Data Mining: Another Tool for Librarians

Lourdes T. David

Presented on Feb. 27-28, 2007 at City Garden Suites, Ermita, Manila as part of the MAHLAP National Congress and Seminar-Workshop “Cutting Edge Information Professional: A Mini-Radical Sabbatical for Medical and Health Librarians.”

Introduction

- Digital library services have changed the way library users look for and access information.
- Library patrons now use both the physical and the virtual libraries.
- Library managers are thus faced with management issues in both types of libraries.
- The question “how can library service be improved to better serve their users,” remains a primary issue for administrators.

Aims of the Presentation

- Explore the concept of data mining and relate it to digital library systems as a tool for measuring how users seek for and access information and how they can be better served.
- Provide some examples where data mining can help the library make decisions/answer queries based on evidence.
- Present some data mining tools

Scope of the Presentation

- Definition and scope of data mining
- Data mining methods
- Data mining tools
- Applications in libraries
- Limitations and challenges

“Introduction to Data Mining for Libraries Online Workshop”

- *A four week workshop that provides an introduction to data mining.*
- *The course defines “Data mining” as the practice of automatically searching large stores of data for discerning meaningful patterns and/or predicting future trends.*
- *The course looks at : sources of data for libraries; basic data mining strategies and tools to extract knowledge from these sources; and current projects and strategies*

Data Mining and Libraries

- Data mining is just emerging as a library practice. Scott Nicholson of Syracuse University's School of Information Studies, NY, even coined the term "bibliomining" because a literature search with the phrase "data mining and libraries" doesn't return hits about data mining in libraries.

Goal of Data Mining

- “The goal of data mining is to explore the dataset for patterns that are novel and useful.”
- Data mining can be directed, where there is a particular goal or topic area in mind, or undirected, where the goal is to find something interesting
- These patterns are then seeds for more thorough explorations (Nicholson).

Goal of Data Mining

- According to Dean Unsworth,”By contrast, the goal of data-mining, including text-mining, is to **produce new knowledge** by exposing unanticipated similarities or differences, clustering or dispersal, co-occurrence and trends…”

Importance of Data Mining

- Data mining uncovers patterns in data using predictive techniques. These patterns play a critical role in decision making because they reveal areas for process improvement.
- Using data mining, organizations can increase the profitability of their interactions with customers, detect fraud, and improve risk management. The patterns uncovered using data mining help organizations make better and timelier decisions.

What is Data Mining?

- data mining (′dad·ə ′mīn·iŋ or dād·ə ′mīn·iŋ)
- (*computer science*) The identification or extraction of relationships and patterns from data using computational algorithms to reduce, model, understand, or analyze data. The automated process of turning raw data into useful information by which intelligent computer systems sift and sort through data, with little or no help from humans, to look for patterns or to predict trends.
- http://www.answers.com/topic/data-mining

What is Data Mining?

- Exploring and analyzing detailed business transactions. It implies "digging through tons of data" to uncover patterns and relationships contained within the business activity and history. Data mining can be done manually by slicing and dicing the data until a pattern becomes obvious. Or, it can be done with programs that analyze the data automatically.

What is Data Mining?

- Data mining is an **artificial intelligence (AI) powered tool** that can discover useful information within a database that can then be used to improve actions.
- Essentially, data mining discovers patterns and relationships hidden in your data. It's part of a larger process called knowledge discovery.

What is Knowledge Discovery?

- **Knowledge Discovery and Data Mining (KDD)** is an interdisciplinary area focusing upon methodologies for **extracting useful knowledge from data** to deliver advanced business intelligence and web discovery solutions.
- The knowledge-discovery process as a whole is essential for successful data mining because it describes the steps you must take to ensure meaningful results." (Edelstein, 1997)

What is Data Mining?

