Data Mining and Data Warehousing

SPEC Kits

Supporting Effective Library Management for Over Thirty Years

Committed to assisting research and academic libraries in the continuous improvement of management systems, OLMS has worked since 1970 to gather and disseminate the best practices for library needs. As part of its committment, OLMS maintains an active publications program best known for its SPEC Kits. Through the OLMS Collaborative Research/Writing Program, librarians work with ARL staff to design SPEC surveys and write publications. Originally established as an information source for ARL member libraries, the SPEC series has grown to serve the needs of the library community worldwide.

What are SPEC Kits?

Published six times per year, SPEC Kits contain the most valuable, up-to-date information on the latest issues of concern to libraries and librarians today. They are the result of a systematic survey of ARL member libraries on a particular topic related to current practice in the field. Each SPEC Kit contains an executive summary of the survey results (previously printed as the SPEC Flyer); survey questions with tallies and selected comments; the best representative documents from survey participants, such as policies, procedures, handbooks, guidelines, Web sites, records, brochures, and statements; and a selected reading list—both print and online sources—containing the most current literature available on the topic for further study.

Subscribe to SPEC Kits

Subscribers tell us that the information contained in SPEC Kits is valuable to a variety of users, both inside and outside the library. Purchasers use the documentation found in SPEC Kits as a point of departure for research and problem solving because they lend immediate authority to proposals and set standards for designing programs or writing procedure statements. SPEC Kits also function as an important reference tool for library administrators, staff, students, and professionals in allied disciplines who may not have access to this kind of information.

SPEC Kits can be ordered directly from the ARL Publications Distribution Center. To order, call (301) 362-8196, fax (301) 206-9789, email <pubs@arl.org>, or go to <http://www.arl.org/pubscat/index.html>.

Information on SPEC Kits and other OLMS products and services can be found on the ARL Web site at <http://www.arl.org/olms/infosvcs.html>. The executive summary or flyer for each kit after December 1993 can be accessed free of charge at the SPEC survey Web site <http://www.arl.org/spec/index.html>.

SPEC Kit 274

Data Mining and Data Warehousing

July 2003

Barbara Mento

Virtual Data Center Manager
Boston College

Brendan Rapple

Collection Development Librarian
Boston College

ASSOCIATION OF RESEARCH LIBRARIES

OFFICE OF LEADERSHIP AND MANAGEMENT SERVICES

Series Editor: Lee Anne George

SPEC Kits are published by the

Association of Research Libraries
Office of Leadership and Management Services
21 Dupont Circle, NW, Suite 800
Washington, D.C. 20036-1118
(202) 296-2296 Fax (202) 872-0884
<http://www.arl.org/olms/infosvcs.html>
<pubs@arl.org>

ISSN 0160 3582

ISBN 1-59407-606-5

SPEC Kit 274

Data Mining and Data Warehousing

July 2003

SURVEY RESULTS

REPRESENTATIVE DOCUMENTS

SELECTED RESOURCES

SURVEY RESULTS

EXECUTIVE SUMMARY

Introduction

Libraries have a history of taking advantage of technological innovation. In the last few years these efforts have been used to push the boundaries of digital libraries and electronic publishing. Most recently, professional librarians are exploring new tools for information management and analysis. The ARL E-Metrics project is exploring ways of standardizing and collecting data on the use and value of electronic resources. Libraries are becoming partners in university repositories of full text documents and other e-works.

Recently there have been a growing number of research articles in library literature in such journals as *Journal of the American Society of Information Science*, *Library Trends* and *Information Technology in Libraries* on data mining, text mining, and Web mining describing how libraries are using associated software for primarily administrative purposes. While these tools have shown a dramatic increase in use in businesses over the last ten years, academia and libraries have only recently recognized the value of these tools as decision support systems in analyzing collection and Web usage. Text mining, in particular, is growing in popularity as a research method at universities.

The goal of this survey was to determine the extent to which data mining technology is being used by ARL member institutions, researchers, libraries, and administrations. The survey also hoped to elicit ideas and opinions concerning the potential role of libraries in supporting data mining and data warehousing in research institutions. The first seven survey questions focus on data mining and data warehousing activities at the institutional level. The remaining questions explore the current library use of data mining technology and opportunities for future use. Since data warehouses are the foundation of data mining, several questions focused on current support and future plans for data warehousing.

