

Bibliomining for Library Decision-Making

Scott Nicholson

Syracuse University School of Information Studies, USA

Jeffrey Stanton

Syracuse University School of Information Studies, USA

INTRODUCTION

Most people think of a library as the little brick building in the heart of their community or the big brick building in the center of a college campus. However, these notions greatly oversimplify the world of libraries. Most large commercial organizations have dedicated in-house library operations, as do schools; nongovernmental organizations; and local, state, and federal governments. With the increasing use of the World Wide Web, digital libraries have burgeoned, serving a huge variety of different user audiences. With this expanded view of libraries, two key insights arise. First, libraries are typically embedded within larger institutions. Corporate libraries serve their corporations, academic libraries serve their universities, and public libraries serve taxpaying communities who elect overseeing representatives. Second, libraries play a pivotal role within their institutions as repositories and providers of information resources. In the provider role, libraries represent in microcosm the intellectual and learning activities of the people who comprise the institution. This fact provides the basis for the strategic importance of library data mining: By ascertaining what users are seeking, bibliomining can reveal insights that have meaning in the context of the library's host institution.

Use of data mining to examine library data might be aptly termed *bibliomining*. With widespread adoption of computerized catalogs and search facilities over the past quarter century, library and information scientists have often used bibliometric methods (e.g., the discovery of patterns in authorship and citation within a field) to explore patterns in bibliographic information. During the same period, various researchers have developed and tested *data-mining techniques*, which are advanced statistical and visualization methods to locate nontrivial patterns in large datasets. Bibliomining refers to the use of these bibliometric and data-mining techniques to explore the enormous quantities of data generated by the typical automated library.

BACKGROUND

Forward-thinking authors in the field of library science began to explore sophisticated uses of library data some years before the concept of data mining became popularized. Nutter (1987) explored library data sources to support decision making but lamented that "the ability to collect, organize, and manipulate data far outstrips the ability to interpret and to apply them" (p. 143). Johnston and Weckert (1990) developed a data-driven expert system to help select library materials, and Vizine-Goetz, Weibel, and Oskins (1990) developed a system for automated cataloging based on book titles (see also Morris, 1992, and Aluri & Riggs, 1990). A special section of *Library Administration and Management*, "Mining your automated system," included articles on extracting data to support system management decisions (Mancini, 1996), extracting frequencies to assist in collection decision making (Atkins, 1996), and examining transaction logs to support collection management (Peters, 1996).

More recently, Banerjee (1998) focused on describing how data mining works and how to use it to provide better access to the collection. Guenther (2000) discussed data sources and bibliomining applications but focused on the problems with heterogeneous data formats. Doszkocs (2000) discussed the potential for applying neural networks to library data to uncover possible associations between documents, indexing terms, classification codes, and queries. Liddy (2000) combined natural language processing with text mining to discover information in digital library collections. Lawrence, Giles, and Bollacker (1999) created a system to retrieve and index citations from works in digital libraries. Gutwin, Paynter, Witten, Nevill-Manning, and Frank (1999) used text mining to support resource discovery.

These projects all shared a common focus on improving and automating two of the core functions of a library: acquisitions and collection management. A few authors have recently begun to address the need to support management by focusing on understanding library users:

Schulman (1998) discussed using data mining to examine changing trends in library user behavior; Sallis, Hill, Jancee, Lovette, and Masi (1999) created a neural network that clusters digital library users; and Chau (2000) discussed the application of Web mining to personalize services in electronic reference.

The December 2003 issue of *Information Technology and Libraries* was a special issue dedicated to the bibliomining process. Nicholson presented an overview of the process, including the importance of creating a data warehouse that protects the privacy of users. Zucca discussed the implementation of a data warehouse in an academic library. Wormell; Suárez-Balseiro, Iribarren-Maestro, & Casado; and Geyer-Schultz, Neumann, & Thede used bibliomining in different ways to understand the use of academic library sources and to create appropriate library services.

We extend these efforts by taking a more global view of the data generated in libraries and the variety of decisions that those data can inform. Thus, the focus of this work is on describing ways in which library and information managers can use data mining to understand patterns of behavior among library users and staff and patterns of information resource use throughout the institution.

MAIN THRUST

Integrated Library Systems and Data Warehouses

Most managers who wish to explore bibliomining will need to work with the technical staff of their Integrated Library System (ILS) vendors to gain access to the databases that underlie the system and create a data warehouse. The cleaning, preprocessing, and anonymizing of the data can absorb a significant amount of time and effort. Only by combining and linking different data sources, however, can managers uncover the hidden patterns that can help them understand library operations and users.

Exploration of Data Sources

Available library data sources are divided into three groups for this discussion: data from the *creation of the library*, data from the *use of the collection*, and data from *external sources* not normally included in the ILS.

ILS Data Sources from the Creation of the Library System

Bibliographic Information

One source of data is the collection of bibliographic records and searching interfaces that represents materials in the library, commonly known as the Online Public Access Catalog (OPAC). In a digital library environment, the same type of information collected in a bibliographic library record can be collected as metadata. The concepts parallel those in a traditional library: Take an agreed-upon standard for describing an object, apply it to every object, and make the resulting data searchable. Therefore, digital libraries use conceptually similar bibliographic data sources to traditional libraries.

Acquisitions Information

Another source of data for bibliomining comes from acquisitions, where items are ordered from suppliers and tracked until they are received and processed. Because digital libraries do not order physical goods, somewhat different acquisition methods and vendor relationships exist. Nonetheless, in both traditional and digital library environments, acquisition data have untapped potential for understanding, controlling, and forecasting information resource costs.

ILS Data Sources from Usage of the Library System

User Information

In order to verify the identity of users who wish to use library services, libraries maintain user databases. In libraries associated with institutions, the user database is closely aligned with the organizational database. Sophisticated public libraries link user records through zip codes with demographic information in order to learn more about their user population. Digital libraries may or may not have any information about their users, based upon the login procedure required. No matter what data are captured about the patron, it is important to ensure that the identification information about the patron is separated from the demographic information before this information is stored in a data warehouse; doing so protects the privacy of the individual.

4 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/chapter/bibliomining-library-decision-making/10574

Related Content

API Standardization Efforts for Data Mining

Jaroslav Zendulka (2005). *Encyclopedia of Data Warehousing and Mining* (pp. 39-43).

www.irma-international.org/chapter/api-standardization-efforts-data-mining/10562

Finding Non-Coincidental Sporadic Rules Using Apriori-Inverse

Yun Sing Koh, Nathan Rountree and Richard O'Keefe (2008). *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications* (pp. 3222-3234).

www.irma-international.org/chapter/finding-non-coincidental-sporadic-rules/7830

Discovering Ranking Functions for Information Retrieval

Weiguo Fan and Praveen Pathak (2005). *Encyclopedia of Data Warehousing and Mining* (pp. 377-381).

www.irma-international.org/chapter/discovering-ranking-functions-information-retrieval/10626

Two Rough Set Approaches to Mining Hop Extraction Data

Jerzy W. Grzymala-Busse, Zdzislaw S. Hippe, Teresa Mroczek, Edward Roj and Boleslaw Skowronski (2008). *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications* (pp. 963-973).

www.irma-international.org/chapter/two-rough-set-approaches-mining/7682

Physical Data Warehousing Design

Ladjel Bellatreche and Mukesh Mohania (2005). *Encyclopedia of Data Warehousing and Mining* (pp. 906-911).

www.irma-international.org/chapter/physical-data-warehousing-design/10725