- The use of computers to find **patterns in masses of corporate data** to provide greater customer satisfaction and lower-cost service.
- Example: The IBM centers in Kawasaki, Japan, and Raleigh, North Carolina, are trying out a **"text mining" system** that uses natural-language algorithms to sift through written summaries that call operators make after each call. Developed by a group of researchers under Tetsuya Nasukawa at IBM's Tokyo Research Laboratory (TRL), the system is unearthing trends in customers' questions and complaints.

Text Mining System

- "Ordinary data mining simply looks for keywords, but the text-mining system -- dubbed **TAKMI (an abbreviation for Text Analysis and Knowledge Mining but also a Japanese word meaning 'skilled craftsman')** -- spots grammatical relationships, as well. Knowing which word is the subject, which the verb, and which the object, TAKMI can categorize calls according to whether they are, say, complaints or questions and according to the product that is causing difficulty.“ (IBM, 1999)

The Theoretical Basis of Data Mining

- The challenge of extracting knowledge from data draws upon research in statistics, databases, pattern recognition, machine learning, data visualization, optimization, and high-performance computing.
- Using a combination of machine learning, statistical analysis, modeling techniques and database technology, data mining finds patterns and subtle relationships in data and infers rules that allow the prediction of future results. (Twocrows)

What is Data Mining?

- Data mining is the automatic discovery of patterns, associations, anomalies and changes in data. (Grossman)
- Initial work in digital libraries focused on the archiving, searching, and retrieval of documents.
- As digital libraries supporting this basic functionality become widely available, algorithms and software for the analysis and mining of information available within digital libraries were developed

What is Data Mining?

- A class of database applications that look for hidden patterns in a group of data that can be used to predict future behavior. The term is commonly misused to describe software that presents data in new ways. True data mining software doesn't just change the presentation, but actually discovers previously unknown relationships among the data.
- It is utilized increasingly by marketers trying to distill useful consumer data from Web sites.

Data Mining Parameters

- Association - looking for patterns where one event is connected to another event
- Sequence or path analysis - looking for patterns where one event leads to another later event
- Classification - looking for new patterns (May result in a change in the way the data is organized but that's ok)
- Clustering - finding and visually documenting groups of facts not previously known
- Forecasting - discovering patterns in data that can lead to reasonable predictions about the future (This area of data mining is known as predictive analytics.)

The Data Warehouse

- In both the traditional and digital library settings, data is collected and stored.
- In an Integrated Library system, libraries often find that their transactional data is locked up in the ILS "black box" (, **the data warehouse)** and can only be viewed with vendor-supplied reports
- In general report generation is done using ready made report formats.
- With data mining tools, reporting becomes flexible and faster.

Mining Tabular Data

- Prediction--Given enough tabular data, a predictive model can be generated to predict the numerical value of a designated field, given a new row with that field element missing.
- Classification--With enough data, a classification model can be generated to predict the satisfaction of a customer, given the other attributes.
- Clustering--This provides a natural means of grouping together similar rows
- Anomalies--These types of data mining queries return rows which are in some specified sense different than the other rows, such as those representing a statistically significant change from the previous rows.

Mining Textual Data

- Conceptual clustering--Queries can return documents which conceptually related to a specified reference document or to a specified topic. Unlike key word searches, a specified word need not appear for the document to be retrieved, simply a specified concept.
- Attribute based associations--association queries can look for patterns involving the author, subject, date, or type of publication which the owner tends to associate with higher value. These patterns can then be used to customize the articles retrieved in response to a query.
- Anomaly detection--A report which is significantly different can be flagged.

Challenges: Issues and Problems

- An important challenge is to develop algorithms and software to work with more complex data, including time series, random fields, unstructured data (such as text), and semi-structured data (such as textural files containing embedded formatting commands).

Challenges: Issues and Problems

- *Scalability* . Most general purpose data mining algorithms are limited to mining data which can fit in memory.
- *Distributed and Agent-based mining.* Today, data mining is generally done by collecting data into a centralized digital library or repository and mining it for information. The third challenge is to develop appropriate distributed and agent-based data mining algorithms to mine distributed data.

Applications

- Whether we like it or not, our daily lives are influenced by data mining applications.
- For example, almost every financial transaction is processed by a data mining application to detect fraud.

