Preprint version of

Nicholson, S. & Stanton, J. (2003). Gaining strategic advantage through bibliomining: Data mining for management decisions in corporate, special, digital, and traditional libraries. In Nemati, H. & Barko, C. (Eds.). Organizational data mining: Leveraging enterprise data resources for optimal performance. Hershey, PA: Idea Group Publishing. 247-262. Updated on 2/16/04

Gaining Strategic Advantage through Bibliomining:

Data Mining for Management Decisions in Corporate, Special, Digital, and Traditional Libraries

Scott Nicholson

Jeffrey Stanton

Syracuse University School of Information Studies

Address all correspondence to: Scott Nicholson, School of Information Studies, Syracuse University, 4-127 Center for Science and Technology, Syracuse NY, 13244-4100, scott@scottnicholson.com, (315) 443-1640, Fax: (315) 443-5806.

Abstract

Library and information services in corporations, schools, universities, and communities capture information about their users, circulation history, resources in the collection, and search patterns (Koenig, 1985). Unfortunately, few libraries have taken advantage of these data as a way to improve customer service, manage acquisition budgets, or influence strategic decision-making about uses of information in their organizations. In this chapter, we present a global view of the data generated in libraries and the variety of decisions that those data can inform. We describe ways in which library and information managers can use data mining in their libraries, i.e. bibliomining, to understand patterns of behavior among library users and staff members and patterns of information

resource use throughout the institution. The chapter examines data sources and possible applications of data mining techniques and explores the legal and ethical implications of data mining in libraries.

Gaining Strategic Advantage through Bibliomining:

Data Mining for Management Decisions in Corporate, Special, Digital, and Traditional Libraries

For several decades, library and information services in corporations, schools, universities, and communities have had the ability to capture information about their users, circulation history, resources in the collection, and search patterns (Koenig, 1985). Collectively, these data can provide library managers more information about common patterns of user behavior to aid in decision-making processes. Unfortunately, few libraries have taken advantage of these data as a way to improve customer service, manage acquisition budgets, or influence strategic decision-making about uses of information in their organizations. The application of advanced statistical and data mining techniques to these kinds of data may provide useful ways of supporting decision-making at every library where user, cataloging, searching, and circulation interfaces are automated.

Use of data mining to examine library data records 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 techniques to plumb the enormous quantities of data generated by the typical automated library.

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, & Oskins (1990) developed a system for automated cataloging based on book titles (also see Morris, 1992; Aluri and 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, 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. What these projects did not discuss was the use of library data to support strategic management decisions for libraries and their host institutions. A few authors have recently begun to address this need 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. 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 chapter 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. The chapter will examine data sources and possible applications of data mining techniques as well as explore the legal and ethical implications of bibliomining.

Data Mining in Library and Information Services

Background

Most people think of libraries 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, e.g., people interested in health and medicine, industry and world news, law, and business. 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 need to know and how well those needs are served, bibliomining can reveal insights that have meaning in the context of the library's host institution.

Because some readers may not have a detailed sense of the behind-the-scenes activities in contemporary libraries, we begin by providing a data-focused overview of the internal workings of libraries. Workflow in a traditional "bricks and mortar" library creates a number of data sources appropriate for bibliomining. Before a library obtains new information resources (e.g., books, databases, reference tools, electronic access, etc.), a librarian assesses the needs of the existing collection in light of available and upcoming publications. Next, acquisitions personnel obtain the information resources specified from this needs assessment. Once the library obtains requested new resources, cataloging personnel either create or purchase a catalog record for the new resource. The circulation department then makes the resource available to end-users. Depending upon the size of the library and the scope of its operations, these activities fall within the purview of one, a dozen, or possibly hundreds of different employees organized into specialized departments.

After an information resource appears in the library's collection, users can locate it using catalog search systems and bibliographic databases. Although there is little uniformity with regard to the specifics of the user interfaces for these systems, most catalogs and bibliographic databases do support the use of a standard Web browser client as the front end of the system. Increasingly, catalogs and databases are all cross-linked, and each user's search record and traversal of links can appear in log files. When users find resources that they wish to borrow, the circulation department records their selection in a database that tracks the location of each resource owned by the library.

As this overview suggests, all functional processes of the library – collection assessment, acquisition, cataloging, end user searching, and circulation – generate large reserves of available data that document information resource acquisition and use. Library information systems frequently use large relational databases to store user information, resource information, circulation information, and possibly bibliographic search logs.

Although the processes outlined above adequately describe the traditional library with its physical infrastructure and tangible resources (e.g., books, maps, etc.), an increasing proportion of libraries offer many or

most of their information resources in digital form over local or wide area networks. These "digital libraries" do, however, employ many similar techniques for selecting and acquiring information as those described above. While the personnel involved may have different titles and slightly different roles, the activities remain the same: identify resources needed for the collection, acquire the resources, make the resource available to users, and assist users in locating the resource through electronic and virtual reference aids. Most digital libraries do not have a circulation function; if users need something, they just save it or print it with to their own equipment. Unlike the traditional library, the information systems of a digital library can track and log the entire visit of a user to a digital library.

