A Study of Data Mining Productivity Review from SSCI Database Using Bibliometric Methodology

Hsu-Hao Tsai , Jiann-Ming Yang

Department of Management Information System, National Chengchi University
No.64, Sec. 2, Zhinan Rd., Wenshan District, Taipei City 11605, Taiwan
98356512@nccu.edu.tw
jmyang@mis.nccu.edu.tw

***Abstract*—The study is mainly concentrating on the data mining literatures productivity review which topic as "data mining" in SSCI database from 1989 to 2009. The result indicated that the literature productivity which topic as "data mining" is still increasing extremely and will demonstrate the trend by different categories including author, country/territory, institution name, document type, language, publication year, and subject area. The paper will perform by the following steps to verify the reliability of Lotka's Law: (1) collect data; (2) list author & literature distribution table; (3) calculate n value (slope); (4) calculate c value; (5) utilizing K-S (Kolmogorov-Smirnov, K-S) test to evaluate if matched Lotka's Law. After checked by K-S test, the distribution of frequency indexes of author productivity is suitable for Lotka's Law. The applications of data mining are primary following by research aspects which in term of information science & library science, computer science & information systems, operations research & management science, management, computer science & artificial intelligence, economics, computer science & interdisciplinary applications, public, environmental & occupational health and engineering, electrical & electronic.**

***Keywords*—Data Mining; Lotka's Law; Bibliometric Methodology**

摘要

本研究主要是探討從 1989 年到 2009 年在 SSCI 資料庫中以資料探勘為主題的文獻發表趨勢。研究結果顯示以資料探勘為主題的文獻仍然在成長當中，同時研究結果會將研究的趨勢以不同的文獻作者、發表的國家/地區、發表的研究機構名稱、發表的文獻類型、發表的語言、發表的發表年份與發表的期刊名稱來呈現。本文將依照以下步驟來驗證洛卡定律的可靠性: (1) 收集數據; (2) 依作者與文獻分布列表; (3) 計算 n 值 (斜率); (4) 計算常數 c 值; (5) 利用柯史(Kolmogorov-Smirnov, K-S)檢測其與洛卡定律是否相符。經過柯史檢測，作者文獻生產力的分布是符合洛卡定律。從研究結果可以看出資料探勘應用的相關研究主要是發表於資訊科學與圖書館學、電腦科學與資訊系統學、作業研究與管理科學、管理學、電腦科學與人工智慧、經濟學、電腦科學與各學科間的應用、公共環境與職業健康及工程,電機與電子等研究主題領域當中。

關鍵字：資料探勘**;** 洛卡定律**;** 文獻計量學

1. INTRODUCTION

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

This analysis uses a bibliometric methodology for forecasting data mining productivity. The result indicated that the literature productivity topic as "data mining" is still increasing extremely and will demonstrate the trend by different categories including author, country/territory, institution name, document type, language, publication year, and subject area. This implies that the phenomenon "success breeds success" is more common in higher quality publications.

For verifying the analysis result, the paper will proceed K-S (Kolmogorov-Smirnov, K-S) test to check Lotka's Law on literature record count versus accumulated authors between 1989 and 2009 to perform author productivity inspection for discovering historical review and collecting the results for research tendency forecast in the near future.

2. LITERATURE REVIEW

2.1. Data Mining

Data mining is the process of extracting patterns from data. As more data are gathered, with the amount of data doubling every three years [1], data mining is becoming an increasingly important tool to transform these data into information. It is commonly used in a wide range of profiling practices, such as marketing, surveillance, fraud detection and scientific discovery.

The term data mining has also been used in a related but negative sense, to mean the deliberate searching for apparent but not necessarily representative patterns in large numbers of data. To avoid confusion with the other sense, the terms data dredging and data snooping are often used. Note, however, that dredging and snooping can be used as exploratory tools when developing and clarifying hypotheses.

