Expert Systems with Applications 39 (2012) 8172–8181

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Global data mining: An empirical study of current trends, future forecasts and technology diffusions

Hsu-Hao Tsai *

Department of Management Information System, National Chengchi University, No. 64, Sec. 2, Zhinan Rd., Wenshan District, Taipei City 11605, Taiwan, ROC

ARTICLE INFO

Keywords:
Data mining
Research trends and forecasts
Technology diffusions
Bibliometric methodology

ABSTRACT

Using a bibliometric approach, this paper analyzes research trends and forecasts of data mining from 1989 to 2009 by locating heading "data mining" in topic in the SSCI database. The bibliometric analytical technique was used to examine the topic in SSCI journals from 1989 to 2009, we found 1181 articles with data mining. This paper implemented and classified data mining articles using the following eight categories—publication year, citation, country/territory, document type, institute name, language, source title and subject area—for different distribution status in order to explore the differences and how data mining technologies have developed in this period and to analyze technology tendencies and forecasts of data mining under the above results. Also, the paper performs the K-S test to check whether the analysis follows Lotka's law. Besides, the analysis also reviews the historical literatures to come out technology diffusions of data mining. The paper provides a roadmap for future research, abstracts technology trends and forecasts, and facilitates knowledge accumulation so that data mining researchers can save some time since core knowledge will be concentrated in core categories. This implies that the phenomenon "success breeds success" is more common in higher quality publications.

1. Introduction

Data mining is an interdisciplinary field that combines artificial intelligence, database management, data visualization, machine learning, mathematic algorithms, and statistics. Data mining, also known as knowledge discovery in databases (KDD) (Chen, Han, & Yu, 1996; Fayyad, Piatetsky-Shapiro, & Smyth, 1996a), is a rapidly emerging field. This technology provides different methodologies for decision-making, problem solving, analysis, planning, diagnosis, detection, integration, prevention, learning, and innovation

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. The major data mining 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 (Fayyad, Piatetsky-Shapiro, & Smyth, 1996b).

Data mining was defined by Turban, Aronson, Liang, and Sharda (2007, p.305) 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, data mining was used to investigate improvement programs, strategic priorities, environmental factors, manufacturing performance dimensions and their interactions (Hajirezaie, Husseini, Barfourosh, et al., 2010). Berson, Smith, and Thearling (2000), Lejeune (2001), Ahmed (2004) and Berry and Linoff (2004) also defined data mining as the process of extracting or detecting hidden patterns or information from large databases. With an enormous amount of customer data, data mining technology can provide business intelligence to generate new opportunities (Bortiz & Kennedy, 1995; Fletcher & Goss, 1993; Langley & Simon, 1995; Lau, Wong, Hui, & Pun, 2003; Salchenberger, Cinar, & Lash, 1992; Su, Hsu, & Tsai, 2002; Tam & Kiang, 1992; Zhang, Hu, Patuwo, & Indro, 1999).

Recently, a number of data mining applications and prototypes have been developed for a variety of domains (Brachman, Khabaza, Kloesgen, Piatetsky-Shapiro, & Simoudis, 1996) including marketing, banking, finance, manufacturing and health care. In addition, data mining has also been applied to other types of data such as time-series, spatial, telecommunications, web, and multimedia data. In general, the data mining process, and the data mining technique and function to be applied depend very much on the application domain and the nature of the data available.

* Tel.: +886 2 27929728; fax: +886 2 29393754.
E-mail addresses: simontsai@yahoo.com, 98356512@nccu.edu.tw

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2012.01.150

Using a bibliometric approach, the paper analyzes technology trends and forecasts of data mining from 1989 to 2009 by locating heading "data mining" in topic in the SSCI database. This paper surveys and classifies data mining articles using the following eight categories – publication year, citation, document type, country/territory, institute name, language, source title and subject area – for different distribution status in order to explore the difference and how technologies and applications of data mining have developed in this period and to analyze technology trends and forecasts of data mining under the above results. Besides, the analysis also reviews the historical literatures to come out technology diffusions of data mining.

The analysis provides a roadmap for future research, abstracts technology trends and forecasts, and facilitates knowledge accumulation so that data mining researchers can save some time since core knowledge will be concentrated in core categories. This implies that the phenomenon "success breeds success" is more common in higher quality publications.

2. Material and methodology

2.1. Research material

Weingart (2003, 2004) pointed at the very influential role of the monopolist citation data producer ISI (Institute for Scientific Information, now Thomson Scientific) as its commercialization of these data (Adam, 2002) 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, Philadelphia, PA, USA. In this study, we discuss the papers published in the period from 1989 to 2009 because there was no data prior to that year. The Social Sciences Citation Index is a multidisciplinary index to the journal article of the social sciences. It fully indexes over 1950 journals across 50 social sciences disciplines. It also indexes individually selected, relevant items from over 3,300 of the world's leading scientific and technical journals.

2.2. Research methodology

Pritchard (1969, p. 349) defined bibliometrics as "the application of mathematics and statistical methods to books and other media of communication." Broadus (1987, p. 376) 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 (Moed & Van Leeuwen, 1995; Van Raan, 1996; Van Raan, 2000), bibliometrics can be applied as a powerful support tool to peer review. Also for interdisciplinary research fields this is certainly possible (Van Raan & Van Leeuwen, 2002).One common way of conducting bibliometric research is to use the Social Science Citation Index (SSCI), the Science Citation Index (SCI) or the Arts and Humanities Citation Index (A&HCI) to trace citations.

There are some research using bibliometric methodology to analyze the trends and forecasts, such as e-commerce, supply chain management, data mining, CRM, and energy management. (Chen, Chen, & Lee, 2010; Tsai, 2011; Tsai & Chang, 2011; Tsai & Chi, 2011).

2.2.1. Lotka's law

Lotka's law describes the frequency of publication by authors in a given field. It states 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%" (Lotka, 1926). Lotka's law is stated by the following formula: $x^n y = c$ where y is the number of authors with x publications, the exponent n is suggested by a value of 0.6079 and the constant c is suggests by a value of 2. This means that out of all the authors in a given field, about 60% will have just one publication, about 15% will have two publications ($1/2^2$ times 0.60), about 7% 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 (Potter, 1988).

