International Journal of Hybrid Information Technology
Vol.8, No. 5 (2015), pp. 105-112
http://dx.doi.org/10.14257/ijhit.2015.8.5.11

Analysis of Scientific Papers Research Trend Based on Granular Computing

Wang Xiaodan, Yu Guang and Li Xueting

Library of Harbin Institute of Technology
wangxd@hit.edu.cn

Abstract

The purpose of this paper is to deal with the data in large document databases, analyzes the research trend of scientific papers and research hot spots. There is practical significance to promote the development of science and technology in China in order to provide effective methods and tools for the researchers.

Keywords: *Granular Computing, Scientific Papers Research, Research Trend*

1. Research Status on Analysis of the Research Trend of the Scientific Paper

Scientific thesis embodies the research results of various disciplines in the field, is a scientific summary of various theories and methods. With the popularization and ripeness of the internet, the era of the great explosion of the information has already come. With the growing of the number of scientific papers at an alarming rate, that leads to the emergence of various digital libraries and literature database (such as Chinese science paper online). Mining and analysis related to these massive data has been the focus of many scientific researchers. [15] The literature from a variety of database, the analysis of dynamic interdisciplinary science paper changed in each historical period (research trends) over time.

At present, it is well known that America ITDT (Interactive Topic Detection and Tracking) is a research project in the field of analysis and discovery [1], in order to deal with the increasingly severe for Internet information explosion problem. The domestic good system is that public sentiment warning assistant decision Support System (ZHISI) in Beijing University FANGZHENG Institute of technology. It successfully realizes automatic and real-time monitoring for Internet public sentiment. This aspect of the theory and technology is mainly applied in the hot news, hot network information finding and tracking, but applied in the analysis of technology research hot spot and trend of research on the less.

2. The Significance of the Research Trend Analysis of Scientific Papers

According to domestic and foreign literature research, the analysis of the literature research trend mainly based on keywords frequency statistics of literature metrology. The Metrology is an important theoretical tool [16] to analyze the literature trend, which uses mathematical and statistical methods to describe the current situation and development trend, evaluation and forecast of science and technology.

At present, the research trends of scientific papers have many problems at home and abroad. It mainly includes the following aspects:

1. Somewhat lacking in large-scale data processing;
2. Can not be effective for large-scale literature database processing;

ISSN: 1738-9968 IJHIT

3. Limited to specific fields;

4. The lack of interdisciplinary collaboration in research theory and method;

5. The analysis method is single, the more based on keywords, the less for the content.

In the development of science and technology, the traditional natural science has been further subdivision and development, the new disciplines constantly produce. The integration trend of interdisciplinary gradually increase, the discipline development tends to be integration. In this paper, it is based on the granular computing [2] technology, we adopted the related technologies (such as automatic keyword extraction, text clustering[3], text classification [4] and so on) in text mining and processed data for all kinds of large-scale science and technology literature database (such as Chinese science paper online). Thus found the prediction of hot research topics of scientific papers, and the formation of automatic analysis and forecasting system.

3. The Research Content of the Research Trend Analysis of Scientific Papers

This research intends to combine theory with experiment, which based on granular computing, adopted the idea of modularization. We constructed test knowledge base based on the electronic resources of Harbin Institute of Technology library, which contains all the metadata including the title, source, author, keywords, abstract, text, references, *etc*. The foundation based on the metadata of the knowledge base, which research the papers of development trend and hotspot detection

3.1. Granular Computing (GrC)

As Granular computing (GrC) is a newly developed technique which has been drawn attention by researchers. [5] It is an umbrella term to cover any theories, methodologies, tools and techniques that make use of granules in problem solving. [17] Basic ingredients of granular computing, *i.e*., granules, are subsets, classes, and clusters of a universe. They have been considered either explicitly or implicitly in many other fields, such as data and cluster analysis, database and information retrieval, concept formation, and machine learning. [5-6] in data mining and classification problems, we can view equivalence classes as granules [18].

3.2. The Four Models of Information Retrieval Systems

Following the study of probability of relevance information retrieval can be formalized into four models as shown in Figure 1. [19] Let u be a single user, U a class of users, d a document and D a class of documents, these four models and relationships among them can be depicted as in Figure 1. [19] Granulation of a universe involves the decomposition of the universe into parts, or the grouping of individual elements into classes, based on available information and knowledge. Elements in a granule are drawn together by in distinguishability, similarity, proximity or functionality. [6] Using techniques and principles of granular computing, one can study these models from granulation point of view, *i.e*., grouped and personalized views should be also studied. Model 1 and Model 2 are considered as finer granulation than Model 0. The elements of these two models are granules of the elements of Model 0. The same principle applies to other models. Model 3 is the finest granulation; Model 0 is the coarsest granulation.