2.2. Lotka's Law

Lotka's law [2], named after Alfred J. Lotka, is one of a variety of special applications of Zipf's law [3]. It describes the frequency of publication by authors in any given field. It states that the number of authors making n contributions is about $1 / n^a$ of those making one contribution, where *a* nearly always equals two. More plainly, the number of authors publishing a certain number of articles is a fixed ratio to the number of authors publishing a single article. As the number of published articles increases, authors producing that many publications become less frequent. There are 1/4 as many authors publishing two articles within a specified time period as there are single-publication authors, 1/9 as many publishing three articles, 1/16 as many publishing four articles, etc. Though the law itself covers many disciplines, the actual ratios involved (as a function of 'a') are very discipline-specific. The general formula says:

$$X^n Y = C \quad (1)$$

or

$$Y = \frac{C}{X^n} \quad (2)$$

Where X is the number of publications, Y the relative frequency of authors with X publications, and n and C are constants depending on the specific field ($n \approx 2$). This law is believed to have applications in other fields for example in the military for fighter pilot kills.

For example, there are 100 authors writing one article each over a specific period, we assume that C=1 and n=2, shown as Table 1.

TABLE 1

THE DISTRIBUTION OF 100 AUTHORS WRITING ARTICLES

Number of articles written	Number of authors writing that number of articles	Number of Article
10	100/102 = 1	10
9	100/92 ≈ 1 (1.23)	9
8	100/82 ≈ 2 (1.56)	16
7	100/72 ≈ 2 (2.04)	14
6	100/62 ≈ 3 (2.77)	18
5	100/52 = 4	20
4	100/42 ≈ 6 (6.25)	24
3	100/32 ≈ 11 (11.111...)	33
2	100/22 = 25	50
1	100	100
Total	155	294

That would be a total of 294 articles with 155 writers with an average of 1.9 articles for each author. This is an empirical observation rather than a necessary result. This form of the law is as originally published and is sometimes referred to as the "discrete Lotka power function" [4].

3. RESEARCH FINDING AND DISCUSSION

This research is accessing the Social Science Citation Index (SSCI) on Web of Science created by ISL. The result is summarizing those 1165 paper indexes which topic are "data mining" from 1989 to 2009, shown as Fig. 1. Obviously,

the literature production of data mining is rising since 1997 and citation is also increasing steadily and gradually by every year. It shows the research of data mining is very popular in the highly exploration period, referred to Fig. 2. The research of data mining reached the highest record in 2009.

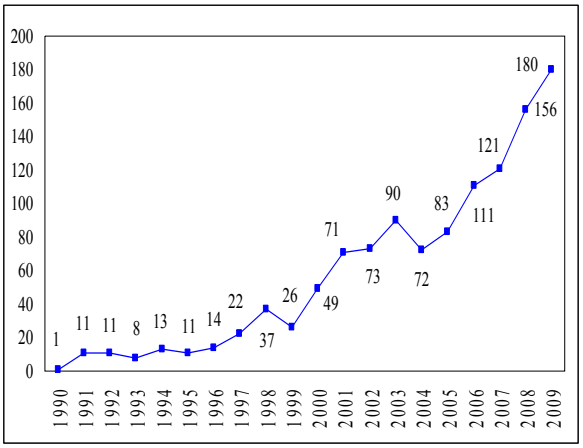

Fig. 1 The tendency chart of literature growth of data mining

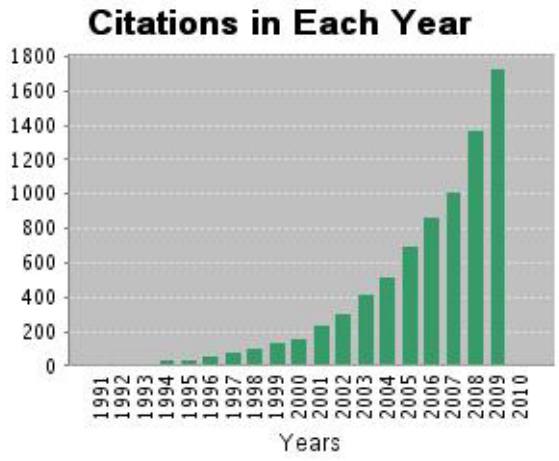

Fig. 2 Citation of data mining in each tear (Source: SSCI database)

