Bibliomining for Library Decision-Making

Scott Nicholson
Syracuse University, USA

Jeffrey Stanton
Syracuse University, 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 campus. These notions greatly oversimplify the world of libraries, however. Most large commercial organizations have dedicated in-house library operations, as do schools, non-governmental organizations, as well as local, state, and federal governments. With the increasing use of the Internet and the World Wide Web, digital libraries have burgeoned, and these serve 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—advanced statistical and visualization methods to locate non-trivial patterns in large data sets. 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 (also see Aluri & Riggs, 1990; Morris, 1991). 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, Banerjeree (1998) focused on describing how data mining works and ways of using 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, Jance, Lovetter, 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 (2003) presented an overview of the process, including the importance of creating a data warehouse that protects the privacy of users. Zucca (2003) discussed an implementation of a data warehouse in an academic library. Wormell (2003), Suárez-Balseiro, Iribarren-Maestro, and Casado (2003), and Geyer-Schultz, Neumann, and Thede (2003) used bibliomining in different ways to understand 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.

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 that system to create a data warehouse. The cleaning, pre-processing, 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 to understand library operations and users.

EXPLORATION OF DATA SOURCES

Available library data sources are divided in 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 represent 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 as traditional libraries.

Acquisitions Information

Another source of data for bibliomining comes from acquisitions, where items are ordered from suppliers and tracked until 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 is captured about the patron, it is important to ensure that the identification information about the patron is separated from the demographic information before storing this information in a data warehouse; this will protect the privacy of the individual.

Circulation and Usage Information

The richest sources of information about library user behavior are circulation and usage records. Legal and ethical issues limit the use of circulation data, however. This is where a data warehouse can be useful, in that basic demographic information and details about the circulation could be recorded without infringing upon the privacy of the individual.

Digital library services have a greater difficulty in defining circulation, as viewing a page does not carry the same meaning as checking a book out of the library, although requests to print or save a full text information resource might be similar in meaning. Some electronic full-text services already implement server-side capture of such requests from their user interfaces.

Searching and Navigation Information

The OPAC serves as the primary means of searching for works owned by the library. Additionally, because most OPACs use