View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by AIS Electronic Library (AISeL)

Association for Information Systems

AIS Electronic Library (AISeL)

ICEB 2016 Proceedings

International Conference on Electronic Business (ICEB)

Winter 12-4-2016

Taiwan Vs Globalization In Data Mining: A Study Of Research Trends, Forecasts And Citations

Hsu-Hao Tsai

Chihlee University of Technology, Taiwan, simon705@mail.chihlee.edu.tw

Follow this and additional works at: https://aisel.aisnet.org/iceb2016

Recommended Citation

Tsai, Hsu-Hao, "Taiwan Vs Globalization In Data Mining: A Study Of Research Trends, Forecasts And Citations" (2016). *ICEB 2016 Proceedings*. 59.
https://aisel.aisnet.org/iceb2016/59

This material is brought to you by the International Conference on Electronic Business (ICEB) at AIS Electronic Library (AISeL). It has been accepted for inclusion in ICEB 2016 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

Taiwan Vs Globalization In Data Mining：
A Study Of Research Trends, Forecasts And Citations

Hsu-Hao Tsai, ChihleeUniversity of Technology Taiwan, Taiwan, simon705@mail.chihlee.edu.tw

ABSTRACT

Although data mining (DM) has already become more important recently,there are few comprehensive studies and categorization schemes to discuss the characteristics for DM. Applying bibliometric method, this paper explores research potential of DM in Taiwan through comparing globalization DM trends, forecasts and citations from 1993 to 2016 by locating heading "data mining" in topic in the Web of Science (WoS) database. The bibliometric analytical technique was used to examine the topic in WoS journals from 1993 to 2016, we found of 245 articles ofTaiwan and 3053 articles of globalization. This paper surveys and classifies DM articles between Taiwan and globalizationusing the following eight categories – publication year, citation, document type, country/territory, institute name, language, source title and research area – for different distribution status in order to find the difference and how DM technologies and applications have developed in this period. Finally, the study will analyze DM technology trends, forecasts and citations based on the above results. Also, the paper performs the K-S test to check whether the distribution of author article productionof Taiwan and globalization follows Lotka's law or not. According to the analyzingresults, this paper provides a roadmap for future researches, abstracts technology trend information and facilitates knowledge accumulation. Therefore, the researches of DM in Taiwancan follow and concentrate the globalization categories, and create the potential in the near future.

Keywords: Data mining, Research trend analysis, Bibliometric methodology

INTRODUCTION

DM is treated as an interdisciplinary domain that combines artificial intelligence, database management, data visualization, machine learning, mathematic algorithms, and statistics. DM, also being as knowledge discovery in databases (KDD) [1-2], is a rapidly emerging domain. This technology introduces different methodology for decision-making, problem solving, analysis, planning, diagnosis, detection, integration, prevention, learning, and innovation.

Since the 1960s, DM techniques have formed a branch of appliedartificial intelligence. During the interveningdecades, important innovations in computer systems have initializedthe introduction of new technologies [3], forweb-based education. Data mining allows a search, for valuableinformation, in large volumes of data [4].The explosive growth in databases has created a need to developtechnologies that use information and knowledge intelligently. Therefore, DM technique has become an increasingly important research area [5].

This technology is motivated by the need of new techniques to help analyze, understand or even visualize the huge amounts of stored data gathered from business and scientific applications. It is the process of discovering interesting knowledge, such as patterns, associations, changes, anomalies and significant structures from large amounts of data stored in databases, data warehouses, or other information repositories. It can be used to help companies to make better decisions to stay competitive in the marketplace. It is suggested by [6] that different social science methodologies, such as psychology, cognitive scienceand human behavior might implement DM technique, as an alternative to the methodologies already on offer.The major DM functions that are developed in commercial and research communities include summarization, association, classification, prediction and clustering. These functions can be implemented using a variety of technologies, such as database-oriented techniques, machine learning and statistical techniques [7].

DM was defined by [8] as a process that uses statistical, mathematical, artificial intelligence and machine-learning techniques to extract and identify useful information and subsequently gain knowledge from large databases. In an effort to develop new insights into practice-performance relationships, [9] indicated that DM was used to explore improvement programs, strategic priorities, environmental factors, manufacturing performance dimensions and their interactions. DM was also defined as the process of extracting or detecting hidden patterns or information from large databases [10-13]. With an enormous amount of customer data, DM technology can support business intelligence to generate new opportunities [14-21].

Recently, [22] demonstrated a number of DM applications and prototypes have been developed for a variety of domains including marketing, banking, financing, manufacturing and health care. In addition, DM has also been applied to other types of data such as time-series, spatial, telecommunications, web, and multimedia data. In general, the DM process, and the DM technique and function to be applied depend very much on the application domain and the nature of the data available.
Based on prior research, this study tries to explore the research potential of DM in Taiwan by bibliometric through comparing globalization trends. In section two, this paper presents the material and methodology. In section three, this study presents the research results, while the section four and five is devoted to an analysis and a discussion of the study, and also describes research limitsand makes suggestions for future research.

MATERIAL AND METHODOLOGY

Research material

The very influential role of the monopolist citation data producer ISI (Institute for Scientific Information, now Thomson

Scientific) was pointed by as its commercialization of these data [25] rapidly increased the non-expert use of bibliometric analysis such as rankings. The materials used in this study were accessed from the database of the Social Science Citation Index (SSCI), obtained by subscription from the ISI, Web of Science (WoS), Philadelphia, PA, USA. In this research, we discuss the papers published in the period from 1993 to 2016 because there was only few data prior to that year.

Research methodology

Bibliometrics was defined by [26]as the application of mathematics and statistical methods to books and other media of communication. [27]also defined bibliometrics as the quantitative study of physical published units, or of bibliographic units, or of the surrogates for either. Bibliometric techniques have been used primarily by information scientists to study the growth and distribution of the scientific article. Researchers may use bibliometric methods of evaluation to determine the influence of a single writer, for example, or to describe the relationship between two or more writers or works. Besides, properly designed and constructed [28-30], bibliometrics can be applied as a powerful support tool to peer review. This is certainly possible for interdisciplinary research fields[31].One common way of conducting bibliometric research is to use the Social Science Citation Index (SSCI), the Science Citation Index (SCI), the Arts and Humanities Citation Index (A&HCI), the Conference Proceedings Citation Index-Science (CPCI-S) or the Conference Proceedings Citation Index-Social Science& Humanities (CPCI-SSH) to trace citations.

Lotka's law

Lotka's law indicates the frequency of publication by authors in a given field. It mentions that the number of authors making n contributions is about $1/n^2$ of those making one; and the proportion of all contributors, that make a single contribution, is about 60 percent [32]. Lotka's law is stated by the following formula is where y is the number of authors making x contributions, the exponent n and the constant c are parameters to be estimated from a given set of author productivity data. This means that out of all the authors in a given field, about 60 percent will have just one publication, about 15 percent will have two publications ($1/2^2$ times 0.60), about 7 percent of authors will have three publications ($1/3^2$ times 0.60), and so on. Lotka's law, when applied to large bodies of article over a fairly long period of time, can be accurate in general, but not statistically exact. It is often used to

estimate the frequency with which authors will appear in an online catalog [33].