Lotka's law is generally used for understanding the productivity patterns of authors in a bibliography (Coille, 1977; Gupta, 1987; Nicholls, 1989; Pao, 1985; Rao, 1980; Vlachy, 1978). In this article, Lotka's law is chosen to perform bibliometric analysis to check the number of publications versus accumulated authors between 1989 and 2009 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.

2.2.2. Research architecture

Using a bibliometric approach, the paper analyzes technology trends and forecasts of data mining from 1989 to 2009 by locating heading "data mining" in topic in the SSCI database. The bibliometric analytical technique was used to examine the topic in SSCI journals from 1989 to 2009, we found of 1181 articles with data mining. This paper surveys and classifies data mining articles using the following eight categories – publication year, citation, document type, country/territory, institute name, language, source title and subject area – for different distribution status in order to explore the difference and how technologies and applications of data mining have developed in this period and to analyze technology trends and forecasts of data mining under the above results. Besides, the analysis also reviews the historical literatures to come out technology diffusions of data mining.

As a verification of its analysis, the paper implements the Kolmogorov-Smirnov (K-S) test by 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 $x^n y = c$ the suggested value of n is 2. The exponent n of applied field is calculated by the least square-method using the following formula (Pao, 1985):

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

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 for the slope of a regression line which is the exponent n for Lotka's law (Pao, 1985). The slope is usually calculated without data points representing authors of high productivity. Since values of the slope change with different number of points for the same set of data, we have made several computations of n. The median or the mean values of n can also be identified as the best slope for the observed

distribution (Pao, 1985). Different values of n produce different values of the constant c.

(4) Calculation the value of c

According to Lotka's law, the generalized formula is $x^n y = c$ the suggested value of c is 0.6079. The parameter c of applied field is calculated using the following formula (Pao, 1985):

$$C = \frac{1}{\left[\sum_{1}^{p-1} + \frac{1}{(n-1)(p^{n-1})} + \frac{1}{p^n} + \frac{n}{24(p-1)^{n+1}}\right]} \quad (2)$$

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

Pao (1985) 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 (Pao, 1985):

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

$\text{Fo}(x)$ is the theoretical cumulative frequency, $\text{Sn}(x)$ is the 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 (Pao, 1985):

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

$\sum y =$ 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.

The analysis provides a roadmap for future researches, abstracts technology trend information and facilitates knowledge accumulation so that data mining researchers can save some time since core knowledge will be concentrated in core categories. This implies that the phenomenon "success breeds success" is more common in higher quality publications.

3. Results

3.1. Distribution by publication year

As Fig. 1 shows, the article production on data mining has been rising since 1996. The article distribution can be divided into three segments to show the trends of development: from 1989 to 1998, from 1999 to 2003 and from 2004 to 2009. From 1989 to 1998, data mining 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 very sharp growth in 2006, and rapidly peaked in 2009.

3.2. Distribution by citation

From Fig. 1, we can see that the citation distribution of data mining is not easy to recognize between 1989 and 1999, followed by a dramatic growth and rapidly peaked in 2009. The result indicates that data mining will keep popular in the future.

3.3. Distribution by country/territory

Table 1 shows the US at the top with 551 (46.66%), following by England with 108 (9.14%) respectively. Taiwan ranks third with 104 (8.81%). Behind them, Australia, Canada, the PRC and Germany are also major academic providers in the field.

In Table 1, we can find the article distribution of the top 25 countries/territories in each year for data mining. The US leads in the field. Taiwan ranks third starting from 2000, and has risen to second by 2008, indicating its potential to increase the production in the near future.

Regarding the relationship between article production and citations, there are only ten articles from Finland in data mining, its citations, however, are 474 times in the domain (Table 1). The others almost follow the article production ranking accordingly.

3.4. Distribution by institution name

In Table 1, Noish, Penn State University and the University of Wisconsin are all no.1 author affiliation in data mining research with 17 record counts (1.44%). After analyzing the locations of

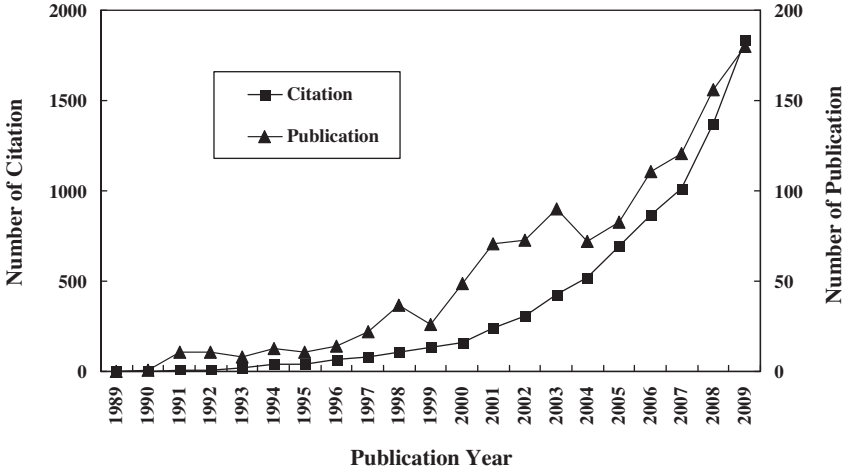

Fig. 1. Publication and citation status in each year. *Source*: SSCI database.

Table 1
Distribution of top 25 countries/territories and institutions from 1989 to 2009.

Rank	Country/territory	NP	% of 1181	Citation	Institution name	NP	% of 1181	Citation	Country
1	The US	551	46.66	4781	NIOSH	17	1.44	76	The US
2	England	108	9.14	997	Pennsylvania State University	17	1.44	202	The US
3	Taiwan	104	8.81	436	University of Wisconsin	17	1.44	122	The US
4	Canada	67	5.67	547	University of Illinois	13	1.10	125	The US
5	The P.R.C.	54	4.57	187	Columbia University	12	1.02	65	The US
6	Australia	47	3.98	350	National Central University	12	1.02	41	Taiwan
7	Germany	32	2.71	177	University of Pennsylvania	12	1.02	76	The US
8	South Korea	32	2.71	232	National Chiao Tung University	11	0.93	22	Taiwan
9	Spain	27	2.29	79	Purdue University	11	0.93	91	The US
10	Netherlands	21	1.78	135	Monash University	10	0.85	85	Australia
11	Belgium	20	1.69	96	University of Texas	10	0.85	100	The US
12	France	20	1.69	105	Duke University	9	0.76	60	The US
13	Japan	18	1.52	49	Tamkang University	9	0.76	87	Taiwan
14	Italy	17	1.44	78	University of North Carolina	9	0.76	113	The US
15	Brazil	13	1.10	33	University of Western Ontario	9	0.76	119	Canada
16	Scotland	13	1.10	45	Yale University	9	0.76	717	The US
17	South Africa	13	1.10	69	Virginia Commonwealth University	9	0.76	25	The US
18	Sweden	12	1.02	11	City University of Hong Kong	8	0.68	15	The PRC
19	Turkey	12	1.02	53	Harvard University	8	0.68	55	The US
20	India	11	0.93	30	NanYang Technology University	8	0.68	90	Singapore
21	Slovenia	11	0.93	4	National Sun Yat-Sen University	8	0.68	29	Taiwan
22	Austria	10	0.85	30	ONR	8	0.68	90	The US
23	Finland	10	0.85	474	Syracuse University	8	0.68	58	The US
24	Singapore	10	0.85	105	University of Arizona	8	0.68	62	The US
25	Wales	10	0.85	117	University of Hong Kong	8	0.68	26	The PRC

