Nicholson, S. (2003) The Bibliomining Process: Data Warehousing and Data Mining for Library Decision-Making. *Information Technology and Libraries 22* (4).

The Bibliomining Process: Data Warehousing and Data Mining for Library Decision-Making

Scott Nicholson, Ph.D., MLIS, Syracuse University School of Information Studies, 4-127 Center for Science and Technology, Syracuse, NY 13026 srnichol@syr.edu
http://bibliomining.org

Introduction

One of the challenges put forth to the library profession by Michael Buckland (2003) is to gain a better understanding of library user communities. Most current library evaluation techniques focus on frequencies and aggregate measures; these statistics hide underlying patterns. Discovering these patterns is key in understanding the communities that use library services. In order to tailor services to meet the needs of different user groups, library decision-makers can use the bibliomining process to uncover patterns of data-based artifacts of use. The bibliomining process, which consists of data warehousing and data mining, will be explored in this brief article.

The term “bibliomining” was first used by Nicholson and Stanton (2003) in discussing data mining for libraries. In the research literature, most works that contain the terms “library” and “data mining” are not talking about traditional library data, but rather using library in the context of software libraries, as data mining is the application of techniques from a large library of tools. In order to make it more conducive for those concerned with data mining in a library setting to locate other works and other researchers, the term “bibliomining” was created. The term pays homage to bibliometrics, which is the science of pattern discovery in scientific communication.

Bibliomining is 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 bibliomining process consists of:

- determining areas of focus;
- identifying internal and external data sources;
- collecting, cleaning, and anonymizing the data into a data warehouse;
- selecting appropriate analysis tools;
- discovery of patterns through data mining and creation of reports with traditional

analytical tools; and

- analyzing and implementing the results.

The process is cyclical in nature: as patterns are discovered, more questions will be raised which will start the process again. As additional areas of the library are explored, the data warehouse will become more complete, which will make the exploration of other issues much easier.

Determining Areas of Focus

The first step in the bibliomining process is to determine the area of focus. This area might come from a specific problem in the library or may be a general area requiring exploration and decision-making. The first decision is to do *directed* or *undirected* data mining (Barry& Linoff 1997). Directed data mining is problem-focused: there is a specific problem that drives the exploration, e.g., “Budget cuts have reduced the staff time for contacting patrons about delinquent materials. Is there a way to predict the chance patrons will return material once it is one week late in order to prioritize our calling lists? ” Undirected data mining is used when a library manager wants to get a better idea of a general topical area; one possible area of focus would be: “How are different departments and types of patrons using the electronic journals?”

There are several dangers with undirected data mining, as this type of exploration involves the use of many tools repeatedly to locate any available patterns. Many data mining techniques are based on probability, and therefore there is a small chance that each pattern found is incorrect: the more times this type of tool is used, the greater the chance of finding an invalid pattern. Another problem arises with data mining tools that locate the most frequently occurring patterns: in a complex data set, the most frequent pattern may only occur a few times. Therefore, all patterns that are found will need to be explored to make sure they are valid, make sense to the librarians involved, and are not trivial. Undirected data mining may produce an overwhelming number of patterns to explore. Given the time it takes to collect and clean the data, undirected data mining should be considered only when a strong data warehouse is in place.

Identifying Data Sources

After determining the areas of focus, the next step is to identify appropriate data sources. The bibliomining process requires transactional, non-aggregated, low-level data. This can be made much more difficult, if not impossible, if librarians delete the operational data in an attempt to protect user privacy. For example, the *New York Times* recently reported the daily destruction of library records by public library staff members. (“Librarians Use Shredder to Show Opposition to New F.B.I. Powers”, April 4, 2003). However, the data warehouse can be used to both maintain a data-based history of the library while protecting the privacy of patrons; discarding everything makes it impossible to provide the data-based justification of the use of library services. As funding for public services continues to be cut, maintaining the ability to justify and defend the existence of library services is essential.

There are two types of data sources that should be considered. Internal data sources are those already within the library system. As most library systems have their data stored in different data silos around the organization (patron database, transactional data, Web server logs), discovering and extracting the internal data might be challenging.

