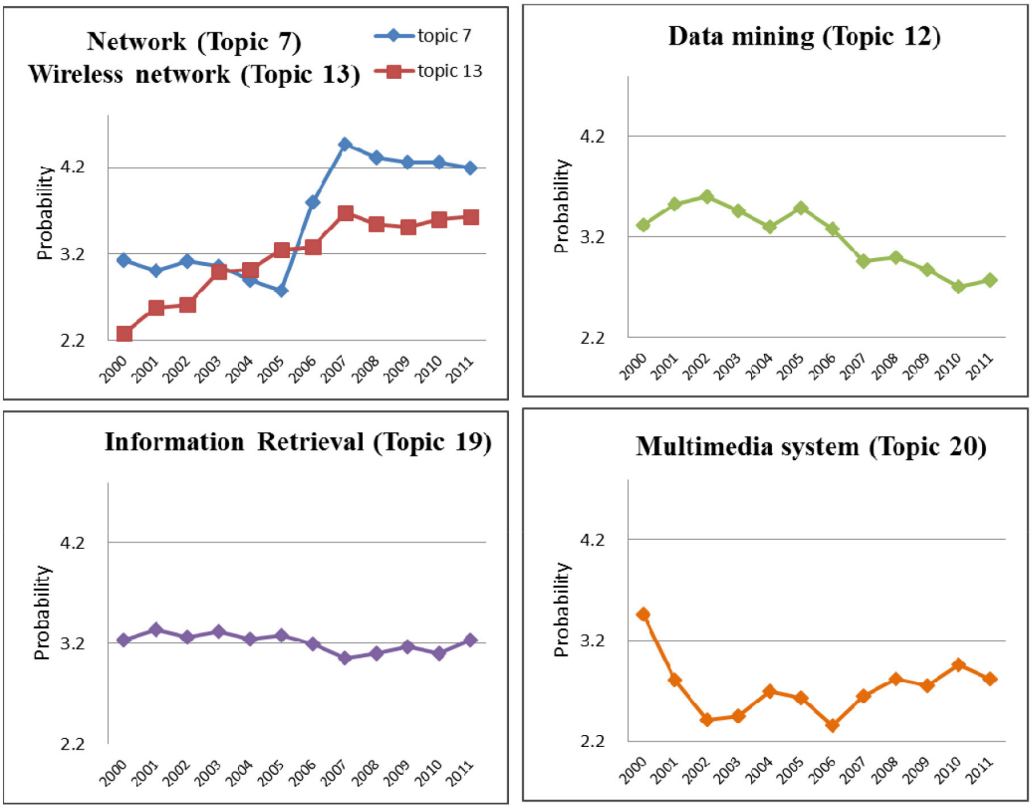
topic	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	61	2	1													10				
2		103																4	1	
3			39																	
4				31																1
5			1		29															
6						20														
7				1			85							26						
8						3		51				14				1				
9									33				11							
10										103					4		11			
11				1				3			62									
12					5	3		**16**				69			1					
13					**63**		5		**32**	1	10		138	11			2		3	
14							**36**		10		5		18	117						
15										1					52					
16	3															15				
17																	46			
18		4										6						55		7
19																		3	47	
20	2																			69

〈Fig. 5〉 Topic similarity based on Publication of Top Authors



〈Fig. 6〉 Topic Trends in Computer and Information Science over Time

Fig. 7 (a) - (d) shows topic trends over time for topic 7, 12, 13, 19 and 20. X-axis is year and y-axis is the topic probability in a particular year. Topic probability represents the importance of a topic among 20 topics in a particular year. Topic 7 shows the increase in the topic value after 2006 (Growing). The same trend is observed in topic 13 and topic 14. This trend may be influenced by the following two factors. First, the conference, GLOBECOM, was first held in 2006. GLOBECOM is a conference dealing with telecommunication that publishes over 1000 papers each year. The number of published papers was increased from 1,008 in 2006 to 1,324 in 2011. Based on the analysis of the published papers in GLOBECOM from 2006 to 2011, we found that topic 7 shows the highest similarity between published papers and topic terms in each topic category. Second, the main topic terms in topic 7 are MIMO and OFDM that are telecommunication techniques, and these terms appeared first in DBLP around year 2006. These two terms were found two times in 2005, 26 times in 2006, and 58 times in 2007 (and after 2007, there is a slight decreases in the number of times appearing in the title) in the entire DBLP database by the title search. Topic 13 started with a low interest in 2000 and made a steady growth until 2007, and it became saturated after 2007. Key terms in topic 13 are wireless, network, sensor, routing, and protocol, and the underlying main topic is Wireless Sensor Network (WSN). As wireless network technology is developed, researches of applying WSN to the ubiquitous environment has proliferated. Topic 13 seems to reflect this research trend. From a DBLP perspective, MASS and GLBOECOM, two core conferences in WSN, were first indexed by DBLP in 2005 and 2006 respectively and 455 papers among papers published by both conferences are closely related to topic 13. Birth and proliferation



〈Fig. 7〉 (a) - (d) Topic Trends of Topic7, Topic12, Topic13, Topic19 and Topic20 in Period of 2000 to 2011

of these two conferences led to the growth trend of topic 13 in mid-2000's. The topic trend shown in topic 13 coincided with the aforementioned explanation. These findings drawn from DMR analysis imply that the emergence of a conference and the development of related technologies result in the quantity increase of a particular topic.

