

Applying Association Rule Analysis in Bibliometric Analysis

—A Case Study in Data Mining

Fang Li, Chengyao Li, and Yangge Tian*(correspondence author)
International School of Software, Wuhan University, 129 Luoyu Road, Wuhan, China
tiandebox@126.com

Abstract—Scientific research needs lots of literature searches that cost a large amount of time and energy. Bibliometric techniques could help us find research hotspots and grasp the research direction yet can't reveal huge hidden information in massive literature since the existing bibliometric analysis techniques employ mainly simple statistical analysis techniques. Hence, we propose to introduce data mining analysis techniques into bibliometrics analysis, and expect to reach some instructive conclusions by mining relations among information like keywords, authors, research institutions, publications and so on. We take the subject 'data mining' as our research object and analyze the records by the method of association rule analysis. Finally, we get some valuable conclusions that verify the idea's feasibility. The results of this research would not only provide a reference for data mining research but also be applied to other research fields.

Index Terms—Terms: data mining; bibliometric; association rule; SCI

I. INTRODUCTION

Large amounts of literature always need to be queried during conducting scientific research, thus it is very crucial to quickly find out research hotspots and grasp the main development directions of current academic research from those documents. Manual analyses which mainly employs simply methods that are not only time-consuming and laborious but also heavily dependent on researcher's personal experience and research interests, often fail to fully extract implicit information and inherent laws.

Nowadays, more and more researchers begin to use bibliometric techniques with statistical analysis of literature content information, citation information, author information, external features of documentation and other related information, so as to provide useful guidelines for research work[1][2]. However, existing bibliometric techniques, while emphasizing using mathematical and statistical methods to describe, evaluate and predict the status and development trend of science and technology, generally use relatively simple mathematical methods (which are basically elementary mathematics methods), derive relatively plain conclusions, and provide limited guidance for scientific research. Hence, we introduce data mining techniques into bibliometric methods.

Data Mining, also known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful

information from data in databases [3]. With the rapid growth in a variety of data, data mining has become an important research topic and is receiving substantial interest from both academia and industry [4].

Association rule mining is one of the most active research directions in data mining. It reflects the interdependence and relevance between things. If there are certain associations between things, then one's information could be predicted by analyzing others'. An association rule problem between itemsets in mining customer transaction database was first proposed by Agrawal in 1993[5], and since then many researchers have conducted a lot of researches regarding mining problems of association rules in the future. At present the technology is already widely applied in business, medicine, earth sciences and other fields.

It is easy to discover some similarities between shopping basket data (which are commonly used in association analysis) and keywords, authors as well as other data sets in bibliometric analysis, so we consider using the application of association rules in bibliometric study and make use of such data mining methods to find hidden information and laws from a large number of the literature data.

II. METHODS AND MATERIAL

In bibliometric study, the Science Citation Index (SCI), provided by the Institute for Scientific Information (ISI) Web of Science databases, is the most important and frequently used source database for the review of scientific achievement in all research fields [6][7].

The data we used was extracted from the SCI online database. Data mining and its synonyms (such as data mine, KDD, exploratory data analysis, information discovery, information extraction, intelligent data analysis etc.) were used as the search topics. All the information was obtained on July, 16th, 2008 when the SCI search process for this study was conducted. The total number of papers related to data mining research in the ISI web database was 10286. After deleting the repeated records, the total number was 9808, which were published between 1962 and 2008. These were published with 13 document types with the distribution analysis. There were 8930 regular/research articles, which accounted for 91.07% of the total production, followed by reviews (303; 3.09%), meeting abstract (214; 2.18%), editorial materials (194; 1.97%), book review(78; 0.8%), letter (26; 0.27%), news

items (14; 0.14%), note(14; 0.14%), correction (9; 0.09%), Biographical-Item (5; 0.05%)and reprint (2; 0.02%).

We extracted from SCI database key words, authors, source, time cited, author address and some other records as its original form, and then transformed them to formats that fit for association rule analysis. Finally, all these records were imported into SQL SERVER 2005.