The vast data stored in the databases of traditional and digital libraries represent the behavioral patterns of two important constituencies: library staff and library users. In the case of library staff, mining available acquisitions and bibliographic data could provide important clues to understanding and enhancing the effectiveness of the library's own internal functions. Perhaps more importantly, however, mining user data for knowledge about what information library users are seeking, whether they find what they need, and whether their questions are answered, could provide critical insights useful in customer relations and knowledge management. Whether the library users are the public in a local community or the internal staff of a large corporation, understanding their search, borrowing, and related behavior patterns can indicate whether they have obtained the information resources they need, what information resources they find most useful, and insights into their future needs. These kinds of information can have strategic utility within the larger organization in which the library is situated.

As a closing note, many companies have information services that are not explicitly called digital libraries, but can nonetheless be analyzed using some of the techniques listed here. These sources may be called knowledge bases, and are commonly associated with online help applications. These services are similar to digital libraries in that they have a collection, i.e. the set of searchable documents gathered to allow users to help themselves, and a virtual reference service, i.e. the mechanism provided to connect users with employees to aid in the resolution of problems. The reference service may produce another collection of previous questions and answers. Once distinction between a help desk and most virtual reference services is that the help desk attempts to resolve the problem directly which the virtual reference services point the user to some resource that may help solve the problem.

Integrated Library Systems and Data Warehouses

Although the system used in most parts of the library is commonly known as an *Integrated* Library System (ILS), very few ILS vendors make it easy for the library to access the data generated by the system in an integrated fashion. Instead, most librarians conceptualize their system as a set of separate data sources. While a relational database stands at the heart of most ILS systems, few system vendors provide sophisticated analytical tools that would promote useful access to this raw data. Instead, vendors encourage library staff to use pre-built front ends to access their ILS databases; these front ends typically have no features that allow exploration of patterns or findings across multiple data sets. As a first step, most managers who wish to explore bibliomining will need to work with the technical staff of their ILS vendors to gain access to the databases that underlie their system.

Once the vendor has revealed the location and format of key databases, the next step in bibliomining is the creation of a data warehouse. As with most data mining tasks, the cleaning and pre-processing of the data can absorb a significant amount of time and effort. A truly useful data warehouse requires a procedure or methods for integration to permit queries and joins across multiple heterogeneous data sources. Only by combining and linking different data sources can managers uncover the hidden patterns that can help understand library operations and users. Two studies have documented the processes needed here. The first was at the University of Florida library system, where a relational database was created for analysis by gathering information directly from screen images on the integrated system because the underlying database was inaccessible. While the authors documented traditional library statistics gathered from this data, they also discussed the future possibility of data mining through a neural network (Su & Needamangala, 2000). Another study in this area occurred at Kansas State University, where a prototype for a decision support system based on information in the library automation system was presented (Bleyberg et. Al., 1999).

Once the data warehouse is set up, it can be used for not only traditional SQL-based question-answering, but also online analytical processing (OLAP) and data mining. Figure 1 provides a schematic overview of the complete process from the collection of the data, through creation of the data warehouse, to the application of various analysis techniques. Multidimensional analysis tools for OLAP (e.g., Cognos) would allow library managers to explore their traditional frequency-based data in new ways by looking at statistics along easily changeable dimensions. The same data warehouse that supports OLAP also sets the stage for data mining. The remainder of this chapter builds on the assumption that this data warehouse is available.

Exploration of Data Sources

Available library data sources are divided in three groups for this discussion: ILS data sources from the creation of the library information system, ILS data sources from the use of the collection, and external data sources not normally included in the ILS. In order to systematically explore these data sources, we define each source, explain current analysis techniques used in libraries, and describe applications of that source to a digital library environment.

ILS Data Sources from the Creators of the Library System

Bibliographic Information. The process of developing a library information system creates several types of data that may be useful in understanding the activities of the library staff. One source of data is the collection of records that represent materials in the library, commonly known as the Online Public Access Catalog (OPAC). The OPAC is created by catalogers and used to locate materials in the library. This typically contains records in MARC (MAchine-Readable Cataloging) format, which contains information about the title, authorship, publication, physical description, access method, unique call number, subject headings, and other bibliographic metadata about the work. Recent efforts on the Dublin Core Metadata definition have standardized a set of bibliographic metadata for use on web-based information resources.

In order to understand the collection, librarians commonly analyze the bibliographic database for basic frequency information. The number of works in the collection is counted both overall and by classification groups (using the call number; e.g., literature; science, etc.). Librarians then compare these breakdowns to other libraries (either individual libraries or among a set of similar peer libraries) in order to evaluate the size and relative proportions of different areas of the collection. Libraries will compare individual items to some type of list -- usually normed for a subject and age group -- to evaluate the quality of the collection.

In a digital library environment, the same type of information collected in a bibliographic library record can be gathered as metadata. Terms such as cataloging, AACR2 (a standard for writing bibliographic records in libraries), and MARC (the protocol for putting these records into a machine-readable format) are replaced by metadata acquisition, XML, DTDs (Document Type Definition), and Dublin Core. 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 do; for bibliomining the same bibliographic data is available from both types of library.

