Applied Mathematics & Information Sciences
An International Journal

A study of data mining trend through the optimized bibliometric methodology based on SCI database from 1993 to 2011

Lu Dai[1], *Lixin Ding*[1], *Yunwen Lei*[1] *and Yangge Tian*[2]

[1] State Key Lab of Software Engineering, Wuhan University, Wuhan 430072, China
[2] International School of Software engineering, Wuhan University, Wuhan 430072, China

Received: January 10, 2012; Revised March 16, 2012; Accepted May 7, 2012
Published online: August 5, 2012

Abstract: An optimized bibliometric method was applied in this work to evaluate global scientific production of data mining papers of the Science Citation Index (SCI). In compared with traditional bibliometric keyword analysis, a semantic words class was established by applying the text extraction mode to remove noise in the abstract and combining with the core relative phrases retrieved from keywords to get the sample for further experiment. The analysis shows a high correlation between title and keywords, and the title reports less information than keywords does. Also, keywords provide more positive guidance to know and be familiar with the status and trends of this field. In addition, there are distinctions among those semantic words used in publications from the ten most productive countries in data mining research. Generally speaking, the research results can be extended to investigate the roadmap for future research, and this innovative propose is provided with instructive meaning for valuable information retrieval.

Keywords: Data mining, semantic words, research trend analysis, bibliometric methodology

1. Introduction

Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets [1]. These tools employ algorithms and techniques from statistics, machine learning, artificial intelligence, databases and data warehousing etc. As an important research topic, it has led to an expeditious increase in quantity of scientific articles on data mining research over the past. The number of publications based on Science Citation Index (SCI) the most frequently-used index in scientific output analysis explorded from several in the early 1990s to 1833 in 2009, and the total 9746 articles were published in a variety of 1822 journals [2]. Despite of the phenomenal growth of data mining papers, there have been few attempts at gathering systematic data on scientific production of data mining research by common research tool bibliometric methods.

Bilbliometrics are made of a set of metholds for measuring the production and dissemination scienfic knowledge. Some of the methods serve to measure quantiative research exercise of academic output which is startiong to threaten practice based research [3]. Bibliometric analysis have been conducted to reveal the global trends of various research fields [4]. It is virtually indispensable for mapping and benchmarking research output and identifing emerging fields. Our learning group has already applied traditional bibliometric metholds to evaluate the data mining research output by the keyword analysis. The analysis relative to manual reading could add an invaluable empirical content and objectivity to the process of understanding and evaluating the strengths and weaknesses of the knowledge production system in the data mining. However, it is hard to find the overall situation of the subject, the analysis of key words represents only a part of information and some dynamics would otherwise be overlooked.

2. Methods and Materials

It is widely accepted that through the SCI database to evaluate the trend and advancing direction of data mining

* Corresponding author: e-mail: qjdxyxx@yahoo.cn

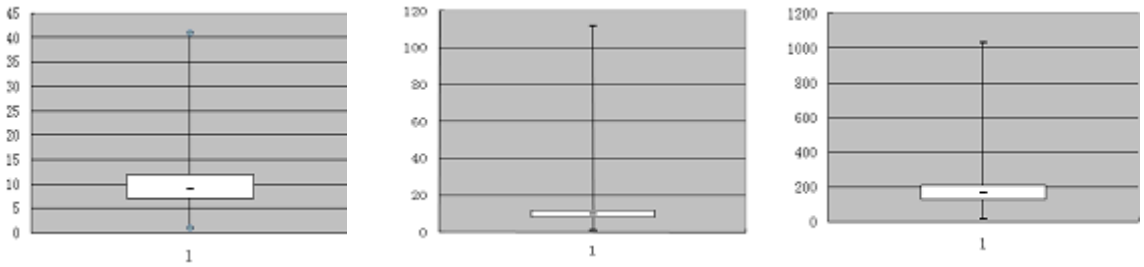

Figure 1 box plots from the numerical feature of title, abstract and keyword

is a good way. Given that, document used in this research was derived from the online database of the SCI retrieved from the ISI Web of Science, Philadelphia, PA, USA. Data mining was used as the search topic to search keywords, title, abstract from 1993 to 2009. Document information involved document type, author information, title, keywords, abstract, language, contact addresses, cited reference count, citation report, page count, publisher information, subject categories. After deleting the repeated records, there were 9747 publications in total that met the selection criteria mentioned, including 3 document types. Since journal articles which occupied the majority of the document are of timeliness and the convenience of comparison, we focused on 7004 articles for further analysis.