Lotka's law is generally used for understanding the productivity patterns of authors in a bibliography [34-39]. In this article, Lotka's law is chosen to perform bibliometric analysis to check the number of publications versus accumulated authors between 1993 and 2016 to perform an author productivity inspection to collect the results for research tendency in the near future. To verify the analysis, the paper implements the K-S test to evaluate whether the result matches Lotka's law.

Research architecture

Using a bibliometric approach, the paper explores research potential of DM in Taiwan through comparing globalization DM trends, forecasts and citationsfrom 1993 to 2016by locating heading "data mining" in topics in the WoS database. The bibliometric analytical technique was used to examine the topic in WoS journals from 1993 to 2016, we found of 245 articles in Taiwan and 3053 articles in globalization. This paper surveys and classifies DM articles between Taiwan and globalization using the following eight categories – publication year, citation, document type, country/territory, institute name, language, source title and research area – for different distribution status in order to find the difference and how DM technologies and applications have developed in this period. Finally, the study will analyze DM technology trends, forecasts and citations under the above results.

As a verification of its analysis, the paper implements the following steps to check whether the analysis follows Lotka's law:

(1) Collect data

(2) List author & article distribution table

(3) Calculation the value of n (slope)

According to Lotka's law, the generalized formula is x^ny=c, the value of n is -2. The parameter n of applied field is calculated by the least square-method using the following formula [39]:

$$n = \frac{N\sum XY - \sum X \sum Y}{N\sum X^2 - (\sum X)^2} \quad (1)$$

N is the number of pairs of data, X is the logarithm of publications(x) and Y is the logarithm of authors (y).

The least-square method is used to estimate the best value forthe slope of a regression line which is the exponent n for Lotka's law[39]. The slope is usually calculated without data pointsrepresenting authors of high productivity. Since values of the slopechange with different number of points for the same set of data, wehave made several computations of n. The median or

the mean valuesof n can also be identified as the best slope for the observed distribution [39]. Different values of n produce different valuesof the constant c.

(4) Calculation the value of c

According to Lotka's law, the generalized formula is x^ny=c, the value of c is 0.6079. The parameter c of applied field is calculated using the following formula [39]:

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

p is the 20, n is the value obtained in (3) Calculation the value of n,and x is the number of publications.

(5) Utilizing the K-S (Kolmogorov-Smirnov, K-S) test to evaluate whether the analysis matches Lotka's law[39]suggests the K-S test, a goodness-of-fit statistical test to assert that the observed author productivity distribution is not significantly different from a theoretical distribution. The hypothesis concerns a comparison between observed and expected frequencies. The test allows the determination of the associated probability that the observed maximum deviation occurs within the limits of chance. The maximum deviation between the cumulative proportions of the observed and theoretical frequency is determined by the following formula [39]:

$$D = \text{Max} \mid Fo(x) - Sn(x) \mid \quad (3)$$

Fo(x) = theoretical cumulative frequency

Sn(x) = observed cumulative frequency

The test is performed at the 0.01 level of significance. When sample size is greater than 35, the critical value of significance is calculated by the following formula [39]:

$$\text{The critical value at the 0.01 level of significance} = 1.63 / \sqrt{(\sum y)} \quad (4)$$

$\sum y$ = the total population under study

If the maximum deviation falls within the critical value the null hypothesis that the data set conforms to Lotka's law can be accepted at a certain level of significance. But if it exceeds the critical value the null hypothesis must be rejected at a certain level of significance and concluded that the observed distribution is significantly different from the theoretical distribution.Based on the analyzingresults, this paper provides a roadmap for future researches, abstracts technology trend information and facilitates knowledge accumulation. Therefore, the researches of DM in Taiwancan follow and concentrate the globalization categories, and create the potential in the near future.

RESEARCH RESULTS

Distribution by Publication Year

As Fig. 1 shows, the article production of Taiwan and globalization on DM has been rising since 1995. The article distribution can be divided into three segments to show the trend of development: (1) from 1993 to 2004, (2) from 2005 to 2008 and (3) from 2009 to 2016 in Taiwan, and (1) from 1993 to 2000, (2) from 2001 to 2005 and (3) from 2006 to 2016 in globalization. From 1993 to 1998, DM did not draw many researchers' attention. After 1998, the publication productivity per annum steadily increased, was followed by fast growth between 1999 and 2003, and rapidly peaked in 2015. The publication in 2016 is still increase. The status implicates that DM has great potential to grow in the future.

Regarding the relationship between article production and citations in DM of Taiwan, thecitations almost follow the article production ranking accordingly (Fig. 2). On the other side, the citations also follow the article production ranking accordingly (Fig. 3).

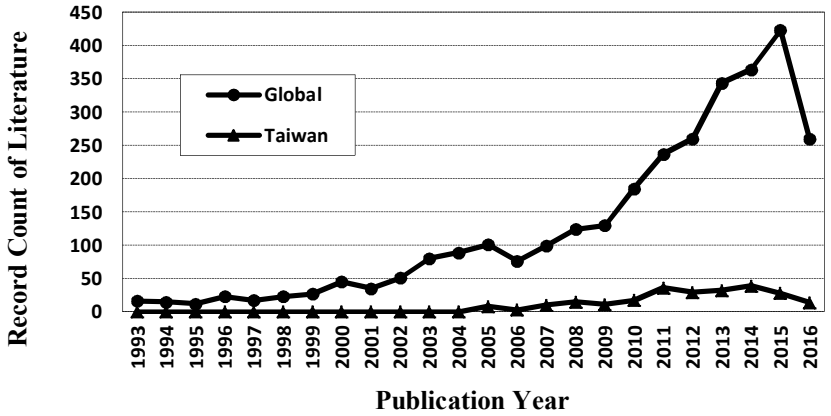

Figure 1: The tendency chart of publication growth of DMin Taiwan and globalization area
(Source: WoS database on 2016/06/28)

Distribution by Citation

From Fig. 4, we can see that the citation distribution of Taiwan and globalizationon DM is not easy to recognize between 1993 and 1999, followed by a dramatic growth and rapidly peaked in 2015 of DM. The citation in 2016 is still increase. The result indicates that DM will become more popular in the future.

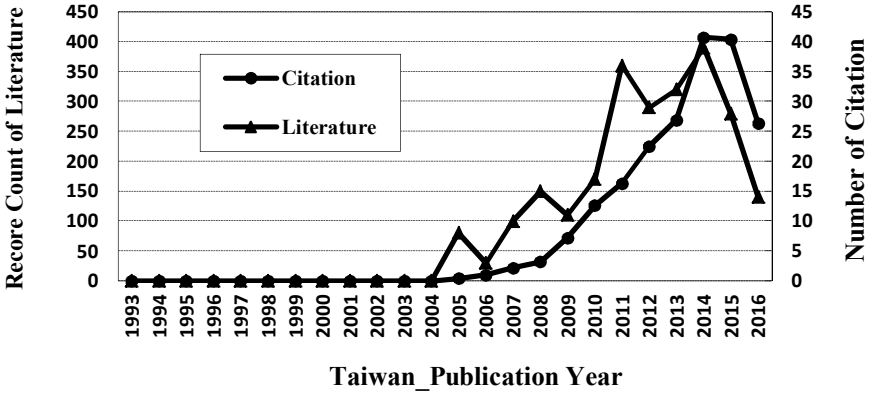

Figure2: Publication and citation distribution of publication year in Taiwan of DM
(Source: WoS database on 2016/06/28)

