Bibliometrical Analysis on the Big Data Research in China

Yang Ruixian
Department of Information Management
Zhengzhou University
Zhengzhou 450001, China
yrx@zzu.edu.cn

JDIM
Journal of Digital Information Management

ABSTRACT: *The author investigated the big data papers included China Academic Journal Network Publishing Database of CNKI from 2002 to 2012, with the help of multivariate statistical analysis tools SPSS17.0 of cluster analysis, multidimensional scaling analysis and visualization tools Ucinet6 from posting growth law, distribution of authors, research institutes distribution, periodical distribution, and high-frequency words such as multiple perspectives bibliometrical analysis and visualization analysis, the data reveal a large field of research status, and summarizes research characteristics,in order to study big data in-depth development and the future development of big data provide reference information.*

Categories and Subject Descriptors:
H.2.8 [**Database Applications**]; **I.2.7** [**Natural Language Processing**]; **G.3 [Probability and Statistics]**: Multivariate Statistics

General Terms: Big Data Research, Text Analysis, Multivariate Analysis

Keywords: Big Data, Visualization, Co-word analysis, Cluster Analysis, Multidimensional Analysis

Received: 3 July 2013, Revised 14 August 2013, Accepted 19 August 2013

1. Introduction

Along with Facebook, QQ, representing the rise of social networks, Twitter, microblogging and other social media, the rapid rise of location-based services LBS represented a new way of information dissemination continue to emerge, as well as cloud computing, networking and other technology rise at an unprecedented speed in constant growth and accumulation of big data era has arrived. McKinsey & Company published "*Big Data: The Next Frontier for Innovation*" [1] in May 2011, and the report is the first use the concept of "*Big data*", followed by industry, technology and government attentions. "*Big Data*", a general software tool is difficult to capture, manage and analyze large amounts of data, generally "*terabyte*" as a unit. "*Big Data*" and "*big*", not only is the "*large capacity*", the greater significance lies in: through the massive data exchange, integration and analysis, the discovery of new knowledge, create new value, bringing the "*big knowledge*", "*big profits*" and "*big development*" [2].

2. Data sources and data processing methods

2.1 Data sources

This paper selected Chinese Text (CNKI) published in the Chinese academic journals Network main library as a searchable database, the "*subject* " as the search term, "*big data* " or "*Big data*" as a search term, select all data from 2002 to 2012 (all topics about "*big data*" articles before 2012, the data of 2013 is not complete, so it has not analyzed temporarily), select the range of journals "*All journals* " in order to improve the recall ratio; and set matching as the "*exact* " in order to improve the pertiency factor. Data acquisition time is 10 April 2013, and a total number of papers are 2898.

2.2 Data Processing Methods

Mainly through multivariate statistical analysis tools

Spss17.0, Excel as a data analysis and processing tools and Ucinet6 visualization tools, use the bibliometrics summary and visualization methods, qualitatively and quantitatively, statistical analysis analyses the law of development of China's large data papers in the past 11 years.

3. Article growth law analysis

The growth of scientific knowledge and its laws and the growth of the scientific literature and its laws are closely linked, and the number of scientific literature directly reflects changes in scientific knowledge, so the number of scientific literature is important yardstick to measure the amount of scientific knowledge [3]. Domestic big data study start late, it has been developed relatively slow in the early years, but it has been in a growth trend now. (See Table 1).

Age (years)	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
issued (articles)	102	113	130	225	194	221	216	248	275	334	840

Table 1. 2002-2012 the number of research papers and Statistics

In Table 1 we have the statistics from 2002 to 2012 in these 11 years of big data research papers shouts case, through the analysis, since 2002, China has been in large data growth (in 2005, despite a little decline, but still have 194 research outputs), especially growth in 2012 is more obvious, we can say that 2012 is a great year for the development of large data. At the same time, we also used multivariate analysis software SPSS17.0 depicts the exponential growth curve literature Price curve (see Figure 1).