By viewing on Table 2 and Fig. 3, they displayed that the distribution of country/territory from 1989 to 2009, USA is a champion with 545 record counts (46.78%), following by England**,** Taiwan**,** Canada and People R. China which achieved the record counts as 106(9.10%), 101(8.67%)**,** 67(5.75%) and 53(4.55%) oppositely. Taiwan is ranking as No.3 (101 paper counts, 8.67%) in this research domain. For the distribution of institution name as shown in Table 3 and Fig. 4, the result indicates that USA is still the most productive country within the research aspect of data mining in the world.

TABLE 2

DISTRIBUTION OF TOP 25 COUNTRY/TERRITORY FROM 1989 TO 2009

Ranking	Country/Territory	Record Count	% of 1165
1	USA	545	46.78%
2	ENGLAND	106	9.10%
3	TAIWAN	101	8.67%
4	CANADA	67	5.75%
5	PEOPLES R CHINA	53	4.55%
6	AUSTRALIA	47	4.03%
7	GERMANY	31	2.66%
8	SOUTH KOREA	31	2.66%
9	SPAIN	26	2.23%
10	NETHERLANDS	21	1.80%
11	BELGIUM	20	1.72%
12	FRANCE	19	1.63%
13	ITALY	17	1.46%
14	JAPAN	17	1.46%
15	BRAZIL	13	1.12%
16	SCOTLAND	13	1.12%
17	SOUTH AFRICA	13	1.12%
18	SWEDEN	12	1.03%
19	TURKEY	12	1.03%
20	INDIA	11	0.94%
21	SLOVENIA	11	0.94%
22	AUSTRIA	10	0.86%
23	FINLAND	10	0.86%
24	SINGAPORE	10	0.86%
25	WALES	10	0.86%
	Others	31	2.66%