III. METHODS

In our research, we primarily used data mining techniques and adopted the classical association rules algorithm in analysis. The definition and algorithm of association rule are introduced below:

A. Definition

Let $I = \{i_1, i_2, \dots, i_m\}$ be a set of items. Let D , the task relevant data, be a set of database transactions where each transaction T is a set of items such that $T \subseteq I$. Each transaction is associated with an identifier, called TID. Let A be a set of items. A transaction T is said to contain A if and only if $A \subseteq T$. An association rule is an implication of the form $A \Rightarrow B$, where $A \subset I$, $B \subset I$ and $A \cap B = \phi$. The rule $A \Rightarrow B$ holds in the transaction set D with support s , where s is the percentage of transaction in D that contain $A \cup B$. The rule $A \Rightarrow B$ has confidence c in the transaction set D if c is the percentage of transactions in D containing A which also contain B . That is,

$$\text{support}(A \Rightarrow B) = \text{Prob}(A \cup B)$$

$$\text{confidence}(A \Rightarrow B) = \text{Prob}(B | A) [8].$$

Support for association rule reflects the frequency of the rule while confidence indicates the accuracy of the rules. In this study, we set the min_support as 5 and the min_confidence as 20%. (In some case, we set the min_support as 3 in order to reach better results.)

B. Algorithm

Association rule mining generally has two steps[8]:

Step 1: Find all frequent itemsets. By definition, each of these itemsets will occur at least as frequently as a pre-determined minimum support count.

Step 2: Generate strong association rules, which are bigger than or equal with the minimum support and minimum confidence, from the frequent itemsets.

During data mining, we use Apriori Algorithm which is the most influential algorithm for mining association rules[9]. Apriori employs an iterative approach, where k-itemsets are used to explore (k+1)-itemset. First, find the set of frequent 1-itemsets which is denoted L_1 . L_1 is used to find L_2 , the frequent 2-itemsets, which is used to

find L_3 , and so on, until no more frequent k-itemsets can be generated. The following two steps are responsible for the process of finding L_k through L_{k-1} :

The join step: C_k is generated by joining L_{k-1} with itself. C_k Here stands for Candidate itemsets of size k.

The prune step: C_k is a superset of L_k , that is, its members may or may not be frequent, but all of the frequent k-itemsets are included in C_k . Any (k-1)-itemsets that is not frequent cannot be a subset of a frequent k-itemsets, hence should be removed.

IV. RESULTS

Through this method, we get some interesting results. After analysis, we reached the following conclusions:

A. Keywords analysis

In traditional keywords analysis, we generally find research hotspots and research directions through statistics on the number and changes of papers' common keywords at different times, as shown in Table I:

TABLE I. FREQUENCY OF AUTHOR KEYWORDS USED IN PUBLICATIONS—TOP 30(EXCERPT FROM REFERENCE[10])

DE	1999-2008	P (%)	1999-2003	P (%)	2004-2008	P (%)
data mining	2472	26.65	809	25.74	1663	27.11
clustering↑	279	3.01	68	2.16	211	3.44
classification↑	266	2.87	81	2.58	185	3.02
machine learning	254	2.74	97	3.09	157	2.56
Knowledge Discovery ↓	219	2.36	110	3.50	109	1.78
association rules ↑	206	2.22	73	2.32	133	2.17
information extraction↑	198	2.13	57	1.81	141	2.30
exploratory data analysis ↓	156	1.68	76	2.42	80	1.30
bioinformatics↑	146	1.57	36	1.15	110	1.79
neural networks ↓	136	1.47	62	1.97	74	1.21
decision trees↓	91	0.98	38	1.21	53	0.86
association rule↑	80	0.86	21	0.67	59	0.96
feature selection	78	0.84	24	0.76	54	0.88
decision tree	73	0.79	18	0.57	55	0.90

Yet, mere statistical analysis cannot fully reveal the implicit rules among keywords. Sometimes, some research directions were neglected by researchers simply because the authors chose different keywords. The association analysis can just make up for this deficiency by analyzing the correlations between keywords and thus presenting the relations between different research fields. As shown below:

TABLE II. ASSOCIATION RULES IN KEYWORDS

Confidence	Rules
1.000	motifs , patterns--> clusters
0.857	signal detection--> pharmacovigilance
0.833	Multidimensional scaling, feature extraction--> Sammon mapping
0.429	workflow management--> Petri nets
0.429	workflow management--> process mining
0.417	combinatorial chemistry--> nonlinear mapping
0.357	granular computing--> rough set
0.353	fuzzy set--> quantitative value
0.294	fuzzy set--> rough set
0.278	Web usage mining--> clustering
0.263	outlier detection--> clustering
0.250	discretization--> machine learning
0.250	discretization--> classification
0.240	decision trees--> classification
0.240	decision trees--> classification
0.238	entropy--> machine learning
0.231	unsupervised learning--> clustering
0.231	unsupervised learning--> clustering
0.227	regression--> classification
0.219	rule induction--> machine learning
0.219	rule induction--> machine learning
0.200	support vector machines--> machine learning
0.200	feature selection--> classification
0.200	self-organizing map--> clustering

Table III presents part of the important rules we discovered through association rules analysis, sorting them according to each rule’s confidence. The majority of these rules indicate research methods or techniques in certain area, such as, ‘multidimensional scaling, feature extraction -> Sammon mapping’, ‘workflow management -> Petri nets’ and so on, while some rules show the correlations between two methods or techniques, like ‘fuzzy set--> rough set’. These findings are of great instructive importance to researchers.

