A Study of Data-Mining in University Libraries

Dr. SANGITA PUROHIT
Head Librarian, Amrut Mody School of Business Management -
Ahmedabad University, India

ABSTRACT
As University Libraries transform their role from passive data collection to a more active exploration and exploitation of information, they face a serious challenge: how can they handle a massive amount of data that the institutions generate, collect and store. There is a need to have a technology that can access, analyze, summarize, and interpret information intelligently and automatically. Responding to this challenge, the field of data mining has emerged in the field of Information Technology.

Data Mining is defined as the process wherein large volumes of data are automatically converted from unknown data to useful data which can be utilized in a manner one wants to utilize. Recent advances using neural network and other statistical data extraction tools have resulted in significance increase in the input data size of Data mining utilization. However it is observed that data mining systems is unable to maintain same rate of growth.

KEYWORDS –
Data mining , University libraries , Information Technology, Software tools.

INTRODUCTION
Data mining technology is a powerful tool which make raw unknown data into an useful and understandable and actionable form data which can further utilized for future trends. Earlier this technique was limited to scientific research and for other medical diseases diagnostic. Today this technique is more popular in marketing and business. It is also becoming popular in university libraries for managing the needs of the library users. It is therefore becoming more complex incorporating additional functions compared to earlier. As a result there is a need for faster execution of these algorithms for its optimization need . Since the change in complexity of data mining algorithms is on higher side, large numbers of data are now collected each year in an exponential way. Data mining is essential to extract useful information from such large amounts of data. Execution time is equally important to extract data and converting them into useful data as a result a need have arises to redesign and mining data to customize systems as required by a customer. This technique is new and hence a very little knowledge is available to librarians. These performance gap between data mining systems and algorithms can be reduced by thorough understanding the system applications , followed by design of computer technology to fulfill the primary need of data mining work in university libraries.
CONCEPTS OF DATA MINING IN UNIVERSITY LIBRARIES

In order to find out actionable information from large quantities of data which are raw/not actionable, Data mining technique is being employed in the university libraries. It utilizes mathematical/statistical analysis to obtain trends/charts/tables etc that exists in the collected data. It is not possible to get these patterns using traditional data exploration as the relationships are much more complex and since there are too much data. These type of data analysis in the form of charts/trends/tables can be obtained and defined as data mining model. It can be applied mainly to specific situations, such as

- Finding sequences
- Forecasting:
- Grouping:
- Recommendations:
- Risk and probability

The model for data mining in university libraries to built is a part of large process which includes everything from asking required questions about the data creation and a model to answer all required questions, involving model into a working environment. These process of building model can be defined by using following steps.

- **Defining the Problem** undertaken
- **Preparing Data** for same
- **Exploring Data** for it
- **Building Models** for the problem
- **Exploring and Validating Models** prepared
- **Deploying and Updating Models** prepared

CATEGORIES OF DATA MINING TOOLS

Most data mining tools can be classified into one of three categories: traditional data mining tools, dashboards, and text-mining tools. Below is a description of each.

TRADITIONAL DATAMINING SOFTWARE.

The software program made in the Traditional form helps university library to establish data pattern and data trends by using number of complex algorithms and data mining techniques. In order to highlight available trends and other captured information not residing in the database are installed on the computer screen. These are available in many operating systems including Microsoft Windows version. Most of the software can handle any data using online analytical processing or similar other technology utilized.

DASHBOARDS TYPE DATAMINING SOFTWARE.

In order to enable user to find out how the university library is performing, the software installed in the computer desktop to monitor information in database, the dashboards indicates data changes and the data is being updated on the monitor screen in the form of chart/table/graphs. This type of function makes dashboards easy to utilize for over viewing the performance of the university library by higher authorities of the library.

TEXT MINING SOFTWARE TOOLS.

This is the third type data mining tool which has capabilities of data mining from various types of data in text form. These data could be a simple Microsoft word, Acrobat PDF or some other operating systems like Mozilla Firefox. This text mining tool first scan the available data which could be either structured or unstructured and convert the selected data into the format which is compatible with the tools database. As a result the librarian
Data mining can access data easily and in a convenient way for utilizing in a manner required. Collecting these available inputs university librarian gets wealth of information that can be minded to find trends, concepts and attitudes.

When evaluating data mining strategies, universities may decide to acquire several tools for specific purposes, rather than purchasing one tool that meets all needs. Although acquiring several tools is not a mainstream approach, a company may choose to do so if, for example, it installs a dashboard to keep managers informed on business matters, a full data-mining suite to capture and build data for its marketing and sales arms, and an interrogation tool so auditors can identify fraud activity.