Examples of Applications

- Corporations employ data mining to analyze operations, find trends in recorded information, and look for new opportunities. Libraries are no different. Librarians manage large stores of data—about collections and usage, for example—and we also want to analyze this data to serve our users better. (Cullen, 2005)

Examples of Applications

- Recently, Katharine Treptow Farrell and Marc Truitt wrote that the Princeton University Library could not answer the question, "What journals do we subscribe to that [a specific company] publishes and what exactly are their price increases over previous years?" It's a fundamental management question, yet the answer is elusive. (Farrel and Truitt)

Examples of Applications

- Dan Walters, executive director of the Las Vegas–Clark County Library District, NV, points out the need to take this kind of question further by mapping data from the ILS against other sources. For example, Walters would like to know what kinds of ebooks are moving and to what types of users. But currently he can't map circulation by media formats against types of users.

Examples of Applications

- Sam Clay, director of the Fairfax County Public Library, VA, also believes that libraries need to use more quantitative decision-making tools. Working as assistant to a city manager convinced him. "[We] have to apply business models to what we do," Clay says. "Don't think, feel, or intuit. Do it because of what you know." His library has gathered roughly 20 years of trend data that it uses to make decisions in an aggressive, entrepreneurial way.

Data Mining in the LIbrary

- Librarians want a solution that will integrate many types of information into a system that allows them to analyze their usage, expenditures, customer base, collections, and more. The ILS is just one source of that information, though it is clearly the largest source—and ILS vendors are the best situated to create data-mining solutions

Vendor Solutions

- Some vendors see the need for data mining, while others are more cautious.
- Dynix, now SirsiDynix
  - Developed **Web Reporter**, which runs as a standalone application for both Horizon and Corinthian users.
  - Teamed up with SwiftKnowledge to create its **Director's Station** product, a data-mining and analysis tool that enables directors to make management decisions re funding and expenses
  - Normative Data Project (NDP), a way to merge data about many different libraries into a normalized database to see industry norms.

Vendor Solutions

- Innovative Interfaces customers are limited to Report Writer's use of data within the ILS. Innovative's XML Server has some potential to ease the process of automated data extraction, but it is currently limited to bibliographic and authority data.
- Endeavor's partnership with business intelligence vendor Cognos is focused on its Meridian electronic resources management (ERM) product.

Vendor Solutions

- Voyager uses
  - PowerPlay cubes which are static data, so staff don't hit the live Voyager system and reduce response time for users. Because the system is optimized for analytical data rather than transactions, they usually get subsecond response time on highly complex queries.
  - PowerPlay's "Drill Through" allows you to view the items that actually meet the criteria, rather than just the number."

Institutional Solutions

- The University of Pennsylvania's Data Farm tries to have a body of unprocessed data available for "when the question comes up, without knowing the question in advance. "
- The Penn Data Farm contains a wide variety of information about the organization and its activities, transactions, and users.
- It allows the library to find elusive answers that allow the library to serve its users better, spend funds more efficiently, or just do more with less. In the future it could be used to perform more workflow analysis.

Commercial Solutions: Quest Group

- The Intelligent Information Systems Research (aka. Quest) group designs information systems that enable the preservation of the privacy and ownership of data while not impeding the flow of information.

Projects

- The Andrew W. Mellon Foundation is funding the two-year, nearly $600,000 multi-institutional project, which John Unsworth, dean of Illinois' Graduate School of Library and Information Science (GSLIS), will lead. In his winning project, titled "Web-based Text-Mining and Visualization for Humanities Digital Libraries"

Aims of the Project

- The Project expects to produce software “for discovering, visualizing and exploring significant patterns across large collections of full-text humanities resources in digital libraries and collections.”
- This is in line with the goal of data-mining, including text-mining, is to produce new knowledge by exposing unanticipated similarities or differences, clustering or dispersal, co-occurrence and trends.”