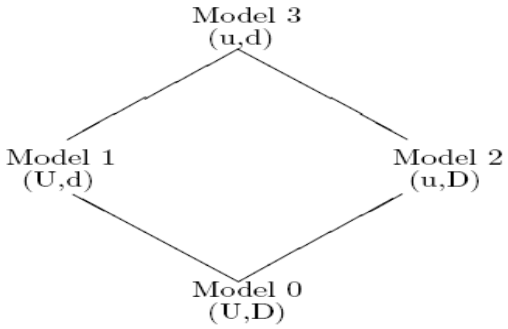

Figure 1. Information Retrieval Models

3.3. The Keyword Extraction of Scientific Papers

In a variety of scientific paper database, there are large amounts of data to he managed by order. Wherein the absence of a certain amount of paper keyword, so that the data is not complete, and cannot carry on the statistical analysis based on keywords, then that affects the accuracy of the analysis result. At the same time, wherein also exist some noisy keywords or keywords quantities too small, so the paper characteristics are not well represented. For these cases, this paper will research papers trend based on keyword extraction techniques.

3.4. Scientific Papers Research Trend Analysis

Each time the same disciplines have different hotspot of study. The paper will study about the automated analysis techniques of development trend of thesis. So forming an analysis report provides a useful reference for the analytical work of scientists. During the same period in various disciplines, the scientific papers will express in hotspot that a large number of published scientific papers similar research this time. In this period of discovery research papers hotspot, may be used text clustering techniques. The class which contains the number of documents can be considered the hotspot of this period, the appropriate category features express in the hotspots.

3.5. Technology Thesis Trend Analysis and Forecasting System

We will accomplish the automatic analysis forecasting system of scientific papers research trends, this system has the following characteristics:

1. The system is stable and efficient；
2. The system can automatically run；
3. Suitable for a variety of literature database；

4. Analysis of Research Process

This research project sets up a test knowledge base based on the electronic resources of the Harbin Institute of Technology library, which contains all the metadata including the title, source, author, keywords, abstract, text, references, etc. On the basis of the papers metadata in the knowledge base, we have studied the paper trend and hotspot detection. The automatic analysis system is based on the keywords. The system flow chat is as follows:

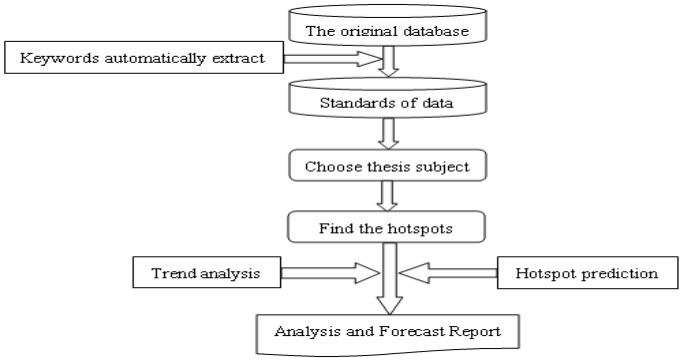

Figure 2. The System Flow Chat

4.1. Create a Metadata Database of Papers

The all databases in have been integrated into a metadata knowledge base, the following is the management platform of knowledge base.