External data sources are those that are not located within the library system. In an academic setting, these might include demographic information related to a specific ID number that is located in the computer center or personnel management system. For a public library setting, this might include demographic information for zip codes from census data.

Creating the Data Warehouse

Assuming that the data do exist, the challenge remains of combining different data sources into a single data warehouse that does not contain publicly identifiable information. A data warehouse is a database that is separate from the operational systems and contains a cleaned and anonymized version of the operational data reformatted for analysis. To create the warehouse, the librarian writes queries to extract the data from the identified sources, combines those data using common fields, cleans the data, and writes the resulting records into either a flat file or a relational database designed specifically for analysis (for more information on this data structure, see *Building the Data Warehouse* by Inmon, 2002). Once this procedure has been tested, it can be automated to pull data from the operational systems into the data warehouse on a regular basis.

Protecting Patron Privacy

On the surface, it may seem that the data warehouse will destroy patron privacy by combining different library systems into one source. The *New York Times* reported a recommendation that libraries delete all transactional data; this also makes it very difficult to evaluate and justify library services ("Librarians Receive Advice on Law and Reader Privacy”, Dec. 12, 2002). Others take no action, in which case the data lingers in operational systems and on backup tapes. Going through the data warehousing process requires the library to examine their data sources; by explicitly determining what to keep and what to destroy, libraries can save the demographic information needed to evaluate communities of users without keeping records of the individuals in those communities.

This extraction and cleaning process is the key to protecting patron privacy during data warehousing. As the records are drawn from internal and external systems, matches are made to connect data and then the personally identifiable information is discarded. This personal information should never be put into the data warehouse, so it will not be backed up, saved, or otherwise archived. After the data warehouse is created, the original data can then be deleted, in accordance with current advice to protect the privacy of patrons. The goal of data warehousing is to create a data source that contains decision-making information that cannot be used to recreate the original transactional records.

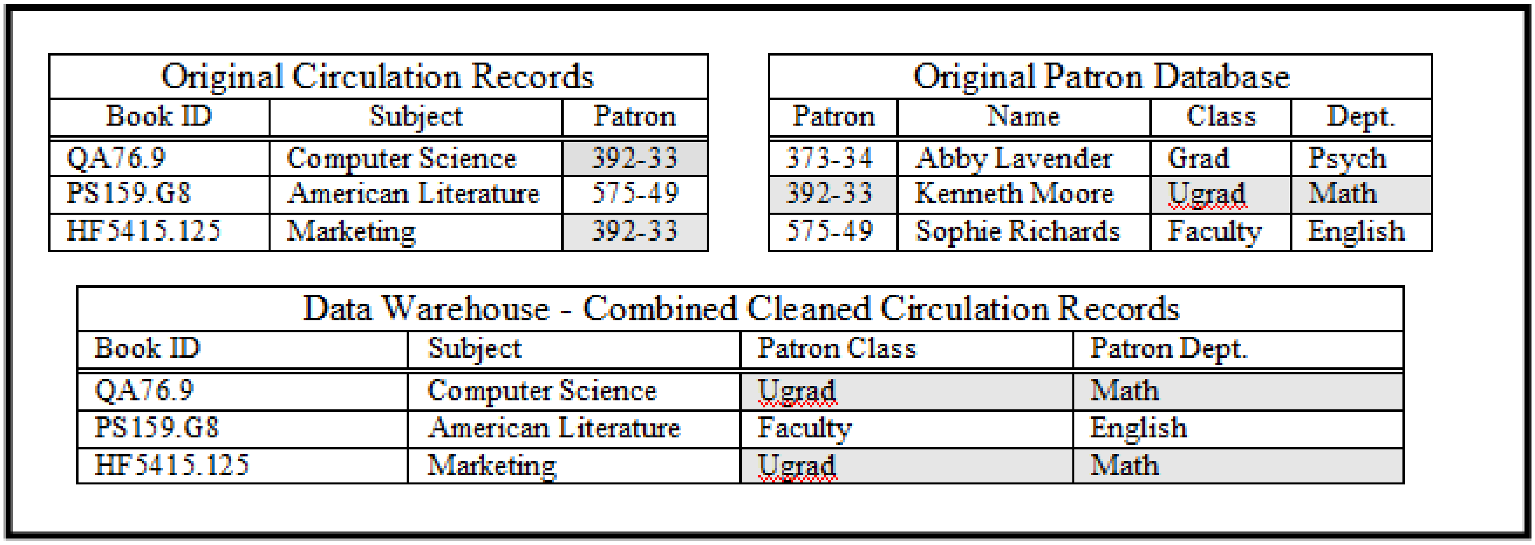
Here are two examples of the cleaning process that maintain important information for decision-making without keeping personally identifiable information:

- When an item is returned, many libraries delete all information about that transaction. However, there is valuable decision-making information that is lost. While the operational system is a user-focused data source, the data warehouse is an item-focused data source. Therefore, before deleting the transactional information, a record should be created in the data warehouse that combines information about the item with demographic information about the patron. This will capture the important information about the transaction without identifying the patron involved, and will still allow the discovery of patterns about patrons without compromising personally identifiable information (See Figure 1 – follow the shaded cells to see how data are replaced).

- The server transaction log contains invaluable information about how the Web-based services of the library are used. These logs contain IP addresses, which could be used to track the individual computers used to access these services. The useful information that does not track personally identifiable information are what pages were looked at in the same session. Instead of deleting the IP address field, some type of code should be used to replace each unique IP address. One way to protect privacy is to replace identical IP addresses within a certain time window with a code generated by the first time/date stamp associated with the IP. For example, an IP in the log registered at 10:32/10-29-02 could be coded as 102902-1032-A. Activity coming from that IP without a 15 minute break would then get the same code.
- In addition, some transaction logs will keep login information; just as with circulation records, these logins should be replaced with the matching demographic information about the user in the data warehouse (See Figure 2).

Whenever capturing demographic information, librarians must ensure that no combination of the demographic variables will lead to the identification of an individual. This should be explored in the development of these capturing tools; demographic groups should be combined or dropped until all combinations of the saved demographics lead to groups of patrons. (See figure 3) These demographic categories should be audited on a regular basis and adjusted as needed to represent the current user population.

Librarians are advised to consult with legal counsel before starting a data warehousing project. Privacy of patron records is currently legislated on a state-by-state basis in the United States; therefore, librarians may have to adjust these techniques to meet the laws of the state. Libraries attached to a research institution may need to get clearance from their Institutional Research Board before engaging in this type of examination.



Original Circulation Records

Book ID	Subject	Patron
QA76.9	Computer Science	392-33
PS159.G8	American Literature	575-49
HF5415.125	Marketing	392-33