The survey was sent to 124 ARL member libraries. Sixty-five (52%) responded to the survey.

Data Mining and Warehousing at Institutions

Based on the definitions of data mining and data warehousing described in the survey, twenty-six (40%) respondents indicated that data mining technology was used at their institution. A majority of the respondents at libraries where data mining was not used (38 or 90%) believe that data mining could be a valuable tool to facilitate library users in the future. Thoughts about possible future roles for libraries in this area centered around three areas: research support, library administration, and repository management. Eight of the responding libraries are currently involved in or planning for repositories at their institutions. Repositories were described as including administrative data, research data, and full text sources such as dissertations, books, archival documents, and local newspapers. Several institutions concluded that these large repositories of full text and numeric data would offer data mining opportunities that would benefit from the expertise found in libraries. Most frequently, future roles for librarians were described

as data warehouse managers and experts in content analysis.

The survey results also indicate that data mining has been integrated into the curriculum at many academic institutions. The scope of disciplines utilizing data mining is very broad, with the heaviest concentrations in the social sciences. Forty-two (71%) of the fifty-nine universities who replied to this question noted that data mining or data warehousing courses are currently being offered. Forty-three respondents listed the departments/schools where data mining is taught. The two predominant academic departments that host these courses are business/management (29 respondents or 67%) and computer science (25 or 58%). Statistics (11 or 26%) and information science (10 or 23%) were mentioned next, followed by economics (4 or 9%) and sociology (3 or 7%). Several areas were mentioned in the "Other" category including biology and various engineering specialties. Education and psychology were also added, while several institutions listed these classes as sponsored by other units on campus such as information technology services and research centers.

Operational Responsibility

Only fifteen (62%) of the twenty-four respondents to this question listed units with operational responsibility for data mining. Most of the data mining and data warehousing operations report to traditional information technology units of various titles. Data mining at institutions was described several times as a distributed responsibility under several units or academic departments. Other unit titles responsible for data mining use were: Data Administration, Data Warehousing Team, Office of Institutional Research and Planning, Administrative Applications, and a Statistical Consulting Center. One institution indicated that joint operational responsibility belonged to both the library and the computing group.

Data warehousing was listed by twenty-one (87%) respondents as under the purview of Information Technology Services or Administrative Information Systems. One institution is part of a multi-university data resources system. Several institutions have very specialized units such as Data Warehouse/Data Administration and a Data Warehouse Team.

Reporting Structures

Operational responsibility was most frequently listed as being under the auspices of the chief information officers, chancellors (or vice chancellors), and provosts (or associate provosts). Other operations report to offices of the academic vice-presidents, financial affairs & information technology, president, Sr. vice-president, business & finance division, and, at one institution, the university librarian.

Support Issues

Support issues for two aspects of data mining and data warehousing were explored by the survey: technical support and content support. Thirty respondents supplied details regarding the group(s) that provide support for data mining at their institution. The number of institutions identified as providing technical support for data mining is fairly small. At eighteen institutions (60%) support is provided by information technology services, at ten (33%) by the library, and at nine (30%) by academic departments. Five (17%) respondents indicated other groups. Sixteen (53%) institutions noted that data mining was not supported. Content support revealed even smaller numbers with ten (33%) using information technology services, eight (27%) the library, eleven (37%) academic departments, and four (13%) other groups. Fifteen respondents (50%) indicated no data mining content support.

A larger group, thirty-nine respondents, described support for data warehousing. Technical support was identified by twenty-nine respondents (74%) as the responsibility of information technology services, fourteen (36%) the library, eight (21%) academic departments, and nine (23%) other groups. Eight (21%) said it was not supported. Content support was identified by eighteen respondents (46%) as the responsibility of information technology services, eighteen (46%) the library, fourteen (36%) academic departments and nine (23%) other groups. Eight (21%) said it was not supported.

Users and Software

Thirty-three respondents provided details regarding the users of data mining technology at their institution. The survey shows faculty as the most frequent

users at 76% (25). They are followed closely by researchers (70%), graduate students (64%), administrative staff (56%), library staff (39%), undergraduates (30%), and others (9%).

