Jointly published by Akadémiai Kiadó, Budapest
and Springer, Dordrecht

Scientometrics, Vol. 81, No. 1 (2009) 33–46
DOI: 10.1007/s11192-009-2063-x

An author co-citation analysis of information science in China with Chinese Google Scholar search engine, 2004–2006

RUIMIN MA, QIANGBIN DAI, CHAOQUN NI, XUELU LI

School of Information Management, Wuhan University, Wuhan, Hubei, 430072, P. R. China

Author co-citation analysis (ACA) is an important method for discovering the intellectual structure of a given scientific field. Since traditional ACA was confined to ISI Web of Knowledge (WoK), the co-citation counts of pairs of authors mainly depended on the data indexed in WoK. Fortunately, Google Scholar has integrated different academic databases from different publishers, providing an opportunity of conducting ACA in wider a range. In this paper, we conduct ACA of information science in China with the Chinese Google Scholar. Firstly, a brief introduction of Chinese Google Scholar is made, including retrieval principles and data formats. Secondly, the methods used in our paper are given. Thirdly, 31 most important authors of information science in China are selected as research objects. In the part of empirical study, factor analysis is used to find the main research directions of information science in China. Pajek, a powerful tool in social network analysis, is employed to visualize the author co-citation matrix as well. Finally, the resemblances and the differences between China and other countries in information science are pointed out.

Introduction

Author co-citation analysis (ACA) first introduced by WHITE & GRIFFITH [1981] has attracted much attention from different fields since then. Many authors have been devoting themselves to testing the practices of ACA [MCCAIN, 1990; AHLGREN & AL., 2003, 2004A, 2004B; BENSMAN, 2004; WHITE, 2003; WHITE & MCCAIN, 1998; LEYDESDORFF &VAUGHAN, 2006]. In addition, ACA has frequently been utilized to analyze the intellectual structure of a given scientific field [MCCAIN, 1990; QIU, 2007]. For example, WHITE & MCCAIN [1998] analyzed strictly on Library and Information Science (LIS)[1] and stated that LIS was mainly divided into two research directions: domain analysis and information retrieval.

[1] Please note that in China, there are comparatively distinct divisions between library science and information science, LIS mentioned here is similar to information science in China.

Received March 10, 2008; Published online March 18, 2009

Address for correspondence:
RUIMIN MA
E-mail: ruimin.ma@yahoo.com.cn

0138–9130/US $ 20.00

In the past few years, ACA were mostly conducted on the basis of primary databases such as SCI, SSCI, the number of indexed journals in which affected the co-citation frequencies between pairs of authors directly. Though quite difficult for various reasons, the goal of integrating resources from different databases has been achieved by Google Scholar which can help gain academic resources from different databases on one retrieval platform and interface. It identifies the same document embodied in different databases into one for free, which is very beneficial to ACA researches. LEYDESDORFF & VAUGHAN [2006] started an exploratory research on ACA by selecting 24 authors of information science under web environment with Google Scholar. However, the raw co-citation data were gained directly from the advanced search window of Google Scholar without any manual control.

On Nov 1st, 2006, Google Company announced that Google Scholar would extend service to Chinese academia (the English version was issued on Nov, 2004). Chinese Google Scholar has integrated three databases including VIP Information, Wanfang Data and Ilib. VIP Information is the largest and earliest academic information provider in China with 12000 journals. Wanfang Data is a company that mainly provides information on technology and engineering, whose database on internet is Ilib that provides a wider range of information than Wanfang Data by storing the documents of social science. In this paper, we manage to conduct the ACA of information science in China on the basis of the Chinese Google Scholar.

Methods

Author selection

Many experts have already analyzed the most important authors in information science in China, whose efforts offer us evidences to select research objects. Two kinds of authors are selected: One is reported to be the highly cited authors in information science by SU [2006]; the other is reported to be the core authors of information science in China by MA [2006]. Considering the actual research condition, 31 most important authors were identified for this research finally. Table 1 lists the basic information of these authors.