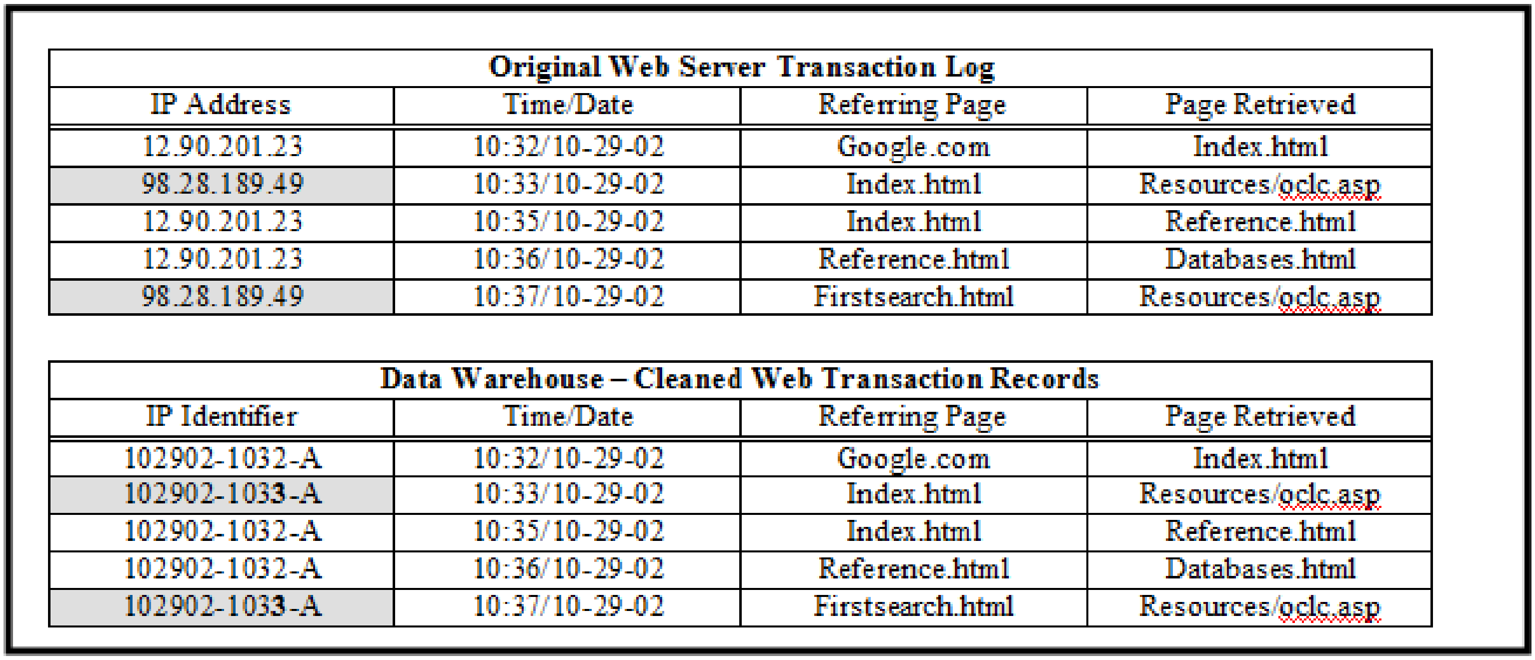
Original Patron Database

Patron	Name	Class	Dept.
373-34	Abby Lavender	Grad	Psych
392-33	Kenneth Moore	Ugrad	Math
575-49	Sophie Richards	Faculty	English

Data Warehouse - Combined Cleaned Circulation Records

Book ID	Subject	Patron Class	Patron Dept.
QA76.9	Computer Science	Ugrad	Math
PS159.G8	American Literature	Faculty	English
HF5415.125	Marketing	Ugrad	Math

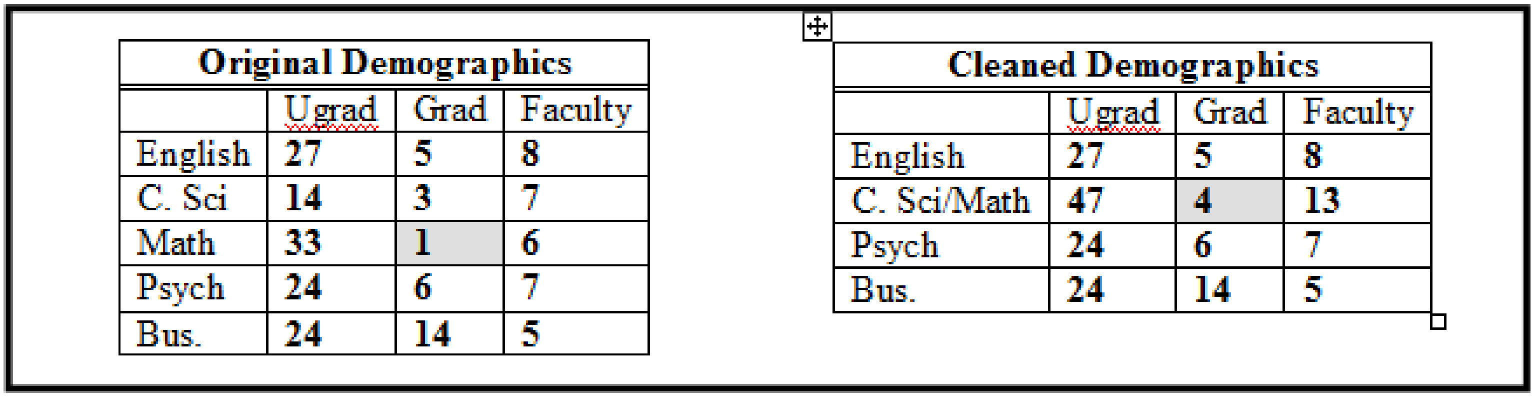
Figure 1: Cleaning Transactional Records