In this research, a semantic word class was established, involving 5878 unique words or phrases, which appeared 541728 times in total. Keywords elaborates the intention of the writer were given priority to build initial semantic class and the semantic words within were extracted directly. Due to the appropriate valuable information in abstract, the text extraction mode was applied to remove noise in the abstract with the aim to get the synonymous unit. Finally, the situation that related semantic words appeared in the title and keywords need to be analyzed on the basis of semantic thesaurus. Semantic words can be used to identify research fronts, which are represented by clusters of articles sharing a key term in abstract, keyword and title. However, this method is based on the assumption that article using the same terms as cognitive content [5]. The unstandardization among words more or less hampered our analysis since the use of synonymous terms, spelling variations and unspecific term. Subsequently, our resulting study based on traditional bibliometric keyword analysis mainly aims at providing an alternative demonstration of research advancements from the multiple perspectives in the evolution of sentiment words we retrieved. This innovative propose is provided with instructive meaning for valuable information retrieval.

The emphasis of the assessment below was to describe global scientific production on data mining during the past 19 years from the following perspectives: the characteristics of semantic words in abstract, keywords and title, the application of outputs in different disciplines and the characteristics of semantic words distributed in source counties. Besides, if the words appeared in title, the frequencies as well as in abstracts and keywords were counted as one-no matter how many times occurred. The misspelled or different words with identical meaning were all considered as separate single keywords.

3. Experiment

3.1. The Analysis of the frequencies of semantic words in abstracts, keywords and titles

Figure 2 shows a high correlation between title and keywords, but the title reports less subject information than keywords does, partly attribute to the limitation of title content, especially when we establish sentiment words mainly based on the keywords. We chose 60 valuable words from our sample set at random and outlined their frequencies in title, abstract and keyword respectively in the bubble diagram above. Specifically, the various bubbles were scaled in size to roughly reflect the frequencies of the words appearance in abstract. A cluster of bubbles located near the original of coordinates were of similarly lower frequencies in title and keyword, while a handful of bigger bubbles with relatively more times of appearance in abstract represented words classification, machine learning, clustering, association rules and knowledge discovery, etc. In order to elaborate the information those bubbles indicated in detail, a computational index was adduced for computing the relative proportion of words appearance in keyword, title and abstract and was demonstrated in table 1. We drew the box plots to provide a precise and accurate description of numerical feature of title, abstract and keyword through their five number summaries: the smallest observation, lower quartile, median, upper quartile, and largest observation.(figure 2) Except for the max length, title and keyword share the similar spread. Specifically, their average length stayed at the similar level 9 and 10, also the statistical distribution of length was between 12 and 7. Abstract offers research purpose, methods, results and conclusion. Thereforethe average length of abstract was much more longer than title and keyword, and the length fluctuated mainly between 208 and 127.

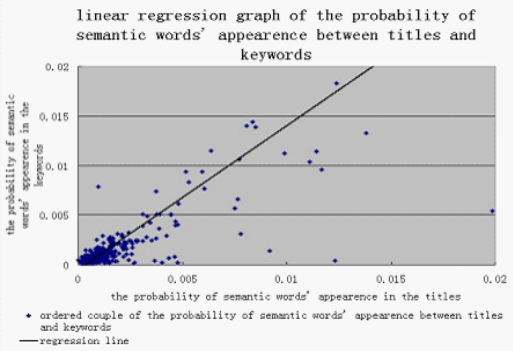

Figure 2 linear regression analysis

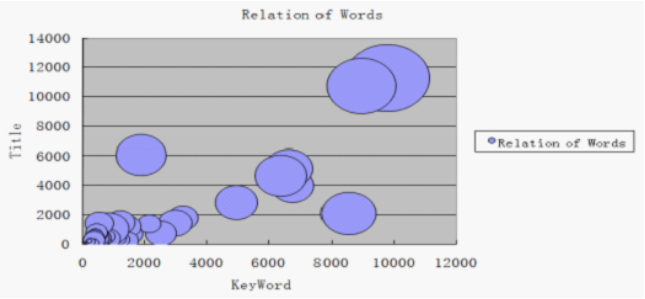

Figure 3 bubble diagram on the frequencies of the 60 words firstly mentioned in title, abstract and keyword