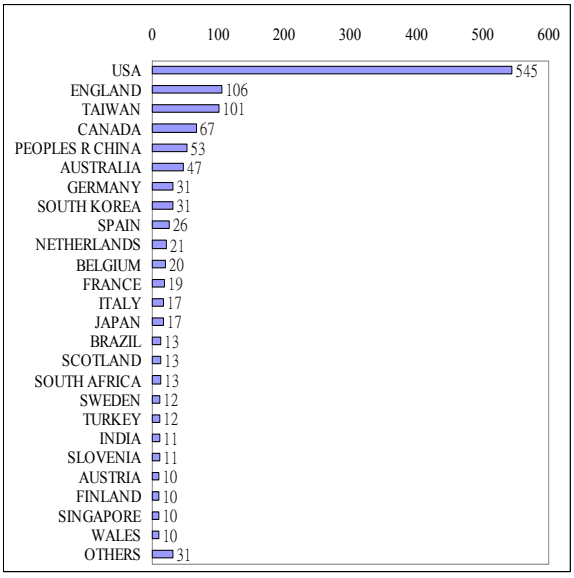

Fig. 3 Distribution of top 25 country/territory from 1989 to 2009

TABLE 3

DISTRIBUTION OF TOP 25 INSTITUTION NAME FROM 1989 TO 2009

Ranking	Institution Name	Record Count	% of 1165
1	NIOSH	17	1.46%
2	PENN STATE UNIV	17	1.46%
3	UNIV WISCONSIN	17	1.46%
4	UNIV ILLINOIS	13	1.12%
5	COLUMBIA UNIV	12	1.03%
6	NATL CENT UNIV	12	1.03%
7	UNIV PENN	12	1.03%
8	PURDUE UNIV	11	0.94%
9	MONASH UNIV	10	0.86%
10	NATL CHIAO TUNG UNIV	10	0.86%
11	UNIV TEXAS	10	0.86%
12	DUKE UNIV	9	0.77%
13	TAMKANG UNIV	9	0.77%
14	UNIV N CAROLINA	9	0.77%
15	UNIV WESTERN ONTARIO	9	0.77%
16	YALE UNIV	9	0.77%
17	CITY UNIV HONG KONG	8	0.69%
18	HARVARD UNIV	8	0.69%
19	NANYANG TECHNOL UNIV	8	0.69%
20	NATL SUN YAT SEN UNIV	8	0.69%
21	OFF NAVAL RES	8	0.69%
22	SYRACUSE UNIV	8	0.69%
23	UNIV ARIZONA	8	0.69%
24	UNIV HONG KONG	8	0.69%
25	UNIV KENTUCKY	8	0.69%
	Others	34	2.92%
	Total	1165	100%

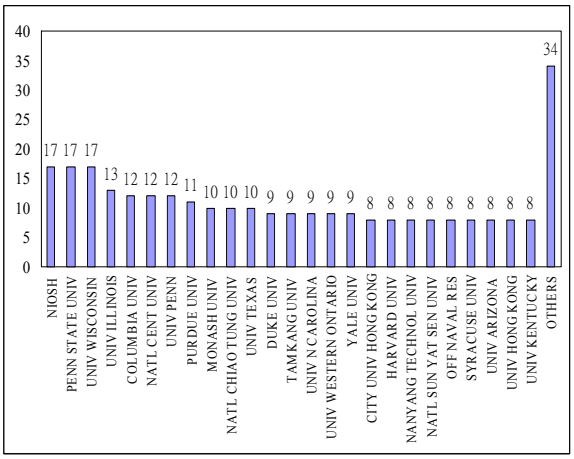

Fig. 4 Distribution of top 25 institution name from 1989 to 2009

From the Table 4, it indicated that the most popular publication document type is "Article" (920 record counts, 78.97%), and the most popular language for literature is using "English" (1133 record counts, 97.25%) in the research domain of data mining. See the following Table 5.

TABLE 4

DISTRIBUTION OF DOCUMENT TYPE FROM 1989 TO 2009

Document Type	Record Count	% of 1165
ARTICLE	920	78.97%
PROCEEDINGS PAPER	106	9.10%
BOOK REVIEW	50	4.29%
REVIEW	41	3.52%
MEETING ABSTRACT	23	1.97%
EDITORIAL MATERIAL	19	1.63%
NEWS ITEM	2	0.17%
CORRECTION	1	0.09%
NOTE	1	0.09%
REPRINT	1	0.09%
SOFTWARE REVIEW	1	0.09%

TABLE 5

DISTRIBUTION OF LANGUAGE FROM 1989 TO 2009

Language	Record Count	% of 1165
ENGLISH	1133	97.25%
SPANISH	12	1.03%
GERMAN	5	0.43%
SLOVAK	4	0.34%
JAPANESE	3	0.26%
CZECH	2	0.17%
FRENCH	2	0.17%
PORTUGUESE	2	0.17%
RUSSIAN	1	0.09%
SLOVENE	1	0.09%
Total	1165	100%

In the Table 6 and Fig. 7, it is important to summarize the trend information for data mining researchers to get better understanding about the distribution of top 25 subject areas in future research trend and forecast. The top three ranking of research domains are information science & library science (259 record counts, 22.23%), following by computer science & information system (249 record counts, 21.37%) and operations research & management science (162 record counts, 13.91%). Moreover, it also discovered that there are a lot of research domains for data mining literature production such as management、computer science & artificial intelligence、economics、computer science & interdisciplinary applications、public, environmental & occupational health and engineering, electrical & electronic and so on.