Original Web Server Transaction Log			
IP Address	Time/Date	Referring Page	Page Retrieved
12.90.201.23	10:32/10-29-02	Google.com	Index.html
98.28.189.49	10:33/10-29-02	Index.html	Resources/oclc.asp
12.90.201.23	10:35/10-29-02	Index.html	Reference.html
12.90.201.23	10:36/10-29-02	Reference.html	Databases.html
98.28.189.49	10:37/10-29-02	Firstsearch.html	Resources/oclc.asp

Data Warehouse – Cleaned Web Transaction Records			
IP Identifier	Time/Date	Referring Page	Page Retrieved
102902-1032-A	10:32/10-29-02	Google.com	Index.html
102902-1033-A	10:33/10-29-02	Index.html	Resources/oclc.asp
102902-1032-A	10:35/10-29-02	Index.html	Reference.html
102902-1032-A	10:36/10-29-02	Reference.html	Databases.html
102902-1033-A	10:37/10-29-02	Firstsearch.html	Resources/oclc.asp

Figure 2: Cleaning Web Server Transactional Records



Original Demographics			
	Ugrad	Grad	Faculty
English	27	5	8
C. Sci	14	3	7
Math	33	1	6
Psych	24	6	7
Bus.	24	14	5

Cleaned Demographics			
	Ugrad	Grad	Faculty
English	27	5	8
C. Sci/Math	47	4	13
Psych	24	6	7
Bus.	24	14	5

Figure 3: Collapsing Demographic Variables

Building the Data Warehouse

Building the data warehouse takes much more time than mining the data (Barry and Linoff 1997). This can make the process seem never-ending. It is essential that all involved realize that this portion of the process is an investment; once the programs for collection and cleaning are written, they will require only minor maintenance in the future as long as the library uses the same system. Because of this initial time investment, it is suggested to start with a narrowly defined bibliomining topic and work through the entire process. This iterative process also has the advantage of allowing those developing the data warehouse to improve their collection and cleaning algorithms early in the life of the bibliomining project.

Selecting appropriate analysis tools

Once the data warehouse has been developed, the stage is set for analysis. As mentioned earlier, traditional aggregations and ratios can be easily calculated to create traditional reports. However, interesting and useful patterns may be hiding behind these aggregate measures; these patterns allow library managers to better understand their individual user groups. For example, use of Interlibrary Loan data for collection development can be problematic due to the short-term needs of some ILL users. Bibliomining could be used to look for patterns in ILL use in order to remove the short-term ILL needs from consistent collection deficiencies.

Management Information Systems (MIS)

These systems provide a manager with the ability to ask basic questions of the data. Many Integrated Library System (ILS) packages have some type of basic MIS built in. However, the data powering these system comes from the operational systems and have not been cleaned nor matched to any external data. In addition, most databases attached to libraries are treated as independent, unconnected data silos. Therefore, an MIS built on top of a data warehouse made for the library will be much more powerful and provide information that the library needs to see (instead of what the ILS vendor wants them to see).

Another addition to the MIS is a critical factor alert system. As managers and administrators work with the reports produced from the data, key variables that provide the “pulse” of the library can be identified. Automatic notification programs can be attached to these variables, so that if they stray outside of a specified zone, the managers can be automatically notified of a potential issue. For example, if hourly circulation is below or above a certain level, a manager could be immediately notified so staffing changes could be made.

On-Line Analytical Processing (OLAP)

OLAP puts an interactive view of the data on the decision-makers desktop. Under the surface, the OLAP tool has run thousands of database queries to combine all of the selected variables along with all of the selected measures. These reports are presented in a easy-to-understand menu-driven system. The user will pick one of many variables from a list to examine, such as use of e-journals. They can then select basic dimensions, such as time and subject. The OLAP system will then present a high-level view of this data in a tabular report, perhaps by year and general classification. The user can then click one of the dimensions to expand the report – in our example, if the user clicked on a year, the tool would expand the year into quarters, leaving the subject headings the same and recalculating the data. The user can then click on another field to “drill down” into the data.

OLAP tools are powered by a data warehouse. All of the fields are defined ahead of time, and the system runs many queries before anyone uses it. Therefore, response to the manager using the OLAP front end for reports is instant, which encourages exploration. During exploration, the manger can capture any view of the data and turn it into a regular report.