We ordered the words by the frequencies of their appearance in keywords, then presented them within each of the 5-year intervals during 1994-2009 in table2. The proportion of each words appearance in keyword, title, and abstract was based on the computational index:

$$P\% = (A, K, T : \text{abstract, keywords, title})$$

Due to the abundant information containing, basically abstract could offer all of the semantic words we chose and the most proportion, we omitted the preparation of abstract. Compared to abstract, keyword and title could offer the information 65% and 38%, respectively. From this tablewe concluded that keywords is much more accurate in the issue of algorithm expression. For example, the proposion of traditionally classical technology in data mining is higher in keywords, basically over 50%. Such as clustering, data segmentation, classification, association rule mining, hierachical clustering, supervised learning,rough sets, feature extraction, dimensionality reduction, distance-based outliners, linear discriminants, the key technology and algorithm in pattern recognition and data processing. As long as they have been studied in various areas of computer science, the number of application for this class of techniques has been steadily growing. In some special caseswith the relevance to the application and evolving research, the frequencies of related semantic words are more

Word	Key word		Title		Abstract
	Num	P%	Num	P%	Num
Classification	9791	24.93	11245	28.63	39223
Clustering	8946	33.66	10710	40.30	26578
Machine learning	8533	52.80	2086	12.91	16044
Bioinformatics	6757	68.38	3983	40.31	9881
Association rules	6639	52.39	5151	40.65	12630
Knowledge discovery	6376	43.09	4660	31.49	14702
neural networks	4952	49.83	2826	28.44	9918
feature selection	3227	64.89	1778	35.75	4963
decision trees	2983	49.24	1429	23.59	6058
pattern recognition	2521	47.06	725	13.53	5357
rough sets	2136	73.68	1380	47.60	2899
Gene expression	1886	13.65	6044	43.75	13814
artificial neural networks	1533	47.34	679	20.97	3238
support vector machines	1475	41.48	1182	33.24	3556
information retrieval	1450	56.95	255	10.02	2538
feature extraction	1244	75.21	320	19.35	1654
association rule mining	1207	25.24	1483	31.01	4782
text mining	1116	68.26	387	23.67	1612
pattern discovery	1083	48.20	705	31.38	2247
web mining	951	100.	266	27.97	903
Evolutionary computation	916	64.69	492	34.75	1416
decision support systems	907	64.88	107	7.65	1398
evolutionary algorithms	853	73.03	255	21.83	1168
hierarchical clustering	835	38.69	548	25.39	2158
simulation	828	8.83	1028	10.97	9281
artificial intelligence	787	21.05	354	9.47	3743
dimensionality reduction	780	51.25	510	33.51	1522
Knowledge extraction	741	100	228	30.77	739
Information visualization	695	100	122	17.55	579
frequent pattern mining	628	48.68	483	37.44	1290
data visualization	617	51.25	227	18.85	1204
Computational biology	608	57.04	1066	100	621
unsupervised learning	596	43.99	201	14.83	1349

Table 1 The proportion of words appearance in keyword, title and abstract

in keywords, such as web mining, graph mining. multi-objective optimization, knowledge extraction, information visualization, etc. Since authors are indicated to interpret the research point and research purpose, the key information is usually found in keywords, in order to attract the researchers intension. Generally speaking, abstract could apply more abundant information and positive guidance meaning to known and familiar with the status and trends of this field. Otherwise, keyword has more relation with the special application and evolving research.