TABLE 6

DISTRIBUTION OF TOP 25 SUBJECT AREAS FROM 1989 TO 2009

Ranking	Subject Area	Record Count	% of 1165
1	INFORMATION SCIENCE & LIBRARY SCIENCE	259	22.23%
2	COMPUTER SCIENCE & INFORMATION SYSTEMS	249	21.37%
3	OPERATIONS RESEARCH & MANAGEMENT SCIENCE	162	13.91%
4	MANAGEMENT	147	12.62%
5	COMPUTER SCIENCE & ARTIFICIAL INTELLIGENCE	126	10.82%
6	ECONOMICS	111	9.53%
7	COMPUTER SCIENCE & INTERDISCIPLINARY APPLICATIONS	103	8.84%
8	PUBLIC, ENVIRONMENTAL & OCCUPATIONAL HEALTH	85	7.30%
9	ENGINEERING, ELECTRICAL & ELECTRONIC	78	6.70%
10	ENVIRONMENTAL STUDIES	67	5.75%
11	BUSINESS	56	4.81%
12	GEOGRAPHY	51	4.38%
13	MEDICAL INFORMATICS	49	4.21%
14	ENVIRONMENTAL SCIENCES	37	3.18%
15	SOCIAL SCIENCES, MATHEMATICAL METHODS	35	3.00%
16	ENGINEERING, INDUSTRIAL	33	2.83%
17	ERGONOMICS	33	2.83%
18	PLANNING & DEVELOPMENT	31	2.66%
19	EDUCATION & EDUCATIONAL RESEARCH	30	2.58%
20	SOCIAL SCIENCES, INTERDISCIPLINARY	30	2.58%
21	SOCIOLOGY	30	2.58%
22	MATHEMATICS, INTERDISCIPLINARY APPLICATIONS	26	2.23%
23	GEOGRAPHY, PHYSICAL	24	2.06%
24	COMPUTER SCIENCE, CYBERNETICS	22	1.89%
25	STATISTICS & PROBABILITY	21	1.80%

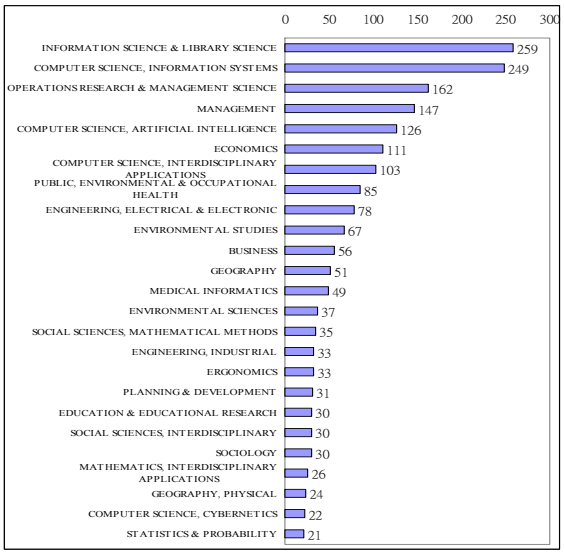

Fig. 7 Distribution of top 25 subject areas from 1989 to 2009

4. THE LITERATURES PRODUCTIVITY ANALYSIS OF DATA MINING BY LOTKA'S LAW

The section will perform by the following steps to verify the reliability of Lotka's Law: (1) collect data; (2) list author & literature distribution table; (3) calculate n value (slope); (4) calculate c value; (5) utilizing K-S (Kolmogorov-Smirnov, K-S) test to evaluate if matched Lotka's Law [5].

TABLE 7

DISTRIBUTION OF AUTHOR PRODUCTIVITY OF DATA MINING FROM 1989 TO 2009

Record Count	Author (s)	Record Count%	Subtotal Record	Subtotal Record %	Subtotal Author (s)	Subtotal Author (s)%
9	1	9	9	0.31%	1	0.04%
8	0	0	9	0.31%	1	0.04%
7	2	14	23	0.79%	3	0.12%
6	3	18	41	1.42%	6	0.24%
5	6	30	71	2.45%	12	0.48%
4	12	48	119	4.11%	24	0.95%
3	37	111	230	7.95%	61	2.42%
2	206	412	642	22.18%	267	10.60%
1	2252	2252	2894	100.00%	2519	100.00%

(1) Collect data & (2) List author & literature distribution table:

It calculated the author quantity by the equality method from 1165 literatures which retrieved by index on SSCI. Thus, it is obtained altogether 2519 of authors on research aspect of data mining. See the table 7 for reference.