NP = number of publication.

these affiliations, the US is still the most productive country in the world in data mining research.

Regarding the relationship between article production and citations, there are only nine articles from Yale University in data mining, but it has the largest amount of citations (717 times) in the domain (Table 1). The others almost follow the article production ranking accordingly.

3.5. Distribution by document type

In Table 2, the distribution of document types from 1989 to 2009 indicates that the most popular publication document type is "Article" (936 articles, 79.25%). The result demonstrates that the article is the major tendency of document type in data mining research.

3.6. Distribution by language

In Table 2, the majority language for data mining is English with 1149 articles (97.29%). Clearly, English is still the main trend in data mining research.

3.7. Distribution by subject area

Table 3 offers critical information for future research tendencies in data mining, allowing researchers a better understanding of the distribution of the top 25 subjects in future research. The top three subjects for data mining research are information science & library science (260 articles, 22.01%), followed by computer science & information system (251 articles, 21.25%) and operations research & management science (168 articles, 14.23%). Besides, this paper's analysis suggests that there are other important research disciplines for data mining article production such as management, computer science & artificial intelligence, economics, computer science & interdisciplinary applications, public environmental & occupational health and engineering, electrical & electronic.

As Table 3 illustrates, data mining citations follow article production ranking in the top 25 subjects, except for statistics & probability (57.48 average citations per article), social sciences & mathematical methods (32.09 average citations per article), economics (12.26 average citations per article), computer science & artificial intelligence (10.79 average citations per article), engineering, electrical & electronic (9.05average citations per article) and computer science & information systems (7.73 average citations per article).

Table 2
Distribution of document type and language from 1989 to 2009.

Document type	NP	% of 1181	Language	NP	% of 1181
Article	936	79.25	English	1149	97.29
Proceedings paper	106	8.98	Spanish	12	1.02
Book review	50	4.23	German	5	0.42
Review	41	3.472	Slovak	4	0.34
Meeting abstract	23	1.95	Japanese	3	0.25
Editorial material	19	1.61	Czech	2	0.17
News item	2	0.17	French	2	0.17
Correction	1	0.08	Portuguese	2	0.17
Note	1	0.08	Russian	1	0.08
Reprint	1	0.08	Slovene	1	0.08
Software review	1	0.08	Total	1181	100
Total	1181	100			

NP = number of publication.

3.8. Distribution by source title

Table 3 highlights information on trends for data mining, allowing researchers to closely approach the distribution of the top 25 sources in future research. The top three research journals of data mining are *Expert Systems with Applications* (69 articles, 5.84%), followed by *Journal of the American Medical Informatics Association* (35 articles, 2.96%) and *Journal of Operation Research Society* (26 articles, 2.20%). In addition, there are a significant number of research sources for data mining article production such as *Journal of the American Society for Information and Technology*, *Information Processing & Management*, *International Journal of Geographical Information Science*, *Journal of Information Science*, *Online Information Review*, *Information & Management,* and *Decision Support Systems*.

Table 3
Distribution of top 25 subjects and sources from 1989 to 2009.

Rank	Subject area	NP	% of 1181	Citation	Source title	NP	% of 1181	Citation
1	Information Science & Library Science	260	22.02	1508	*Expert Systems with Applications*	69	5.84	447
2	Computer Science, Information Systems	251	21.25	1941	*Journal of the American Medical Informatics Association*	35	2.96	147
3	Operations Research & Management Science	168	14.23	1096	*Journal of the Operational Research Society*	26	2.20	44
4	Management	149	12.62	864	*Journal of the American Society for Information Science and Technology*	22	1.86	164
5	Computer Science, Artificial Intelligence	132	11.18	1424	*Information Processing & Management*	21	1.78	142
6	Economics	112	9.48	1373	*International Journal of Geographical Information Science*	20	1.69	194
7	Computer Science, Interdisciplinary Applications	103	8.72	713	*Journal of Information Science*	19	1.61	114
8	Public, Environmental & Occupational Health	85	7.20	588	*Online Information Review*	17	1.44	12
9	Engineering, Electrical & Electronic	82	6.94	742	*Information & Management*	16	1.35	236
10	Environmental Studies	68	5.76	367	*Decision Support Systems*	15	1.27	46
11	Business	56	4.74	350	*Resources Policy*	15	1.27	300
12	Geography	52	4.40	348	*Computers & Education*	11	0.93	52
13	Medical Informatics	49	4.15	239	*Journal of the American Society for Information Science*	11	0.93	137
14	Environmental Sciences	38	3.22	378	*International Journal of Forecasting*	10	0.85	47
15	Social Sciences, Mathematical Methods	35	2.96	1123	*Journal of Safety Research*	9	0.76	26
16	Ergonomics	34	2.88	146	*Safety Science*	9	0.76	34
17	Engineering, Industrial	33	2.79	147	*Scientometrics*	9	0.76	81
18	Planning & Development	31	2.62	201	*Society & Natural Resources*	8	0.68	38
19	Education & Educational Research	30	2.54	97	*Technological Forecasting and Social Change*	8	0.68	63
20	Social Sciences, Interdisciplinary	30	2.54	92	*American Journal of Industrial Medicine*	7	0.59	56
21	Sociology	30	2.54	197	*Educational Technology & Society*	7	0.59	15
22	Mathematics, Interdisciplinary Applications	26	2.20	221	*Electronic Library*	7	0.59	15
23	Geography, Physical	24	2.03	212	*Journal of Biomedical Informatics*	7	0.59	56
24	Computer Science, Cybernetics	23	1.95	114	*Social Work in Health Care*	7	0.59	15
25	Statistics & Probability	21	1.78	1207	*European Journal of Operational Research*	6	0.51	14

NP = number of publication.