Data collection

It is necessary to know about the retrieval principle and data format of Chinese Google Scholar before collecting data. Figure 1 illustrates the retrieval interface of the advanced scholar search in which we can type pairs of authors in turn into the first data cell and define time span. Although the option “where the two authors occur” is provided, it is notable that its function is limited. Only the retrieval in title or full-text is available here.

Table 1. Concise introduction to 31 important authors

Name	Institution	Research fields
C.H. Bao	Commission of Science Technology and Industry for National Defense	Competitive intelligence
G.Z. Chen	Wuhan University	Information retrieval; Electronic publication
S.N. Chen	East China University of Science and Technology	Information retrieval
J.L. Chu	National Science Library	Application and theory of library science
X.Y. Dong	Peking University	Information organization; Knowledge management; Competitive intelligence system
S.L. Fan	Shanghai University	Competitive intelligence
C.P. Hu	Wuhan University	Theory of information science; Information service and support
X.B. Huang	Zhongshan University	Knowledge management; Web resource management; Competitive intelligence
G.Q. Huo	National Science Library	Information resources management; Knowledge management
M.S. Lai	Peking University	Theory of information science; Information retrieval
F.H. Leng	National Science Library	Knowledge management; Competitive intelligence; high-tech information analysis
J. Liu	Peking University	Information organization
T.H. Lu	Zhongshan University	Methods of information science
F.C. Ma	Wuhan University	Information resource management; Basic theory of information science; Information economics
H.Q. Ma	Heilongjiang University	Knowledge management; Intellectual property
G.J. Meng	National Science Library	Information resource management
B. Ni	Nanjing University	Information resource management; Competitive intelligence
J.P. Qiu	Wuhan University	Bibliometrics–Science evaluation
B. Wang	National Science Library	Basic theory of information science; scientometrics; Digital library
C.D. Wang	Tianjin Normal University	Bibliometrics; scientometrics
Z.J. Wang	Nankai university	Basic theory of information science; knowledge organization–competitive intelligence
R.S. Wen	National Science Library	Document index
J.P. Wu	State information Center	Information economics
Y.C. Xu	National Science Library	Information resource management
Y.M. Yan	Wuhan University	Basic theory of information science
J.B Yue	Peking University	Basic theory of information science
M.Z. Zeng	Commission of Science Technology and Industry for National Defense	Methods of information science; Information automation
Q.Y. Zhang	Nanjing Institute of Politics	Document index and retrieval
X.L. Zhang	National Science Library	Digital library
J.H. Zhao	Zhejiang University	Digital library; Information service
Z.R.Zou	Nanjing University	Theory and Method of information science; Bibliometrics

Since an ACA research is feasible on the precondition that two authors appear in the references of a paper simultaneously, fuzzy data in a retrieval result is unavoidable under three conditions: Firstly, A is the author of a paper, B appears in the references of the paper and A is not self-cited; secondly, A and B are both authors of a paper, and

neither of them appear in the references; thirdly, A and B appear in the text of an article rather than references. The co-citation frequencies of the two authors may be magnified under any of these conditions. Consequently more reliable co-citation data has to be screened from inaccurate results (see Figure 2). Fortunately, Google Scholar, which allows us to crawl through web pages with routine, is able to filter such a huge amount of data, which is impossible by hand. However, the three integrated databases by Chinese Google Scholar still organize data in different formats until today, as are shown in Figure 3, 4, and 5. It means that different labels have to be defined to distinguish the exact positions of references, in other words, to find out where references are located in a page. Accordingly, we write a routine with Java programming language to collect author co-citation data automatically and time span is defined between 2004 and 2006.