Acquisitions Information. Another source of data for bibliomining comes from the process of acquisitions, in which items are ordered from suppliers and tracked until received and processed by the library. This data source contains information about the orders (e.g., the vendor or publisher), the acquisitions librarian who ordered the work, the cost of the work, the order date, problems with the order, and the delivery date. Sometimes it may also contain information about the person who selected and/or requested the work. Other than simple analyses to reveal problem vendors, libraries have rarely analyzed and/or shared these data for competitive advantage (Nutter, 1987). Instead, these data mainly support logistics, e.g., to track missing or late orders. In general, because digital libraries do not order tangible, physical goods such as books, somewhat different acquisition methods and vendor relationships are the norm. Nonetheless, in both traditional and digital library environments, acquisition data have much untapped potential for examining, understanding, controlling, and forecasting information resource costs.

ILS Data Sources from Usage of the Library System

User Information. In order to track users who wish to check out materials (and in some cases, even allow users to enter the institutions), libraries maintain user databases users. In libraries associated with institutions, e.g. academic, school, and corporate libraries, the user database is closely aligned with the organizational database, so user data may come from enrollment or personnel databases rather than directly from users. When library managers wish to understand their user population better, they often perform frequency analyses to explore the demographics of their user base. In public libraries, however, user databases often contain only sparse information (e.g., name, date, and contact information) gathered directly from the user by the library. In the public library environment, analysis of the user population often requires outside data, such as census data or user surveys, and is therefore more time-consuming and expensive. Sophisticated libraries link user records through zip codes with demographic information and learn more about their user population.

Digital libraries may or may not have any information about their users. If the library requires no login to use, then the users are only accesses to a Web page. Libraries that do require a login procedure can also collect demographic information about their user base and this may provide a better idea of different user groups that access the library.

Circulation Information. One of the richest untapped sources of information about library user behavior is circulation records. Notably, however, important legal and ethical issues limit the uses of circulation data and make it the most protected information source in the library (see the final section of this paper for an analysis of legal and

ethical issues). When a user checks out a work, the checkout date, due date, call number, and user identification is recorded. When the work is returned, borrower information disappears from the librarian's perspectives, but the borrowing record may remain hidden in the database. A similar data source is a hold/request database, where requests for works either not in the library or already checked out are recorded. Again, once the hold/request has been dealt with, the user information is removed from the view of the librarian. In most cases, however, the user identifiers are still kept in the underlying database, locked away in inaccessible tables.

Librarians use this information to guide purchasing and removal of materials from the library. Items with a high number of circulations, holds, and/or requests help selectors choose new works to bring into the library. In addition, when collection personnel make decisions about which books to remove from the shelves, the circulation information gives them an idea of which works have rarely circulated.

With the exception of some "e-book" services, most digital library information resources do not provide circulation information for mining. Viewing a page by clicking a link does not necessarily carry the same message as checking a book out of the library, although it is easy to imagine capturing meaningful information in a database that records requests to print or save a full text information resource. Current information retrieval services (e.g., Proquest), already implement server-side capture of such requests from their browser-based user interfaces.

Searching and Navigation Information. The Online Public Access Catalog (OPAC) serves as the primary means of searching for works owned by the library. Additionally, because most OPACs use a browser interface, users may also access bibliographic databases, the World Wide Web, and other online resources during the same session. Search records and activity logs can be captured for the entire session for users at the library, or, for users outside of the library, for those elements of the session that use the library's servers. Note that defining which set of searches comprise a single searching session and linking individual user to their searching sessions can be difficult. While libraries have not routinely conducted investigations of these logs, they have been a data source explored by researchers. Online searching has been an active area of research in Library and Information Science for studies in information seeking behavior and human-computer interaction. See Harter and Hert (1997) for a review of this research area.

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. Through the use of browser cookies, one can circumvent some of the difficulties in determining which set of searches constitutes a single session. In

addition, the combination of a login procedure and cookies allow connecting specific users to the services and searches they used in a session.

External Data Sources

Reference Desk Interactions. The reference librarian's primary responsibility is to assist users in defining information needs and locating resources to satisfy them. In the typical face-to-face or telephone interaction with a library user, the reference librarian records very little information about the interaction. Some libraries keep limited records about the interaction (e.g., the type of question; directional, factual, research, etc.). If librarians wish to gain knowledge from reference transactions, they must capture more information from these types of transactions from not only the reference desk but from other areas of the library, because users direct questions toward most library staff. This process has recently become considerably easier because many reference librarians use a desktop computer as the first resource they consult to answer user inquiries.

Library managers can examine reference interaction data to help make staffing decisions, both in what type of staff (librarians or paraprofessionals) and the amount of staffing needed. Some research has examined the success of reference librarians from the perspective of a user; a frequently-cited study in this area Hernon and McClure's (1986) , which reported that reference librarians are correct only 55 percent of the time.

Digital reference transactions occur through an electronic format such as e-mail, Web-based forms, message boards, instant messaging, or live chat. The text of the questions and answers can be captured for later analysis. Even digital reference work done through synchronous interfaces can be recorded. These data provides analysts a much richer record of transactions than is available in traditional reference work. Some traditional libraries are beginning to offer e-mail and other virtual reference services, and thus may also have the ability to capture and analyze these transactions. The utility of these data can be increased if demographic information about the user can be captured as well.