WHAT CAN BE ACHIEVED

Data mining can be used to achieve many types of problems. Based on the type of the knowledge to be discovered, it can be broadly divided into supervised discovery and unsupervised discovery, it can be broadly divided into supervised discovery and unsupervised discovery. The former requires the data to be pre-classified. Each item is associated with a unique label, signifying the class in which the item belongs. In contrast, the latter does not require pre classification of the data and can form groups that share common characteristics.. To achieve these two main tasks, four data mining approaches are commonly used: classification, clustering, association rules, and visualization.

These tasks have been implemented in many application domains. The main application domains that data mining can support in the field of information science.

<table>
<thead>
<tr>
<th>Section</th>
<th>Application</th>
<th>Contributions</th>
</tr>
</thead>
</table>
| 1       | Personalized Environments | • To adapt content presentation and navigation support based on each individual’s characteristics.  
• To understand users’ access patterns by mining the data collected from log files.  
• To tailor to the users’ perceived preferences by matching usage and content profiles. |
| 2       | Electronic Commerce | • To divide the customers into several segments based on their similar purchasing behaviour.  
• To explore the association structure between the sales of different products.  
• To discover patterns and predict future values by analysing time series data. |
| 3       | Search Engine | • To identify the ranking of the pages by analysing the interconnections of a series of related pages.  
• To improve the precision by examining textual content and user’s logs.  
• To recognize the intellectual structure of works by analysing how authors are cited together. |
VARIOUS SOFTWARE TOOLS FOR DATAMINING

The development and application of data mining algorithms requires the use of powerful software tools. As the number of available tools continues to grow, the choice of the most suitable tool becomes increasingly difficult, however following is the list of various software tools available for data mining.

ADAPA (Zementis)  www.zementis.com
Alice (d’Isoft)  www.alice-soft.com
Bayesia Lab  www.bayesia.com
C5.0  www.rulequest.com
CART  www.salford-systems.com
Data Applied  data-applied.com
DataDetective  www.sentient.nl/?dden
DataEngine  www.dataengine.de
Datascope  www.cygron.hu
DB2 Data Warehouse  www.ibm.com/software/data/infosphere/warehouse
DeltaMaster  www.bissantz.com/deltamaster
Forecaster XL  www.alyuda.com
GhostMiner  www.fqs.pl/business-intelligence/products/ghostminer
IBM SPSS Modeler  www.spss.com/software/modeling/modeler
IBM SPSS Statistics  www.spss.com/software/statistics
iModel  www.biocompsystems.com/products/imodel
InfoSphere Warehouse  www.ibm.com/software/data/infosphere/warehouse
JMP  www.jmpdiscovery.com
KnowledgeMiner  www.knowledgeminer.net
KnowledgeStudio  www.angoss.com
KXEN  www.kxen.com
Magnum Opus  www.giwebb.com
MATLAB  www.mathworks.com
MATLAB Neural Network Toolbox  www.mathworks.com
Model Builder  www.fico.com
ModelMAX  www.asacorp.com/products/mmxxover.jsp
Molegro Data Modeler  www.molegro.com
NAG Data Mining Components  www.nag.co.uk/numeric/DR/DRdescription.asp
NeuralWorks Predict  www.neuralware.com/products.jsp
Neurofusion  www.alyuda.com
Neuroshell  www.neuroshell.com
Oracle Data Mining (ODM)  www.oracle.com/technology/products/bi/odm/index.html
Partek Discovery Suite  www.partek.com/software
Partek Genomics Suite  www.partek.com/software
PolyAnalyst  www.megaputer.com/polyanalyst.php
PolyVista  www.polyvista.com
Random Forests  www.salford-systems.com
RapAnalyst  www.raptorinternational.com/rapanalyst.html
R-PLUS  www.experience-rplus.com
SAP Netweaver Business Warehouse (BW)  www.sap.com/platform/netweaver/components/businesswarehouse
SAS Enterprise Miner  www.sas.com/products/miner
See5  www.rulequest.com
SPAD Data Mining  eng.spadsoft.com
SQL Server Analysis Services  www.microsoft.com/sql
SuperQuery  www.azmy.com
Teradata Database  www.teradata.com
Think Enterprise Data Miner (EDM)  www.thinkanalytics.com
TIBCO Spotfire  spotfire.tibco.com
Unica PredictiveInsight  www.unica.com
WizRule and WizWhy  www.wizsoft.com
XAffinity  www.exclusiveore.com
BETTER UNDERSTANDING OF DATA-MINING.
In my previous part of this research paper, an effort is being made to understand as to how data-Mining in the university Libraries are done, However to better understand the process of Data Mining at large including its usability, I have referred to Oracle data mining tool, which is very much popular in industries segments at large. This data-Mining tool enables users to discover new insights hidden in data and leverage investments in oracle database technology with which one can build and apply predictive models that help target best customers, develop detailed customer profiles and find and prevent fraud. With the use of Oracle data miner, one can not only access data, also give results of predictions, recommendations and discovering.