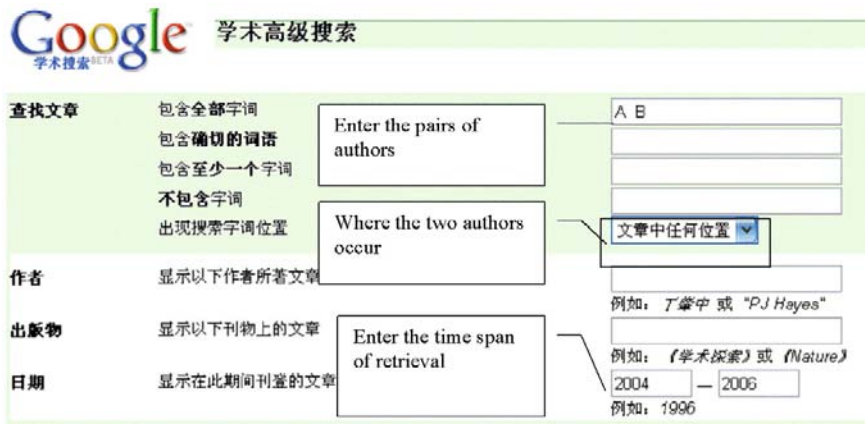

Figure 1. The interface of Google Scholar advanced search

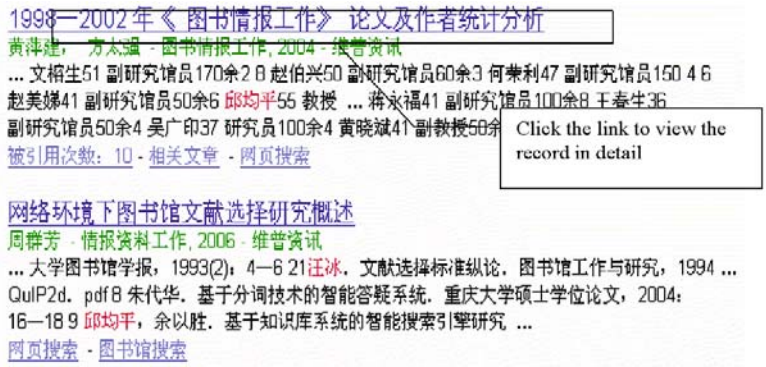

Figure 2. Retrieval result of Google Scholar

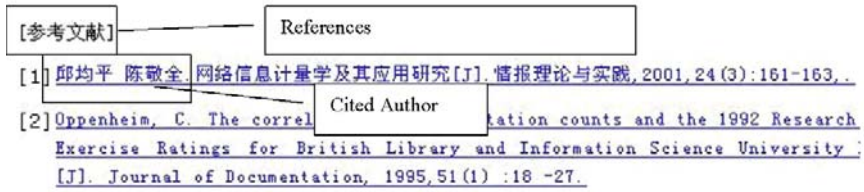

Figure 3. The record format of references of a paper in VIP Information

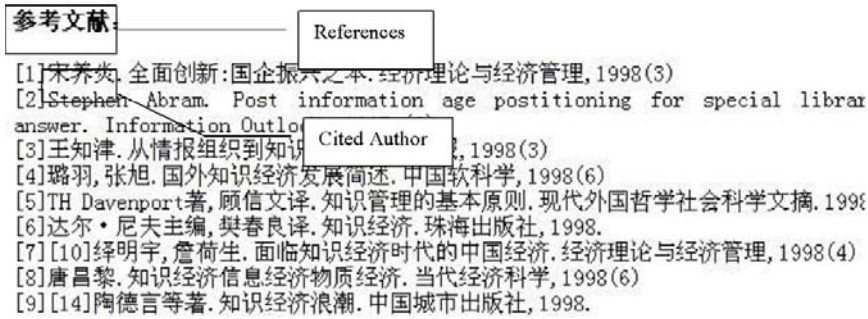

Figure 4. The record format of references of a paper in Wanfang Data

Figure 5. The record format of references of a paper in Ilib

Data processing

The routine gathers a set of raw co-citation data without adding the values of diagonal to it. In fact, the diagonal values make great impact on the correlation coefficient calculation such as Pearson's *r*, however, for which explanations varies and some are even subjective. For example, dividing the sum of any author's top three co-citation counts by two [WHITE & GRIFFITH, 1981]; missing value [MCCAIN, 1990;

WHITE & MCCAIN, 1998]; the maximum, namely any author's highest co-citation count with someone else [WHITE, 2003]. In general, we agree with White's proposal but make an adjustment. In this article, we define maximum+1 as diagonal value in order to emphasize the intimacy degree of an author with himself. [QIU & AL., 2008]. The raw co-citation matrix of 31 authors is sent by the authors at request.