Table 4
Calculation of author productivity of data mining.

NP	Author (s)	(NP) * (Author)	Accumulated record	Accumulated record (%)	Accumulated author(s)	Accumulated 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

NP = number of publication.

Table 5
Calculation of the exponent n for data mining.

x (NP)	y (Author)	$X = \log(x)$	$Y = \log(y)$	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

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

In Table 3, data mining citations follow article production ranking in the top 25 sources, except for *Decision Support Systems* (20.00 average citations per article), *Information & Management* (14.75 average citations per article), *Journal of the American Society for Information Science* (12.45 average citations per article), *International Journal of Geographical Information Science* (9.70 average citations per article) and *Scientometrics* (9.00 average citations per article).

4. Discussion

The section implements the steps which are demonstrated in Section 2.2.2 to verify whether the distribution of author article production follows Lotka's law in data mining research.

4.1. The literatures productivity analysis by Lotka's law

(1) Collect data and
(2) List author & article distribution table

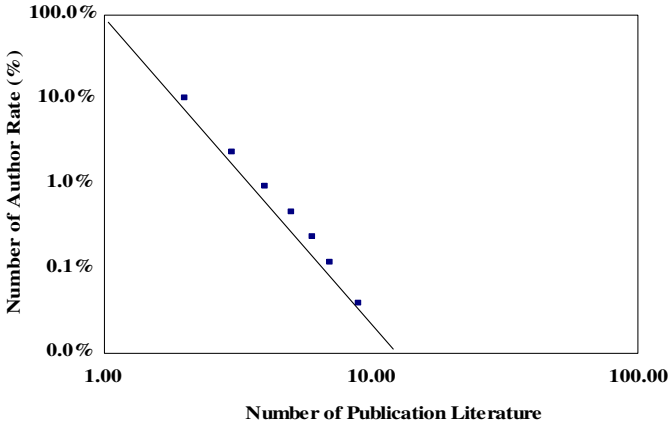

Fig. 2. Distribution of literature productivity of author on data mining research aspect.

Author quantity is calculated by the equality method from 1181 articles retrieved by the SSCI index. Altogether, 2519 authors on data mining are included. See Table 4 for reference.

(3) Calculation the value of n (slope)

In Table 5, we list the number of authors and the number of publications by one author for calculation of the exponent n with topic as "data mining" in SSCI database. The results of the calculations in Table 5 can be brought into Eq. (1) to calculate the value of n:

$$n = \frac{9(3.26) - (5.56)(9.87)}{9(4.22) - (5.56)^2} \quad (5)$$

Table 8
The overview of organization adoptions in data mining.

Adoption	Authors
Database marketing	Forcht and Cochran (1999)
Interface	Lavington et al. (1999)
Semantic indexing	Jiang et al. (1999)
Cancer information system	Houston et al. (1999)
Customer retention and insurance claim patterns	Smith et al. (2000)
Data quality	Feelders et al. (2000); Hand (2000)
Customer service support	Hui and Jha (2000)
Electroencephalography application	Flexer (2000)
Prediction of corporate failure	Lin and McClean (2001)
Network intrusion detection	Zhu et al. (2001)
Knowledge refinement	Park et al. (2001)
Software integration	Chua et al. (2002)
Credit card portfolio management	Shi et al. (2001)
Knowledge warehouse	Nemati et al. (2002)
Grid services	Cannataro et al. (2002)
Library material acquisition budget allocation	Wu (2003)
Selection of insurance sales agents	Cho and Ngai (2003)
Prediction of physical performance	Fielitz and Scott (2003)
Library decision making	Nicholson (2003)

Then we can find n = -3.629488955

(4) Calculation the value of c

The value of c is calculated by using Eq. (2), where $P = 20$, $x = 1, 2, 3, 4, 5, 6, 7, 8$ and $n = 3.629488955$, then we can find c = 0.892795157.

With $n = -3.629488955$ and $c = 0.892795157$, the Lotka's law equation of data mining is:

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

When the result is compared to Table 4, we can see that authors with only one article account for 89.40% (100 − 10.60%=89.40%), which almost matches the primitive c value 89.28% generated by

Table 6
The K-S test for data mining.

NP	Author(s)	Data mining (Observed)	Sn(x)	Data mining (Expected)	Fo(x)	D
1	2252	0.8940	0.8940	0.8928	0.8928	0.0012
2	206	0.0818	0.9758	0.0721	0.9649	0.0109
3	37	0.0147	0.9905	0.0166	0.9815	0.0090
4	12	0.0048	0.9952	0.0058	0.9873	0.0079
5	6	0.0024	0.9976	0.0026	0.9899	0.0077
6	3	0.0012	0.9988	0.0013	0.9913	0.0076
7	2	0.0008	0.9996	0.0008	0.9920	0.0076
8	0	0.0000	0.9996	0.0005	0.9925	0.0071
9	1	0.0004	1.0000	0.0003	0.9928	0.0072

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

Table 7
The overview of technology innovations in data mining.

Innovation	Authors
Correcting results	Markowitz et al. (1994)
Model uncertainty	Chatfield (1995)
Discovery	Trybula (1997)
Asymptotic complexity	McSherry (1997)
Scientific computing	Kral (1997)
DNA intragenic mutation	Evans et al. (1997)
Introduction	Raghavan et al. (1998)
Hospital infection control and public health surveillance	Brossette et al. (1998)
User-guided query construction	Chen and Zhu (1998)
Transforming corporate information into value	Cheng and Chang (1998)
Organizational learning	Dhar (1998)
Natural language understanding	Wilcox and Hripcsak (1998)

Table 9
The overview of market diffusions in data mining.

