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