The Journal of Academic Librarianship 41 (2015) 499–510

Contents lists available at ScienceDirect

The Journal of Academic Librarianship

ELSEVIER

Literature Review of Data Mining Applications in Academic Libraries

Lorena Siguenza-Guzman [a,b,*], Victor Saquicela [a], Elina Avila-Ordóñez [a,b], Joos Vandewalle [c], Dirk Cattrysse [b]

[a] *Department of Computer Science, University of Cuenca, 12 de Abril Av., ECU-010150 Cuenca, Ecuador*
[b] *Centre for Industrial Management Traffic & Infrastructure, KU Leuven, Celestijnenlaan 300, Box 2422, BE-3001 Leuven, Belgium*
[c] *Department of Electrical Engineering ESAT/Stadius, KU Leuven, Kasteelpark Arenberg 10, Box 2440, BE-3001 Leuven, Belgium*

ARTICLE INFO

Article history:
Received 28 January 2015
Accepted 6 June 2015
Available online 25 June 2015

Keywords:
Data mining
Bibliomining
Literature review
Academic libraries

ABSTRACT

This article provides a comprehensive literature review and classification method for data mining techniques applied to academic libraries. To achieve this, forty-one practical contributions over the period 1998–2014 were identified and reviewed for their direct relevance. Each article was categorized according to the main data mining functions: clustering, association, classification, and regression; and their application in the four main library aspects: services, quality, collection, and usage behavior. Findings indicate that both collection and usage behavior analyses have received most of the research attention, especially related to collection development and usability of websites and online services respectively. Furthermore, classification and regression models are the two most commonly used data mining functions applied in library settings.

Additionally, results indicate that the top 6 journals of articles published on the application of data mining techniques in academic libraries are: College and Research Libraries, Journal of Academic Librarianship, Information Processing and Management, Library Hi Tech, International Journal of Knowledge, Culture and Change Management, and The Electronic Library. Scopus is the multidisciplinary database that provides the best coverage of journal articles identified. To our knowledge, this study represents the first systematic, identifiable and comprehensive academic literature review of data mining techniques applied to academic libraries.

INTRODUCTION

Data mining, also known as knowledge discovery in databases, can be defined as the process of analyzing large information repositories and of discovering implicit, but potentially useful information (Han, Kamber, & Pei, 2011). Data mining has the capability to uncover hidden relationships and to reveal unknown patterns and trends by digging into large amounts of data (Sumathi & Sivanandam, 2006). The functions, or models, of data mining can be categorized according to the task performed: association, classification, clustering, and regression (Hui & Jha, 2000; Kao, Chang, & Lin, 2003; Nicholson, 2006b).

Data mining analysis is based normally on three techniques: classical statistics, artificial intelligence, and machine learning (Girija & Srivatsa, 2006). *Classical statistics* is mainly used for studying data, data relationships, as well as for dealing with numeric data in large databases (Hand, 1998). Examples of classical statistics include regression analysis, cluster analysis, and discriminate analysis. *Artificial intelligence* (AI) applies “human-thought-like” processing to statistical problems (Girija & Srivatsa, 2006). AI uses several techniques such as genetic algorithms, fuzzy logic, and neural computing. Finally, *machine learning* is the combination of advanced statistical methods and AI heuristics, used for data analysis and knowledge discovery (Kononenko & Kukar, 2007). Machine learning uses several classes of techniques: neural networks, symbolic learning, genetic algorithms, and swarm optimization. Data mining benefits from these technologies, but differs from the objective pursued: extracting patterns, describing trends, and predicting behavior.

A typical data mining process, as shown in Fig. 1, is an interactive sequence of steps that normally starts by integrating raw data from different data sources and formats. These raw data are cleansed in order to remove noise, and duplicated and inconsistent data (Han et al., 2011). These cleansed data are then transformed into appropriated formats that can be understood by other data mining tools, and filtration and aggregation techniques are applied to the data in order to extract summarized data. In fact, interesting knowledge is extracted from the transformed data. This information is analyzed in order to identify the truly interesting patterns. Eventually, knowledge is visualized to the user. More detailed information regarding a data mining process can be found in Han et al. (2011).

Data mining techniques are applied in a wide range of domains where large amounts of data are available for the identification of unknown or hidden information. In this sense, Girija & Srivatsa (2006)

[*] Corresponding author at: Celestijnenlaan 300, Box 2422, Office: 04.44, BE-3001 Leuven, Belgium. Tel.: +32 16 37 27 65; fax: +32 16 32 29 86, (GSM) +32 484 26 50 03 (mobile).
E-mail addresses: lorena.siguenza@ucuenca.edu.ec, lorena.siguenzaguzman@kuleuven.be (L. Siguenza-Guzman), victor.saquicela@ucuenca.edu.ec (V. Saquicela), elina.avilao@ucuenca.edu.ec (E. Avila-Ordóñez), joos.vandewalle@kuleuven.be (J. Vandewalle), dirk.cattrysse@kuleuven.be (D. Cattrysse).
URL: http://cib.kuleuven.be (L. Siguenza-Guzman).

http://dx.doi.org/10.1016/j.acalib.2015.06.007
0099-1333/© 2015 Elsevier Inc. All rights reserved.