Key terms of topic 19 are information, retrieval, ranking, and evaluation, and those terms represent Information Retrieval (IR). Topic 19 shows the consistent topic probability (Continuing). This indicates that researches in IR have been consistently conducted during 2000 and 2011. According to Fig. 5, topic 19 is a unique topic that is not similar with the rest of 19 topics. This implies that the field of IR has not rigorously interacted with other sub-fields of computer and information science at least from the eyes of DBLP. Topic 12 is an exemplar topic representing shrinking. There are two possible explanations for this shrinking trend. The first explanation is that the topic itself became disinterest, and the second explanation is that the topic became

fragmented and diversified. In fact, the decrease in a research trend is reflected in reduction of the number of papers published or the number of conferences held. However, in case of specialization of a topic, we need to analyze the topic over time to see if key terms in the topic become more similar with terms in other topics. Topic 12 is related to data mining. We found that there is no decrease in the number of papers and the number of conferences in this topic. However, in the key term distribution by year, three different patterns are revealed. In early 2000s, the topic was led by terms like clustering, classification, and pattern. In mid-2000s, terms like web, graph, and protein were shown. In late 2000s, terms like social, network, dynamic, and time were newly introduced, and this implies that an emerging new field can be forked out of data mining. This, in turn, influences on the decrease in direct research on data mining per se. Key terms in topic 12 are not the terms encompassing broad application areas of data mining such as graph theory, social network analysis, and biomedical analysis. Rather, they are the typical research topic terms such as pattern detection, classification, and clustering. Due to this reason, the topic trend by DMR shows that topic 12 is in shrinking.

Topic 20's year-on-year trend shows fluctuating patterns repeating rise and fall. The topic terms corresponding to this topic include image, video, recognition, tracking, and detection, and the topic is about techniques demanded in the application of visual data or system implementation. In Fig 7(d), topic 20's trend shows rise and fall in turn from 2000 until 2006, but after 2006, it tends to increase gradually. As a matter of fact, about CVPR, one of the top conferences corresponding to topic 20, the number of papers published in 2007 increased 69.7% compared with that in the previous year, and about ICPR that is held biennially, too, the number of papers increased in 2006. The rise in 2006 is because of increased interest in image retrieval. For instance, Google selected image search as one of its major services in the early 2000's and has developed the service constantly. This kind of image search service provision and constant development has influenced on maintaining interest in image search. From the aspect that in 2009, Google started to provide 'similar image search' service as a new image search function, we can assume that many studies were conducted about image search during the mid-2000's.

Also, the growth rate was slow from 2008 but resurged between 2009 and 2010. The amount of paper publications in major conferences also tended to increase in general in 2010. This trend is attributed to the fact that various display devices have emerged thanks to the appearance of smartphones. And image processing is required in medicine or biology. For example, it was the time that there was a demand for applications appropriate for a certain environment such as fingerprint verification and image processing technology for augmented reality. Therefore research may have been actively conducted to meet such demand. Meanwhile, one of the reasons why topic 20 tended to show fluctuations in general is that conferences related with ICCV or ECCV are held biennially. Topic 20's trend may repeat rise and fall

within a short period, but biennial conferences either increased or decreased in the number of papers introduced with the cycle of two years. This can be regarded as one of the reasons why it showed the trend of fluctuations.

As shown in the analysis of the results, we ascertain that the topic analysis using state-of-art text mining techniques such as topic modeling can be supportive and supplementary to the traditional bibliometric analysis to discover topic trends (Tang et al., 2013).

6. Conclusions

In this paper, we explore dynamics of the field of computer and information science with the aid of text mining techniques. We collect more than 200,000 documents from 353 conferences in DBLP related to computer and information science. We build the system to automatically map out the field by parsing DBLP data in XML. We adopt DMR-based topic modeling. By applying DMR to the segmented DBLP data by years, we examine the trends of the field. We analyze topic modeling results along with statistics of the number of papers and the number of conferences over the period of 2000 to 2011, top authors per topic, and top conferences per topic. We identify four different patterns in topic trends in the field of computer and information science between 2000 and 2011: growing, shrinking, continuing, and fluctuating.