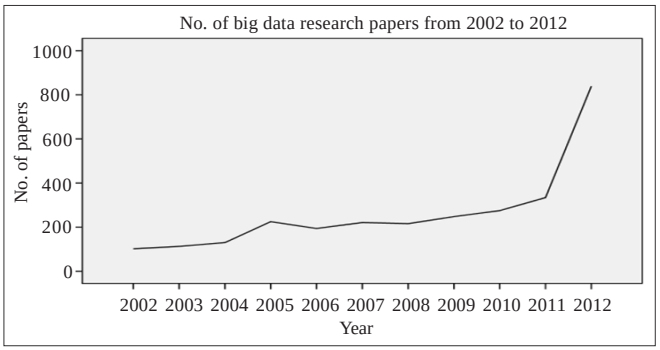

Figure 1. Price curve

One of the founders of bibliometrics Price, a variety of scientific indicators have been a lot of statistical analysis, proposed four-stage theory of growth in scientific literature [4]. As can be seen in Figure 1 through 2002-2012, large data research boom phase in gradually, research papers showed rapid growth, exponential growth. Therefore, in accordance with the Price of logistic curve growth theory predicts large domestic research in the field of data is in a great period of development.

4. Author distribution law

4.1 Information producers Lotka experience law

Lotka, American statistician scientist found there is a certain production capacity of the law, he proposed in 1926 firstly, reflecting the production capacity of Lotka's law [5]. Next we will be based on this principle to a research paper on large data analysis in order to verify the law of Lotka experience and further research topics fitting the Lotka formula.

Lotka's law usually expression, i.e.: $y(x) = C / xn\ C > 0, x = 1, 2, 3, xmax$ where $xmax$ represents the maximum capacity in a certain period of time of author [6]. These studies, a total of statistical correlation of 4766 authors were published in 2898 papers. We use the method of least squares to the calculated value of $n = \frac{N\sum XY - \sum X \sum Y}{N\sum X^2 - (\sum X^2)}$, $X = lgx$, $Y = lgy_x$, N is the total number of all authors. This should be 4766. The calculated $n = 2.290834339$, which Lotka result is basically the same, that is consistent with the inverse square law. And $C = \sum (1 / x^{2.290834339})$, calculates C is 1.400143544. Therefore, we analyzed large data field Lotka formula: $yx = 1.400143544 / x\ 2.290834339$. Relevant statistical data are shown in Table 2.

4.2 Prolific author analysis

The publication of scientific papers is an important indicator of the creative work of evaluation of scientific and

Papers	y_x	X	Y	XY	XX
1	3700	0	3.568201724	0	0
2	862	0.30103	2.935507266	0.88367574	0.090619058
3	136	0.47712125	2.133538908	1.017956761	0.227644692
4	41	0.60205999	1.612783857	0.970992635	0.362476233
5	11	0.69897	1.041392685	0.72790225	0.488559067
6	4	0.77815125	0.602059991	0.468493735	0.605519368
Σ	4754	2.8573325	11.89348443	4.06902112	1.774818419

Table 2. Posting less than six of distributed data

Name	Published articles	Authors	Cumulative number	Percentage
Xu Cuiping / Li Lu	11	2	2	0.076923077
Jiang Qi Ping / Zhang Peng	10	2	4	0.076923077
Li Xin	9	1	5	0.038461538
Li Yan	8	1	6	0.038461538
Liu Jie / Wang Lei / Zhang Jian / Yang Jie / Ren-yi / Nan Liu	7	6	12	0.230769231
Baiyun Chuan / Xudong Qi / Wang Wei / Xu Xiangyang	6	4	16	0.153846154
Chen Hua / Li Yunfei / Hu Feng / Wang Suihua/ Cheng Yimin Yang Chuanjian/Wang Hua / Sun Zhihui / Zhu Dehai /Li Jun /Zhou Shu	5	10	26	0.384615385

***Note:** The data collected in this table does not consider the case of the same name, which is that there is no same name in the large data field of study.