We generated the information about the articles included the 200 semantic words we chose randomly firstly came out in title, keyword and abstract, respectively. As can be seen from the box plot, the articles with the words appeared in keyword originally are cited more. Since the median was over than 40. Meanwhile, the cited number of articles with those words first appearance in title and abstract came after. Although the cited number of articles related

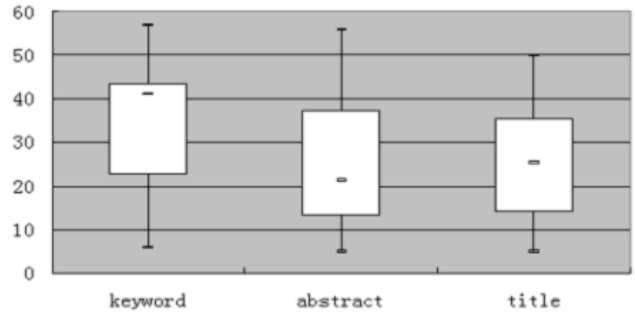

Figure 4 Box plots from the information about the 200 semantic words firstly came out in title, keyword and abstract

that of abstract was marginally less than that of title, they generally bear approximately similar average, maximum and minimum. In contrast, the range of abstract quotation is remarkable.

3.2. The application of outputs in different disciplines

As the study has advanced during the last several years, the publication output data of data mining relates research is distributes in several disciplines.

Genomics	Machine learning
	High throughput screening
	Fuzzy logic
	Hybrid algorithms
	Multiobjective optimization
	predictive models
	Support vector machines
DNA repair	Literature data mining
	Adaptive evolution
Systems biology	Data integration
molecular biology	Evolutionary algorithm
Gene ontology	Domain ontology

Table 2 The main technique of data mining in biology

Traditionally, data mining learning has been viewed as the special preserve of people who major in computer, rather than a necessary part of the field an educated person specialized in. the emphasis on a cross Cdiscipline perspective brought to the solution of academic issue and development. The number of scientific articles per discipline exhibited sustaining growth during the time period covered, which indicated that the research based on data mining has been steadily developed in various fields. Since the data from SCI database has been into microdisciplines, we divided the publication outputs into four different disciplines, biology, medicine, geography, chemistry, according to the corresponding search topic.

Many data mining techniques have already been proposed to deal with the obstacles and constraints in biology. Scientists in biology are facing a growing flood of biological data, data mining is well positioned to help the biology draw meaningful observations and discoveries form the vast arrays of biological data that are now available for analysis. During the past decade, various techniques appear in keywords and abstract most frequently but seldom in topic. This also explains these techniques were used for biology, while they were less applied to enrich the computer science by the knowledge of biological aspects. The time of various techniques appearing was all after 2001 and mainly distributed in 2006-2011, it shows that data mining used in biology subject is a new trend and will obtain more development in the future.

For example, High throughput screening is a method for scientific experimentation especially used in drug discovery and relevant to the fields of biology and chemistry. In the field of biology, High throughput screening allows a researcher to quickly conduct genetic tests. Though this process one can rapidly identify genesadaptive evolution, support vector machine, evolutionary algorithm, multi objective optimization, fuzzy logic shared the same frequency and time in keywords and abstract whereas they were not shown in the topic. These words play an auxiliary role in the biological field, and they have just started to be applied. Moreover, domain ontology only appeared 28 times in keywords during the years from 2006 to 2011, while it neither showed in the abstract nor in the topic. This point indicates that domain ontology is in a preliminary stage when being applied in the biology field and its development is not mature yet.

Support vector machine is a new pattern recognitiontechnique. In statistical learning theory, Support vector machine has been shown to provide higher performance in pattern recognition and functional regression than traditional learning machines [6].It has been introduced as a powerful tool for solving classification problems in recent years. By the year of 2009, this promising method had already appeared in author keywords and abstracts for 384 times. It was more likely that this technology would be fully utilized in biological field.

Machine learning is an international forum for research on computational approaches to learn. Its frequency of occurrence among keywords, abstracts and titles increased during the years from 2006 to 2009, and gained the biggest increase in abstracts. It indicates that this technique is applied in biological field more often, such as genomics which is mature of practical use. It probably has connection with the computation of massive amount and unpredictability of biological information.