Compared to the conclusion reached by traditional bibliometric analysis, as shown in Table III, results of analysis employed association rules are apparently more informative and enlightening. For instance, from Table II, we can only figure out that clustering is an important research direction, yet we discovered that clustering is closely related with many techniques like outlier detection, unsupervised learning, self-organizing map and so on, and could be applied in motifs patterns and web usage mining analysis.

Though some association rules’ confidence is relatively lower than others, it doesn’t necessarily means that it is of little importance. Confidence is influenced by popularity. That is, if a keyword refers to a new technique which was new introduced, it’s thus hard to be discussed widely. Consequently, rules associate with such keywords are not possible to be of high confidence yet still have great value in instructing research. For instance, the rule ‘Web usage mining -> clustering’ possess’ has a low confidence since ‘web usage mining’ is a new application of data mining and appeared only recent years. This rule indicates that clustering methods could be applied in ‘web usage mining’. Though it is not widely been discussed yet, it enlightens other researchers who are interested in this area.

B. Keywords and Journals

Similarly, we studied association rules between keywords and journals. As shown in Table IV, the analysis reflects different journals’ preference of keyword. For example, ACM journals prefer data mining algorithms while International Journal on Document Analysis and Recognition shows particular interest in information extraction. These findings could lead researchers to read some journals correspond to their research interests, or choose the proper publications to submit their research results.

TABLE III. ASSOCIATION RULES BETWEEN KEYWORDS AND JOURNALS

Confidence	Rules
1.000	SURFACE AND INTERFACE ANALYSIS -> chemical analysis
1.000	ACM COMPUTING SURVEYS -> algorithms
0.944	ACM TRANSACTIONS ON DATABASE SYSTEMS -> algorithms
0.750	INTERNATIONAL JOURNAL ON DOCUMENT ANALYSIS AND RECOGNITION -> information extraction
0.750	APPLICATIONS OF BIOINFORMATICS IN CANCER DETECTION -> bioinformatics
0.737	INTERNATIONAL JOURNAL OF DATA MINING AND BIOINFORMATICS -> bioinformatics
0.667	JOURNAL OF COMPUTATIONAL CHEMISTRY -> combinatorial chemistry
0.600	JOURNAL OF STATISTICAL MECHANICS-THEORY AND EXPERIMENT -> data mining (experiment)
0.438	ACM TRANSACTIONS ON INFORMATION SYSTEMS -> algorithms
0.417	KYBERNETES -> cybernetics
0.375	ANNALS OF OPERATIONS RESEARCH -> classification
0.333	DATA WAREHOUSING AND KNOWLEDGE DISCOVERY, PROCEEDINGS -> association rules
0.300	IEEE TRANSACTIONS ON SOFTWARE ENGINEERING -> association rules
0.286	JOURNAL OF PROTEOME RESEARCH -> bioinformatics
0.273	ADVANCED DATA MINING AND APPLICATIONS, PROCEEDINGS -> association rules
0.263	IEEE TRANSACTIONS ON SYSTEMS MAN AND CYBERNETICS PART A-SYSTEMS AND HUMANS -> classification
0.231	AMERICAN STATISTICIAN -> EXPLORATORY DATA ANALYSIS

C. Authors and Keywords

The analysis of relations between authors and keywords revealed the foremost researchers in certain area. For example, as shown in Table VI, ‘Petri nets, workflow management -> van der Aalst, WMP’ suggests that the main researcher about Petri nets and workflow is van der

Aalst and WMP. Hence, other researchers interested in this area could pay more attention to his papers, publications and also research trend.