Applications	Authors
Credit scoring	Liu and Schumann (2005), Huang et al. (2007) and Ince and Aktan (2009)
Finance	Zhang and Zhou (2004)
Customer behaviors/ service	Liu and Shih (2005), Wang and Hong (2006), Cheng et al. (2005), Casillas and Martínez-López (2009), Ngai et al. (2009) and Hayashi et al. (2009)
Healthcare management	Madigan and Curet (2006), Cheng et al. (2006), Ceglowski et al. (2007) and Glowacka et al. (2009)
Prediction of failure	Lin et al. (2009)
Sales/marketing	Imms (2004), Prinzie and Van den Poel (2005), Chen et al. (2006) and Liao et al. (2009)
Hypermedia	Lee (2007)

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 Pao (1989), 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 (=3.629488955) is between 1.2 and 3.8 and is matched the reference data by observation. The detail distribution chart is shown in Fig. 2.

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

We use Eq. (3) to evaluate whether the analysis matches Lotka's law. From Table 6, we can find D ($D = \text{Max}|\text{Fo}(x) - \text{Sn}(x)|$) = 0.0109. According to the K-S test, the critical value at 0.01 level of significance is calculated by using Eq. (4):

$$1.63/\sqrt{2519} = 0.032477 \tag{7}$$

4.2. Discussion

(1) Based on Lotka's methodology, the value of the exponent n for data mining is estimated 3.629488955 and the constant c computed 0. 892795157. Using the K-S test it is found that at the 0.01 level of significance the maximum deviation is 0.0109 which falls within the critical value of 0.032477. Therefore, it can be concluded that the author productivity distribution of data mining fits Lotka's law (Potter, 1981).
(2) The reason why data mining does fit Lotka's law is that the rate of authors who publish only one article is close to constant c; as a result, the difference between the observed value and the expected value becomes smaller than the K-S test critical value. This outcome causes the data mining distribution to fit the slope of Lotka's law.

5. Technology diffusion review and discussion

The section will perform the analysis of technical innovation, adopting organizations and industry diffusions for data mining and reveal the existence of three eras: from 1989 to 1998, from 1999 to 2003, and from 2004 to 2009.

5.1. Technology innovations of data mining (1989–1998)

From a retrospective view of data mining technological innovation, we found that data mining technology was improved and introduced itself before its adoptions and diffusions. These included correcting results due to "luck", data mining introduction and data mining discovery. The significant events during data mining development were to review the effects of model uncertainty, asymptotic complexity, interactive with scientific computing, defining valid or "true" patterns for a specific DNA intragenic mutation, for hospital infection control and public health surveillance, user-guided query construction, transforming corporate information into value, organizational learning and assisting natural language understanding (Chatfield, 1995; Markowitz & XU, 1994; Trybula, 1997; Brossette, Sprague, Hardin, et al., 1998; Chen & Zhu, 1998; Cheng & Chang, 1998; Evans, Lemon, Deters, et al., 1997; Kral, 1997; McSherry, 1997; Raghavan, Deogun, & Sever, 1998; Dhar, 1998; Wilcox & Hripcsak, 1998). These innovation activities are categorized in Table 7.

After this series of innovations, data mining started to grow up and setting the stage for adoption of the technology.

5.2. Organization adoptions of data mining (1999–2003)

After the activities of technological innovation during the period of 1989–1998, a variety of data mining applications were used by different sections, such as database marketing, interface, semantic indexing, analysis of customer retention and insurance claim patterns, data quality, research on a cancer information system, customer service support, material acquisition budget allocation for libraries, electroencephalography application, the prediction of corporate failure, network intrusion detection, knowledge refinement, software integration, credit card portfolio management, knowledge warehouse, grid services, selection of insurance sales agents, Prediction of physical performance and library decision making (Cannataro, Talia, & Trunfio, 2002; Cho & Ngai, 2003; Chua, Chiang, & Lim, 2002; Feelders, Daniels, & Holsheimer, 2000; Fielitz & Scott, 2003; Flexer, 2000; Forcht & Cochran, 1999; Hand, 2000; Houston, Chen, Hubbard, et al., 1999; Hui & Jha, 2000; Jiang, Berry, Donato, Ostrouchov, & Grady, 1999; Lavington, Dewhurst, Wilkins, & Freitas, 1999; Lin & McClean, 2001; Nemati, Steiger, Iyer, & Herschel, 2002; Nicholson, 2003; Park, Piramuthu, & Shaw, 2001; Shi, Wise, Luo, & Lin, 2001; Smith, Wills, & Brooks, 2000; Wu, 2003; Zhu, Premkumar, Zhang, et al., 2001). These adoption activities are summarized in Table 8.

While many enterprises have already applied data mining technology, the technology and its applications must continue to be improved. Data mining faces new challenges as various enterprises prepare for the rapid and large-scale adoptions and diffusions of this technology.

5.3. Market diffusions of data mining (2004–2009)

This section discusses data mining technological diffusions underwent its technical evolution based on the time frame from 2004 to 2009, these diffusions include credit scoring, finance, customer behaviors/services, healthcare management, prediction of failure, sales/marketing and hypermedia (Casillas & Martínez-López, 2009; Ceglowski, Churilov, & Wasserthiel, 2007; Chen, Chen, & Tung, 2006; Cheng, Chang, & Liu, 2005; Cheng, Luo, & Chen, 2006; Glowacka, Henry, & May, 2009; Hayashi, Hsieh, & Setiono, 2009; Huang, Chen, & Wang, 2007; Imms, 2004; Ince & Aktan, 2009; Lee, 2007; Liao, Chen, & Hsu, 2009; Lin, Shiue, Chen, et al., 2009; Liu & Schumann, 2005; Liu & Shih, 2005; Madigan & Curet,

2006; Ngai, Xiu, & Chau, 2009; Prinzie & Van den Poel, 2005; Wang & Hong, 2006; Zhang & Zhou, 2004). The details of diffusions activities are presented in Table 9.

From the diffusion activities, we can easily demonstrate that credit scoring, customer behaviors & service, healthcare management and sales and marketing are the major diffusions of data mining application.

6. Conclusions

Using a bibliometric approach, the paper analyzes technology trends and forecasts of data mining from 1989 to 2009 by locating heading "data mining" in topic in the SSCI database. The bibliometric analytical technique was used to examine the topic in SSCI journals from 1989 to 2009, we found of 1181 articles with data mining. This paper surveys and classifies data mining articles using the following eight categories – publication year, citation, document type, country/territory, institute name, language, source title and subject area – for different distribution status in order to explore the difference and how technologies and applications of data mining have developed in this period and to analyze technology trends and forecasts of data mining under the above results. Also, the paper performs the K-S test to check whether the analysis follows Lotka's law.