Biological networks are naturally represented as graphs of interconnected nodes and edges [7]. These graphs can be studied visually or with powerful analysis methods. Graph mining offers solid supports via computational strategies and tools to deal with the complexities and volume of the

data. Hence, it is reasonable to realize that the sharply increasing trend of its appearing frequency in author keywords and abstracts is a sign of being made most use in the further research.

Visual data mining occurred around 1990s. Although a little bit decline in recent years, its stable frequencies in author keywords, abstracts and titles may illustrate that it had already become a fixed branch of applicable techniques of high value in exploratory data analysis and high potential for exploring large databases in biological researches, which seeks to transform the efficient information or graphs provided by the genomic or biological structure into a more understandable manner visual display [8].

Predictive models appeared in keywords, abstracts as well as topics. The appearing time concentrated in 2001-2005, and it never showed up again. The frequency of Fuzzy logic using in abstracts and keywords was the same for 52 times between 2001 and 2005, and they never appeared subsequently. It was likely that the function of this technique in biology have been excavated completely; perhaps they have been replaced by other technologies. Medicine

Artificial immune systems	Rule extraction
	Hybrid neural networks
	Unsupervised learning
pharmacovigilance	Automated signal detection
	bayesian data mining
	Computed-assisted signal detection algorithm
	Association measures

Table 3 The main technique of data mining in medicine

is a branch of biology. Each word began to be used and appeared from 2006 to 2009, the reduced frequency showed that technologies like data mining were lack of further application in medical science.

Rule extraction, hybrid neural networks, unsupervised learning, association measures appeared from 2006 to 2009. Apart from rule extraction, others merely showed in keywords. These technologies demonstrated the advancing front of medicine. With the development of medical science, these technologies will play an increasing role. Rule extraction appearing in the topic is mainly because it might be a new method that can be applied in medical field.

Sequence matching is a method to find the similar patterns in various real-world domains, e.g., medical (electrocardiogram) [9]. From the table 3,the gradually increasing trends in titles and abstracts between 2006 and 2011 may illustrate that this technique still has potential to be made the most of. Automated signal detection appeared in keywords from 2001 to 2005. There was an slight increase in abstract between 2006 and 2009, which meant that the use of this technology has not been hot spot anymore.

Computer-assisted signal detection algorithm was no more prevalent than automated signal detection, which disappeared in keywords and disappeared after 2005.

Combinatorial chemistry	Virtual screening
	Feature extraction
Biochemistry	Fuzzy clustering
	Dimensionality reduction
	Nonlinear mapping
	Pattern recognition
	neural network
Chemogenomics	Similarity searching

Table 4 The main technique of data mining in chemistry

In the field of chemistry, most of the technical words began to come out in relevant papers around 1996, reaching a peak over the period from 2001 to 2005, before descending dramatically in recent years. It was probably contributes to the use of these technologies have been very mature, and continue to play an important role. Although new ones were of potential in application, while still without breakthroughs up to present. Virtual screening first came out in 2001 and climb obviously in the years to come. Pattern recognition and neural network almost appeared in chemistry corpus at the same time. They both provided feasible methods dealing with classification. Pattern recognition had a long and respectable history within engineering, and neural network have arisen from analysis with models of the way that humans might approach pattern recognition tasks [10]. After all, the stable technical branches in chemistry study have been established, obviously, the concerns of neural networks were of wider applicability. The rapid increase in the number and complexity of data sets in chemistry learning and a corresponding need for appropriate interpretation tools which provides a means of reducing the dimensionality in a chemistry meaningful way [11]. Feature extraction is a category of dimensionality reduction, which seeks to transform the data in the high-dimensional space to a space of fewer dimensions. According to the table 4 it will be further applied in the relatively new chemistry research. By contrast, similarity search and multivariate data analysis may not be the hot techniques in the field, as the disappearance after 2006.