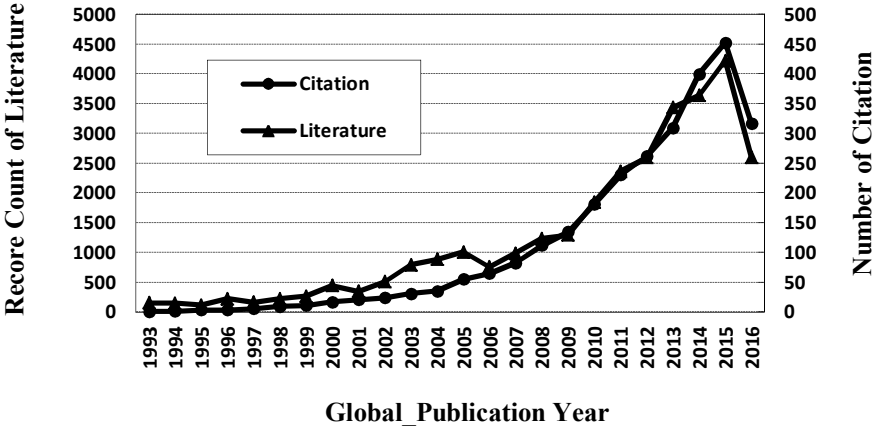

Figure3: Publication and citation distribution of publication year in globalization of DM
(Source: WoS database on 2016/06/28)

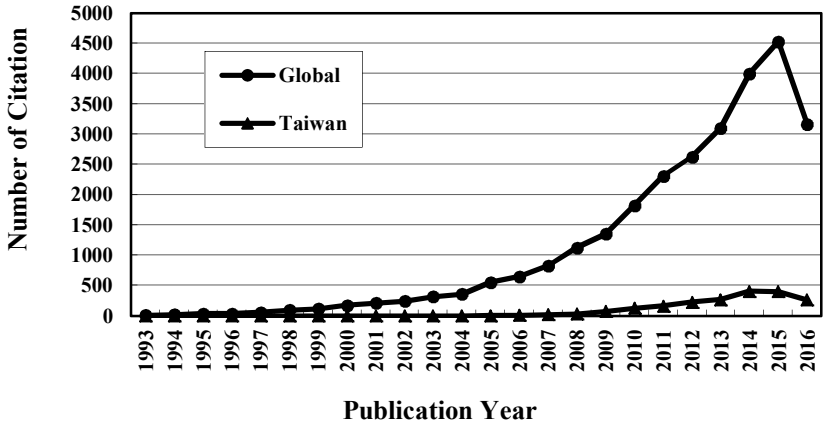

Figure4:Citation status in each year of DMin Taiwan and globalization area
(Source: WoS database on 2016/06/28)

Distribution by Country/Territory

According to Fig. 5, we can figure out that the US is at the top with 1225 (40.12%) in DM of globalization, following by England, with 288 (9.43%), respectively. Taiwan ranks third with 245 (8.03%) in DM. Behind them, the PRC, Australia, Canada, Spain and Germany are also major academic providers in the field of DM.

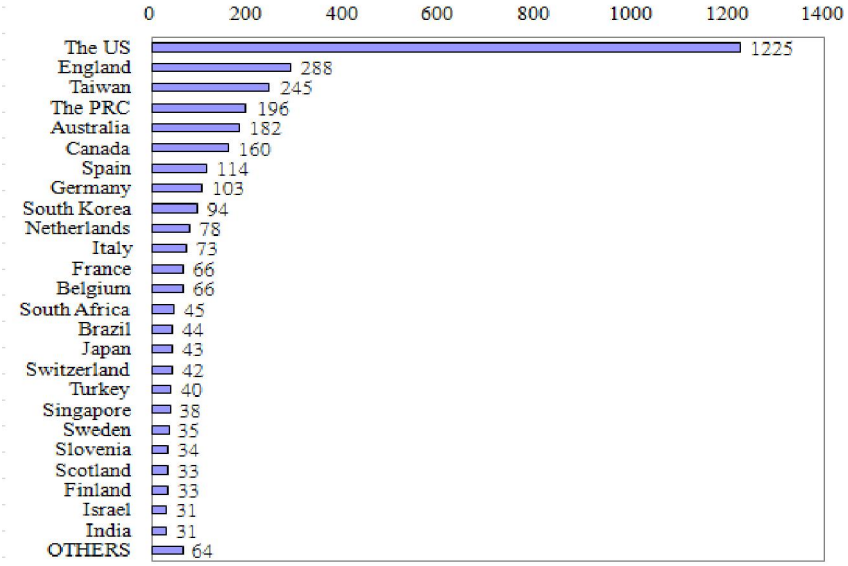

Figure5: Distribution of top 25 countries/territories in DM of globalization from 1993 to 2016
(source: WoS database on 2016/06/28)

Distribution by Institution Name

Table 1 is easy to summarize that National Central University and National ChiaoTung University are the same scholarly affiliations of both Taiwan and globalization in DM research domain. The distribution of institutions shows that the US is still the most productive country within the research aspect of DM in globalization.

Regarding the relationship between article production and citations in DM of Taiwan, there are 25 articles from National ChiaoTung University and 16 articles from TamKang University, but their citations are 411 times and 222 times in the domain. The others almost follow the article production ranking accordingly (Fig. 6). On the other side, there are 29 articles from University of Wisconsin, 29 articles from Pennsylvania State University and 25 articles from National ChiaoTung University in DM of globalization, but their citations are 385 times, 477 times and 411 times in the domain. The others follow the article production ranking accordingly (Fig. 7).

Table 1: Distribution of the top 10 institutions for DM from 1993 to 2016
(source: WoS database on 2016/06/28)

Data Mining of Taiwan			Data Mining of Globalization			
Institution Name	NP	% of 245	Institution Name	NP	% of 3053	Country

National Central University	28	11.43%	Columbia University	31	1.02%	The US
National ChiaoTung University	25	10.20%	University of Wisconsin	29	0.95%	The US
NationalChungChengUniversity	17	6.94%	University of Illinois	29	0.95%	The US
TamKangUniversity	16	6.53%	National Central University	28	0.92%	Taiwan
National Taiwan University	13	5.31%	Pennsylvania State University	27	0.88%	The US
NationalChangHuaUniversityofEducation	13	5.31%	NIOSH	27	0.88%	The US
NationalChengChiUniversity	12	4.90%	University of Pennsylvania	26	0.85%	The US
NationalTsingHuaUniversity	11	4.49%	University of Queensland	25	0.82%	Australia
NationalChengKungUniversity	11	4.49%	National ChiaoTung University	25	0.82%	Taiwan
YuanZeUniversity	10	4.08%	KatholiekeUniversiteit Leuven	24	0.79%	Belgium
NationalYunLinUniversity of Science and Technology	10	4.08%	Hong KongPolytech University	24	0.79%	The PRC
National Taiwan University of Science and Technology	10	4.08%				