Institution-wide the most utilized data mining software are Clementine and Enterprise Miner, both of which work with existing statistical software—SPSS and SAS respectively—which are currently used by many universities and libraries. Also, these two data mining tools are the ones most frequently used by libraries. Other data mining software reported by the respondents include, in descending order by number of users: TextAnalyst, Oracle Data Mining Suite, IBM Intelligent Miner, Arrowsmith, and WebAnalyst. Of the other software that was reported, some fall into the data warehousing category: Survey Documentation Analysis, iVia, Mercator, WebCrawler, Microsoft Analysis Services, Cuadra Star, ArcGis, SAS Text Miner, SPSS Answer Tree, Data Beacon, Phrase Rate, Cognos, Brio Query (under development), mySQL, NCSA D2K (Data to Knowledge), and MicroStrategy. In addition to the Brio Query tool under development, tools are being developed for digital library initiatives such as Project Prism.

Library Applications

Research

Nine libraries (14%) stated that they utilize data mining technology in order to facilitate library users' research. For example, iVia software, an open source Internet subject portal, was created to build INFOMINE, a virtual library of evaluated Internet resources, the content of which is appropriate for the university level. A couple of libraries mentioned that they use data mining technology to analyze GIS data. Another mentioned a data services unit that facilitates users' analysis of social sciences computing data. Yet another referred to its use of a suite of data mining tools to perform text analysis and clustering in large Open Archives Initiative metadata repositories.

Administration

On the other hand, twice as many respondents (18 or 29%) use data mining technology to facilitate their library's own administrative or strategic purposes. Included among the topics or areas mentioned: circulation statistics and patterns; electronic resource usage; collection development and acquisitions decisions; Web site organization; analysis of users' use of the Web; evaluation of the usage of both on-site and off-site resources; door count statistics; subscription database logs; and analysis of interlibrary loan data for copyright compliance.

Library Established Data Warehouses

The survey revealed that a number of libraries have established their own data warehouses which may then be mined for diverse information. Some are research data warehouses and include a digital library of multi-formatted data covering many subjects; a warehouse formed by a collaboration of three universities and composed of statistical, ICPSR, and international financial data; and a warehouse of tens of thousands of Internet resources and associated metadata. The contents of the administrative data warehouses include circulation, acquisitions and other transaction data, patron information, human resource and financial data, and bibliographic information. One institution observed that their library management system was their data warehouse.

Web Content Mining and Software

The survey inquired whether libraries were engaged in any form of Web content mining. Eight libraries (13%) replied positively. More libraries (33 or 53%) declared that they were involved in aspects of Web usage mining. These included identifying Web usage patterns, analyzing Web information architecture, locating linking and other problems, and ascertaining which pages, databases, etc. are being used the most. Five libraries specifically mentioned that they utilized WebTrends software. Other software cited included NetTracker, Sawmill, Cognos, and CREP. A number of libraries mentioned that analysis of statistics of Web usage was helpful in modifying pages in order to render them more user-friendly and effective. Another library stated that they mine their Web, OPAC, and search engine logs to determine how precisely the systems are used, presumably with the aim of making appropriate changes.

Benefits from Data Mining and Warehousing

Many survey respondents replied that a number of benefits have resulted from library data mining and/or data warehousing operation(s). Gains in research productivity were paramount. For example, a consortium of three libraries has developed a data warehouse of social science data to enhance users' learning and research. Another respondent pointed to their library's custom-created software and its crawler/classifiers that greatly improve the gathering and subsequent evaluation of relevant and quality Internet resources. Another stated that its data mining operations have spawned new research. Some libraries mentioned benefits that have resulted in the administrative sphere from their data mining and/or data warehousing operation(s). They have helped in making better serials cancellation, collection development, budget, workflow, collection weeding, OPAC design, and Web development decisions; in evaluating databases and other resources; in determining user needs; in monitoring system performance and usability; in developing forecasts; in making policies; and in improving Web security. One respondent referred to the benefit of gaining technical expertise in organizing a large digital archive. Another mentioned that Web log data mining can point to areas where users might benefit from instruction in using the particular search tools.