TABLE IV. ASSOCIATION RULES BETWEEN KEYWORDS AND AUTHORS

Confidence	Rules
1.000	clusters, patterns -> Parida, L
1.000	clusters, motifs -> Parida, L
1.000	nonlinear mapping, combinatorial chemistry -> Agrafiotis, DK
0.833	Petri nets, workflow management -> van der Aalst, WMP
0.833	quantitative value, fuzzy set -> Hong, TP
0.750	classification problems -> Hu, YC
0.714	information visualization, visual data mining -> Keim, DA
0.556	rough sets, pattern recognition -> Pal, SK
0.500	information fusion -> Wang, ZY
0.500	information fusion -> Leung, KS
0.444	soft computing -> Pal, SK
0.357	decision making -> Kusiak, A
0.316	data clustering -> Chen, MS
0.313	evolutionary computation -> Wong, ML
0.313	evolutionary computation -> Leung, KS
0.294	fuzzy set -> Wang, SL
0.263	Grid computing -> Talia, D
0.261	distributed data mining -> Kargupta, H
0.240	Bayesian networks -> Wong, ML
0.227	customer relationship manageme -> Van den Poel, D

D. Keywords and Research institutions

Similar to analysis of relations between authors and keywords, analysis of relations between keywords and research institutions, as shown in Table VII, indicates institutions' preference in some certain research area. This helps researchers find out relevant research institutions and carry out academic exchanges.

TABLE V.

Confidence	Rules
1.000	chemical analysis -> Tottori Univ
1.000	biomedical literature data min -> Norwegian Univ Sci & Technol
1.000	Petri nets -> Eindhoven Univ Technol
1.000	CLASSIFIER DESIGN -> Ben Gurion Univ Negev
1.000	cascade generalization -> Univ Wisconsin
1.000	Choquet integrals -> Chinese Univ Hong Kong
0.750	information visualization -> Univ Konstanz
0.750	decision making -> Univ Iowa
0.600	cybernetics -> Portland State Univ
0.500	clusters -> Univ Haifa
0.500	classification problems -> Chung Yuan Christian Univ
0.500	distributed data mining -> Univ Calabria
0.333	data cleaning -> Florida Atlantic Univ
0.250	combinatorial chemistry -> 3 Dimens Pharmaceut Inc

Besides, we also conducted research on the relations between authors and journals, keywords and time cited and so forth, and came to some interesting results as well.

V. CONCLUSION

Association analysis is an important data mining technique, widely used in commercial, financial, telecommunications, medical fields and so on, but rarely

applied in the bibliometric analysis. Our research shows that association rules can discover information hidden in the keywords, publications, authors, research institutions and other materials. In particular, it can instruct researchers to find research fields and techniques related to its research direction. At the same time, it helps them broaden research ideas and even discover new research fields by providing relevant publications, authors and research institutions. This has significant instructive value to research works.

Different from general rules, when applying association rules in bibliometrics, we should also pay attention to those rules with relative low confidence because scientific research stressed on innovativeness. Generally, new research directions have a relatively low confidence since researches that focus on them are limited yet. On the other hand, some rules with a relatively high confidence might have little guiding significance. Hence, making the right judgment on this matter requires researchers' research experience.

On the whole, the association rule analysis provides us a new, objective and credible approach to analyze and evaluate scientific literature. It instructs us conducting scientific research by quickly discovering implicit information and rules among numerous literature data and shows considerable research perspectives and application value.

ACKNOWLEDGMENT

This paper is supported by NSFC (Granted No. 40601026).

REFERENCES

- [1] Cronin, 2001 B. Cronin, "Bibliometrics and beyond: Some thoughts on web-based citation analysis," *Journal of Information Science*, vol. 27, pp. 1-7, January 2001.
- [2] H.F. Moed, R.E. Debruin and T.N. Vanleeuwen, "New bibliometric tools for the assessment of national research performance—Database description, overview of indicators and first applications," *Scientometrics*, vol. 33, pp. 381-422, July 1995.
- [3] W. Frawley and G. Piatetsky-Shapiro and C. Matheus. "Knowledge Discovery in Databases: An Overview". *AI Magazine*: pp. 213-228, Fall 1992.
- [4] JF Roddick, M Spiliopoulou. "A bibliography of temporal, spatial and spatio-temporal data mining research". *ACM SIGKDD Explorations Newsletter*, vol. 1(1), 1999, pp. 34 - 38.
- [5] W.T. Chiu and Y.S. Ho, "Bibliometric analysis of homeopathy research during the period of 1991 to 2003," *Scientometrics*, vol. 63, pp. 3-23, March 2005.
- [6] A.E. Bayer and J. Folger, "Some correlates of a citation measure of productivity in science", *Sociology of Education*, vol. 39 (4) , 1966, pp. 381-390.
- [7] R.N. Kostoff, "The underpublishing of science and technology results", *The Scientist*, vol. 14 (9), 2000, p. 6.
- [8] Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, 2000, chap. 6 pp.4-5
- [9] http://en.wikipedia.org/wiki/Apriori_algorithm
- [10] Shuang Deng, Yangge Tian, "Using the bibliometric analysis to evaluate global scientific production of data mining papers"