Item Use Information. While circulation information and computer logs give two views of the user's borrowing and search behavior, 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. Although, some libraries capture information about in-house use of materials, most do not on a regular basis. However, because most libraries have their works bar-coded, this information could be captured easily at the time of reshelving. Bibliomining opportunities may provide the impetus

to begin a regular tracking procedure when reshelving books and other materials. This task becomes much easier in a digital library, as Web logs can be analyzed to discover what sources users examined.

Interlibrary Loan. Most physical libraries offer some type of Interlibrary Loan (ILL) service. This service is unique in that the resulting data source is a combination of information seeking behaviors from both library users and library staff. In most ILL situations, the user seeking an item contacts the ILL department, and the ILL staff members locate the material from another library and borrow the book. Once the requesting library receives the book, they lend it to their user. Therefore, there are two levels of seeking: the user seeking the original material and the staff seeking a source for lending. The cost for lending and the time it takes for delivery vary by the source selected. Many libraries have a number of consortium agreements, each with different response times and costs, which makes this a complex data source. The information tracked varies by library, but usually contains information about the item requested, the user, the lending library, and the essential dates involved.

Bibliomining Applications

By going beyond the standard frequency-based reports provided by the vendors, managers can learn much more about the needs and behaviors of those involved in staffing and using the library. Bibliomining can provide deeper understanding of the individual sources listed above; however, much more information can be discovered when sources are used in conjunction with one another. Most of these data sources contain fields that can be used to link them to other sources. This is where a data warehouse may prove useful; as it stands, many databases in a library system are optimized for searching and tracking instead of reporting and mining.

Each bibliomining analysis can reveal a pattern of activity within the library. Uncovering and reporting these patterns may have potential benefits at three nested levels: benefits for individuals through improved library services, benefits for library management through the provision of improved decision-making information, and benefits for the institution that the library serves through reporting of relevant patterns of user behavior. Additionally, by providing information on the performance and utility of the library as a unit, bibliomining can provide justification for continued financial and institutional support for library operations. These levels serve to structure our presentation of bibliomining opportunities and applications.

Bibliomining to Improve Library Services

The users of library services are one of the most important constituencies in most library organizations. Most libraries exist to serve the information needs of users, and therefore, understanding those needs is crucial to a

library's success. Examining user behavior on an individual level may aid in understanding that individual, but it tells a librarian very little about the larger audience of users. Examining the behaviors of a large group of users for patterns can then allow the library to have a better idea of the information needs of their user base, and therefore better customize the library services to meet those needs.

For many decades libraries have provided readers' advisory services with the help of librarians who know the collection well enough to help a user choose a work similar to other works. Market basket analysis can provide the same function by examining circulation histories to locate related works. In addition, this information could be provided to the OPAC to allow users to see similar works to one they have selected based upon circulation histories. While it is technically possible to build a profile for users based upon their own circulation history (see Amazon.com for example), it may be legally and ethically questionable to do this without a user's permission. Nonetheless, by obtaining and using anonymous data from a large number of users, one can obtain similar or better results.

In order to locate works in the library, users rely on the OPAC. Librarians often examine user comments and surveys to assess user satisfaction with these tools. However, these comments can be deceptive; Applegate (1993) found that users often report satisfaction with a system even when they receive poor results, and Hildreth (2000) confirmed these results on Web-based OPACs. Therefore, librarians may wish to examine the artifacts of those searches for problem areas instead of relying on user comments and surveys in order to improve the user experience. When upgrading or changing library system interfaces, librarians can explore these patterns of common mistakes in order to make informed decisions about system improvements.

Bibliomining can also be used to predict future user needs. By looking for patterns in high-use items, librarians can better predict the demand for new items in order to determine how many copies of a work to order. To prevent inventory loss, predictive modeling can be used to look for patterns commonly associated with lost/stolen books and high user fees. Once these patterns have been discovered, appropriate policies can be put in place to reduce inventory losses. In addition, fraud models can be used to determine the appropriate course of action for users who are chronically late in returning materials.

The library can also better serve their user audience by determining areas of deficiency in the collection. However, such deficiencies can be difficult to discover. The reference desk and the OPAC are two sources of data that can aid in solving problems with the collection. If the topics of questions asked at the reference desk are

recorded along with the perceived outcome of the interaction, then patterns can be discovered to guide librarians to areas that need attention in the collection. In addition, OPAC searches that produce no results can be analyzed for common areas searched that produce few or no results.

With electronic information resources, the same work may be offered by a number of different vendors in different packages. Sometimes these resources are unavailable, and other times the resource is difficult to find on a vendor's site, even if it is available. In addition, when a vendor changes the layout of the Web resource, links from the online catalog to the resource may become invalid. By looking at the access paths by users to electronic resources, libraries can detect repeated patterns of failure with a vendor or a resource. Instead of waiting for user complaints, librarians can be notified if a problem exists and can work to resolve the problem.