The knowledge base of HIT

Create Database | Edit | Display All

Go to page : 1 / 7

Database	Code	State	Supplier
2011 Cambridge Journals Digital Archive - CJDA Full Collection	-2P	Subscribed	Cambridge University Press
Academic Search Premier	EAP	Subscribed	EBSCOhost
ACM Digital Library	ACM	Subscribed	ACM Digital Library
AIAA Conference Proceedings	AEL	Subscribed	American Institute of Aeronautics and Astronautics
Allen Press Journals	Y3D	Subscribed	Allen Press, Inc
American Chemical Society Web Editions	ACS	Subscribed	American Chemical Society
American College of Physicians	TPH	Subscribed	American College of Physicians
American Institute of Aeronautics and Astronautics Publications	RAA	Subscribed	American Institute of Aeronautics and Astronautics
American Institute of Physics (AIP) Publications	RIP	Subscribed	American Institute of Physics
American Physical Society Journals	3MX	Subscribed	American Physical Society
AMS Books Online (Freely Accessible)	9L9	Subscribed	American Mathematical Society
Annual Reviews 1996 - present	1KX	Subscribed	Annual Reviews
Arts & Humanities Citation Index	AKT	Subscribed	Thomson Reuters
arXiv Computer Science	AKY	Subscribed	Cornell University
arXiv Mathematics	AKZ	Subscribed	Cornell University
arXiv Nonlinear Science	ALA	Subscribed	Cornell University
arXiv Physics	ALB	Subscribed	Cornell University
arXiv.org	GOX	Subscribed	Cornell University
ASABE Technical Library	RAE	Subscribed	American Society of Agricultural and Biological Engineers
ASCE Conference Proceedings	RWO	Subscribed	American Society of Civil Engineers
ASCE Library	RAC	Subscribed	American Society of Civil Engineers
ASME Proceedings (Archives)	SDN	Under Review	American Society of Mechanical Engineers
ASME Transactions Journals (Current)	RAI	Subscribed	American Society of Mechanical Engineers
Bentham Science Journals	GH2	Subscribed	Bentham Science Publishers
BioOne Abstracts & Indexes	7QN	Subscribed	ProQuest

Figure 3. The Management Platform of Knowledge Base

4.2. The Standard of Data Format

According to the norms format of the knowledge base, we automatically extract data from each database so that forming a metadata repository. When retrieving data, retrieve metadata repository directly, no longer access to the original database.

4.3. Choose Thesis Subject

In the metadata repository, may select the corresponding discipline, then to retrieval metadata abased on the keywords.

4.4. Research Trends and the Hot Spot Analysis

Analysis of the retrieval results, it is concluded that the analysis report.

4.5. Examples

Such as the search key words in the field of computer: “Java and database”, the retrieval results as follows:

1. The retrieval results include 349 records. The retrieval time is from 1994 to 2014. Figure 4 shows the articles in each issue every year.

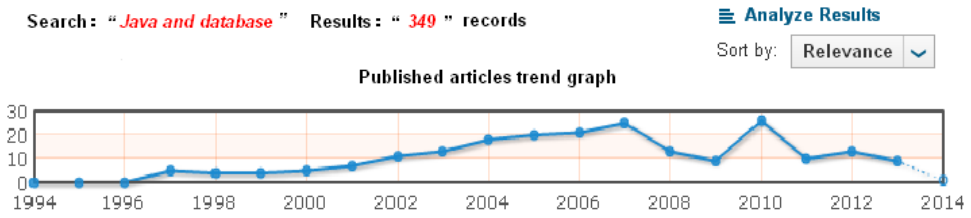

Figure 4. Published Articles Trend Graph

2. Through the analysis, the following analysis results are obtained. Figure 5 shows the knowledge related to the keywords "java and database". Figure 6 shows the article number in journal issue every year. Figure 7 shows the article in dissertation issue every year.