In 1990, McCain published a technical overview of ACA, which has been adopted as a worldwide standard. In the overview, she summarized and listed three multivariate analysis methods suitable for ACA, namely factor analysis, cluster analysis and multidimensional scaling. In view of data characteristics, factor analysis is employed to discover the main research fields in information science in China in this article. The raw co-citation matrix is converted to Pearson's correlation coefficient one by factor routine in SPSS, and factors are extracted by principal components analysis with varimax rotation [HUANG, 2004].

In addition, we use the software Pajek,[1] a powerful tool adopted widely in social network analysis, to visualize the data, Pearson correlation coefficient as similarity measurement and the spring-embedded algorithm of Kamada-Kawai for the representation[2] [NOOY & AL., 2005].

Results

Statistics of the 31 important authors

Figure 6 shows the distribution of institutions for which these 31 important authors work. The authors distribute in 14 institutions: 11 universities with 20 authors, 2 government departments with 3 and 1 public library with 8. Universities play an important role in the research activities of information science in China. In terms of the number of authors, National Science Library, Wuhan University and Peking University are ranked top 3, all of which have the qualification of awarding doctor's degree. A majority of the most important authors work for National Science Library, which is a department of Chinese Academy of Sciences. Chinese Academy of Sciences is the largest academy of science research in China. Wuhan University is the biggest institution of higher education of information science and the first one that is qualified to award doctor's degree of information science in China. Peking University is known as one of the most outstanding universities in China.

[1] Pajek is freely available for academic use at http://vlado.fmf.uni-lj.si/pub/networks/pajek/

[2] Normalized author co-citation matrix is a similarity matrix. Therefore, before emerging the network, one should select the option *Similarities* as the values of lines. This option tells the energy procedures that line values indicate similarity: the higher a line value, the closer two vertices should be drawn [NOOY & AL., 2005, P. 90].

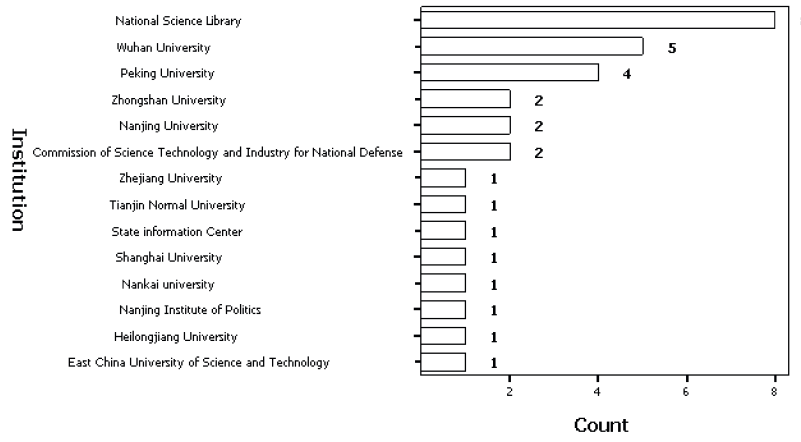

Figure 6. Institution distribution of 31 important authors

Factor analysis of 31 important authors

The scree plot suggests the extraction of six factors explaining 73.76% of the variance. Table 2 provides the six-factor solution using Varimax rotation and Kaiser Normalization. Only loadings greater than or equal to 0.5 will be displayed in each factor.

According to the research directions of these 31 authors, the six factors are explained as follows:[3]

- Factor 1, basic theories and methods of information science, including information economics. Lai, Yue, Ma, Yan, Lu, Wang and Zeng all devote themselves to the research of basic theories and have published related books such as *An Introduction to Information Management* by Ma, *Conspectus of Information Science* by Yan. Lai's interests are relatively broad, including information management, document retrieval, information economics, etc. Ma and Wu have ever written a book named *Information Economics* respectively. Lu is interested in the methods of information science and has published a book named *Information Analysis*.
- Factor 2, Bibliometrics. Zou is one of the founders of Chinese Social Science Citation Index (CSSCI). Wang is known as one of the most excellent bibliometricians and his works receive the largest number of citations in the

[3] MCCAIN [1990, p.440] suggested that only authors with loadings greater than ±0.7 were likely to be used in interpreting the factor. We follow this principle here.

very field in China. Qiu published the first book on bibliometrics in China in 1988, which has received many citations and high appraisal. Leng, a young scientist in Chinese Science Library, mainly devoted himself to information analysis with bibliometric methods.