In order to understand in terms of technology, following diagram is prepared with its workflow using SQL developer.

In order to get details about the data mining software, Oracle data mining concept is being reproduced below for ready reference of the reader of this article.

Oracle Data Miner work flows GUI accelerates development and deployment of predictive analytics methodologies

Also for the use of Data mining in various fields of life, a case study prepared by By Stephen Langdell, PhD, Numerical Algorithms Group for the use of data mining in Financial applications is reproduced below for the ready reference of the reader of this article.

In general, data mining methods such as neural networks and decision trees can be a useful addition to the techniques available to the financial analyst. However, the data mining techniques tend to require more historical data than the standard models and, in the case of neural networks, can be difficult to interpret.

Stock market returns and foreign currency exchange rates

Data can be thought to fall into one of four categories as follows.

1. Five time series: index value at open, index value at close, highest index value, lowest index value and trading volume.
2. Fundamental factors: e.g., the price of gold, retail sales index, industrial production indices, foreign currency exchange rates.
3. Lagged returns from the time series of interest.
4. Technical factors: variables that are functions of one or more time series, e.g., moving averages.
The standard approach to modelling stock market returns or exchange rates is to model the univariate time-series with autoregressive (AR) and moving average (MA) models. A trader can determine an appropriate number of lags for AR and ARMA models based on experience and by analysing the time series data. Similarly, an appropriate number of regimes for SETAR (self-exciting transition AR) and STAR (smooth transition AR) models can be determined. These models are deterministic in the sense that they attempt to use mathematical equations to describe the process that generates the time-series. The advantage of these models lies in their interpretability.

Another approach, drawn from data mining, is to adopt a model that is flexible in the sense that it can approximate a wide class of functions with high accuracy. Such models are non-parametric in the sense that there need not be a direct relationship between the parameter values of a fitted model and the data. The advantages of using such a model include

1. The ability to model highly complex functions.
2. The ability to use a high number of variables in the model and, therefore, to include other data (i.e., fundamental and technical factors) in addition to lagged time series data.

The disadvantage of non-parametric models is that they are not easy to interpret.

In the case of data mining time series data, the model of choice is a neural network. By adjusting the number of free parameters associated with a model, a trader controls its flexibility. Often, cross-validation, or hold-out data, is used to determine a suitable value for the number of free parameters contained in a neural network structure. The neural network most commonly used in financial applications is a multi-layer perceptron (MLP) with a single hidden layer of nodes.

The problem of predicting stock market returns or exchange rates at time \( t+1 \) can be cast as either a regression or classification problem. Whereas the regression problem for exchange rate data involves modelling the actual exchange rate, the classification problem involves predicting whether the exchange rate has increased or decreased.

Applications that involve modelling returns from the stock market include portfolio management and trading futures (see below).

**MLP regression example: portfolio management**

The regression case involves predicting the (raw) return values. Such predictions can be used to manage a portfolio of \( n \) stocks as follows.

Suppose that historical data for \( N \) (\( N > n \)) stocks is used to fit \( N \) multi-layer perceptrons. At the end of each week the MLPs are re-fitted to include the latest historical data. For example, suppose company Z's pension fund has been managing a portfolio of $100 million since December 1993 using multi-layer perceptrons. The fund monitors a pool of 1,000 U.S. stocks on a weekly basis. For each of these stocks there is a MLP which models the future performance of the stock as a function of the stock's exposure to 40 fundamental and technical factors, and gives an estimate of its weekly price change. The company then selects a portfolio of the top \( n \) stocks and allocates the fund proportionately to predicted returns.

**MLP classification example: trading futures**

As an example of a data mining classifier, consider the problem of trading a future of stock A at price B on date C by using a neural network.
Firstly, the historical data is prepared. At each time step, data is classified into one of two categories according to whether it was profitable to buy or sell stock A at price B on date C:

1. Long: buy the stock on date C.
2. Short: sell the stock on date C.

Having fitted a model with this historical data, the model can be used to predict a profitable position at time t+1 (e.g., the next day or week). At the end of each time step the model is updated to include the new historical data.