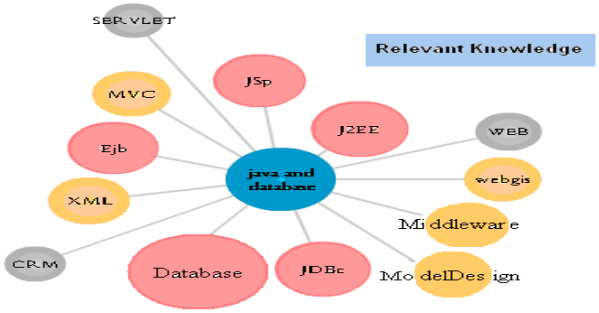

Figure 5. Relevant Knowledge

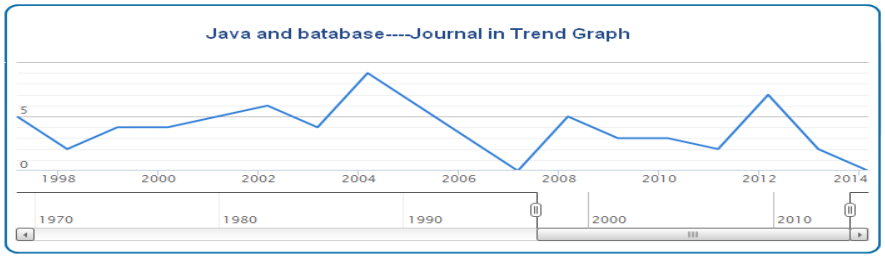

Figure 6. Journal in Trend Graph

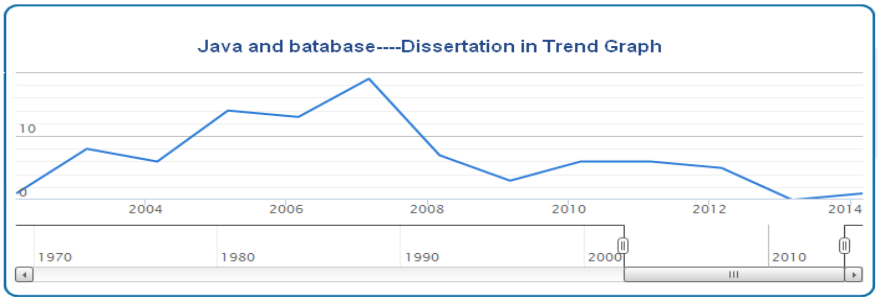

Figure 7. Dissertation in Trend Graph

5. Conclusion

In the paper, this project realizes the automatic extraction of keywords, automatically; get the retrieval results and automatic analysis of the retrieval results. Study on the trend of the research of this project can effectively analyze and predict

the various disciplines of a research trend, for scientific research workers to provide accurate basis for scientific research.

Reference

[1] M. Masnizah, C. Fabio, R. Ian, "Evaluation of an interactive topic detection and tracking interface", Journal of information science, vol. 4, no. 38, **(2012)**.
[2] S. Andrzej, S. Jaroslaw, S. Roman, "Modeling rough granular computing based on approximation spaces", Information sciences. vol. 1, no. 184, **(2012)**.
[3] C. C. Campos, P. B. Galvan, A. V. Coronado, "Improving statistical keyword detection in short texts: Entropic and clustering approaches", Physica A-Statistical mechanics and its Applications, vol. 6, no. 392, **(2013)**.
[4] L. Sheng-Tun and T. Fu-Ching, "A fuzzy conceptualization model for text mining with application in opinion polarity classification", Knowledge-based systems, vol. 39, **(2013)**.
[5] Y. Y. Yao, "Relational interpretations of neighborhood operators and rough set approximation operators", Information Sciences, vol. 111, **(1998)**.
[6] L. A. Zadeh, "Towards a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic", Fuzzy sets and systems, vol. 19, **(1997)**.
[7] A. Cem, C. Fazli, K. Seyit, "Novelty detection for topic tracking. Journal of the American society for information science and technology", vol. 63, **(2012)**.
[8] W. Wei, X. Xin, "Online public opinion hotspot detection and analysis based on document clustering", New Technology of Library and Information Service, vol. 3, **(2009)**.
[9] J. Yongxin, Z. Huaqing, "Trend Analysis of Library and Information Sciences Based on Co-Keyword Statistics", Library and Information Service, vol. 9, **(2008)**.
[10] W. Sizhu, "Analysis of Hot Spots in Field of Data Mining and Knowledge Discovery", Journal of Intelligence, vol. 7, **(2010)**.
[11] S. Shaoting, C. Fang, "Applied study on text mining technique to S＆T management field hot topic extraction direction. Computer applications and software, vol. 7, **(2012)**.
[12] Y. Yan, L. H. Chen, W. C. Tjhi, "Fuzzy semi-supervised co-clustering for text documents", Fuzzy sets and systems, vol. 215, **(2013)**.
[13] A. Skabar and K. J. Abdalgader, "Clustering sentence-level text using a novel fuzzy relational clustering algorithm", IEEE transactions on knowledge and data engineering, vol. 2, **(2013)**.
[14] V. Carlos, S. David, M. Antonio, "Engineering applications of artificial intelligence", vol. 26, **(2013)**.
[15] R. Sanderson and P. Watry, "Integrating Data and Text Mining Processes for Digital Library Applications", Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries, **(2007)**.
[16] M. K. McBurney, P. L. Novak, "What is bibliometrics and why should you care?", IEEE International Proceedings on Professional Communication Conference, **(2002)** September 17-20.
[17] Y. Y. Yao, "Granular computing: basic issues and possible solutions", In Proceedings of the 5th Joint Conference on Information Sciences, **(2001)**.
[18] Y. Y. Yao and J. T. Yao, "Granular computing as a basis for consistent classification problems", In Proceedings of PAKDD'02 Workshop on Toward the Foundation of Data Mining, **(2002)**.
[19] J. T. Yao, Y. Y. Yao, "Information Granulation for Web based Information Retrieval Support Systems", Data Mining and Knowledge Discovery: Tools And Technology v Proceedings of the society of Photo optical Instrumentation Engineers (SPIE), **(2003)**.
[20] C. Wartena and R. Brussee, "Topic Detection by Clustering Keywords. Database and Expert Systems Application", 19th International Conference, **(2008)** September 1-5.
[21] Y. WenBin and L. Ronaldo, "Exploring user feedback of a e-learning system: A text mining approach", 15th International Conference on Human Interface and the Management of Information: Information and Interaction for Learning, Culture, Collaboration and Business. **(2013)** July 21-26.
[22] A. Dinesh, "Finding similar files using text mining. Proceedings of the 8th International Conference on Computer Science and Education", **(2013)** August 26-28.
[23] K. M. Saroj, K. P. Sankar and D. Soumitra, "Granular computing models in the classification of web content data", 2012 IEEE/WIC/ACM International Conference on Intelligent Agent Technology, **(2012)** December 4-7.
[24] A. A. Alsaw and H. A. Hefny, "Solving uncertain shortest path problem based on granular computing", 2013 IEEE International Conference on Computational Intelligence and Computing Research, **(2013)** December 26-28.
[25] I. Bhushan and P. Ujawla, "Operational pattern revealing technique in text mining. 2014 IEEE Students' Conference on Electrical", Electronics and Computer Science, **(2014)** March 1-2.
[26] M. Sukanya and S. Biruntha, "Techniques on text mining. 2012 IEEE International conference on advanced communication control and computing technologies", (**2012)** August 23-25.
[27] N. Haydemar and R. Esmeralda, "Automatic classification of academic documents using text mining techniques", 38th Latin America Conference on Informatics, **(2012)** October 1-5.