- Factor 3, digital library and information service. This is a "hot" topic in information science in China and has even gained much attention from other fields. What's more, some scholars of information science behave excellently in this direction. Most researches focus on the theories of digital library and information service, while few on the applications and technologies of that. Zhang and Chu are colleagues in the same institution and are both good at the theory of digital library. Zhao has strong interest in reference service of digital library and contributes enormously to the development of customer management of library in China.
- Factor 4, information resource management. Huo and Meng have coauthored a book *An Introduction to Information Resource Management.* In addition, Huo and Xu also coauthored a book *Modern Library Theory*. Huo, Meng and Xu are colleagues in Chinese Science Library and Huo has ever been a doctor student of Meng. Information resource management is a large research field with many branches. Authors mainly loading on this factor incline towards the basic theories of information resource management, including its framework, content and applications.
- Factor 5, Document index and theory of information retrieval. The three authors loading on this factor positively were engaged in document index and traditional document retrieval before 1990s, including document cataloging, subject index, etc. Since 1990s, they have embarked on the research of information retrieval on the basis of databases and web resources.
- Factor 6, Competitive intelligence. This direction is of much attraction to many researchers in China. The two authors whose loadings are greater than 0.7 in factor 6 are considered as two of the most influential authors and founders of this direction. Bao has published a highly cited book *Competitive Intelligence of Enterprise*. Fan also published a high-quality book *The Applications of Competitive Intelligence*.

In addition, as Table 2 shows, some authors load on two factors simultaneously, which indicates the diversities of their research directions. For example, Z.J. Wang who loads on both factors 1 and 2 specializes in basic methods of information science and bibliometrics. More importantly, these authors may be the bridge between different research directions. For example, Wang may connect factor 1 and factor 2. This is just a conjecture to be further tested by the visualization with Pajek.

Table 2. Factor solution of the 31important authors
Rotated Component Matrix[a]

	Component					
	1	2	3	4	5	6
M.S. Lai	0.824					
J.B Yue	0.815					
F.C. Ma	0.806					
Y.M. Yan	0.742					
J.P. Wu	0.728					
T.H. Lu	0.709					
Z.J. Wang	0.596	0.526				
M.Z. Zeng	0.558					
Z.R.Zou		0.848				
C.D. Wang		0.843				
J.P. Qiu		0.803				
F.H. Leng		0.717				
H.Q. Ma		0.606				
B. Wang		0.605				
X.L. Zhang			0.879			
J.L. Chu			0.752			
J.H. Zhao			0.736			
J. Liu			0.695			
X.Y. Dong			0.674			
G.Z. Chen			0.617			
C.P. Hu	0.593		0.608			
X.B. Huang			0.594			
G.Q. Huo				0.962		
Y.C. Xu				0.957		
G.J. Meng				0.793		
S.N. Chen					0.946	
Q.Y. Zhang					0.936	
R.S. Wen					0.708	
C.H. Bao						0.848
S.L. Fan						0.783
B. Ni				0.526		0.581

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
[a] Rotation converged in 6 iterations.

Visualization with Pajek

Figure 7 illustrates the co-citation network of the 31 authors. Only the lines whose values are greater than or equal to 0.5 can be displayed in this figure. The strength of relation between them can be observed in terms of distance or thickness of lines. For example, X.L. Zhang is quite close to J.L. Chu in Figure 7, correspondingly the line between them is thick, which indicates their high similarity (r=0.822).