By the time date C arrives, the trader should be in a profitable position (either long or short) given the current market value of stock A.

**Trading rules**

Trading rules can be determined from data with a categorical outcome, e.g., buy or sell, rise or fall. Such rules take the form of a set of conditional statements and an action, e.g.,

\[
\text{IF CONDITION1 AND CONDITION2 THEN ACTION}
\]

and can be found by viewing a fitted decision tree model. Given an appropriate historical data set, this approach could be used to either validate a rule thought to exist, or generate new rules and ideas.

Suppose that a decision is fitted using historical data. Each internal node in the tree is a test on one of the variables used to predict the outcome in the historical data. If the variable, say X1, takes continuous values, this test is either

\[
(X1 \geq VALUE) \quad \text{or} \quad (X1 < VALUE),
\]

where X1 and VALUE are determined by the algorithm that fits the decision tree. If the variable, say X2, can take one of m discrete values, this test is one of

\[
\{(X2 = i)\}, \quad \text{for } i=1, 2, m,
\]

where X2 and i are chosen by the fitting algorithm. Leaf nodes contain the actions, e.g., buy or sell. Thus tracing down the tree from the root to each leaf node will give a set of rules.

For example, historical data could be collected and a decision tree built to test the validity of the following rule.

"When the 10-day moving average crosses above the 30-day moving average and both moving averages are increasing it is time to buy"

**Component Approach**

Analysts seeking new insights from massive databases of tick data and similar market information can now accelerate their development of customized data mining applications by using NAG Data Mining Components as building blocks for their applications. The NAG Data Mining Components are expected to free quantitative analysts from needing to re-invent basic data mining routines, enabling the development of specialized data mining applications more quickly and at lower cost. Such exhaustive exploration of financial databases is expected to give organizations a competitive advantage in designing new derivative products, discovering asset mis-pricings, and in identifying similar routes to maximize portfolio returns.

The NAG Data Mining Components can be used for each stage of the modelling process - data preparation (case wise deletion and dummy variables generation), data transformation (principal component analysis and data scaling), and model building (k-
means and hierarchical clustering; \( k \)-nearest neighbours; decision tree analysis; multi-layer perceptron neural networks; logistic regression; general multiple regression). Application developers can use these components in their own application designs using standard development tools.

**CONCLUSION**

Data mining is the extraction of useful patterns and relationships from data sources, such as databases, texts, the web... Using data mining to understand and extrapolate data and information can reduce the chances of fraud, improve audit reactions to potential business changes, and ensure that risks are managed in a more timely and proactive fashion. Auditors also can use data mining tools to model "what-if" situations and demonstrate real and probable effects to management, such as combining real-world and business information to show the effects of a security breach and the impact of losing a key customer.

**REFERENCES**

1. Frawley, W., Piatetsky-Shapiro, G., Matheus, C. J., “Knowledge Discovery in Database an Overview,” in knowledge Discovery in Databases, G. Piatetsky-Shapiro and W. Frawley, (Eds.), MIT Press, 1991
4. Carbone, Patricia L., Data Mining or “Knowledge Discovery in Databases” : An Overview, MITRE Corporation, 1997

***

**Mrs. Sangita Purohit** is employed by the Ahmedabad Education Society (AES Post Graduate Institute of Business Management) located in the state of Gujarat –INDIA. Having obtained her Masters degree in Library Science from the university of Sagar-INDIA she is now pursuing Ph.D in Library-Science and has 9 years experience in the field of Library-Science in INDIA. Mrs. Purohit has attended several National / International seminars in INDIA, presented several papers in the field of Library-Science and conducted one seminar in DIGITAL-LIBRARY.

**Ahmedabad University** (AU) is a self-financed non-affiliating university promoted by the Ahmedabad Education Society (AES). Established in the year 1936 by visionaries like Kasturbhai Lalbhai, Ganesh Mavalankar and other luminaries, AES has remained a benchmark for high academic standards for more than seven decades. It is the largest education society in Western India which has spread the light of knowledge and wisdom in diverse fields of education and established new benchmarks of excellence.

AU is committed to the pursuit of inquiry and to provide better access to the students of Gujarat to professional and general higher education which is contemporary in scope and career oriented in content. AU will be the new star in the education landscape of Gujarat. The University is committed to delivering global quality education in the Indian context and ethos.