Data Mining Tools/Software

- Most analysts separate data mining software into two groups:
  - data mining tools-- Provide a number of techniques that can be applied to any business problem.
  - data mining applications--Embed techniques inside an application customized to address a specific business problem.
- Organizations are using data mining tools and data mining applications together in an integrated environment for predictive analytics.

Role of Data Mining Tools

- Data mining tools are used to ensure flexibility and the greatest accuracy possible.
- Essentially, data mining tools increase the effectiveness of data mining applications. Since no two organizations or data sets are alike, no single technique delivers the best results for everyone.

Data Mining Standards and Guides

- Because data mining tools are so flexible, a set of data mining guidelines and a data mining methodology have been developed to help guide the process. The Cross-Industry Standard Process for Data Mining (CRISP-DM) ensures your organization's results with data mining tools are timely and reliable. This methodology was created in conjunction with practitioners and vendors to supply data mining practitioners with checklists, guidelines, tasks, and objectives for every stage of the data mining process.

Examples of Data Mining Tools

- You can get a lot of data mining done without much capital outlay if you're willing to dedicate internal IT resources to the task.
- Data mining development kits that will help you build your own data mining applications are available commercially or as free/open source

Examples of Data Mining Tools

- AC2 (from Isoft), a set of librairies for building data mining solutions on the server side.
- Clementine Solution Publisher generates embedded SQL and C code for processing and modeling steps performed in Clementine.
- Data Intelligence Add-In for Excel, integrates the most popular algorithms found in data mining and AI into Excel.
- Decision Science SDK, from Stone Analytics enables developers to embed predictive models in their applications.
- GEPSR, a COM component for integrating Gene Expression Programming into custom applications.

Examples of Data Mining Tools

- KnowledgeSTUDIO SDK, offering enterprise application developers access to an extensive API and library of data mining components
- KXEN Components, based on Vapnik Support Vector Machines theory.
- K.wiz, open Java data mining and knowledge discovery platform providing a full API and extensive range of data mining components.
  K.wiz Application Enterprise enables building packaged analytical applications using HTML, XML and Javascript.
- The LPA Data Mining Toolkit provides an embeddable collection of routines which support the discovery of association rules within RDMS.
- Microsoft OLE DB for Data Mining, providing a low-level interface for data mining operations. Also available are OLE DB Data Mining Sample Provider and Visualization Controls

Examples of Data Mining Tools

- Microsoft OLE DB for Data Mining, providing a low-level interface for data mining operations. Also available are OLE DB Data Mining Sample Provider and Visualization Controls
- MLF: machine learning framework for Mathematica, the multi-method system for creating understandable computational models from data
- NAG Data Mining Components, statistics and machine learning components for data cleaning, transformation and model building -- for creating applications with data mining functionality.
- Neusciences aXi.DecisionTree and aXi.Kohonen, ActiveX Controls for building a decision tree and Kohonen Clustering. Includes a Delphi interface.

Examples of Data Mining Tools

- PolyAnalyst COM, an SDK offering a large selection of machine learning algorithms as separate COM components for simple integration in external applications.
- TextAnalyst COM, an SDK for building intelligent applications with semantic analysis, summarization, clustering, categorization, and retrieval of texts and fragments.
- WizWhy-OCX, ActiveX control includes all the functions of WizWhy rule-finding product
- XAffinity(tm), ActiveX toolkit (DLL) for association and sequential analysis in SQL databases.
- XELOPES, an open platform-independent and data-source-independent library for Embedded Data Mining.

Free or Shareware Data Mining Tools

- ADaM, Algorithm Development and Mining version 4.0 toolkit
- AlphaMiner, open source data mining platform that offers various data mining model building and data cleansing functionality.
- Databionic ESOM Tools, a suite of programs for clustering, visualization, and classification with Emergent Self-Organizing Maps (ESOM).
- *f*Data Mining Template Library (DMTL), an open-source collection of generic algorithms and data structures for mining complex patterns, including Itemsets, Sequences, Trees and graphs.