The results in this paper have several important implications:

(1) Based on the distribution of publication year, data mining has the potential to grow up and becomes more popular in the future.
(2) An existing upward trend of data mining is expected to continue in the future from the distribution of citation.
(3) On the basis of the countries/territories, the US, England and Taiwan are the top three countries/territories and the sum of the research output reaches 64.61% of the total publication. Australia, Canada, the P.R.C. and Germany also become the major academic work providers in the field of data mining research. Regarding to the relationship between article production and citation, there are only ten articles from Finland in data mining, however, its citations are 474 times in the domain. The others almost follow the article production ranking accordingly.
(4) Regarding the institutions, Noish, Pennsylvania State University and the University of Wisconsin are the specific scholarly affiliation in data mining research. After analyzing the locations of these affiliations, the U.S. is still the most productive country within the research aspect in the world as well. Regarding to the relationship between article production and citation, there are only nine articles from Yale University in data mining, their citations, however, are the largest amount in the domain. The others almost follow the article production ranking accordingly.
(5) The article is the main trend of document type in data mining research.
(6) English is still the major tendency of language in data mining research.
(7) Judging from the subjects, the most relevant disciplines for subject category of data mining provided by information science & library science, computer science & information system, operations research & management science, management, computer science & artificial intelligence, economics, computer science & interdisciplinary applications, public, environmental & occupational health, engineering, electrical & electronic and environmental studies and will become the most important categories for data mining researchers. The citation of data mining follows the article production ranking except statistics & probability, social sciences & mathematical methods, economics, computer science & artificial intelligence, engineering, electrical & electronic and computer science & information systems.
(8) Based on the sources, the most enthusiastic supports for scholarly publishing enterprises of data mining come from *Expert Systems with Applications*, *Journal of the American Medical Informatics Association*, *Journal of the Operational Research Society*, *Journal of the American Society for Information and Technology*, *Information Processing & Management*, *International Journal of Geographical Information Science*, *Journal of Information Science*, *Online Information Review*, *Information & Management*, *Decision Support Systems* and *Resources Policy* and will turn into the most critical journals for data mining researchers. The citation of data mining follows the article production ranking except for *Decision Support Systems*, *Information & Management*, *Journal of the American Society for Information Science*, *International Journal of Geographical Information Science* and *Scientometrics*.
(9) According to the K-S test, the result shows that the author productivity distribution predicted by Lotka holds for data mining. The reason why data mining does fit Lotka's law is the rate of authors who published one article is close to constant *c*. The result causes that the difference between observed value and expected value becomes smaller than the K-S test critical value. The outcome causes the data mining distribution to fit the slope of Lotka's law.

The research findings can be extended to investigate author productivity by analyzing variables such as chronological and academic age, number and frequency of previous publications, access to research grants, job status, etc. In such a way characteristics of high, medium and low publishing activity of authors can be identified.

Besides, the research findings can also support governments and enterprises to judge scientific research trends and forecasts of data mining, and to understand the scale of development of research in data mining through analyzing the increases of the article author. The resources are limited, especially for emerging and developing countries, and small and medium enterprises. Based on the above information, governments and enterprises may infer collective tendencies and demands for scientific researcher in data mining to facilitate the decision of appropriate training strategies and policies in the future.

The analysis provides a roadmap for future research, abstracts technology trends and forecasts, and facilitates knowledge accumulation so that data mining researchers can save some time since core knowledge will be concentrated in core categories. This implies that the phenomenon "success breeds success" is more common in higher quality publications.

References

Adam, D. (2002). The counting house. *Nature, 415*, 726–729.
Ahmed, S. R. (2004). Effectiveness of neural network types for prediction of business failure. *Information Technology: Coding and Computing, 2*, 455–459.
Berry, M. J. A., & Linoff, G. S. (2004). *Data mining techniques second edition – for marketing, sales, and customer relationship management*. New York: Wiley.
Berson, A., Smith, S., & Thearling, K. (2000). *Building Data Mining Applications for CRM*. New York: McGraw-Hill.
Bortiz, J. E., & Kennedy, D. B. (1995). Effectiveness of neural network types for prediction of business failure. *Expert Systems with Applications, 9*, 503–512.
Brachman, R. J., Khabaza, T., Kloesgen, W., Piatetsky-Shapiro, G., & Simoudis, E. (1996). Mining business databases. *Communication of the ACM, 39*(11), 42–48.
Broadus, R. N. (1987). Toward a definition of bibliometrics. *Scientometrics, 12*(5/6), 373–379.
Brossette, S. E., Sprague, A. P., Hardin, J. M., et al. (1998). Association rules and data mining in hospital infection control and public health surveillance. *Journal of American Medical Informatics Association, 5*(4), 373–381.