The application of data mining technology in GIS (geographic information system) was a long story. Mining the property data of GIS using property correlative analysis and showing its result were of beneficial to its further development. Visualizing large amount of information interactively is one of the most attractive and useful capabilities of GIS. Correspondingly, from the table, visualization attempts to be a very important part of analysis. Otherwise, artificial neural networks and rough sets were the two most widely used methods. Despite that there was no breakthrough in theory, spatial association rule or decision tree have the potential application. Furthermore, the need

GIS	decision tree classification
	Artificial neural networks
	Spatial knowledge discovery
	Rough sets
	Predictive modelling
Geographical data mining	Large spatial databases
Geographical information system	Fuzzy c-means clustering
	visualization
Digital soil mapping	Spatial data mining

Table 5 The main technique of data mining in geography

for efficient data mining algorithms due to large amounts of spatial data brought about classification evolving to be a main research direction. On the contrary, spatial knowledge discovery had gone out of use.

3.3. The characteristic of sentiment words distributed in top 10 countries

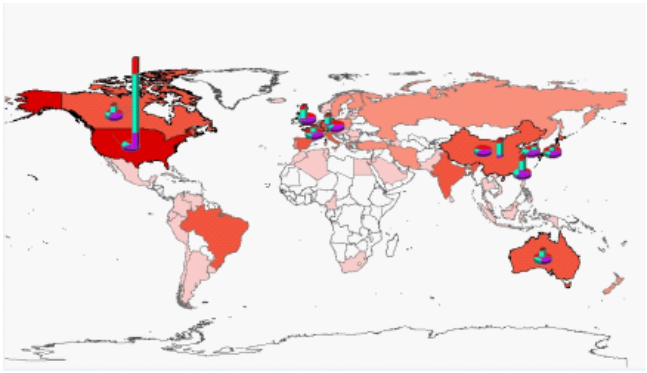

Figure 5 The distribution of sentiment words in top 10 countries

The contribution of different countries/territories was estimated by the location of the affiliation of at least one author of the published papers. Except that these articles without any author address information on the ISI Web of Science. The output was distributed in 73 different countries, we generated the data on the geographic and institutional distributions of publications in each countries/territories, we counted the number of papers belong to each countries, and displayed by different color in order. The most productive 10 countries in recent years are demonstrated in terms of number of publications and semantic words distribution in title, keyword and abstract in the figure 2. Figure 2 illustrates that the 7 major industrial countries Australia, Canada, France, Germany, South Korea, the UK, and the USA are all form developed world. The domination in publication was not surprising from mainstream countries since this pattern has occurred in most scientific fields [12]. To a certain extent, the number of research papers reflected the activity and academic level of these countries were likewise high [13].

It can be seen that there are distinctions among those semantic words used in publications from the ten most productive countries in data mining research. Specifically, all the ten countries found clustering, classification, association rules, rough sets, machine learning, neural networks, decision trees, knowledge discovery and bioinformatics to be hotspots in the current field of data mining research. It indicates, apparently, data mining commonly involves four classes of tasks, classification, clustering, regression and decision rule learning [14]. However, regression is a sophisticated data prediction technique with a wide range of application. Also, data mining is multidisciplinary, combing concepts, methods and analytical frameworks from a variety of research fields. Consequently, the 10 countries with most active academic resources covered machine learning, information science and relied on all of the data mining methods using techniques from other fields, such as neural networks, rough sets. To be mentioned that, in the previous experiment we have already found that almost all of the semantic words retrieved from title and keywords were included in abstract. But clustering, classification, association rules those frequently-used terms in data mining study In Japan the article is rarely in keyword and topic in the Japanese papers. This may due to the Japanese scholars are used to regard the innovative research points as keywords and title to attract the readers attention, not the traditional research topic.

4. Conclusion

In this study on related data mining papers dealing with the SCI database, we obtained some significant points on the global research performance. With the study of national research publications in the last 20 years, the increasing trend in the number of countries worldwide participating in data mining research can be easily observed. To a certain extent, a large numbers of research papers from a country is correlated with this countrys high activity and academic level. Clustering, classification, machine learning, and Knowledge Discovery were the emphasis of data mining research in all study periods. There are clear distinctions among author keywords used in publications from the five most productive countries in data mining research. Bibliometric method could quantitatively characterize the development of global scientific production in a specific research field. Furthermore, we are prepared to use data mining method to study data mining papers themselves.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (Grant No. 60975050, 60903168),

the Research Fund for the Doctoral Program of Higher Education (Grant No. 20070486081).