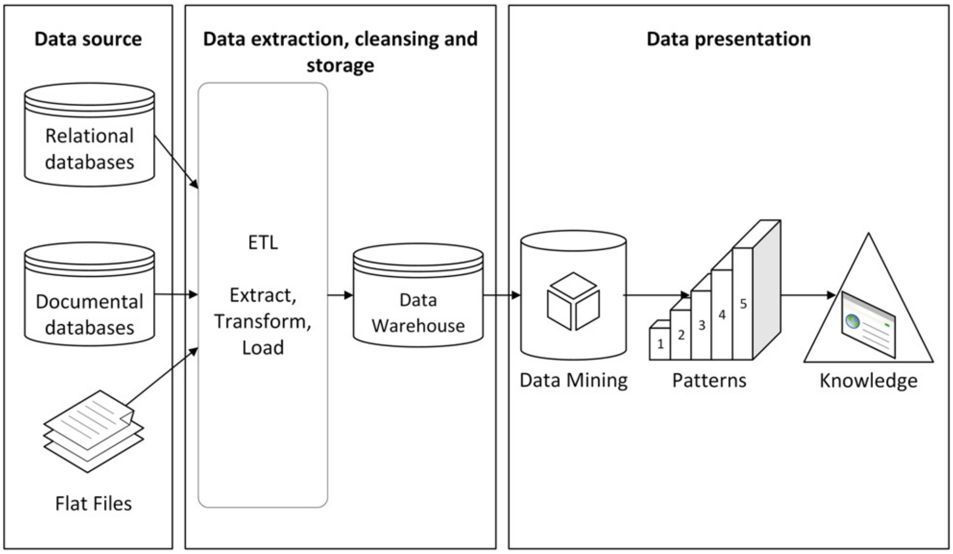

Fig. 1. Data mining process, based on Han et al. (2011).

indicate that data mining techniques used in WWW are called web mining, used in text are called text mining, and used in libraries are called bibliomining.

The term bibliomining, or data mining for libraries, was first used by Nicholson & Stanton (2003) to describe the combination of data warehousing, data mining and bibliometrics. This term is used to track patterns, behavior changes, and trends of library systems transactions. Although the concept is not new, the term bibliomining was created to facilitate the search of the terms “library” and “data mining” in the context of libraries rather than in software libraries. Bibliomining is an important tool to discover useful library information in historical data to support decision-making (Kao et al., 2003). However, to provide a complete report of the library system, bibliomining needs to be used iteratively applied in combination with other measurement and evaluation methods; as strategic information is discovered, more questions may be raised and thus start the process again (Nicholson, 2003b).

Bibliomining, as any knowledge extraction method, needs to follow a systematic procedure in order to allow an appropriate knowledge discovery. The bibliomining process starts by determining areas of focus and collecting data from internal and external sources (Nicholson, 2003b). Then, these data are collected, cleansed, and anonymized into a data warehouse. To discover meaningful patterns in the collected data, the bibliomining process includes the selection of appropriate analysis tools and techniques from statistics, data mining, and bibliometrics (Nicholson, 2006a). Interesting patterns are analyzed and visualized through reports. The mining process will be iterated until the resulted information is verified and proved by key users such as librarians and library managers (Shieh, 2010).

The application of bibliomining tools is an emerging trend that can be used to understand patterns of behavior among library users and staff, and patterns of information resource use throughout the library (Nicholson & Stanton, 2006). Bibliomining is highly recommended to provide useful and necessary information for library management requirements, focusing on the professional librarianship issues, but highly database technical dependent (Shieh, 2010). Bibliomining can also be used to provide a comprehensive overview of the library workflow in order to monitor staff performance, determine areas of deficiency, and predict future user requirements (Prakash, Chand, & Gohel, 2004). The resulting information gives the possibility to perform scenario analysis of the library system, where different situations that need to be taken into account during a decision-making process are evaluated (Nicholson, 2006a). An additional application is to standardize structures and reports in order to share data warehouses among groups of libraries, allowing libraries to benchmark their information (Nicholson, 2006a). Therefore, in order to improve the interaction quality between a library and its users, the application of data mining tools in libraries is worth pursuing (Chang & Chen, 2006).

The aim of this study is to investigate how far academic libraries are pragmatically using data mining tools, and in which library aspects librarians are implementing them. To this end, content and statistical analyses are used to examine articles that include case studies of academic libraries implementing data mining tools. The remainder of the article provides a detailed explanation of the research methodology adopted in this literature study. This is followed by a description of the proposed method for classifying data mining applications in libraries. Classification results are then presented and discussed. The article concludes by presenting limitations of the study, and by outlining research implications and prospects for future research.

RESEARCH METHODOLOGY

The present study follows the methodology employed by Ngai et al. (2009) to analyze and classify data mining techniques applied to customer relationship management. In this study, the analysis and classification are based on the examination of selected search engines and the use of a set of descriptors, all related to their specific interests. Then, the selected articles are reviewed and categorized based on a classification framework. The resulting list and classification is independently verified by research triangulation; finally, findings are reported in order to identify implications and future research directions.

Thus, following the Ngai et al. selection criteria and evaluation framework, a Web-based literature research on practical documents about data mining applications was conducted in order to identify relevant articles. As the nature of research on data mining and libraries is difficult to comprehend within the confines of specific disciplines, the relevant articles are scattered throughout numerous scholarly journals. Consequently, bearing in mind the degree of relevance or specialization to the subject analyzed, a set of four search engines was first selected to perform journal browsing. Based on the specialization degree, two major Library and Information Science (LIS) databases were searched: Library Information Science & Technology Abstracts (LISTA) accessed through EBSCOhost, and Library and Information Science Abstracts (LISA) accessed through ProQuest. In addition, two multidisciplinary databases: Web of Science (WoS) and Scopus were also consulted as complementary databases, as both search engines are among the largest and most common of the multidisciplinary databases available.