From Table 7, that would be a total of 2894 articles with 2519 writers with an average of 0.87 author for each article. The result indicates that the literatures of data mining were usually generated by single author.

TABLE 8
PRODUCTIVITY ANALYSIS OF AUTHOR VIA RECORD COUNT OF DATA MINING FROM 1989 TO 2009 (A)

Record Count (x)	Author(y)	X=log(x)	Y=log(y)	XY	XX
9	1	0.95	0.00	0.00	0.91
8	0	0.90	0.00	0.00	0.82
7	2	0.85	0.30	0.25	0.71
6	3	0.78	0.48	0.37	0.61
5	6	0.70	0.78	0.54	0.49
4	12	0.60	1.08	0.65	0.36
3	37	0.48	1.57	0.75	0.23
2	206	0.30	2.31	0.70	0.09
1	2252	0.00	3.35	0.00	0.00
Total	2519	5.56	9.87	3.26	4.22

(3) Calculate n value (slope):

By the result of calculation on Table 8, it could bring into Lotka's Law's equation as below to calculate n value:

$$n = \frac{N\sum XY - \sum X \sum Y}{N\sum X^2 - (\sum X)^2} \quad (3)$$

We can refer Table 8 for the values in equation:

$$n = \frac{9(3.26) - (5.56)(9.87)}{9(4.22) - (5.56)^2} \quad (4)$$

Then n = -3.629488955

(4) Calculate c value:

After that, we also found c from the equation is shown as below:

$$c = \frac{1}{\sum_1^{p-1} \frac{1}{x^n} + \frac{1}{(n-1)(p^{n-1})} + \frac{1}{2p^n} + \frac{n}{24(p-1)^{n-1}}} \quad (5)$$

P = 8 (max(x)-1), x = 1, 2, 3, 4, 5, 6, 7, 8, 9

Then c = 0.892795157

When we got n = -3.629488955 and c = 0.892795157, it explored:

$$f(x) = 0.892795157/x^{3.629488955} \quad (6)$$

Referring to Table 7, authors with only one literature is 89.40% (100%-10.6%=89.40%), which is almost matched of primitive c value 89.28% generated by Lotka's law. After that, it can follow the calculation to get n and c value by the least squares law, carry onto the further proceeding examination for Lotka's law compliance.

According to Pao [6] suggestion, the absolute value of n should be between 1.2 and 3.8 which formed by the generalized Lotka's Law, the result is matched the reference data by observation. The distribution chart is shown as Fig. 8.

(5) Utilizing K-S (Kolmogorov-Smirnov, K-S) test to evaluate if matched Lotka's Law:

For discussing the value of both n and c, ideal n approximately is -2, c is 0.6079 generated by Lotka's Law, the observation value demonstrated that the data mining literature author distribution and the primitive Lotka's Law are matched approximately. In order to examine whether the theoretical value and the observation value are tallied, the study utilized K-S test to evaluate the reliability of Lotka's Law. Regarding the n and c value which gained by the formula, it is possible to calculate the expected value and the accumulation value of author, following by K-S test examination.

