Bibliomining for Library Decision-Making

Article · January 2006

CITATION
1

READS
130

1 author:

Scott Nicholson
Syracuse University
56 PUBLICATIONS **796** CITATIONS

SEE PROFILE

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

a Web browser interface, users may also access bibliographic databases, the World Wide Web, and other online resources during the same session; all of this information can be useful in library decision-making. Digital libraries typically capture logs from users searching their databases and can track, through "clickstream" analysis, the elements of Web-based services visited by users. In addition, the combination of a login procedure and cookies allow connecting user demographics to the services and searches they used in a session.

External Data Sources

Reference Desk Interactions

In the typical face-to-face or telephone interaction with a library user, the reference librarian records very little information about the interaction. Digital reference transactions, however, occur through an electronic format, and the transaction text can be captured for later analysis, which provide a much richer record than is available in traditional reference work. The utility of these data can be increased if identifying information about the user can be captured as well, but again, anonymization of these transactions is a significant challenge.

Item Use Information

Fussler and Simon (as cited in Nutter, 1987) estimated that 75-80% of the use of materials in academic libraries is in-house. Some types of materials never circulate, and therefore, tracking in-house use is also vital in discovering patterns of use. This task becomes much easier in a digital library, as Web logs can be analyzed to discover what sources users examined.

Interlibrary Loan and other Outsourcing Services

Many libraries using Interlibrary Loan and/or other outsourcing methods to get items on a "just-in-time" basis for users. The data produced by this class of transactions will vary by service, but can provide a window to areas of need in a library collection.

FUTURE TRENDS

Bibliomining can provide understanding of the individual sources listed earlier; however, much more information can be discovered when sources are combined through common fields in a data warehouse.

Bibliomining to Improve Library Services

Most libraries exist to serve the information needs of users, and therefore, understanding those needs of individuals or groups is crucial to a library's success. For many decades, librarians have suggested works; market basket analysis can provide the same function through usage data to aid users in locating useful works. Bibliomining can also be used to determine areas of deficiency and predict future user needs. Common areas of item requests and unsuccessful searches may point to areas of collection weakness. By looking for patterns in high-use items, librarians can better predict the demand for new items.

Virtual reference desk services can build a database of questions and expert-created answers, which can be used in a number of ways. Data mining could be used to discover patterns for tools that will automatically assign questions to experts based upon past assignments. In addition, by mining the question/answer pairs for patterns, an expert system could be created that can provide users an immediate answer and a pointer to an expert for more information.

Bibliomining for Organizational Decision-Making within the Library

Just as the user behavior is captured within the ILS, the behavior of library staff can also be discovered by connecting various databases to supplement existing performance review methods. While monitoring staff through their performance may be an uncomfortable concept, tighter budgets and demands for justification require thoughtful and careful tracking of performance. In addition, research has shown that incorporating clear, objective measures into performance evaluations can actually improve the fairness and effectiveness of those evaluations (Stanton, 2000).

Low use statistics for a work may indicate a problem in the selection or cataloging process. Looking at the associations between assigned subject headings, call numbers and keywords along with the responsible party for the catalog record may lead to a discovery of system inefficiencies. Vendor selection and price can be examined in a similar fashion to discover if a staff member consistently uses a more expensive vendor when cheaper alternatives are available. Most libraries acquire works both by individual orders and through automated ordering plans that are configured to fit the size and type of that library. While these automated plans do simplify the selection process, if some or many of the works they recommend go unused, then the plan might not be cost effective. Therefore, merging the acquisitions and circulation databases and seeking patterns that predict low use can aid in appropriate selection of vendors and plans.

Bibliomining for External Reporting and Justification

The library may often be able to offer insights to their parent organization or community about their user base through patterns detected with bibliomining. In addition, library managers are often called upon to justify the funding for their library when

budgets are tight. Likewise, managers must sometimes defend their policies, particularly when faced with user complaints. Bibliomining can provide the data-based justification to back up the anecdotal evidence usually used for such arguments.

Bibliomining of circulation data can provide a number of insights about the groups who use the library. By clustering the users by materials circulated and tying demographic information into each cluster, the library can develop conceptual "user groups" that provide a model of the important constituencies of the institution's user base which can fulfill some common organizational needs for understanding where common interests and expertise reside in the user community. This capability may be particularly valuable within large organizations where research and development efforts are dispersed over multiple locations.

In the future, organizations that fund digital libraries can look to text mining to greatly improve access to materials beyond the current cataloging / metadata solutions. The quality and speed of text mining continues to improve. Liddy (2000) has researched the extraction of information from digital texts, and implementing these technologies can allow a digital library to move from suggesting texts that might *contain the answer* to just *providing the answer*, extracted from the appropriate text or texts. The use of such tools risks taking textual material out of context and also provides a few hints about the quality of the material, but if these extractions were links directly into the texts, then context could emerge along with an answer. This could provide a substantial asset to organizations that maintain large bodies of technical texts because it would promote rapid, universal access to previously scattered and/or uncataloged materials.

CONCLUSION

Libraries have gathered data about their collections and users for years, but have not always used those data for better decision-making. By taking a more active approach based on applications of data mining, data visualization, and statistics, these information organizations can get a clearer picture of their information delivery and management needs. At the same time, libraries must continue to protect their users and employees from misuse of personally identifiable data records. Information discovered through the application of bibliomining techniques gives the library the potential to save money, provide more appropriate programs, meet more of the user's information needs, become aware of gaps and strengths of their collection, and serve as a more effective information source for its users. Bibliomining can provide the data-based justifications for the difficult decisions and funding requests library managers must make.