Table 3. From 2002 to 2012 of the distribution pattern data

technological personnel, scientific and technical personnel will be published scientific papers as their self-expression and a way to confirm their position in the scientific community [7]. The authors of the sample statistics are 4766. First of all, we have the case of the posting statistics (see Table 3).

As can be seen from Table 3, with a high of published articles Xu Cui-ping and Li Lu prominent in the field of large data, it is likely sometime to become the leader of the domestic large data in the future. According to Price's research in the field of literature distribution, we know that 75% of life scientists published a paper, based on

No.	Name	Published Articles
1	University of Defense Technology	55
2	Graduate School of the Chinese Academy of Sciences	36
3	Huazhong University of Science and Technology	36
4	University of Electronic Science and Technology	33
5	Zhejiang University	33
6	Tsinghua University	32
7	Northwestern Polytechnical University	31
8	Shanghai Jiao Tong University	29
9	Beijing University of Aeronautics and Astronautics	28
10	Wuhan University	28
11	Tongji University	27
12	Chinese University of Science and Technology	26
13	South China University	25
14	Southeast University	24
15	Tianjin University	24
16	China University of Geosciences	23
17	Xi'an University of Electronic Science and Technology	23
18	Xi'an Jiaotong University	22
19	PLA Information Engineering University	21
20	Central South University	20
21	Harbin Institute of Technology	20

Table 4. 2002-2012 Posting TOP21 list of research institutions

the statistical results of Table 3, we have come to the author of a paper published in the data field of study accounting for 77.63%, with Price very close to the statistical results.

5. Research Institutions

Table 4 Statistics 2002-2012 TOP21 most large data posting research institutions, with a strong academic and research capabilities in the data field of study, National University of Defense Technology (55, 9.23%) topped the list. Addition to the Chinese Academy of Sciences, the rest are institutions of higher education, colleges and universities in large data has a leading position in the Chinese Academy of Sciences as the highest level of research institutions in China, the research efforts should not be underestimated (see Table 4) .

6. Journal Analysis

Sample data of the respective journals statistics, have a total of 908 journals. The distribution of literature in journals was discrete state of Bradford, UK study found that there are certain rules, the distribution of literature in journals and Bradford's law[8]. Table 5 lists of published articles in 16 journals (or more) and the Amount of Papers.

mainly based on the amount of data algorithms and techniques inquiry. The field of Library and Information on large data is relatively scarce; a small amount of research has focused on the Library, competitive intelligence, and based on mass Citation data academic research has also taken off.

7. Keywords analysis

7.1 Word frequency statistics

Usually keywords can reflect the transdisciplinary themes and concerns, high-frequency words can well reflect the particular area of concern hot, and co-word characteristics between keywords better summarized in a subject professional focus. So we statistics the keywords and word frequency from 2002 to 2012 in the paper, and summarize the research focus using the multidimensional scaling analysis methods in modern statistical techniques. See Table 7, Figure 2.

Statistics based on Excel, all the research journal articles, a total of 9147 Keywords Figure 2 can be seen from Table 7, "*data mining*", "*database*", "*FPGA*", "*data*", " *clustering*" will be the focus of future research in the field of large data. It is consistent with the design of FPGA-based data acquisition system with a large number of journal articles

Name	Papers	Sum of Periodical: C	Logarithm of Sum of Periodical: lgC	Sum of Papers: R(n)
Computer Engineering and Applications	67	1	0	67
Computer Engineering	63	2	0.301029996	130
Communication World	53	3	0.477121255	183
Microcomputer	53	4	0.602059991	236
Computer Applications	49	5	0.698970004	285
Application Research of Computers	38	6	0.77815125	323
Computer and Network	37	7	0.84509804	360
Computer Engineering and Design	33	8	0.903089987	393
Computer Science	31	9	0.954242509	424
Silicon	31	10	1	455