Data Mining

Another way to analyze the data, and part of the inspiration of the term “bibliomining,” is data mining. The goal of data mining is the discovery of valid, novel, and actionable patterns in large amounts of data using statistical and artificial intelligence tools (Berry and Linoff 1997). Data mining came out of corporate America in the early 1990’s after data warehousing became popular; companies had stored large amounts of data and looked for ways to take advantage of these warehouses.

There are two main categories of data mining tasks: description and prediction. In description, the goal is to understand the data from the past and the present. The patterns discovered are used to seek out “affinity groups” of variables common to different patrons or clusters of demographic groups that exhibit certain characteristics. On the other hand, prediction is used to make a statement about the unknown based upon what is known. It can be used to predict the future or to make statements about the present. The two types of tasks for prediction are classification, where the goal is to place an item into a category, or estimation, where the goal is to produce a numeric value for an unknown variable. The task selected will dictate the tools chosen.

Data Mining Software Packages

There are a number of software packages that provide access to different data mining tools. For most of these packages, one imports the data in a flat file or other common data format. Once the data have been imported, the analyst can implement different tools and explore results. The statistical packages SAS and SPSS both have optional data mining programs, although these can be quite expensive. An open-source data mining tool is Weka (Witten and Frank 1999), which contains many of the same options at a much more reasonable cost; however, it is not as user-friendly.

The simplicity of running model after model holds a danger. Many of the individual tools are based on probabilistic statistical methods. The methods require certain types of data, such as data from a normal distribution, in order to provide accurate results. The danger comes in that the tools do not have any type of check for these requirements built in to the system. Therefore, an unsuspecting user can run the tools on the data and the tools will provide an answer; however, that answer may not be meaningful if the data do not meet certain assumptions. To learn more about the individual algorithms in a data mining package, one can refer to the texts by Barry and Linoff (1997, 2000) or the text by Witten and Frank (1999) that refers to the free Weka data mining suite.

Analysis and Implementation

Once the results, either static reports or decision-making models and aids, have been developed, they must be validated. The first step is to test the data on a sample of the data that was not used to build the model, in order to test the robustness of the model on different data. The most important validation, however, is to have a librarian who is familiar with that library context examine the models. The librarian should concur with the results found; the patterns can be the data-based validation of the tacit knowledge gained from working in a library situation. If the librarian disagrees with a pattern, it is essential to explore the data streams carefully; this type of disagreement usually stems from a flaw in the data or a misapplication of a tool.

The final step is to implement the report/model. It is essential to monitor the variables that power the models over time; if the mean of a variable strays too far because of changes in the library, the model may have to be reevaluated.

Conclusion

The goal of this brief article was to explain the bibliomining process. Emphasis was placed on data warehousing and patron privacy issues because they are required before anything else can begin. It is essential to capture our data-based institutional records but still protect the privacy of users. By using a data warehouse, both goals can be met. Once the data warehouse is in place, the library can use a plethora of reporting and exploration tools to gain a more thorough knowledge of their user communities and resource utilization.

Literature Cited

Buckland, M. 2003. Five grand challenges for library research. *Library Trends* 51(4). [cited 27 June 2003]. Available from http://www.sims.berkeley.edu/~buckland/trends03.pdf.

Berry, M. and G. Linoff. 1997. *Data Mining Techniques For Marketing, Sales, and Customer*

Support. New York: John Wiley & Sons.

Berry, M. and G. Linoff. 2000. *Mastering Data Mining*. New York: John Wiley & Sons.

Inmon, W. 2002. *Building the Data Warehouse, 3rd Edition.* New York: John Wiley & Sons.

Nicholson, S. and J. Stanton. 2003. “Gaining strategic advantage through bibliomining: Data mining for management decisions in corporate, special, digital, and traditional libraries.” In *Organizational Data Mining: Leveraging Enterprise Data Resources for Optimal Performance*, ed. H. Nemati & C. Barko. Hershey, PA: Idea Group Publishing.

Witten, I. and E. Frank. 1999. Practical Machine Learning Tools and Techniques with Java Implementations. San Francisco, CA: Morgan Kaufmann.