Digital libraries, as well as libraries with Web-based services, can use different techniques and tools to examine common patterns in the path taken through their Web site. This could give the library ideas of problem areas of their Web site, good places to post important messages, areas where more guidance is needed, and opportunities to keep people from leaving the site. Just as library managers must observe a physical library for signage needs, digital libraries must also be examined for appropriate areas for guidance. In addition, predictive modeling could be used to present user with the information they are seeking with fewer clicks. Research in the analysis of Web logs is fairly well-established for non-library situations, one example of which is research by Zaiane, Zin, and Han(1998), who applied data mining and OLAP to Web logs to discover trends.

Virtual reference desk services can build a database of questions and expert-created answers. This database 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 while they wait for a human to answer the question.

Bibliomining for Organizational Decision-Making within the Library

Bibliomining can be used to aid library managers in monitoring their organization and making decisions. Just as the user behavior is captured within the ILS, the behavior of library staff can also be discovered by connecting various databases. While monitoring staff through their performance may be an uncomfortable concept for many librarians, 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). The concepts suggested here are not intended to replace more typical methods of library staff performance measurement; rather, they can help to quantitatively justify statements and bring possible problem areas to light.

If items that are not circulated or used are considered to be an inappropriate expenditure, then bibliomining may provide insight as to how those items got into the library. By looking for correlations between low-use items and subject headings, publisher, vendor, approval plan, date, format, acquisitions librarian, collection development librarian, library location, and other items in the data warehouse, managers might discover problem areas in the collection or organization.

Bibliomining can aid in staffing decisions. If the times of transactions are captured in the system, they can be used to build patterns of behavior at the circulation desk. By looking at these patterns, library managers can optimize the number of staff members needed at the circulation desk. In addition, by looking at the time and frequency of different types of reference interactions for patterns, reference desk schedules can be optimized.

Low use statistics for a work may indicate a problem in the cataloging process. By seeking patterns in low-use statistics with the library staff responsible for that work, possible issues could be unearthed. Along these lines, looking at the associations between assigned subject headings, call numbers and keywords along with the responsible party for the catalog record may lead a discovery of "default" subject heading terms or call numbers that library staff members or the outsourcing organization are assigning inappropriately.

Interlibrary loan departments are frequently asked to justify and reduce their costs. Bibliomining can aid the ILL manager in refining policies for lender selection. In interlibrary loan, a library usually participates in a number of consortium and reciprocal agreements. In addition, there are a number of suppliers that offer their services for fees. The higher-cost ILL services are often more convenient and less effortful to use than lower-cost methods. To complicate the matter, other institutions borrow from the library and may or may not return works in a timely fashion. Bibliomining is a tool that will allow ILL librarians to look within these complicated relationships for patterns that are either favorable or unfavorable and create appropriate policies. In addition, these techniques can justify policies and standards by discovering the appropriate patterns to support policies.

In acquisitions, vendor selection and price can be examined in a similar fashion. If a staff member consistently uses a more expensive vendor when cheaper alternatives are available, the acquisitions librarian may need to get involved with clearer policies about vendor selection. Seeking patterns in the time it takes between the

receipt of a book and the placement of that work on the shelf may bring to light some areas that need attention. This is difficult to track using normal means, as different types of works require different processes in preparing them for use. Bibliomining can be used to look for patterns in the complexity, and allow the clustering of types of works by the average time it takes to prepare them for the shelf. This can then be used to set policies for expected turnaround time for preparing works.

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. Outside vendors provide automated ordering plans, for a fee, as a way of reducing the collection assessment and acquisition workload within the 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.

The library also has the option of outsourcing particular subsets of acquisitions and cataloging processes. If the acquisitions database contains details about this outsourcing, and that can be tied into a file that records changes to bibliographic records, library managers can then look for patterns of problems with different types of outsourcing. One example of this involves the process of copy cataloging, where a library purchases or accesses a catalog record from an outside source, copies it into the bibliographic database, and changes it as needed. If some measure of the changes is tracked, either by the number of characters that are changed or time taken to make the change, then it could be matched up with the record sources. By looking for and eliminating problematic sources, copy catalogers could streamline their workflow.

As with all bibliomining projects, more information allows a better chance for analysts to discover meaningful patterns. Therefore, it is important to ensure the identities of the personnel involved with record creation, circulation, ILL, and other interactions with the system are recorded. With this information, library managers can investigate problems and solve issues in order to make the limited budget dollar stretch as far as possible. Data mining also works optimally on very large data sets. This consideration argues for the formation of data sharing consortia, particularly among smaller libraries.

Bibliomining for External Reporting and Justification

The library does not exist independently but rather answers to a parent organization or is embedded within a larger community. The library may often be able to offer insights to the 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. If the library has collected usage data, even more can be ascertained about user group usage patterns. Digital libraries can use Web log information for similar patterns, but care should be taken as to the importance given to these patterns, as it takes much less effort to follow a link than it does to check out a book. In either case, the user group concept can fulfill some common organizational needs for understanding where common interests and expertise reside in the user community. By helping to identify groups of individuals with common interest and/or expertise in a topic, bibliomining can support a key component of knowledge management systems. This capability may be particularly valuable within large organizations where research and development efforts are dispersed over multiple locations.