NP=number of publication

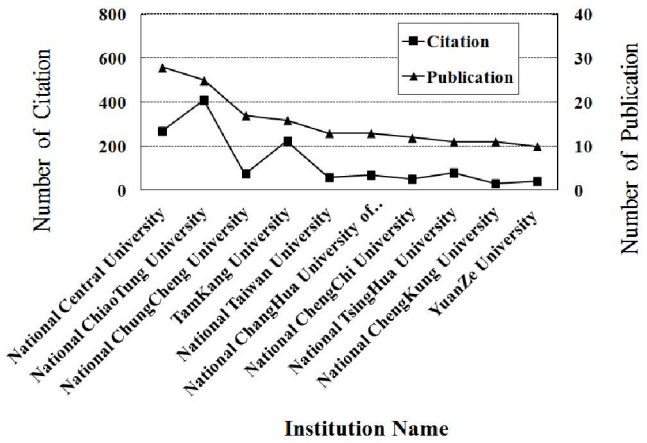

Figure6:Publication and citation distribution of top 10 institutions of DM in Taiwan
(Source: WoS database on 2016/06/28)

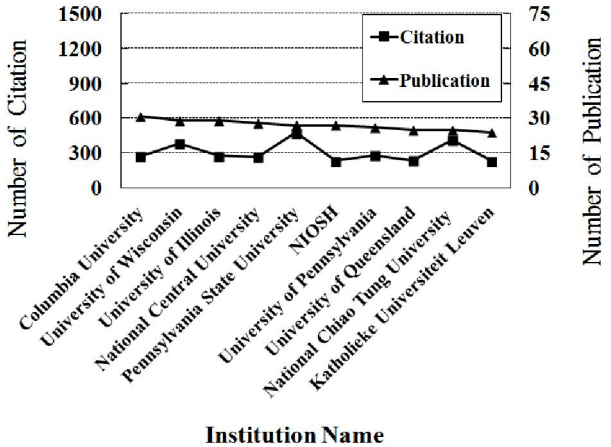

Figure7:Publication and citation distribution of top 10 institutions of DM in globalization
(Source: WoS database on 2016/06/28)

Distribution by Document Type

In Table 2, the distribution of document types from 1993 to 2016 indicates that the most popular publication document type is "Article" (2777 articles, 90.96% in globalization and 240 articles, 97.96% in Taiwan). The result demonstrates that the article is the major tendency of document type in DM research domain.

Table 2: Distribution of document types for DM from 1993 to 2016
(source: WoS database on 2016/06/28)

Data Mining of Taiwan			Data Mining of Globalization		
Document Type	NP	% of 245	Document Type	NP	% of 3053
Article	240	97.96%	Article	2777	90.96%
Proceedings Paper	7	2.86%	Proceedings Paper	135	4.42%
Review	2	0.82%	Review	103	3.37%
Meeting Abstract	2	0.82%	Book Review	76	2.49%
			Editorial Material	49	1.61%
Total	245	100%	Meeting Abstract	38	1.25%
			Note	3	0.10%
			Correction	3	0.10%
			News Item	2	0.07%
			Book Chapter	2	0.07%
			Total	3053	100%

NP=number of publication

Distribution by Language

In Table 3, the majority language for Taiwan and globalization researchers is English, with 245 articles (100.00%) in Taiwan and 2973 articles (97.38%) in globalization. Clearly, English is still the main trend in DM research domain.

Table 3: Distribution of languages for DM from 1993 to 2016
(source: WoS database on 2016/06/28)

Data Mining of Taiwan			Data Mining of Globalization		
Language	NP	% of 245	Language	NP	% of 3053
English	245	100%	English	2973	97.38%
			Spanish	23	0.75%
Total	245	100%	German	13	0.43%
			Portuguese	11	0.36%
			French	6	0.20%
			Slovak	5	0.16%
			Czech	5	0.16%
			Slovene	3	0.10%
			Russian	3	0.10%
			Japanese	3	0.10%
			Italian	3	0.10%
			Korean	2	0.07%
			Total	3053	100%

NP= number of publication

Table 4: Distribution of top 10research areas for DM from 1993 to 2016
(source: WoS database on 2016/06/28)

Data Mining of Taiwan			Data Mining of Globalization		
Research Area	NP	% of 245	Research Area	NP	% of 3053
Computer Science	154	62.86%	Computer Science	998	32.69%
Engineering	78	31.84%	BusinessEconomics	623	20.41%
Operations Research & Management Science	75	30.61%	Information Science & Library Science	483	15.82%

Information Science & Library Science	45	18.37%	Engineering	443	14.51%
BusinessEconomics	37	15.10%	Operations Research & Management Science	400	13.10%
Education & Educational Research	30	12.25%	Environmental Sciences Ecology	228	7.47%
Medical Informatics	16	6.53%	PublicEnvironmental&Occupational Health	213	6.98%
Health Care Sciences& Services	11	4.49%	Psychology	162	5.31%
Social Sciences& Other Topics	9	3.67%	Geography	140	4.59%
Mathematics	6	2.45%	Medical Informatics	136	4.46%
			Neurosciences Neurology	33	1.08%

NP=number of publication

Distribution by ResearchArea

Table 4 offers critical information for future research tendencies in DM, allowing researchers a better understanding of the distribution of the top 10 research areas in future research. The top three research areas in Taiwan for DM research domains are computer science (154 articles, 62.86%), followed by engineering (78 articles, 31.84%) and operations research & management science (75 articles, 30.61%). Besides, this paper's analysis suggests that there are other important research disciplines in Taiwanfor DM article production such as information science & library science, business economics, education & educational research, medical Informatics, healthcare sciences & services, social sciences & other topics and mathematics.
On the other hand, the top three in globalization for DM research domains are computer science (998 articles, 32.69%), followed by business economics (623 articles, 20.41%) and information science & library science (483 articles, 15.82%). Analysis reveals that there are many additional research domains in globalizationfor DM article production, such as engineering, operations research &management science, environmental sciences ecology, public environmental & occupational health, psychology, geography and medical informatics.

As Fig. 8 illustrates, the citations of Taiwanfollow article production ranking in the top 10 research area, except for Social sciences & other topics (14.89 average citations per article). From Fig. 9, we can find that the citations of globalization follow article production ranking in the top 10 research area, except for public environmental & occupational health (9.56 average citations per article).

Analysis of the top five research area—computer science, business economics, information science & library science, engineering, and operations research & management science—shows that these research area are all repeated in DM research for Taiwan and globalization, indicating that these research area will become the most important category for DM researchers.