Cannataro, M., Talia, D., & Trunfio, P. (2002). Distributed data mining on the grid. *Future Generation Computer Systems, 18*, 1101–1112.
Casillas, J., & Martínez-López, F. J. (2009). Mining uncertain data with multiobjective genetic fuzzy systems to be applied in consumer behaviour modelling. *Expert Systems with Applications, 36*(2), 1645–1659.
Ceglowski, R., Churilov, L., & Wasserthiel, J. (2007). Combining data mining and discrete event simulation for a value-added view of a hospital emergency department. *Journal of the Operational Research Society, 58*(2), 246–254.
Chatfield, C. (1995). Model Uncertainty, Data Mining and Statistical-Inference. Journal of the Royal Statistical Society Series A-Statistics in Society, 158, 419–466, Part: Part 3.
Chen, Y. H., Chen, C. Y., & Lee, S. C. (2010). Technology forecasting of new clean energy: The example of hydrogen energy and fuel cell. *African Journal of Business Management, 4*(7), 1372–1380.
Chen, Y. L., Chen, J. M., & Tung, C. W. (2006). A data mining approach for retail knowledge discovery with consideration of the effect of shelf-space adjacency on sales. *Decision Support Systems, 42*(3), 1503–1520.
Chen, M. S., Han, J., & Yu, P. S. (1996). Data mining: an overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering, 8*(6), 866–883.
Chen, Z., & Zhu, Q. (1998). Query construction for user-guided knowledge discovery in database. *Journal of Information Sciences, 109*, 49–64.
Cheng, P. S., & Chang, P. (1998). Transforming corporate information into value through data warehousing and data mining. *ASLIB Proceedings, 50*(5), 109–113.
Cheng, B. W., Chang, C. L., & Liu, I. S. (2005). Enhancing care services quality of nursing homes using data mining. *Total Quality Management & Business Excellence, 16*(5), 575–596.
Cheng, B. W., Luo, C. M., & Chen, K. H. (2006). Using data mining to evaluate patient-oriented medical services for chronic senility outpatients. *Quality & Quantity, 40*(6), 1079–1087.
Cho, V., & Ngai, E. W. T. (2003). Data mining for selection of insurance sales agents. *Expert Systems with Applications, 20*(3), 123–132.
Chua, C. E. H., Chiang, R. H. L., & Lim, E. P. (2002). An intelligent middleware for linear correlation discovery. *Decision Support Systems, 32*, 313–326.
Coille, R. C. (1977). Lotka's frequency distribution of scientific productivity. *Journal of American Society for Information Science, 28*, 366–370.
Dhar, V. (1998). Data mining in finance: Using counterfactuals to generate knowledge from organizational information systems. *Information Systems, 23*(7), 423–437.
Evans, S., Lemon, S. J., Deters, C., et al. (1997). Using data mining to characterize DNA mutations by patient clinical features. Journal of American Medical Informatics Association, 253–257. Supplement, Suppl. S.
Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996a). From data mining to knowledge 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.
Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996b). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM, 39*(11), 27–34.
Feelders, A., Daniels, H., & Holsheimer, M. (2000). Methodological and practical aspects of data mining. *Information and Management, 37*, 271–281.
Fielitz, L., & Scott, D. (2003). Prediction of physical performance using data mining. *Research Quarterly for Exercise and Sport, 74*(1). A25-A25.
Fletcher, D., & Goss, E. (1993). Forecasting with neural networks: An application using bankruptcy data. *Information and Management, 24*(3), 159–167.
Flexer, A. (2000). Data mining and electroencephalography. *Statistical Methods in Medical Research, 9*(4), 395–413.
Forcht, K. A., & Cochran, K. (1999). Using data mining and datawarehousing techniques. *Industrial Mamagement & Data Systems, 99*(5–6), 189–196.
Glowacka, K. J., Henry, R. M., & May, J. H. (2009). A hybrid data mining/simulation approach for modelling outpatient no-shows in clinic scheduling. *Journal of the Operational Research Society, 60*(8), 1056–1068.
Gupta, D. K. (1987). Lotka's law and productivity of entomological research in Nigeria for the period 1900–1973. *Scientometrics, 12*, 33–46.
Hajirezaie, M., Husseini, S. M. M., Barfourosh, A. A., et al. (2010). Modeling and evaluating the strategic effects of improvement programs on the manufacturing performance using neural networks. *African Journal of Business Management, 4*(4), 414–424.
Hand, D. J. (2000). Data mining – New challenges for statisticians. *Social Science Computer Review, 18*(4), 442–449.
Hayashi, Y., Hsieh, M. H., & Setiono, R. (2009). Predicting consumer preference for fast-food franchises: a data mining approach. *Journal of the Operational Research Society, 60*(9), 1221–1229.
Houston, A. L., Chen, H. C., Hubbard, S. M., et al. (1999). Medical data mining on the internet: Research on a cancer information system. *Artificial Intelligence Review, 13*(5–6), 437–466.
Huang, C. L., Chen, M. C., & Wang, C. J. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications, 33*(4), 847–856.
Hui, S. C., & Jha, G. (2000). Data mining for customer service support. *Information and Management, 38*, 1–13.
Imms, M. (2004). Optimal database marketing – Strategy, development and data mining. *International Journal of Market Research, 46*(2), 259–261.
Ince, H., & Aktan, B. (2009). A comparison of data mining techniques for credit scoring in banking: A managerial perspective. *Journal of Business Economics and Management, 10*(3), 233–240.
Jiang, J., Berry, M. W., Donato, J. M., Ostrouchov, G., & Grady, N. W. (1999). Mining consumer product data via latent semantic indexing. *Intelligent Data Analysis, 3*, 377–398.
Kral, E. R. (1997). IBM research in interactive data mining and scientific computing. *Behavior Research Methods Instruments & Computers, 29*(1), 119–121.
Langley, P., & Simon, H. A. (1995). Applications of machine learning and rule induction. *Communication of the ACM, 38*(11), 54–64.
Lau, H. C. W., Wong, C. W. Y., Hui, I. K., & Pun, K. F. (2003). Design and implementation of an integrated knowledge system. *Knowledge-Based Systems, 16*, 69–76.
Lavington, S., Dewhurst, N., Wilkins, E., & Freitas, A. (1999). Interfacing knowledge discovery algorithms to large database management systems. *Information and Software Technology, 41*, 605–617.
Lee, C. S. (2007). Diagnostic, predictive and compositional modeling with data mining in integrated learning environments. *Computers & Education, 49*(3), 562–580.
Lejeune, M. A. P. M. (2001). Measuring the impact of data mining on churn management. *Internet Research: Electronic Networking Applications and Policy, 11*, 375–387.
Liao, S. H., Chen, J. L., & Hsu, T. Y. (2009). Ontology-based data mining approach implemented for sport marketing. *Expert Systems with Applications, 36*(8), 11045–11056.
Lin, F. Y., & McClean, S. (2001). A data mining approach to the prediction of corporate failure. *Knowledge-Based Systems, 14*, 189–195.
Lin, S. W., Shiue, Y. R., Chen, S. C., et al. (2009). Applying enhanced data mining approaches in predicting bank performance. A case of Taiwanese commercial banks. *Expert Systems with Applications, 36*(9), 11543–11551.
Liu, Y., & Schumann, M. (2005). Data mining feature selection for credit scoring models. *Journal of the Operational Research Society, 56*(9), 1099–1108.
Liu, D. R., & Shih, Y. Y. (2005). Integrating AHP and data mining for product recommendation based on customer lifetime value. *Information & Management, 42*(3), 387–400.
Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences, 16*(12), 317–324.
Madigan, E. A., & Curet, O. L. (2006). A data mining approach in home healthcare: outcomes and service use. *BMC Health Services Research, 6*(18), 1–10.
Markowitz, H. M., & XU, G. L. (1994). Data mining corrections. *Journal of Portfolio Management, 21*(1), 60–69.
McSherry, D. (1997). Knowledge discovery by inspection. *Decision Support Systems, 21*, 43–47.
Moed, H. F., & Van Leeuwen, TH. N. (1995). Improving the accuracy of the Institute for Scientific Information's Journal Impact Factors. *Journal of the American Society for Information Science, 46*, 461–467.
Nemati, H. R., Steiger, D. M., Iyer, L. S., & Herschel, R. T. (2002). Knowledge warehouse: an architectural integration of knowledge management, decision support, artificial intelligence and data warehousing. *Decision Support Systems, 33*, 143–161.
Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications, 36*(2), 2592–2602. Part 2.
Nicholls, P. T. (1989). Bibliometric modeling processes and empirical validity of Lotka's law. *Journal of American Society for Information Science, 40*(6), 379–385.
Nicholson, S. (2003). Bibliomining for automated collection development in a digital library setting: Using data mining to discover web-based scholarly research works. *Journal of American Society for Information Science and Technology, 54*(12), 1081–1090.
Pao, M. L. (1985). Lotka's law, a testing procedure. *Information Processing and Management, 21*, 305–320.
Pao, M. L. (1989). *Concept of information retrieve*. Colorado: Libraries Unlimited.
Park, S. C., Piramuthu, S., & Shaw, M. J. (2001). Ynamic rule refinement in knowledge-based data mining systems. *Decision Support Systems, 31*, 205–222.
Potter, W. G. (1981). Lotka's law revisited. *Library Trends, 30*(1), 21–39.
Potter, W. G. (1988). 'Of Making Many Books There is No End': Bibliometrics and Libraries. *Journal of Academic Librarianship, 14*, 238a-c.
Prinzie, A., & Van den Poel, D. (2005). Constrained optimization of data-mining problems to improve model performance. A direct-marketing application. *Expert Systems with Applications, 29*(3), 630–640.
Pritchard, A. (1969). Statistical Bibliography or Bibliometrics. *Journal of Documentation, 25*(4), 348–349.
Raghavan, V. V., Deogun, J. S., & Sever, H. (1998). Special topic issue: Knowledge Discovery and Data Mining – Introduction. *Journal of American Society for Information Science, 49*(5), 397–402.
Rao, I. K. R. (1980). The distribution of scientific productivity and social change. *Journal of American Society for Information Science, 31*, 111–122.
Salchenberger, L. M., Cinar, E. M., & Lash, N. A. (1992). Neural networks: A new tool for predicting thrift failures. *Decision Sciences, 23*, 899–916.
Shi, Y., Wise, M., Luo, M., & Lin, Y. (2001). Data mining in credit card portfolio management: A multiple criteria decision making approach. *Multiple Criteria Decision Making in the New Millennium, Book Series: Lecture Notes in Economics and Mathematical Systems, 507*, 427–436.
Smith, K. A., Wills, R. J., & Brooks, M. (2000). An analysis of customer retention and insurance claim patterns using data mining: A case study. *Journal of the Operational Research Society, 51*, 532–541.
Su, C. T., Hsu, H. H., & Tsai, C. H. (2002). Knowledge mining from trained neural networks. *Journal of Computer Information Systems, 42*, 61–70.