The result of visualization is similar to that of factor analysis. The map can be divided into six parts according to the six factors given in Table 2. For example, Wen, Chen, and Zhang who are situated at the right-of-center of the map constitute a group corresponding to factor 5. However, it is noticeable that the boundary between group 1

(at the center) and group 2 (at the left-of-center) is not so clear. The distance between these two groups is very short and some author in a certain group has strong associations with the authors in the other group. For example, Yan in group 1 has strong associations with C.D. Wang, Qiu, and Zou in group 2, so do Z.J. Wang, Qiu, C.D. Wang, Zou, and F.C. Ma. In fact, Qiu, F.C. Ma, Yan, Z.J. Wang, C.D. Wang, and Zou are the grand old men in information science in China. In the earliest stage of their researches, all the attention was paid to the basic laws and specific methods of information science such as Lotka's law, Bradford's Law, Zipf's Law, Price's Law and citation analysis. The authors in these two groups take up about 50% of all the 31 important authors and are situated in the center of the figure.

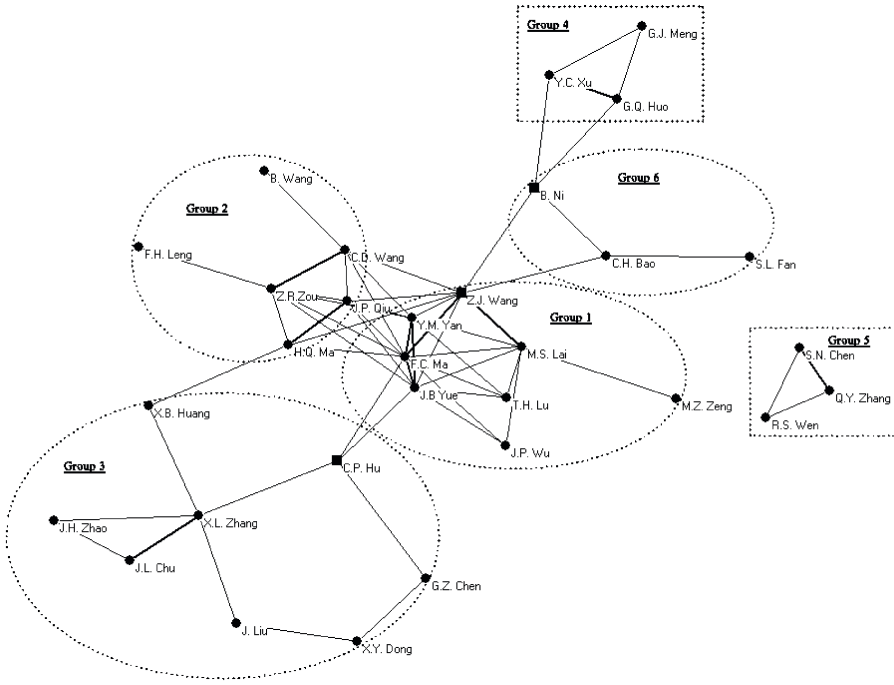

Figure 7. Visualization result of 31 authors with Pajek (Pearson's $r \geq 0.5$)

In addition, the vertices of three authors who load on two factors are drawn with boxes. Through Hu, who loads on both factor 1 and factor 3, group 1 and group 3 are connected. If this vertex was removed, group 1 and group 3 would obviously be disconnected. Ni connects group 4 with group 6. Ni, who is interested in competitive intelligence, basic methods of information science and information resource

management, has strong associations with authors in group 4, group 1 and group 6. Z.J. Wang, whose research interests concentrate on competitive intelligence and basic methods of information science, connects group 6 with group 1. These three authors are essential to the knowledge communication between different groups. In addition, it can be found that X.L. Zhang, the most highly cited author of information science in China according to the latest investigation reported by SU [2006], is a key author of group 3. If his vertex and the lines incident with it are removed, group 3 would break up into 3 new branches.

The visualization result makes a more intuitionistic impression on us and contains richer information. Note that its classification must be based on the result of factor analysis.