The searches employed and reference questions asked by users can be mined for patterns. This has been an area explored by researchers (Harter & Hert, 1997), although few advanced data mining tools have been applied to such data and even fewer applications of the technique have yet emerged. Clustering techniques could be used to explore common search topics in order to help determine instructional or training needs. This strategy would be particularly appropriate in corporate settings, where training needs assessment is a critical element in developing and maintaining a skilled workforce. Tracking the frequency of questions on these topics before and after training can also allow library managers to justify the cost for training courses or instructional materials.

In reviewing organizational policies, managers can use bibliomining to explore the appropriateness of current policies. Seeking patterns between circulation time, renewals, and holds in conjunction with format, user classification, topic, and other variables may help library managers set circulation policies that will better match the information needs of users. Bibliomining can also provide data that justify these policies to the parent organization or community.

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 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.

Because users can access information in a variety of formats – print, electronic, microform, ILL – mining these transactions for patterns of use can help library managers to better understand how information is used in their library. This will allow managers to make more appropriate collection development decisions. Most importantly, it will allow managers to justify their collection decisions to their funding bodies.

Legal and Ethical Implications of Bibliomining

To close the chapter, we discuss the legal and ethical implications of bibliomining. User privacy may possibly present a somewhat less thorny problem in corporate and special libraries, but in many states, strict policies govern the privacy of user records for publicly funded libraries. In fact, many automation systems hide or discard circulation records after the material has been returned for precisely this reason. Organizations must strike a balance between discovering patterns at a higher level and connecting those patterns to particular individuals in the system.

Legal Issues

User information used to be unprotected; for example, library circulation cards used to list the names of those who had previously checked out a book (Estabrook, 1996). In the 1940's, FBI director J. Edgar Hoover initiated the Library Awareness program, an FBI program which monitored the circulation records of library users. This activity continued on through to the 1980's, when library records were opened further as several states passed open record laws making the records of the state government (and therefore libraries that were part of this system) open to the public (Seaman, 2001).

The American Library Association (ALA) reacted and librarians across the country lobbied for laws that would protect user records in the libraries. Now, all 50 states have some type of law protecting library records

(Seaman, 2001). These laws vary in their scope and protection; however, internal use of user information for library management is usually permitted. Libraries can use patron records to support the mission of the library but third parties are usually proscribed from reviewing this information.

This stipulation causes a problem when groups of libraries work together, as is commonly done through consortium agreements. One of the advantages to data mining is that it can take advantage of large amounts of data. Library consortium groups can concentrate large amounts of information about behavior, far beyond what any individual library can collect. If state laws prohibit the sharing of user information with a third party, then consortium sharing of information is also prohibited. Laws may specify particular types of protected data. For example, data from acquisitions and bibliographic catalogs can usually be shared freely, but data connected to a user's behavior must remain protected. When a consortium crosses state lines, the problem compounds and each library must only share information allowed in their state.

Restrictions may also govern the granularity of analysis and distribution of analytical results. Illinois' state law, for example, allows the sharing of "reasonable statistical reports where those reports are presented so that no individual is identified" (Library Records Confidentiality Act, 1983). Conversely, Virginia's regulations for public libraries state that after the user returns an item, the record of that transaction is deleted. The record of a transaction may stay attached to an item, however, until no circulation of that item has taken place for a year (General Schedule No. 22, 1996). If this type of regulation exists, the library will lose valuable circulation information for good once the item is returned. If bibliomining becomes more popular in libraries, it is possible that librarians, libraries, and professional organizations may lobby to permit more comprehensive administrative analysis of circulation records and common federal legislation to govern these activities.

While bibliomining may aid librarians in the management of their library, the state-by-state patchwork of laws can make bibliomining tricky to implement. These laws may extend beyond public libraries, as well, into any library that is open to the public. As many university libraries participate in a Government Documents program, they may fall under the same legal restrictions as public libraries. Therefore, before engaging in bibliomining, librarians must ensure that their bibliomining activities fall within the boundaries of lawful activity.

Ethical Issues

"We protect each library user's right to privacy and confidentiality with respect to information sought or received and resources consulted, borrowed, acquired, or transmitted" (Code of Ethics of the American Library Association, 1995).

For years, librarians have looked at the behavior of users. In order to track in-house use, library employees scurry around behind users, tracking which books were used during a library visit. Researchers have explored OPAC searching logs for decades. Today, libraries still keep Web logs, proxy records, and other artifacts left behind from a user's visit (Pace, 2001).

Despite this history of "gentle intrusion," libraries are known as a place to conduct research privately. Librarians have kept prying eyes of outsiders away from the circulation records of users for years. As previously mentioned, librarians spearheaded the counterattack against Hoover's intrusive Library Awareness program. Users do research in the library with a sense of security that their research will not be scrutinized. Therefore, even when it is legal, it is important to scrutinize the ethicality of delving into user behavior with these tools. As a general guide, we recommend the Code of Fair Information Practices (HEW, 1972), which specifies principles and practices for the ethical handling of personally identifiable data about people. Appendix A provides a list of the five principles described in the Code.

In the past, the librarian had to depend upon surveys for gathering user information, but with bibliomining, they can discover similar patterns without wasting the user's time or the taxpayer's money on surveys (Estabrook, 1996). To do this, however, may require the suggestion and support of new legislation that will allow the circulation histories to be kept for library management purposes only.