Tam, K. Y., & Kiang, M. Y. (1992). Managerial applications of neural networks: The case of bank failure predictions. *Management Science, 38*, 926–947.
Trybula, W. J. (1997). Data mining and knowledge discovery. *Annual Review of Information Science and Technology, 32*, 197–229.
Tsai, H. H. (2011). Research trends analysis by comparing data mining and customer relationship management through bibliometric methodology. *Scientometrics, 87*(3), 425–450.
Tsai, H. H., & Chang, J. K. (2011). E-Commerce research trend forecasting: A study of bibliometric methodology. *International Journal of Digital Content Technology and its Application, 5*(1), 101–111.
Tsai, H. H., & Chi, Y. P. (2011). Trend analysis of supply chain management by bibliometric methodology. *International Journal of Digital Content Technology and its Application, 5*(1), 285–295.
Turban, E., Aronson, J. E., Liang, T. P., & Sharda, R. (2007). *Decision support and business intelligence systems* (8th ed.). Taiwan: Pearson Education.
Van Raan, A. F. J. (1996). Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. *Scientometrics, 36*, 397–420.
Van Raan, A. F. J. (2000). The Pandora's box of citation analysis: Measuring scientific excellence, 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: ASIS Monograph Series.
Van Raan, A. F. J., & Van Leeuwen, TH. N. (2002). Assessment of the scientific basis of interdisciplinary, applied research. Application of bibliometric methods in nutrition and food research. *Research Policy, 31*, 611–632.
Vlachy, J. (1978). Frequency distribution of scientific performance. A bibliography of Lotka's law and related phenomena. *Scientometrics, 1*, 109–130.
Wang, H. F., & Hong, W. K. (2006). Managing customer profitability in a competitive market by continuous data mining. *Industrial Marketing Management, 35*(6), 715–723.
Weingart, P. (2003). Evaluation of research performance: the danger of numbers. In: Bibliometric Analysis in Science and Research. Applications, Benefits and Limitations. Second Conference of the Central Library, Forschungszentrum Jülich. pp. 7–19.
Weingart, P. (2004). Impact of bibliometrics upon the science system: Inadvertent consequences? In H. F. Moed, W. Glanzel, & U. Schmoch (Eds.), *Handbook on Quantitative Science and Technology Research*. The Netherlands: Kluwer Academic Publishers.
Wilcox, A., & Hripcsak, G. (1998). Knowledge discovery and data mining to assist natural language understanding. Journal of American Medical Informatics Association, 835–839, Supplement, Suppl. S.
Wu, C. H. (2003). Data mining applied to material acquisition budget allocation for libraries: design and development. *Expert Systems with Applications, 25*(3), 401–411.
Zhang, G., Hu, M. Y., Patuwo, B. E., & Indro, D. C. (1999). Artificial neural networks in bankruptcy prediction: General framework and cross validation analysis. *European Journal of Operational Research, 116*, 16–32.
Zhang, D. S., & Zhou, L. (2004). Discovering golden nuggets: Data mining in financial application. *IEEE Transactions on Systems MAN and Cybernetics Part C – Applications and Reviews, 34*(4), 513–522.
Zhu, D., Premkumar, G., Zhang, X. N., et al. (2001). Data mining for network intrusion detection: A comparison of alternative methods. *Decision Sciences, 32*(4), 635–660.