The comparison between China and foreign countries in the field of information science

In our opinions, it is necessary to discuss the resemblances and the differences of information science between China and foreign countries according to our understanding of this subject. WHITE & MCCAIN [1998] drew the conclusion that there were 12 specialties in information science by analyzing the co-citation data of 120 authors from 12 international journals. However, the data in their article were collected in 1996, whereas ours was in 2007. The intellectual structure of LIS in foreign must have changed during the past ten years. However, few studies have dealt with the period after 1995. ÅSTRÖM [2007] utilized paper co-citation analysis to discover the main topics during the period between 1990 and 2004, and found that the main intellectual structure of LIS had even not changed and still contained two directions: information seeking and retrieval and informetrics, however webometrics has become a dominant area. Consequently, we regard webometrics as one of the research directions of LIS in order to exhibit the structure of LIS abroad as comprehensively as possible on the basis of the intellectual structure proposed by White and McCain. In our opinions, despite continuous changes, the main structural of LIS is stable. Figure 8 provides a comparison of the research directions between China and foreign countries.

As Figure 8 shows, the differences of research directions of information science between China and foreign countries are not obvious (the directions marked gray are independent). In fact, “imported ideas” also exists in information science in China, however, the authors of this direction are few and the related researches are mostly introductory. Information resource management is a new field that aims at promoting the management standard of libraries and companies more scientifically and more efficiently. Competitive intelligence is another field whose target is to discover the underlying information of the competitors, which is regard as one of the most important directions of information science in China by many authors. With the rapid

development of Chinese economy, the cooperation between scholars in academic world and enterprisers is wider and more frequent, which will greatly enhance scholars' abilities of discovering and analyzing knowledge gradually.

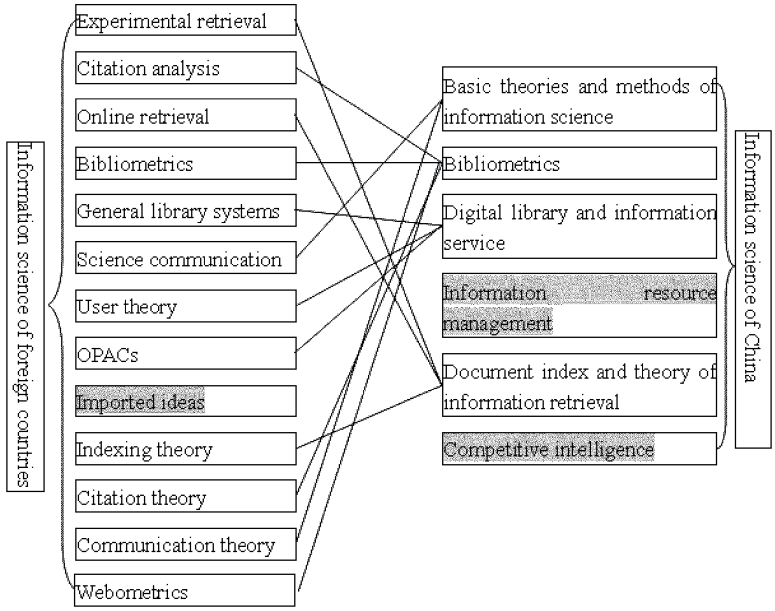

Figure 8. Comparison of information science between China and foreign countries

However, the researches on information science in China are comprehensive but not intensive in contrast to that of foreign countries. For example, bibliometrics in China contains the directions of citation analysis, citation theory, bibliometrics and webometrics, all of which are independent areas in foreign countries. In fact, the researches on bibilometrics in China have lagged behind that of foreign countries for several decades and are scattered in several subjects besides information science, such as philosophy of science and technology, and science of science. The authors from different subjects rarely communicate with each other, which hinders the development of bibliometrics.

In addition, we find that the authors in information science in China prefer theoretical studies to empirical studies. There have been a number of papers introducing concepts or developments of related researches in foreign countries, whereas further quantitative studies are scarce. As a result, information science is regarded to be minor and unfamiliar to many scholars and students in other subjects.