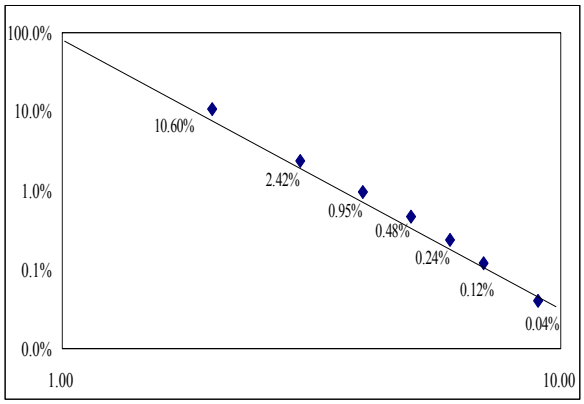

Fig. 8 Distribution of literature productivity of author on data mining research aspect

From Table 9, we can find D_{max} (D_{max} = ABS Value Fo(x)-Sn(x)) = 0.0109. According to K-S test, the threshold value is:

$$1.63/\sqrt{2519} = 0.032477 \quad (7)$$

Because D_{max} is smaller than the K-S test threshold value, the result also indicated that the distribution of author productivity is matched by the Lotka's Law. The result means the Lotka's Law is suitable for the literature author productivity distribution in data mining research domain [7].

TABLE 9
PRODUCTIVITY ANALYSIS OF AUTHOR VIA RECORD COUNT OF DATA MINING FROM 1989 TO 2009 (B)

Record Count	Observation by Author(s)	Subtotal Value Sn(x)	Expected Value by Author(s)	Subtotal Value Fo(x)	ABS Value Fo(x)-Sn(x)
1	0.8940	0.8940	0.8928	0.8928	0.0012
2	0.0818	0.9758	0.0721	0.9649	0.0109
3	0.0147	0.9905	0.0166	0.9815	0.0090
4	0.0048	0.9952	0.0058	0.9873	0.0079
5	0.0024	0.9976	0.0026	0.9899	0.0077
6	0.0012	0.9988	0.0013	0.9913	0.0076
7	0.0008	0.9996	0.0008	0.9920	0.0076
8	0.0000	0.9996	0.0005	0.9925	0.0071
9	0.0004	1.0000	0.0003	0.9928	0.0072

5. CONCLUSION

Data mining is one of fast growing research topics in recently years, this historical review and trend forecast of this research field by each kind of literature characteristic and author productivity distribution is also in growing period. In this analysis, it demonstrated that the current data mining literatures are still continuously to grow, the main research development facility with delivered the largest production is USA, but England, Taiwan, Canada and People R. China also have potential to deliver more literatures in the future. The study verified by K-S test to demonstrate frequency indexes of data mining author productivity distribution were certainly followed by Lotka's Law. The distribution of author productivity result indicates that the literatures of data mining were usually generated by single author. The applications of data mining are mainly following by research aspects which in term of information science & library science, computer science & information systems, operations research & management science, management, computer science & artificial intelligence, economics, computer science & interdisciplinary applications, public, environmental & occupational health and engineering, electrical & electronic. This implies that the phenomenon "success breeds success" is more common in higher quality publications.

REFERENCES

[1] Lyman, Peter; Hal R. Varian, "How Much Information", Retrieved 2008-12-17, 2003

[2] Lotka, Alfred J., "The frequency distribution of scientific productivity", *Journal of the Washington Academy of Sciences*, vol. 16, issue 12, pp. 317–324, 1926.

[3] Zipf, G. K., *Human Behavior and the Principle of Least Effor: An Introduction to Human Ecology*, Cambridge, Mass Addision-Wesley, 1949

[4] Egghe, L., "Relations between the continuous and the discrete Lotka power function", *Journal of the American Society for Information Science and Technology*, vol. 56, issue 7, pp. 664–668, May 2005.

[5] Tsai, Ming-Yue, *The characteristic of informetric and bibliometric*, Taiwan: Hwa-Tai, ch. 6, 2003

[6] Pao, M. L., *Concept of Information Retrieve*, Englewood, Colorado: Libraries Unlimited, pp. 23-25, 1989.

[7] Potter, W. G., "Lotka's Law Revisited", *Library Trends*, 30(1), 21-393, Summer, 1981.