References

[1] Two Crows Corporation, Introduction to Data Mining and Knowledge Discovery, Third Edition (Potomas. MD: Two Crows Corporation, 1999): Pleter Adriaans and Dolf Zantinge, Data Mining (New York: Addison Wesley, 1996).

[2] Braun, T., Glanzel, W., & Grupp, H .. The sxientometric weight of 50 nations in 27 science areas, 1989-1993. Part 1. All fields combined, mathematics, engineering, chemistry and physics. Sciento-metrics, 33 (1995).

[3] HO, Y. S., Bibliometric analysis of adsorption technology in environmental science. Jounal of Enviro Protection Science (2007).

[4] Colman,A. M., Dhillon, D., & Coulthard, B. A bibliometric evaluation of the research performance of British university politics departments: Publications in leading journals. Scientometrics (1995).

[5] A.E. Bayer and J. Folger, "Some correlates of a citation measure of productivity in science", Sociology of Education, 39(1966).

[6] C.J.C.Burges, "A tutorial on support vector machines for pattern recognition", Data Mining and Knowledge Discovery **2**, 2(1998).

[7] George Chin Jr., Grant C. Nakamura, Daniel G. Chavarria and Heidi J. Sofia, "Graph Mining of Networks from Genome Biology", Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering, BIBE(2007)

[8] Keim. Daniel A, "Information visualization and visual data mining", IEEE Transactions on Visualization and Computer Graphics **8**, 1(2002).

[9] Niennattrakul. Vit, Wanichsan. Dechawut, Ratanamahatana. Chotriat Ann, "Accurate subsequence matching on data stream under time warping distance", Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 5669(2010).

[10] Ferreira, Pedro Gabriel, Azevedo, Paulo J, "Protein sequence pattern mining with constraints", Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),3721(2005).

[11] Ivosev G, Burton L, Bonner R. Dimensionality reduction and visualization in principal component analysis. Anal Chem, 80(2008).

[12] R.N. Kostoff, "The underpublishing of science and technology results", The Scientist 1, 4(2000).

[13] M. Rahman, T.L. Haque and T. Fukui, "Research articles published in clinical radiology journals: Trend of contribution from different countries", Academic Radiology, 12(2005).

[14] S.Curteanu, F.Leon,and D.Galea, "Alternatives for mutiobjective optimization of a polymerization process," J.Applied Polymer Science (2006).

Lu Dai received the M.S. degree in Computer Science from Wuhan University , Wuhan, China, in 2009. She is currently working toward the Ph.D. degree from State Key Laboratory of Software Engineering, Wuhan University, Wuhan, China. Her current research interests include evolutionary computation, text mining.

Lixin Ding received the B.S. and M.S. degrees from the Department of Applied Mathematics, Hunan University, Hunan, China, in 1989 and 1992, respectively, and the Ph.D. degree from State Key Laboratory of Software Engineering from Wuhan University, Wuhan, China, in 1998.

From 1998 to 2000, he was a Postdoctoral Fellow at the Department of Armament Science and Technology, Naval University of Engineering, Wuhan, China. He is currently a Professor at the State Key Laboratory of Software Engineering, Wuhan University, Wuhan, China. His main research interests include basic theory of evolutionary computation, intelligent information processing, and machine learning.

Yunwen Lei received the B.S. degree from the College of Mathematics and Econometrics, Hunan university, Hunan, China, in 2008. He is currently a Ph.D. student at State Key Laboratory of Software Engineering, Wuhan University, Wuhan, China. His research interests include computational learning theory, statistical learning, evolutionary computation and optimization theory.

Yangge Tian received his Ph.D. degree from Wuhan university, Wuhan, 430072, in 1998. He is now an associate professor at Wuhan University, P.R. China. His current research include spatial data mining, Social Network, machine learning, evolutionary computation and optimization theory. He is also a reviewer of Science in China (Information), Journal of Computers and Information Processing, Letter,

the IEEE, the Trans on System, Man & Cybernetics (Part B), IEEE Trans on Evolutionary Computation, and other authoritative publications at home and abroad.