Table 5. References and periodicals distribution data

As can be seen from the periodical distribution table, the core journals of the field of big data: "*Computer Engineering and Applications*", "*Computer Engineering*", "*Communication World* ", "*Microcomputer Information*", "*computer*", "*computer applied research*" and so on. 30 journals, only "*successful marketing*" does not belong to the IT journals, and others belong to the IT journals. It is the proportion of the total as high as 97.99% in the total journals, most of the domestic large data are concentrated in the IT sector, phenomenon. Keywords tables and data reveals a perspective view of "*data mining*", "*big data*" hot spots, but also does not reflect the relationship between the various keywords, so it needs to define keyword analysis, and to visualize the way of presentation. Next, I will construct a high-frequency words co-word analysis matrix, using a SPSS17.0 cluster analysis and multidimensional scaling analysis of the relationship between the high frequency keywords and related degrees.

No.	Keywords	Frequency	No.	Keywords	Frequency
1	Data Mining	110	11	Rough Set	26
2	Database	52	12	cloud computing	24
3	FPGA	51	13	Data Processing	24
4	Big Data	51	14	DSP	23
5	Clustering	45	15	association rules	20
6	Support Vector Machine	37	16	algorithm	20
7	Data Collection	35	17	huge amounts of data	20
8	Large Amount of Data	30	18	neural network	19
9	Data Warehouse	29	19	Attribute Reduction	19
10	Data Centers	28	20	multi-threaded	18

Table 7. High frequency Keywords (Top 20)