Table 5: Distribution of top 10 source titles for DM from 1993 to 2016
(source: WoS database on 2016/06/28)

Data Miningof Taiwan			Data Mining of Globalization		
Source Title	NP	% of 245	Source Title	NP	% of 3053
Expert Systemswith Applications	55	22.45%	*Expert Systemswith Applications*	177	5.80%
Electronic Library	10	4.08%	*Journalof the American Medical Informatics Association*	58	1.90%
Computers & Education	9	3.67%	*Decision Support Systems*	50	1.64%
Knowledge Based Systems	8	3.27%	*International Journal of Geographical Information Science*	46	1.51%
Journalof Information Science	8	3.27%	*Journalof the American Society for Information Science and Technology*	44	1.44%
Educational Technology & Society	7	2.86%	*Resources Policy*	41	1.34%
Decision Support Systems	6	2.45%	*Information Processing &Management*	39	1.28%
Online Information Review	5	2.04%	*Journalof the Operational Research Society*	32	1.05%
Information Processing &Management	5	2.04%	*Journal of Information Science*	28	0.92%
Electronic Commerce Research and Applications	5	2.04%	*Scientometrics*	27	0.88%

NP=number of publication

Distribution by Source Title

Table 5 highlights information on trends for DM, allowing researchers to closely approach the distribution of the top 25 source

titles in future research. The top three DM research journals in Taiwanare *Expert Systems with Applications* (55 articles, 22.45%), followed by *Electronic Library* (10 articles, 4.08%) and *Computers &Education*(9 articles, 3.67%). In addition, there are a significant number of research sources for DM article production in Taiwansuch as *Knowledge Based Systems*, *Journalof Information Science*, *Educational Technology & Society*, *Decision Support Systems*, *Online Information Review*,*Information Processing &Management* and *Electronic Commerce Researchand Applications*.

In the meantime, the top three DM research journals in globalization are *Expert Systems with Applications* (177 articles, 5.80%), followed by *Journal of the American Medical Informatics Association* (58 articles, 1.90%) and *Decision Support Systems* (50 articles, 1.64%). Moreover, it also find out that there are a lot of research sources for DM article production such as *International Journal of Geographical Information Science, Journal of the American Society for Information and Technology*, *Resources Policy, Information Processing & Management*,*Journalof the Operational Research Society*, *Journalof Information Science* and *Scientometrics*.

In Fig. 10, the citations of Taiwanfollow article production ranking in the top 10 sources, except for *Electronic Library*(4.90 average citations per article), *Decision Support Systems* (11.83 average citations per article), *Information Processing &Management*(8.20 average citations per article), and *Electronic Commerce Researchand Applications*(8.00 average citations per article).

As Fig. 11 shows, we can find that the citations of globalization follow article production ranking in the top 10 sources, except for *Decision Support Systems* (16.74 average citations per article), *International Journal of Geographical Information Science*(13.87 average citations per article), *Information Processing &Management*(9.69 average citations per article), *International Journal of Geographical Information Science* (9.70 average citations per article) and *Scientometrics*(9.63average citations per article).

Analysis of the top ten journal sources—*Expert Systems with Applications*, *Decision Support Systems, Information Processing &Management* and *Journalof Information Science*—shows that these source title are all the same at DM research for Taiwanand globalization, indicating that these source title will become the most important category for DM researchers.

DISCUSSION

The section implements the steps which are demonstrated in Section 2.2.2 to verify whether the distribution of author article production of Taiwan and globalization follows Lotka's law in DM research domain.3

Taiwan

(1) Collect data and

(2) List author & article distribution table

Author quantity is calculated by the equality method from 245 articles retrieved by the WoS index. Altogether, 495 authors on DM in Taiwanare included. See Table 6 for reference.

(3) Calculation the value of n (slope)

In Table 7, we list the number of authors and the number of publications by one author for calculation of the exponent n with the topic as "data mining" in WoS database. The results of the calculations in Table 7 can be brought into the Equation (1) to calculate the value of n:

$$n = \frac{12(2.59)-(8.68)(7.83)}{12(7.46)-(8.68)(8.68)} \quad (5)$$

Then we can find n = -2.591496952

Table 6: Calculation of author productivity of DM in Taiwan

NP	Author (s)	(NP) * (Author)	Accumulated Record	Accumulated Record %	Accumulated Author(s)	Accumulated Author(s)%
12	1	12	12	1.75%	1	0.20%
11	0	0	12	1.75%	1	0.20%
10	0	0	12	1.75%	1	0.20%
9	0	0	12	1.75%	1	0.20%
8	1	8	20	2.92%	2	0.40%

7	0	0	20	2.92%	2	0.40%
6	4	24	44	6.42%	6	1.21%
5	3	15	59	8.61%	9	1.82%
4	10	40	99	14.45%	19	3.84%
3	22	66	165	24.09%	41	8.28%
2	66	132	297	43.36%	107	21.62%
1	388	388	685	100.00%	495	100.00%

NP=number of publication

(4) Calculation the value of c

The value of c is calculated by using the Equation (2), where P = 20, x = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 and n = 2.591496952, then we can find c = 0.777715357.

With n = -2.591496952 and c = 0.777715357, the Lotka's law equation of DM in Taiwan is:

$$f(x) = 0.777715357/\mathrm{x}^{2.591496952} \quad (6)$$

When the result is compared to Table 6, we can see that authors with only one article account for 78.38% (100%-21.62%= 78.38%), which almost matches the primitive c value 77.77% generated by Lotka's law. The values for n and c can be calculated by the least squares law and then brought into further analysis for Lotka's law compliance.

According to [40], the absolute value of n should be between 1.2 and 3.8, as given by the generalized Lotka's law. The result indicates that n (=2.591496952) is between 1.2 and 3.8 and is matched the reference data by observation. The detail distribution chart is shown in Fig. 12.

Table 7: Calculation of the exponent n for DM in Taiwan

x (NP)	y (Author)	X=log(x)	Y=log(y)	XY	XX
12	1	1.08	0.00	0.00	1.16
11	0	1.04	-	-	1.08
10	0	1.00	-	-	1.00
9	0	0.95	-	-	0.91
8	1	0.90	0.00	0.00	0.82
7	0	0.85	-	-	0.71
6	4	0.78	0.60	0.47	0.61
5	3	0.70	0.48	0.33	0.49
4	10	0.60	1.00	0.60	0.36
3	22	0.48	1.34	0.64	0.23
2	66	0.30	1.82	0.55	0.09
1	388	0.00	2.59	0.00	0.00
Total	495	8.68	7.83	2.59	7.46

x= number of publication
y=author
X = logarithm of x
Y = logarithm of y

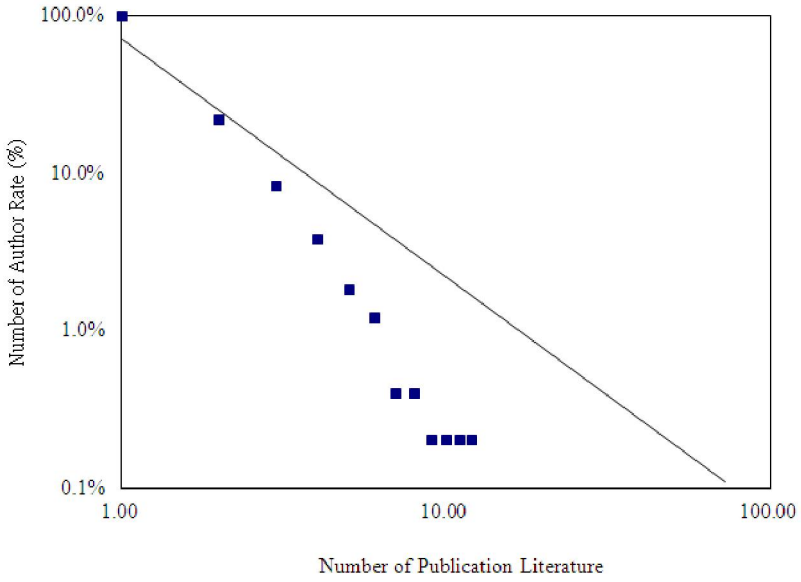

Figure12:Distribution of literature productivity of author on DM research aspect in Taiwan