The results of analysis indicate that computer and information science is a fast growing cross-disciplinary field. Several topics such as topic 5, 6, 8, 9, 11, 17, 18, and 19 became matured and show a steady pattern (Continuing). Those topics did not show a noticeable increase or decrease in topic probability over time. Topics such as topic 2, 4, 10, 12, 15, and 16 became unpopular (Shrinking). On the contrary, topics like topic 7, 13, and 14 became popular over time (Growing). Unlike the aforementioned topics, some topics such as 1, 3, and 20 do not show consistent patterns (Fluctuating).

In the follow-up study, we plan to explore what topics are merged, shrunken, remain to be the same with a more precise and statistical approach. We are also interested in who are the main players of having an impact on topic shift in depth. In addition we develop the web-based interactive trend tracking system.

References

Adamic, L., & Adar, E. (2005). How to search a social network. Social Networks, 27(3), 187-203.

Blei, D., & Lafferty, J. (2006). Dynamic topic models. In Proceedings of the 23rd International Conference on Machine Learning, 113-120.

Blei, D., Ng A., & Jordan, M. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993-1022.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems, 30, 107-117.

Buckland, M. (2012). What kind of science can information science be?. Journal of the American Society for Information Science and Technology, 63(1), 1-7.

Chen, C., & Carr, L. (1999). Visualizing the evolution of a subject domain: A case study. In Proceedings of the conference on Visualization '99: celebrating ten years, 449-452.

Cutting, D., Karger, D., & Pederson, J. (1993). Constant interaction-time scatter/gather browsing of very large document collections. In Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 126-134.

Frank, E., Paynter, G., Witten, I., Gutwin, C., & Nevill-Manning, C. (1999). Domain-specific keyphrase extraction. In Proceeding of 16th International Joint Conference on Artificial Intelligence, 668-673.

Glänzel, W. (2012). Bibliometric methods for detecting and analysing emerging research topics. El profesional de la información, 21(2), 194-201.

Griffiths, T., & Steyvers, M. (2004). Finding scientific topics. Proceedings of the National Academy of Sciences, 101(suppl. 1), 5228-5235.

HaCohen-Kerner, Y., Gross, Z., & Masa, A. (2005). Automatic extraction and learning of keyphrases from scientific articles. In Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing, 657-669.

He, D., & Parker, S. (2010). Topic dynamics: an alternative model of 'bursts' in streams of topics. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, 443-452.

Hulth, A. (2003). Improved automatic keyword extraction given more linguistic knowledge. In Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, 216-223.

Janssens, F., Glänzel W., & De Moor, B. (2008). A hybrid mapping of information science. Scientometrics, 75(3), 607-631.

Kleinberg, J. (2003). Bursty and hierarchical structure in streams. Data Mining and Knowledge Discovery, 7(4), 373-397.

Liu, F., Liu, F., & Liu, Y. (2008). Automatic keyword extraction for the meeting corpus using supervised approach and bigram expansion. In Proceedings of 2008 IEEE Workshop on Spoken Language Technology, 181-184.

Liu, Z., Huang, W., Zheng, Y., & Sun, M. (2010). Automatic keyphrase extraction via topic decomposition. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 366-376.

McCallum, A. (2002). MALLET: A Machine learning for language toolkit. Retrieved from http://mallet.cs.umass.edu

Matsuo, Y., & Ishizuka, M. (2004). Keyword extraction from a single document using word co-occurrence statistical information. International Journal on Artificial Intelligence Tools, 13(1), 157-169.

Merriam-Webster and American Heritage Dictionary. Retrieved from http://www.britannica.com/EBchecked/topic/19759/The-American-Heritage-Dictionary

Mimno, D., & McCallum, A. (2008). Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. Retrieved from http://arxiv.org/abs/1206.3278v1

Tang, X., Yang, C. C., & Song, M. (2013). Understanding the evolution of multiple scientific research domains using a content and network approach. Journal of the American Society for Information Science and Technology, 64(5), 1065-1075.

Treeratpituk, P., & Callan, J. (2006). Automatically labeling hierarchical clusters. In Proceedings of the 2006 International Conference on Digital Government Research, 167-176.

Wan, X., Yang, J., & Xiao, J. (2007). Towards an iterative reinforcement approach for simultaneous document summarization and keyword extraction. Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, 552-559.

Wang, C., Blei, D., & Heckerman, D. (2012). Continuous time dynamic topic models. Retrieved from http://arxiv.org/abs/1206.3298v1

Wang, X., Mohanty, N., & McCallum, A. (2005). Group and topic discovery from relations and text. The 11th ACM SIGKDD International conference on Knowledge Discovery and Data Mining Workshop on Link Discovery: Issues, Approaches & Applications, 28-35.

White, D., & McCain, W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. Journal of American Society of Information Science and Technology, 49(4), 327-355.

Xu, J., Marshall, B., Kaza, S., & Chen, H. (2004). Analyzing and visualizing criminal network dynamics: A case study. In H.Chen, R.Moore, D.D.Zeng, & J.Leavitt (Eds.), Lecture Notes in Computer Science, 3073: Intelligence and Security Informatics, 359-377. Berlin: Springer.