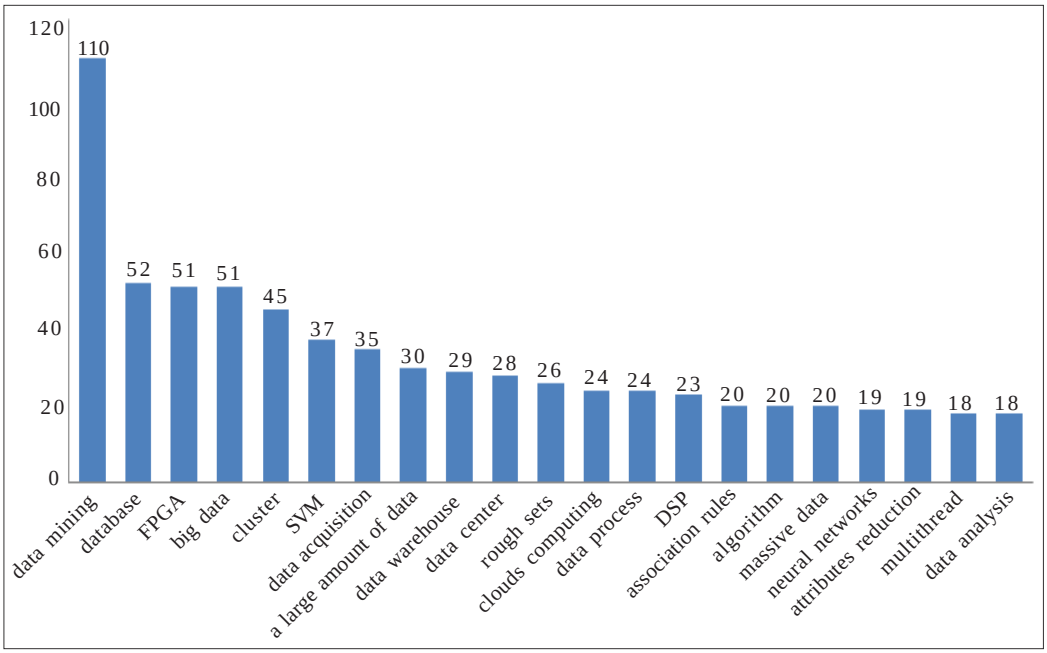

Figure 2. High-frequency keywords PivotChart

7.2 Construction co-word analysis matrix of high-frequency words

Co-word analysis of the principle of the method is mainly these words as a basis for them in the same literature a number of occurrences of a set of pairwise statistical cluster analysis, multidimensional scaling analysis, so as to reflect between words kinship, and then analyzes these words represent the disciplines and topics structural changes[9]. According to the principle of co-word analysis, for further processing of high-frequency keywords, pairwise statistics account the number of occurrences in the same paper. If the two key words, the higher the frequency is, the closer the relationship between them. Final matrix is shown in Table 8.

As can be seen from the co-word of matrix, the keywords appear direction, so that the matrix is a symmetric matrix, i.e. elements corresponding equal to the diagonal axis of symmetry. Keywords co-occurrence frequency is generally low, even a large part of zero indicates that the field of big data as an independent discipline its development is not mature enough, is not stable enough, the other with large data involves a multidisciplinary cross the actual situation is consistent, from different disciplines of the focus of the data is bound to be different, less contact. It is also because of their research broad distribution of large data research papers in journals scattered journals involving management, economics, computer science and other disciplines.

7.3 Cluster Analysis

Cluster analysis is the number of samples (or variables)

	Data Mining	Data base	FPGA	Big Data	clustering	support vector machine	Data Collecting	large amount of data	data warehouse	data centers
Data Mining	11	2	0	5	10	5	2	0	6	2
Database	2	4	0	1	0	0	0	1	2	3
FPGA	0	0	4	0	0	0	3	0	0	0
Big Data	5	1	0	6	0	0	0	0	0	1
clustering	10	0	0	0	11	0	0	1	0	0
support vector machine	5	0	0	0	0	6	0	0	0	0
DataCollecting	2	0	3	0	0	0	4	0	0	0
large amount of data	0	1	0	0	1	0	0	2	1	0
data warehouse	6	2	0	0	0	0	0	1	7	1
data centers	2	3	0	1	0	0	0	0	1	4

Table 8. 2002 to 2012 data research papers frequency words co-word matrix

data according to its many features, the degree of can closeness in nature in the case of no prior knowledge be automatically classified, resulting in more than one classification results [10]. Similar individuals are very similar to the individual differences between the different types. The system cluster, clustering method select the between-groups linkage, measurement method using Squared Euclidean distance, standardized method for Z-scores" to draw a dendrogram, as shown in Figure 3.

From the dendrogram can be seen that horizontal distance indicates differences in size, you can clearly see that the variable clustering process. At the same time, if the keyword is divided into two categories, then from the clustering process can be seen keywords can be divided into five categories: database, data centers, large amount of data, FPGA and data collection should be classified as a class; big data support vector machines for the second class; data warehouse, data mining and clustering as a class respectively.

7.4 Multidimensional Scaling Analysis

Multidimensional scaling analysis is used to study the degree of similarity (dissimilarity) between multiple things, through appropriate dimensionality reduction methods, indicates this similarity (dissimilarity) the degree of distance between points in the low-latitude space [11]. Results of the analysis shown in Figure 5, each investigator distribution reflects the relationship and strength among the investigators, and a high degree of similarity of the Investigator together, the formation of the academic community, indicates that the closer the distance between the research direction, to marginalized or no researchers classified to the research community