(5) Utilize the K-S test to evaluate whether the analysis matches Lotka's law

We use Equation (3) to evaluate whether the analysis matches Lotka's law. From Table 8, we can find D= 0.0104(D = Max | Fo(x)-Sn(x) |). According to the K-S test, the critical value at 0.01 level of significance is calculated by Equation (4):

$$1.63/\sqrt{495} = 0.07326305 \quad (7)$$

Table 8: The K-S test for DM in Taiwan

NP	Author (s)	Taiwan (Observed)	Sn(x)	Taiwan (Expected)	Fo(x)	D
1	388	0.7838	0.7838	0.7777	0.7777	0.0061
2	66	0.1333	0.9172	0.1290	0.9067	0.0104
3	22	0.0444	0.9616	0.0451	0.9519	0.0097
4	10	0.0202	0.9818	0.0214	0.9733	0.0085
5	3	0.0061	0.9879	0.0120	0.9853	0.0026
6	4	0.0081	0.9960	0.0075	0.9928	0.0032
7	0	0.0000	0.9960	0.0050	0.9978	0.0018
8	1	0.0020	0.9980	0.0036	1.0013	0.0034
9	0	0.0000	0.9980	0.0026	1.0040	0.0060
10	0	0.0000	0.9980	0.0020	1.0060	0.0080
11	0	0.0000	0.9980	0.0016	1.0075	0.0095
12	1	0.0020	1.0000	0.0012	1.0088	0.0088

NP: number of publication; Taiwan: author productivity of Taiwan; Sn(x) = observed cumulative frequency; Fo(x) = theoretical cumulative frequency; D = maximum deviation

Globalization

(1) Collect data and

(2) List author & article distribution table

Author quantity is calculated by the equality method from 3053 articles retrieved by the WoS index. Altogether, 7354 authors on DM in globalization are included. See Table 9 for reference.

Table 9: Calculation of author productivity of DM in globalization

NP	Author (s)	(NP) * (Author)	Accumulated Record	Accumulated Record %	Accumulated Author(s)	Accumulated Author(s)%

15	1	15	15	0.17%	1	0.01%
14	1	14	29	0.33%	2	0.03%
13	0	0	29	0.33%	2	0.03%
12	2	24	53	0.60%	4	0.05%
11	4	44	97	1.09%	8	0.11%
10	1	10	107	1.21%	9	0.12%
9	4	36	143	1.61%	13	0.18%
8	5	40	183	2.06%	18	0.24%
7	8	56	239	2.69%	26	0.35%
6	15	90	329	3.71%	41	0.56%
5	20	100	429	4.84%	61	0.83%
4	57	228	657	7.41%	118	1.60%
3	175	525	1182	13.32%	293	3.98%
2	629	1258	2440	27.50%	922	12.54%
1	6432	6432	8872	100.00%	7354	100.00%

NP = number of publication

(3) Calculation the value of n (slope)

In Table 10, we list the number of authors and the number of publications by one author for calculation of the exponent n with the topic as "data mining" in WoS database. The results of the calculations in Table 10 can be brought into the Equation (1) to calculate the value of n:

$$n = \frac{15(7.08)-(12.12)(15.49)}{15(11.4)-(12.12)^2} \quad (8)$$

Then we can find n = -3.36257094

Table 10: Calculation of the exponent n for DM in globalization

x (NP)	y (Author)	X=log(x)	Y=log(y)	XY	XX
15	1	1.18	0.00	0.00	1.38
14	1	1.15	0.00	0.00	1.31
13	0	1.11	-	-	1.24
12	2	1.08	0.30	0.32	1.16
11	4	1.04	0.60	0.63	1.08
10	1	1.00	0.00	0.00	1.00
9	4	0.95	0.60	0.57	0.91
8	5	0.90	0.00	0.00	0.82
7	8	0.85	0.90	0.76	0.71
6	15	0.78	1.18	0.92	0.61
5	20	0.70	1.30	0.91	0.49
4	57	0.60	1.76	1.06	0.36
3	175	0.48	2.24	1.07	0.23
2	629	0.30	2.80	0.84	0.09
1	6432	0.00	3.81	0.00	0.00
Total	7354	12.12	15.49	7.08	11.40

x = number of publication
y = author
X = logarithm of x
Y = logarithm of y

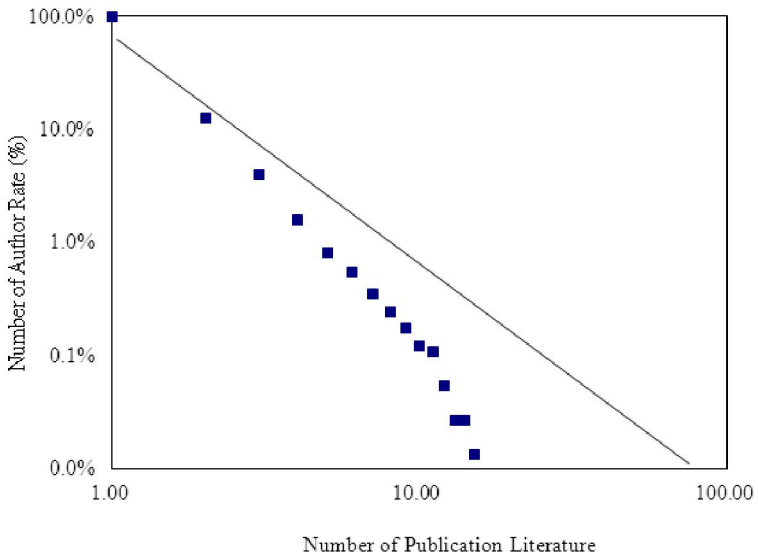

Figure13: Distribution of literature productivity of author on DM research aspect in globalization

(4) Calculation the value of c

The value of c is calculated by using the Equation (2), where P = 20, x = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15 and n = 3.36257094, then we can find c = 0.87243801.

With n = -3.36257094and c = 0. 87243801, the Lotka's law equation of DM is:

$$f(x) = 0.87243801/x^{3.36257094} \tag{9}$$

When the result is compared to Table 9, we can see that authors with only one article account for 87.46% (100%-12.54%=87.46%), which almost matches the primitive c value 87.24% generated by Lotka's law. The values for n and c can be calculated by the least squares law and then brought into further analysis for Lotka's law compliance.

According to [40], the absolute value of n should be between 1.2 and 3.8, as given by the generalized Lotka's law. The result indicates that n (=3.36257094) is between 1.2 and 3.8 and is matched the reference data by observation. The detail distribution chart is shown in Fig. 13.