Staffing, Training, and Budget

The survey revealed that the number of staff allotted to data mining and/or data warehousing operations is generally small. Only twenty-four libraries responded that they had any staff assigned to these areas. These ranged from four libraries that declared that their allotted staff was less than 0.5 FTE to one library with 6.5 FTE. As might be expected, the position titles of staff working in these areas ranged widely. Among the titles were: systems analyst, systems librarian, analyst programmer, Web applications developer, project manager, library software engineer, serials assistant, collection development coordinator, budget analyst, and inventory copy cataloguer. A complete list of the title(s) of these staff, the departments in which they work, and the title of the person to whom they report may be found in the Survey Questions and Responses.

The survey revealed that library staff have gained their knowledge of or expertise in data mining and/or data warehousing from a variety of sources and experiences. 92% specified on-the-job experiences; 54% attendance at conference presentations; 54% product demonstrations; 27% parent institution-sponsored workshop(s). A smaller proportion, 15%, mentioned library-sponsored workshop(s), 12% undergraduate degree classes, and 12% graduate degree classes respectively. Other respondents gained their knowledge or expertise from Data Warehouse Institute classes; library literature; interviews with practitioners; work experience in system development; courses offered by Cognos; usage of SPSS software for data statistics; the WebTrends data analysis tool; and Iplanet Indexing and Metadata retrieval software.

Only 10 libraries, or 15% of the respondents, devote a specific budget to support data mining and/or data warehousing operations. The amount ranges from a low of $150 to a high of $250,000. The average was $57,000, the median $30,000.

Evaluation

The evaluation of a service or activity is generally an important consideration for libraries. The survey respondents reported using the following techniques and/or measures to evaluate the effectiveness of their library's data mining and/or data warehousing activities. Fifteen libraries mentioned informal feedback; twelve, usage data; ten, Web logs; eight, user surveys; seven, focus groups. Five libraries reported other evaluation techniques ranging from software such as WebTrends Web log analysis, PERL script utilization counts, and NIST METRICS software to usability studies. Twelve libraries stated that they were not presently engaged in data mining or data warehousing activities.

Additional Comments about the Survey

Several comments indicated that some institutions had difficulty in answering the questions. One institution is so decentralized that they found it difficult to attribute responsibilities for some areas. There was also a sense that some reported projects were not technically data mining. Some responses may be referring

to more traditional statistical analysis rather than data mining.

Six of the seventeen institutions which supplemented their answers with additional comments described current and planned university projects for warehousing administrative records for data in such areas as finance, human resources, enrollments, and user surveys. One of these efforts involves a three-university collaborative using PeopleSoft software to gather financial, human resources, payroll, and student systems data.

A few libraries clarified that library data warehousing referred to integrated library systems collecting circulation and other data. One library discontinued analysis of collection use patterns because of concerns for patron confidentiality. WebTrends is used by several libraries to collect and analyze Web site data.

There appears to be a trend in developing partnerships, with university units, other universities, and with external organizations, primarily government agencies such as the Department of Education and NOAA/NASA. These partnerships include developing warehouses and digital collections, which has led to collaboration in developing data mining tools. The goal of institutions in adopting data mining technology is reported as primarily for the purpose of supporting research and improving the quality of service to patrons.

Conclusion

This survey confirmed some hypotheses developed by the survey authors that were based on a review of data mining literature and on interviews with authors of some key articles treating data mining and libraries. Libraries are discovering, as businesses have, the value of merging existing data or full text resources to form a very large data warehouse that can be mined for analytical purposes. Libraries that are using data mining are primarily doing so for such administrative purposes as facilitating the collection and analysis of, for example, circulation, acquisition, Web usage, and other diverse patron data. The aim is generally to strengthen library decision-making and the library's own internal operations. The survey highlights a growing participation by libraries in creating such data warehouses.

While the major activities in data mining reside in academic departments and involve academic research, there is an awareness of and enthusiasm for the possibilities of data mining as a tool for ARL libraries. Moreover, a few libraries discussed a growing vision of how data mining technology can be used as an effective resource to facilitate scholarly research as well as administrative processes.

In conclusion, libraries are taking a leadership role in creating and managing data warehouses for both administrative and research purposes. Based on this survey, librarians recognize data mining techniques as offering new approaches to analyzing content and knowledge discovery within these very large databases and the Web. In addition, more widespread availability of data mining software provides a new avenue for libraries to explore data mining's potential in both academic research and decision-making.