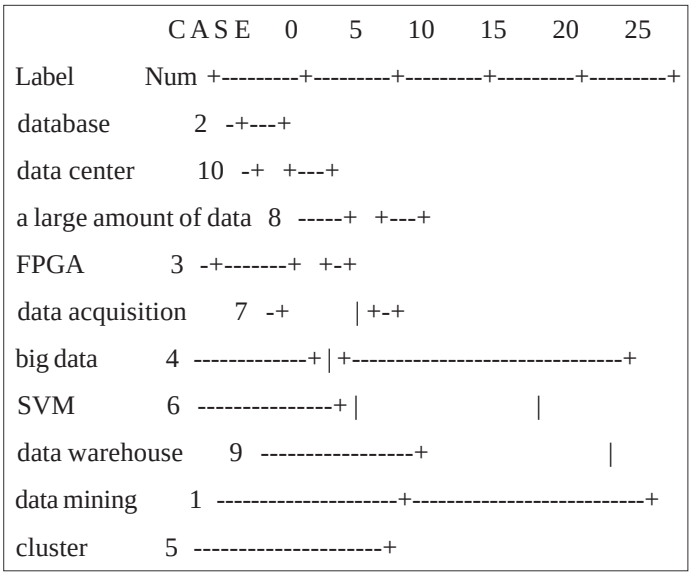

Figure 3. Data field of high-frequency words dendrogram

that the research direction is narrow, or the transition to the other direction [12].

Multidimensional the scale analysis Figure keywords can be shown that the distribution of the degree of association between them. High similar keywords together form a community, the closer the distance the greater of their associated. Such as data acquisition and data centers; marginalized or not classified within the group of keywords, its affiliates is very small, or are in transition to a certain direction, such as clustering, data warehousing. Figure 4 using multidimensional scaling to analyze the results from Figure 3 clustering analysis of the results are consistent.

Data usage is still shown in Table 9, no missing for all analysis. Stress and RSQ are two multidimensional scaling analysis of the reliability and validity of the

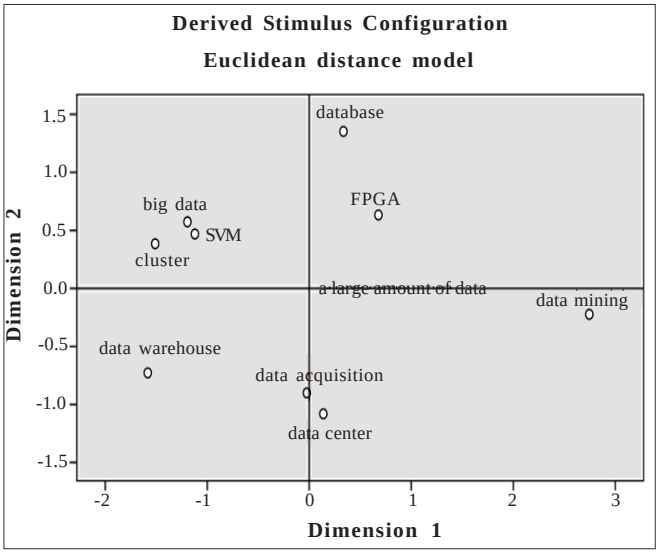

Figure 4. Data field of high-frequency words multidimensional scaling analysis

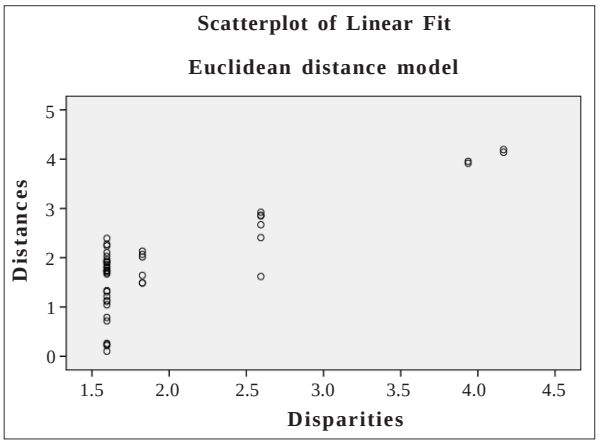

Figure 5. Linear fit scatter plots

estimated value, RSQ bigger and more desirable, usually in the 0.60 is acceptable. The multidimensional scaling analysis Stress = 0.26697 RSQ = 0.65665, description fits the data better. From Figure 6, points in the figure are substantially distributed around a straight line, showing a significant linear trend analysis also shows that the effect is better. In general, below multidimensional scaling diagram is our big data research and development focus.