REFERENCES

Atkins, S. (1996). Mining automated systems for collection management. *Library Administration & Management, 10*(1), 16-19.

Chau, M.Y. (2000). *Mediating off-site electronic reference services: Human-computer interactions between libraries and Web mining technology*. Fourth International Conference on Knowledge-based Intelligent Engineering Systems & Allied Technologies (vol. 2, pp.695-699). Piscataway, NJ: IEEE.

Chaudhry, A.S. (1993). Automation systems as tools of use studies and management information. *IFLA Journal, 19*(4), 397-409.

Doszkocs, T.E. (2000). Neural networks in libraries: The potential of a new information technology. Retrieved October 24, 2001, from *http://web.simmons.edu/~chen/nit/NIT%2791/027~dos.htm*

Geyer-Schulz, A., Neumann, A., & Thede, A. (2003). An architecture for behavior-based library recommender systems. *Information Technology and Libraries, 22*(4), 165-174.

Guenther, K. (2000). Applying data mining principles to library data collection. *Computers in Libraries, 20*(4), 60-63.

Gutwin, C., Paynter, G., Witten, I., Nevill-Manning, C., & Frank, E. (1999). Improving browsing in digital libraries with keyphrase indexes. *Decision Support Systems, 2I,* 81-104.

Johnston, M., & Weckert, J. (1990). Selection advisor: An expert system for collection development. *Information Technology and Libraries, 9*(3), 219-225.

Lawrence, S., Giles, C.L., & Bollacker, K. (1999). Digital libraries and autonomous citation indexing. *IEEE Computer, 32*(6), 67-71.

Liddy, L. (2000, November/December). Text mining. *Bulletin of the American Society for Information Science*, 13-14.

Mancini, D.D. (1996). Mining your automated system for systemwide decision making. *Library Administration & Management, 10*(1), 11-15.

Morris, A. (Ed.) (1991). *Application of expert systems in library and information centers*. London: Bowker-Saur.

Nicholson, S. (2003). The bibliomining process: Data warehousing and data mining for library decision-making. *Information Technology and Libraries, 22*(4), 146-151.

Nutter, S.K. (1987). Online systems and the management of collections: Use and implications. *Advances in Library Automation Networking, 1*, 125-149.

Peters, T. (1996). Using transaction log analysis for library management information. *Library Administration & Management, 10*(1), 20-25.

Sallis, P., Hill, L., Jance, G., Lovetter, K., & Masi, C. (1999). A methodology for profiling users of large interactive systems incorporating neural network data mining techniques. *Proceedings of the 1999 Information Resources Management Association International Conference* (pp. 994-998). Hershey, PA: Idea Group Publishing.

Schulman, S. (1998). Data mining: Life after report generators. *Information Today, 15*(3), 52.

Stanton, J.M. (2000). Reactions to employee performance monitoring: Framework, review, and research directions. *Human Performance, 13*, 85-113.

Suárez-Balseiro, C.A., Iribarren-Maestro, I., & Casado, E.S. (2003). A study of the use of the Carlos III University of Madrid library's online database service in scientific endeavor. Information Technology and Libraries, 22(4), 179-182.

Wormell, I. (2003). Matching subject portals with the research environment. *Information Technology and Libraries, 22*(4), 158-166.

Zucca, J. (2003). Traces in the clickstream: Early work on a management information repository at the University of Pennsylvania. *Information Technology and Libraries, 22*(4), 175-178.

KEY TERMS

Bibliometrics: The study of regularities in citations, authorship, subjects, and other extractable facets from scientific communication using quantitative and visualization techniques. This allows researchers to understand patterns in the creation and documented use of scholarly publishing.

Bibliomining: The application of statistical and pattern-recognition tools to large amounts of data associated with library systems in order to aid decision-making or justify services. The term "bibliomining" comes from the combination of bibliometrics and data mining, which are the two main toolsets used for analysis.

Data Mining: The extraction of non-trivial and actionable patterns from large amounts of data using statistical and artificial intelligence techniques. Directed data mining starts with a question or area of interest, and patterns are sought that answer those needs. Undirected data mining is the use of these tools to explore a dataset for patterns without a guiding research question.

Data Warehousing: The gathering and cleaning of data from disparate sources into a single database, optimized for exploration and reporting. The data warehouse holds a cleaned version of the data from operational systems, and data mining requires the type of cleaned data that lives in a data warehouse.

Integrated Library System: The automation system for libraries, combining modules for cataloging, acquisition, circulation, end-user searching, database access, and other library functions through a common set of interfaces and databases.

Online Public Access Catalog (OPAC): The module of the integrated library system designed for use by the public to allow discovery of the library's holdings through the searching of bibliographic surrogates. As libraries acquire more digital materials, they are linking those materials to the OPAC entries.

ENDNOTE

[1] This work is adapted from: Nicholson, S. & Stanton, J. (2003). Gaining strategic advantage through bibliomining: Data mining for management decisions in corporate, special, digital, and traditional libraries. In H. Nemati, & C. Barko (Eds.), *Organizational data mining: Leveraging enterprise data resources for optimal performance (pp.247-262)*. Hershey, PA: Idea Group Publishing.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 272-277, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).