In order to do this ethically, however, the library should develop, implement, and disseminate a privacy policy. The library must inform users of the intended use of the records, and at the very least, have a procedure available where users can opt out of the analysis. It would be more appropriate to get the permission of users before using their circulation records. It would take some time and money to gather the permissions through mail, the telephone, or in person, but would help to make the public aware of what is happening and ensure that only those users that wish to participate in the analysis are used. In addition, it would avoid a costly lawsuit if users felt that their rights were infringed in the analysis of their behaviors.

Another option, and perhaps a safer path on which to start, is to not examine any information that ties a particular user to a circulation history. This would reduce the power and the personalization available from bibliomining, but may be needed to avoid ethical and legal information. Public institutions may find that this is the only solution, while private libraries may have more flexibility in what they can do. Each library must balance their legal and ethical situation with the desire to better manage their library and better serve their users.

Conclusion

Libraries have gathered data about their collections and users for years, but have not often 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. In "Sacred trust or competitive opportunity: Using user records," Estabrook (1996) discussed this moral dilemma. She pointed that librarians must balance information protection with the need to create new library services (e.g., personalization functions). Now that libraries must compete against online booksellers, downloadable audio books, and the vast supply of "free" information of varying quality from the Internet, librarians must begin to take the initiative in using their systems and data for competitive advantage and to justify continued support and funding of libraries.

The process of using library data more effectively begins by discovering ways to connect the disparate sources of data most libraries create. Connecting these disparate sources in data warehouses can facilitate systematic exploration with different tools to discover behavioral patterns of the libraries primary constituencies. These patterns can help enhance the library experience for the user, can assist library management in making decisions and setting policies, and can assist the library's parent organization or community in understanding the information needs of its members.

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. Finally, bibliomining can inform the processes and products of knowledge management that have grown in importance within contemporary organizations.

References

Applegate, R. (1993). Models of user satisfaction: understanding false positives. *RQ*, *32*(4), 525-539.

American Library Association. (1995). *Code of ethics of the American Library Association.* Retrieved January 27, 2002 from http://www.ala.org/alaorg/oif/ethics.html

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

Banerjee, K. (1998). Is data mining right for your library? *Computers in Libraries, 18*(10), 28-31.

Bleyberg, M. Z., Zhu, D., Cole, K., Bates, D., Zhan, W. (1999). Developing an integrated library decision support warehouse. *IEEE international conference on systems, man, and cybernetics* (Vol. 2, pp. 546-551). Piscataway, NJ: IEEE.

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. (n.d.). *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

Estabrook, L. (1996). Sacred trust or competitive opportunity: Using patron records. *Library Journal, 121*(2), 48-49.

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.

Harter, S. P. & Hert, C. A. (1997). Evaluation of information retrieval systems: Approaches, issues, and methods. In M.E. Williams (Ed.), *Annual review of information science and technology* (Vol. 32, pp. 3-94). Medford, NJ: Information Today.

Hernon, P. & McClure, C. R. (1986). Unobtrusive reference testing: The 55 percent rule. *Library Journal, 111*(7), 37–41.

HEW (U.S. Dep't. of Health, Education and Welfare) (1973). Secretary's Advisory Committee on automated personal data systems, records, computers, and the rights of Citizens. Washington, DC: Government Printing Office.

Hildreth, C. (2001). Accounting for users' inflated assessments of on-line catalogue search performance and usefulness: An experimental study. *Information Research, 6*(2). Retrieved January 25, 2002, from http://InformationR.net/ir/6-2/paper101.html,

Jansen, B.J. & Spink, A. (2000, November/December). Methodological approach in discovering user search patterns through Web log analysis. *Bulletin of the American Society for Information Science,* 15-17.

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

Koenig, M. E. D. (1985). Bibliographic Information Retrieval Systems and Database Management Systems. Information Technology and Libraries, 4, 247-272.

Library of Virginia Records Management Division. (1996). *General Schedule No. 22*. Retrieved January 27, 2002 from http://www.lva.lib.va.us/state/records/schedule/gs%2D22.htm

Library Records Confidentiality Act, 75 ILCS §70(1983).

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.

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

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.

Pace, A. K. It's a matter of privacy. *Computers in Libraries, 21*(6), 50-52.

Patron confidentiality, millennium style [Electronic version]. (1999, June/July). *American Libraries, 30*, 86.

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

Sallis, P., Hill, L., Janee, G., Lovette, 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.

Seaman, S. (2001, October 27). *Confidentiality of library records.* Presentation at the Colorado Library Association Annual Meeting. Retrieved January 27, 2002 from http://spot.colorado.edu/~seaman/confidentialitylaws.htm

Sprain, M. (2001). Confidentiality in libraries. *Colorado Libraries, 27*(1), 36-38.

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

Su, S. & Needamangala, A. (2000). Harvesting information from a library data warehouse. *Information Technology and Libraries, 19*(1), 17-28.

Wormell, I. (2000). Informetrics – a new area of quantitative students. *Education for Information 18*, 131-138.