8. Conclusions

In summary, we learned of large data in the 2002-2012 research has made some achievements, specific performance as follows:

• **Growth:** Research Papers posting large data volume increase year by year, and the statistics from Table 1, big data research papers has been in a growth trend since 2002, while in 2012 issued documents have exceeded 800, and these data indicate studies of large data exploration period and preliminary stages of development have passed and it is in the boom period now.

• **Regularity:** Domestic research papers posting large data volume followed Price growth law. The law of author's distribution accorded Lotka and periodical distribution showed Bradford law. Lotka distribution formula which is $yx = 1.400143544 / x2.290834339$.

• **Concentration:** Big data field's productive authors have Xu Cui-ping, Li Lu, Jiang Qi, Zhang Peng and Li Xin, etc. Most influential research institutions are University of Defense Technology, Chinese Academy of Sciences, Huazhong University of Electronic Science and Technology University and Zhejiang University; core journals "*Computer Engineering and Applications*", "*computer Engineering*", "*communication World*", "*Microcomputer information*" and "*computer Applications*" and so on.

This study showed that: domestic demand for big data research has focused on algorithms and techniques in the field of IT applications, and Library and Information field is still in its initial stage of exploration. Conforming to large data trends, big data in library and information field of research has attracted a number of research institutions and researcher's attention in recent years. It has been the National Social Science Foundation and the National Natural Science Foundation as a research topic; research based on big data has taken shape. But now the field of big data LIS related researches less, and most of these papers in the theoretical stage of exploration, lack adequate practical support. To this end, the field of Library and Information Studies for large data must intensify our efforts based on theory and practice, and continuously inject new research forces to form the core of the group is the leading force in the research team.

9. Acknowledgements

I am so grateful to Hu Yu, my student for computing data of this article, and also grateful to Qi Feiei for reviewing the English format of the manuscript.

References

[1] Big Data: The next frontier for innovation, competition, and productivity [EB / OL]. http://www.mckinsey.com/Insights/MGI/Research/Technologu_and_Innovation/Big_data_the_next_frontier for innovation. 2013-04-18.

[2] Zipei, Xu . (2012). Large data: data revolution is coming, and how to reform government, business and our lives. *Guilin: Guangxi Normal University Press*, 7, p. 57.

[3] Junping, Qiu., Jinyan, Su., Zunyan, Xiong (2008). Information Resource Management: a Comparative Study Based on Bibliometrics. *Journal of Library Science in China,* (5) 37-38.

[4] Jing'an, Pang. (1999). Scientometrical Research Methodology. *Beijing: Scientific and Technical Documentation Press,* p. 299-301.

[5] Feicheng, Ma . (2002). Information Management Science [M]. *Wuhan: Wuhan University Press,* 12, p. 80-85.

[6] Junping, Qiu . (2007). Informatics. *Wuhan: Wuhan University Press*, p. 45-55.

[7] Derek de Solla Price. (1963). Little Science, Big Science. *New York: Columbia Press*.

[8] Bradford, S. C. (1934). Sources of Information on Specific Subjects. *Engineering*, 26 (1) 85-86.

[9] Lu, Feng., Fuhai, Leng (2006). Development of Theoretical Studies of Co-word Analysis. *Journal of Library Science in China*, (2) 88-92.

[10] Wei, Xue . (2009). SPSS Statistical Analysis Method and Application. *Beijing: Publishing House of Electronics Industry*.

[11] Qiang, Du., Liyan, Jia . (2009). Breakthrough of Understanding SPSS statistical analysis from entry to master. *Beijing: People's Posts and Telecommunications Press*.

[12] Jiming, Hu. (2009). Author Co-Citation Analysis of Information Service Research in China. *Journal of Intelligence,* (10) 170-174.