Free or Shareware Data Mining Tools

- Gnome Data Mining Tools, including apriori, decision trees, and Bayes classifiers.
- IBM Intelligent Miner. University scholars can now receive free copies of DB2 UDB and Intelligent Miner for educational or research purposes.
- KNIME, extensible open source data mining platform implementing the data pipelining paradigm (based on eclipse).
- MiningMart, a graphical tool for data preprocessing and mining on relational databases; supports development, documentation, re-use and exchange of complete KDD processes. Free for non-commercial purposes.

Free or Shareware Data Mining Tools

- MLC++, a machine learning library in C++.
  See also Kansas State U. port of MLC++: Binary (tar.gz), and Linux source
- Machine Learning in Java (MLJ), an open-source suite of Java tools for research in machine learning.
- Orange, C++ components for data mining,includes preprocessing, modelling and data exploration techniques.
- Rattle, a data mining suite based on open source statistical language R, includes graphics, clustering, modeling, and more.

Free or Shareware Data Mining Tools

- Sav Z, Java(TM) API language for developing high-performance mobile object-relational database applications (an improved JDBC). Free download!
- StarProbe, Web-based multi-user server available for academic institutions.
- Superinduction, based on SPSS Clementine and other methods.
- TANAGRA, offers a GUI interface and methods for data access, statistics, feature selection, classification, clustering, visualization, association and more.

Free or Shareware Data Mining Tools

- VFML (Very Fast Machine Learning) library for mining very large databases and data streams. Written in C, it includes highly scalable implementations of several widely used machine learning algorithms and tools for data preparation, testing, and rapid development of stream mining systems.
- Weka, collection of machine learning algorithms for solving real-world data mining problems (in Java).

Free or Shareware Data Mining Tools

- XML Miner, offers a class library for mining data and text expressed in XML, extracting knowledge and re-using that knowledge in products and applications in the form of fuzzy logic expert system rules. Weka, collection of machine learning algorithms for solving real-world data mining problems. It is written in Java and runs on almost any platform.
- YALE (Yet Another Learning Environment), flexible and modular environment for machine learning, including classification, regression, and clustering; XML config; and more

Conclusion

- Until vendors come up with relatively easy solutions, data mining probably isn't for every library. When the time comes to consider a system, Fairfax's Clay offers some advice. Think about what you need to know more about. "What are your needs and requirements? Use that as a template for judging products. Finally, is it worth the investment?" Many, however, would argue that it's an investment we can no longer ignore.

References

- Nicholson, Scott. The Basis for Bibliomining: Frameworks for Bringing Together Usage-Based Data Mining and Bibliometrics through Data Warehousing in Digital Library Services.
- The Economist Technology Quarterly (June 10, 2004).
- Boyd, John. Mining for trends at the help desk. IBM Think Research (1999).
- Knowledge Discovery & Data Mining Research at IBM.

  http://www.twocrows.com/glossary.htm
- Edelstein, Herb Data Mining: Exploiting the Hidden Trends in Your Data. DB2 Online Magazine (Spring 1997).

References

- Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From Data Mining to Knowledge Discovery in Databases. AI Magazine 17(3): Fall 1996, 37-54.
- **Robert L. Grossman** Mining Challenges for Digital Libraries
- Nancy Cohen Data Mining: Nagging That It Really Adds Up
- Cullen, Kevin. Delving into Data. Library Journal August 15, 2005
- http://www.pcmag.com/encyclopedia_term/0,2542,t=data+mining&i=40813,00.asp

References

- "The Case for Acquisitions Standards in the Integrated Library System," *Library Collections, Acquisitions, and Technical Services*, 27.(4), p. 483–492.
- Lyn, Andrea. News bureau, University of Illinois Urbana Champaigne.
- California Computer News (October 27, 2004).
- http://www.webopedia.com/TERM/D/data_mining.html (January 20, 2007) http://searchcrm.techtarget.com/sDefinition/0,,sid11_gci211901,00.html (July 21, 2007)
- http://www.answers.com/topic/data-mining