(5) Utilize the K-S test to evaluate whether the analysis matches Lotka's law

We use Equation (3) to evaluate whether the analysis matches Lotka's law. From Table 11, we can find D (D = Max | Fo(x)-Sn(x) |) = 0.0050. According to the K-S test, the critical value at 0.01 level of significance is calculated by using the Equation (4):

$$1.63/\sqrt{7354} = 0.01900753 \tag{10}$$

Table 11: The K-S test for DM in globalization

NP	Author (s)	Globalization (Observed)	Sn(x)	Globalization (Expected)	Fo(x)	D
1	6432	0.8746	0.8746	0.8724	0.8724	0.0022
2	629	0.0855	0.9602	0.0848	0.9573	0.0029
3	175	0.0238	0.9840	0.0217	0.9790	0.0050
4	57	0.0078	0.9917	0.0082	0.9872	0.0045
5	20	0.0027	0.9944	0.0039	0.9911	0.0033
6	15	0.0020	0.9965	0.0021	0.9932	0.0033
7	8	0.0011	0.9976	0.0013	0.9945	0.0031
8	5	0.0007	0.9982	0.0008	0.9953	0.0030
9	4	0.0005	0.9988	0.0005	0.9958	0.0030

10	1	0.0001	0.9989	0.0004	0.9962	0.0027
11	4	0.0005	0.9995	0.0003	0.9965	0.0030
12	2	0.0003	0.9997	0.0002	0.9967	0.0031
13	0	0.0000	0.9997	0.0002	0.9968	0.0029
14	1	0.0001	0.9999	0.0001	0.9969	0.0029
15	1	0.0001	1.0000	0.0001	0.9970	0.0030

NP = number of publication
Data Miming = author productivity of data mining
Sn(x) = observed cumulative frequency
Fo(x) = theoretical cumulative frequency
D = maximum deviation

Discussion

(1) Constant c computed 0.777715357. Using the K-S test it is found that at the 0.01 level of significance the maximum deviation is 0.0104 which is smaller than the critical value of 0.07326305. Therefore, it can be concluded that the author productivity distribution of DM in Taiwan conforms to Lotka's law.

(2) Based on Lotka's methodology, the value of the exponent n for DM is estimated 3.36257094 and the constant c computed 0.87243801. Using the K-S test it is found that at the 0.01 level of significance the maximum deviation is 0.0050 which falls within the critical value of 0.01900753. Therefore, it can be concluded that the author productivity distribution of DM fits Lotka's law.

CONCLUSION

Although DM has already become more important recently,there are few comprehensive studies and categorization schemes to discuss the characteristics for DM. Using a bibliometric approach, the paper explores research potential of DM in Taiwan through comparing globalization DM trends, forecasts and citationsfrom 1993 to 2016by locating heading "data mining" in topic in the WoS database. The bibliometric analytical technique was used to examine the topic in WoS journals from 1993 to 2016, we found of 245 articles ofTaiwan and 3053 articles of globalization. This paper surveys and classifies DM articles between Taiwan and globalizationusing the following eight categories – publication year, citation, document type, country/territory, institute name, language, source title and research area – for different distribution status in order to find the difference and how DM technologies and applications have developed in this period. Finally, the study will analyze DM technology trends, forecasts and citations under the above results. Also, the paper performs the K-S test to check whether the distribution of author article productionof Taiwan and globalization follows Lotka's law or not.

The results in this study have several important conclusions:

(1) From the distribution of publication year, DM has more potential to grow up and becomes more popular.

(2) The article is the main tendency of document type in DM research both for Taiwan and globalization area.

(3) Clearly, English is still the major trend of language in DM research both for Taiwan and globalization area.

(4) An existing upward trend of DM is expected to continue in the future from the distribution of citation.

(5) On the basis of the countries/territories, the US, England, and Taiwan are the top three countries/territories in DM research. Besides, the PRC, Australia, Canada, Spain and Germany are also the major academic article providers in DM.

(6) Regarding the institutions, National Central University and National ChiaoTung University are the specific scholarly affiliations in DM research both for Taiwan and globalization. Analysis of the locations of these affiliations shows that the US is still the most productive country within the research aspect of DM in Taiwan and globalization. Regarding the relationship between article production and citations in DM of Taiwan, there are 25 articles from National ChiaoTung University and 16 articles from TamKang University, but their citations are 411 times and 222 times in the domain. The others almost follow the article production ranking accordingly. On the other side, there are 29 articles from University of Wisconsin, 29 articles from Pennsylvania State University and 25 articles from National ChiaoTung University in DM of globalization, but their citations are 385 times, 477 times and 411 times in the domain. The others follow the article production ranking accordingly.

(7) Judging from the research area, the most relevant disciplines in DM research category provided by computer science, business economics, information science & library science, engineering, and operations research & management science will become the most important categories for DM researchers in Taiwan and globalization. The citations of Taiwan follow article production ranking in the top 10research area, except for Social sciences& other topics. In the meantime,

we can find that the citationsof globalization follow article production ranking in the top 10research area, except for public environmental& occupational health.

(8) Based on the sources, the most enthusiastic supports for DM scholarly publishing enterprise come from *Expert Systems with Applications*, *Decision Support Systems*, *Information Processing &Management* and *Journal of Information Science* which are repeated in DM research domain and will turn into the most critical journals for DM researchers in Taiwan and globalization. The citations of Taiwan follow article production ranking in the top 10 sources, except for *Electronic Library*, *Decision Support Systems*, *Information Processing &Management*, and *Electronic Commerce Research and Applications*. On the other hand, we can find that the citations of globalization follow article production ranking in the top 10 sources, except for *Decision Support Systems*, *International Journal of Geographical Information Science*, *Information Processing &Management*, *International Journal of Geographical Information Science* (9.70 average citations per article) and *Scientometrics*.

(9) According to the K-S test, the result shows that the author productivity distribution predicted by Lotka holds for both in Taiwan and globalization of DM. This means that out of all the authors in DM, about 60 percent will have just one publication, about 15 percent will have two publications, about 7 percent of authors will have three publications, and so on.

Besides, these findings can support to recognize scientific research tendencies and realize the scale of researching development on DM domain in Taiwan by analyzing the increases of the article authors. The resources are always limited, especially for emerging and developing countries, and small and medium enterprises. According to these above information, Taiwan government and enterprises may infer collective trends and demands for scientific researchers on DM domain to organize and create proper training strategies and policies in the future.

Based on the analyzing results, this paper provides a roadmap for future researches, abstracts technology trend information and facilitates knowledge accumulation. Therefore, the researches of DM in Taiwan can follow and concentrate the globalization categories, and create the potential in the near future.

Limitation of the study

The results and conclusion are limited and not intended to be exclusive. WoS journalsadopt stringent journal reviewing criteria,the articles might take one or two years from submissionto publication. Therefore, findings in this study may not reflect the most recent researchtrends.

Research on DM has been carried out since the 1960s and 1970s, and even before thatdate.However, this study used only one search term each ("data mining") to analyze DM publications from 1993 to 2016collected in the WoS databases at that time. Future studies with greater resources, usingmore search terms, are needed to expand these findings.

REFERENCES

[1] Chen, M.S., Han, J., & Yu, P. S. (1996)'Data mining: an overview from a database perspective',IEEE Transactions on Knowledge and Data Engineering, Vol. 8, No. 6, pp. 866–883.

[2] Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996a) 'From data mining toknowledge discovery: an overview'. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy (Eds.) Advances in Knowledge Discovery and Data Mining (pp. 1–34). Cambridge, MA: AAAI Press.

[3] Ha, S., Bae, S., & Park, S. (2000) 'Web mining for distance education'. In IEEE international conference on management of innovation and technology (pp. 715–719).