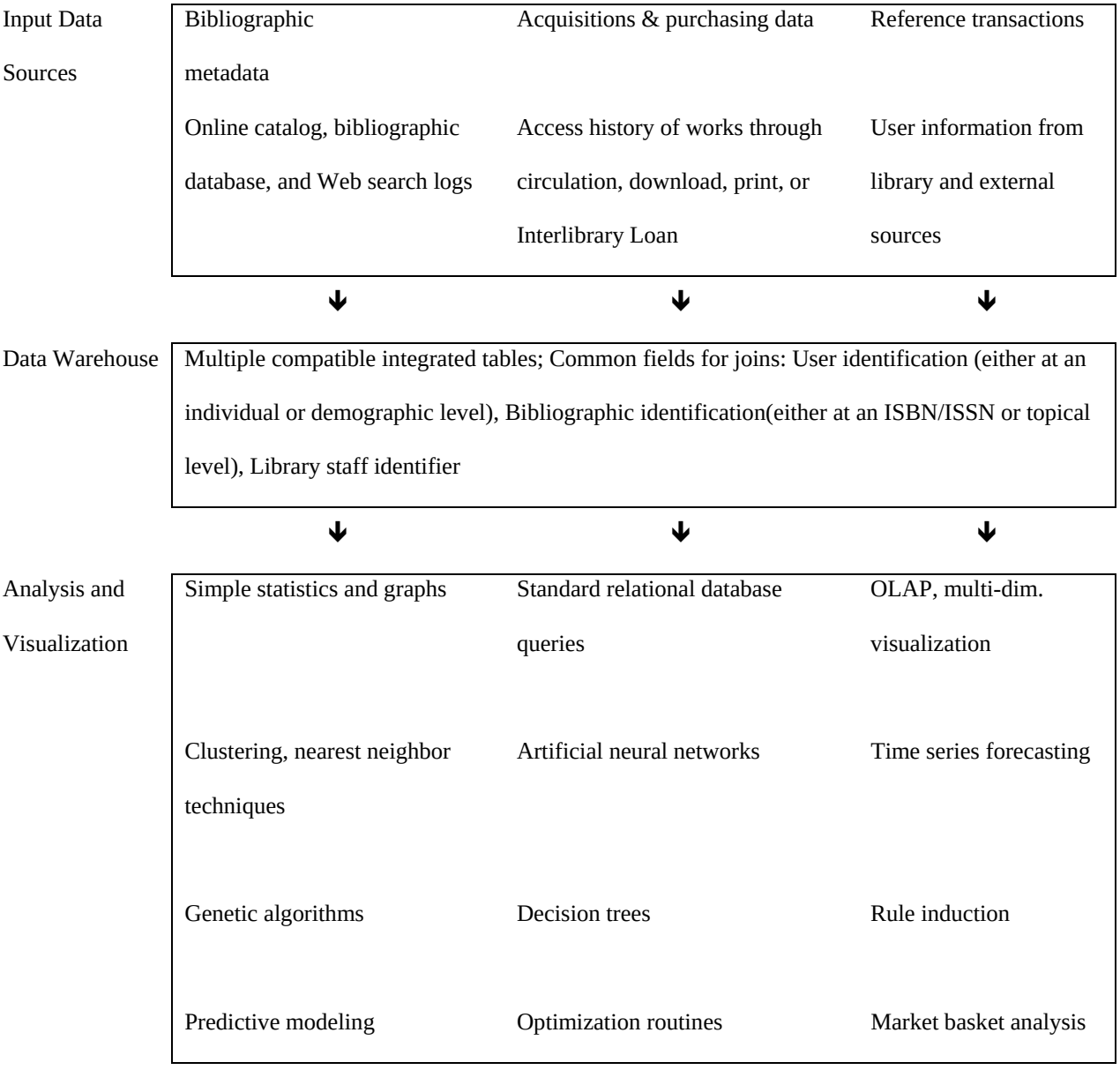
Zaiane, O. R., Xin, M., & Jiawei, H. (1998). Discovering Web access patterns and trends by trends by applying OLAP and data mining technology on Web logs. *IEEE international forum on research and technology advances in digital libraries* (pp. 19-29). Los Alamitos, CA: IEEE Computer Society.

Author Biographical Sketches:

Scott Nicholson is an assistant professor at Syracuse University's School of Information Studies. Dr. Nicholson has an extensive background both in data mining and library and information science, and has worked to combine librarianship with statistical and artificial intelligence methods from decision science in order to improve library management.

Jeffrey Stanton is an assistant professor at Syracuse University's School of Information Studies. Dr. Stanton is an organizational psychologist with two decades of professional and consulting experience putting technology to work in organizational settings. Dr. Stanton has published more than 20 peer reviewed journal articles and book chapters, about half of which pertain to research methods and statistics.

Figure 1: Typical Data Flow for Bibliomining Applications



Appendix: Principles from the Code of Fair Information Practices (HEW, 1972)

1. There must be no personal data record-keeping systems whose very existence is secret.
2. There must be a way for a person to find out what information about the person is in a record and how it is used.
3. There must be a way for a person to prevent information about the person that was obtained for one purpose from being used or made available for other purposes without the person's consent.
4. There must be a way for a person to correct or amend a record of identifiable information about the person.
5. Any organization creating, maintaining, using, or disseminating records of identifiable personal data must assure the reliability of the data for their intended use and must take precautions to prevent misuses of the data.