[28] S. Neeraj, M. K. Kumar and G. S. Thakur, "Document clustering using message passing between data points", 2013 International Conference on Communication Systems and Network Technologies, **(2013)** April 6-8.
[29] X. D. Wang, G. Yu, T. C. Wang, "Granular Computing for Web Applications", International conference on artificial intelligence and soft computing, **(2012)** March 15-16.
[30] G. B. Orgaz and C. David, "Comparative study of text clustering techniques in virtual worlds", 3rd International conference on web intelligence, mining and semantics, **(2013)** June 12-14.
[31] S. C. Punitha, M. Punithavalli, "Performance evaluation of semantic based and ontology based text document clustering techniques. International Conference on Communication Technology and System Design 2011, **(2011)** December 7-9.
[32] E. A. Calvillo, A. Padilla, J. Munoz, J. Ponce and J. T. Fernandez, "Searching Research Papers Using Clustering and Text Mining", 23rd Annual International Conference on Electronics, Communications and Computing, **(2013)** March 11-13.
[33] D. Munkova, M. Munk, M. Vozar, "Data Pre-Processing Evaluation for Text Mining: Transaction/Sequence Model", 2013 International conference on computational science, **(2013)** June 5-7
[34] W. Black, R. Procter, S. Gray, S. Ananiadou, "A data and analysis resource for an experiment in text mining a collection of micro-blogs on a political topic", 8th International Conference on Language Resources and Evaluation, **(2012)** May 21-27.
[35] J. Manimaran and T. Velmurugan, "A survey of association rule mining in text applications", 2013 IEEE International conference on computational intelligence and computing research, **(2013)** December 26-28.
[36] M. Pandey and V. Ravi, "Detecting phishing e-mails using text and data mining", 3rd IEEE International conference on computational intelligence and computing research. **(2012)** December18-20.

Authors

Xiaodan Wang, received the B. Eng and M. Eng degree in Harbin Institute of Technology (HIT), China in 1997 and 2003 respectively. She is currently working at Harbin Institute of Technology Library. Her current research interests on digital library and data mining.

Guang Yu, received her B. Eng, M. Eng, and Ph.D. Man in Harbin Institute of Technology (HIT), China in 1985, 1990 and 2007 respectively. She is currently working at the school of Economics and Management of Harbin Institute of Technology as a professor. Her current research interests on complex network theory and the applications of pattern recognition theory in social network, information dissemination network and financial data management.

Xueting Li, received the B.A and M.A degree in English Language and Literature from Northeast Forestry University (NFC), China in 2004 and 2007 respectively. She is currently working at Harbin Institute of Technology Library. Her current research interests on digital library and bibliometrics.