[4] Weiss, S.M. & N. Indurkhya (1998) 'Predictive Data Mining: A Practical Guide', MorganKaufmann Publishers, Inc.

[5] Fayyad, U. M., Djorgovski, S. G., & Weir, N. (1996)'Automating the analysis and cataloging of sky surveys'. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy (Eds.), Advances in Knowledge Discovery and Data Mining(pp. 471–494). Cambridge, MA: MIT Press.

[6] Liao, S. H., Chu,P. H., &Hsiao, P. Y. (2012) 'Data mining techniques and applications-A decade review from 2000 to 2011', Expert Systems with Applications, Vol. 39, pp. 11303-11311.

[7] Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996b)'The KDD process forextracting useful knowledge from volumes of data', Communications of the ACM, Vol. 39, No. 11, pp. 27–34.

[8] Turban, E., Aronson, J. E., Liang, T. P., &Sharda, R. (2007)'Decision support and business intelligence systems' (8th Ed.). Taiwan: Pearson Education.

[9] Hajirezaie, M., Husseini, S. M. M., &Barfourosh, A. A., et al. (2010) 'Modeling andevaluating the strategic effects of improvement programs on the manufacturingperformance using neural networks', African Journal of Business Management, Vol. 4, No. 4, pp. 414–424.

[10] Berson, A., Smith, S., &Thearling, K. (2000) 'Building Data Mining Applications forCRM', New York: McGraw-Hill.

[11] Lejeune, M. A. P. M. (2001) 'Measuring the impact of data mining on churnmanagement', Internet Research: Electronic Networking Applications and Policy, Vol. 11, pp. 375–387.

[12] Ahmed, S. R. (2004) 'Effectiveness of neural network types for prediction of business failure', Information Technology: Coding and Computing, Vol. 2, pp. 455–459.

[13] Berry, M. J. A., &Linoff, G. S. (2004) 'Data mining techniques second edition – formarketing, sales, and customer relationship management', New York: Wiley.
[14] Bortiz, J. E., & Kennedy, D. B. (1995) 'Effectiveness of neural network types forprediction of business failure', Expert Systems with Applications, Vol. 9, pp. 503–512.
[15] Fletcher, D., & Goss, E. (1993) 'Forecasting with neural networks: An applicationusing bankruptcy data', Information and Management, Vol. 24, No. 3, pp. 159–167.
[16] Langley, P., & Simon, H. A. (1995) 'Applications of machine learning and ruleinduction', Communication of the ACM, Vol. 38, No. 11, pp. 54–64.
[17] Lau, H. C. W., Wong, C. W. Y., Hui, I. K., & Pun, K. F. (2003) 'Design andimplementation of an integrated knowledge system', Knowledge-Based Systems, Vol. 16, pp. 69–76.
[18] Salchenberger, L. M., Cinar, E. M., & Lash, N. A. (1992) 'Neural networks: A new toolfor predicting thrift failures', Decision Sciences, Vol. 23, pp. 899–916.
[19] Su, C. T., Hsu, H. H., & Tsai, C. H. (2002) 'Knowledge mining from trained neuralnetworks', Journal of Computer Information Systems, Vol. 42, pp. 61–70.
[20] Tam, K. Y., & Kiang, M. Y. (1992) 'Managerial applications of neural networks: Thecase of bank failure predictions', Management Science, Vol. 38, pp. 926–947.
[21] Zhang, G., Hu, M. Y., Patuwo, B. E., &Indro, D. C. (1999) 'Artificial neural networks inbankruptcy prediction: General framework and cross validation analysis',European Journal of Operational Research, Vol. 116, pp. 16–32.
[22] Brachman, R. J., Khabaza, T., Kloesgen, W., Piatetsky-Shapiro, G., &Simoudis, E.(1996) 'Mining business databases', Communication of the ACM, Vol. 39, No. 11, pp. 42–48.
[23] Weingart, P. (2003) 'Evaluation of research performance: the danger of numbers'. In:Bibliometric Analysis in Science and Research. Applications, Benefits andLimitations. Second Conference of the Central Library, ForschungszentrumJülich. pp. 7–19.
[24] Weingart, P. (2004) 'Impact of bibliometrics upon the science system: Inadvertentconsequences?' In H. F. Moed, W. Glanzel, & U. Schmoch (Eds.), Handbook onQuantitative Science and Technology Research. The Netherlands: KluwerAcademic Publishers.
[25] Adam, D. (2002) 'The counting house', Nature, Vol. 415, pp. 726–729.
[26] Pritchard, A. (1969) 'Statistical Bibliography or Bibliometrics', Journal ofDocumentation, Vol. 25, No. 4, pp. 348–349.
[27] Broadus, R. N. (1987) 'Toward a definition of bibliometrics',Scientometrics, Vol. 2, No. 5/6,pp. 373–379.
[28] Moed, H. F., & Van Leeuwen, TH. N. (1995) 'Improving the accuracy of the Institutefor Scientific Information's Journal Impact Factors', Journal of the AmericanSociety for Information Science, Vol. 46, pp. 461–467.
[29] Van Raan, A. F. J. (1996) 'Advanced bibliometric methods as quantitative core ofpeer review based evaluation and foresight exercises',Scientometrics, Vol. 36,pp. 397–420.
[30] Van Raan, A. F. J. (2000) 'The Pandora's box of citation analysis: Measuring scientificexcellence, the last evil?' In B. Cronin & H. B. Atkins (Eds.),The Web of Knowledge:A Festschrift in Honor of Eugene Garfield (pp. 301–319), New Jersey: ASISMonograph Series.
[31] Van Raan, A. F. J., & Van Leeuwen, TH. N. (2002) 'Assessment of the scientific basis ofinterdisciplinary, applied research. Application of bibliometric methods innutrition and food research', Research Policy, Vol. 31, pp. 611–632.
[32] Lotka, A. J. (1926) 'The frequency distribution of scientific productivity', Journal of theWashington Academy of Sciences, Vol. 16, No. 12,pp. 317–324.
[33] Potter, W. G. (1988) 'Of Making Many Books There is No End: Bibliometrics andLibraries', Journal of Academic Librarianship, Vol. 14, pp. 238a-c.
[34] Gupta, D. K. (1987) 'Lotka's law and productivity of entomological research inNigeria for the period 1900–1973',Scientometrics, Vol. 12, pp. 33–46.
[35] Nicholls, P. T. (1989) 'Bibliometric modeling processes and empirical validity ofLotka's law', Journal of American Society for Information Science, Vol. 40, No. 6,pp. 379–385.
[36] Coille, R. C. (1977) 'Lotka's frequency distribution of scientific productivity', Journalof American Society for Information Science, Vol. 28, pp. 366–370.
[37] Vlachy, J. (1978) 'Frequency distribution of scientific performance: A bibliography ofLotka's law and related phenomena',Scientometrics, Vol. 1, pp. 109–130.
[38] Rao, I. K. R. (1980) 'The distribution of scientific productivity and social change',Journal of American Society for Information Science, Vol. 31, pp. 111–122.
[39] Pao, M. L. (1985) 'Lotka's law, a testing procedure', Information Processing andManagement, Vol. 21, pp. 305–320.
[40] Pao, M. L. (1989) 'Concept of information retrieve'. Colorado: Libraries Unlimited.