Conclusions

This paper takes advantage of Chinese Google Scholar to conduct an ACA on the subject of information science in China. Chinese Google scholar is a powerful tool to carry out ACA for its scientific organization of data and free access. However, the reported results of the co-citation counts of pairs of authors directly from the search engine are mixed with much fuzzy data unavoidably. It is necessary to write a routine to realize data filtering and collecting from such a huge scale of data. In the empirical part of this paper, the method of factor analysis is used to find out the six research directions of information science in China, the process of which has been explained in detail above. In order to observe the relationship between pairs of authors, Pajek, widely adopted in social network analysis, is employed to visualize the normalized data. We find that there are many similarities between the first group of the authors and the second group and some special authors functioning as bridges between different groups exist. Combining factor analysis with visualization by Pajek, the exploration of the intellectual structure of information science can be better conducted.

Through the comparisons between China and foreign countries in information science, the resemblances and the differences are pointed out. However, we find ourselves in a serious situation that studies in China are comprehensive but not intensive because most scholars focus on theoretical studies rather than empirical studies.

In the end, we hope this paper will provide everyone with a rudimentary grasp of the development of information science in China and also expect more extensive and in-depth academic communications between China and foreign countries and regions.

*

The authors would like to acknowledge the support of National Natural Science Foundation of China (70673071/G0309) and enlightening comments from reviewers. We appreciate Doctor Ronald Rousseau for his sincere comments and selfless provision of materials.

References

AHLGREN, P., JARNEVING, B., ROUSSEAU, R. (2003), Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficient, *Journal of the American Society for Information Scienceand Technology*, 54 (6) : 550–560.

AHLGREN, P., JARNEVING, B., ROUSSEAU, R. (2004a), Author cocitation and Pearson's r, *Journal of the American Society for Information Science and Technology*, 55 (9) : 843.

AHLGREN, P., JARNEVING, B., ROUSSEAU, R. (2004b), Rejoinder : In defense of formal methods. *Journal of the American Society for Information Science and Technology*, 55 (10) : 936.

BENSMAN, S. J. (2004), Pearson's r and author cocitation analysis : A commentary on the controversy, *Journal of the American Society for Information Science and Technology*, 55 (10) : 935–936.

GOOGLE SCHOLAR (2007), Retrieved December 20, 2007, from http://scholar.google.com

HUANG, R. L., *The Advanced Tutorial for SPSS*. Beijing: Higher Education Press, 2004.
LEYDESDORFF, L, VAUGHAN L. (2006), Co-occurrence matrices and their applications in information science: extending ACA to the web environment, *Journal of the American Society for Information Science and Technology*, 57 (12) : 1616–1628.
MA, F. C., SONG, E. M. (2006), An author co-citation analysis of information science in China, *Journal of the China Society for Scientific and Technical Information*, 25 (3) : 259–268.
MCCAIN, K. W. (1990), Mapping authors in intellectual space: A technical overview, *Journal of the American Society for Information Science*, 41 (1) : 433–443.
DE NOOY, W., MRVAR, A, BATAGELJ, V., *Exploratory Social Network Analysis with Pajek*. London: Cambrige University Press, 2005.
QIU, J. P., *Informatrics*. Wuhan: Wuhan University Press, 2007.
QIU, J. P., MA, R. M., LI, Y. J. (2008). New thoughts about co-citation analysis, *Journal of the China Society for Scientific and Technical Information* (in Chinese), 27 (1) : 69–74.
SU, X. L. (2006), Report on academic influence in library, information and documentation science (2000–2004), *Journal of the China Society for Scientific and Technical Information*, 25 (2) : 131–153.
THELWALL, M., VAUGHAN, L. BJÖRNEBORN, L. (2005), Webometrics, *Annual Review of Information Science and Technology*, 39 : 81–135.
WHITE, H. D. (2003), Author cocitation analysis and Pearson's r, *Journal of the American Society for Information Science and Technology*, 54 (13) : 1250–1259.
WHITE, H. D., GRIFFITH, B. (1981). Author cocitation : A literature mesaure of intellectual structures, *Journal of the American Society for Information Science*, 32 (3) : 163–171.
WHITE, H. D., MCCAIN, K. W. (1998), Visualizing a discipline : An author cocitation analysis of information science, 1972–1995, *Journal of the American Society for Information Science*, 49 (4) : 327–355.