Data Mining

Ihssan Alkadi, (Email: ixa6880@louisiana.edu), University of Louisiana at Lafayette,

ABSTRACT

Recently data mining has become more popular in the information industry. It is due to the availability of huge amounts of data. Industry needs turning such data into useful information and knowledge. This information and knowledge can be used in many applications ranging from business management, production control, and market analysis, to engineering design and science exploration. Database and information technology have been evolving systematically from primitive file processing systems to sophisticated and powerful databases systems. The research and development in database systems has led to the development of relational database systems, data modeling tools, and indexing and data organization techniques. In relational database systems data are stored in relational tables. In addition, users can get convenient and flexible access to data through query languages, optimized query processing, user interfaces and transaction management and optimized methods for On-Line Transaction Processing (OLTP). The abundant data, which needs powerful data analysis tools, has been described as a data rich but information poor situation. The fast-growing, tremendous amount of data, collected and stored in large and numerous databases. Humans can not analyze these large amounts of data. So we need powerful tools to analyze this large amount of data. As a result, data collected in large databases become data tombs. These are data archives that are seldom visited. So, important decisions are often not made based on the information-rich data stored in databases rather based on a decision maker's intuition. This is because the decision maker does not have the tools to extract the valuable knowledge embedded in the vast amounts of data. Data mining tools which perform data analysis may uncover important data patterns, contributing greatly to business strategies, knowledge bases, and scientific and medical research. So data mining tools will turn data tombs into golden nuggets of knowledge.

DEFINITION OF DATA MINING

ata mining refers to extracting or mining knowledge from large amounts of data. Many people treat data mining as a synonym for another popularly used term, Knowledge Discovery in Databases or KDD. Alternatively, others view data mining as simply an essential step in the process of knowledge discovery in databases. Knowledge discovery as a process is depicted in the following figure, and consists of an iterative sequence of the following steps:

1. **data cleaning** to remove noise or irrelevant data.
2. **data integration** where multiple data sources may be combined.
3. **data selection** where data relevant to the analysis task are retrieved from the database.
4. **data transformation** where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance
5. **data mining** an essential process where intelligent methods are applied in order to extract data patterns
6. **pattern evaluation** to identify the truly interesting patterns representing knowledge based on some interesting measures.
7. **knowledge presentation** where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

The data mining step may interact with the user or a knowledge base. The interesting patterns are presented to the user, and may be stored as new knowledge in the knowledge base. We adopt a broad view of data mining

functionality: data mining is the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories.

Data Mining as a step in the process of Knowledge discovery

MAJOR COMPONENTS OF DATA MINING SYSTEM

Database, Data Warehouse, Or Other Information Repository

This is one or a set of databases, data warehouses, spread sheets, or other kinds of information repositories. Data cleaning and data integration techniques may be performed on the data.

Database Or Data Warehouse Server

The database or data warehouse server is responsible for fetching the relevant data, based on the user's data mining request.

Knowledge base

This is the domain knowledge that is used to guide the search, or evaluate the interestingness of resulting patterns. Such knowledge can include concept hierarchies used to organize attributes or attribute values into different levels of abstraction. Knowledge such as user beliefs, which can be used to assess a pattern's interestingness based on its unexpectedness, may also be included. Other examples of domain knowledge are additional interestingness constraints or thresholds, and metadata (e.g., describing data from multiple heterogeneous sources).

Data Mining Engine

This is essential to the data mining system and ideally consists of a set of functional modules. These modules perform tasks such as characterization, association analysis, classification, evolution and deviation analysis.

Pattern Evaluation Module

This component typically employs interestingness measures and interacts with the data mining modules to focus the search towards interesting patterns. It may access interestingness thresholds stored in the knowledge base. Alternatively, the pattern evaluation module may be integrated with the mining module, depending on the implementation of the data mining method used. For efficient data mining, we need to perform the evaluation of pattern interestingness as deep as possible into the mining process. As a result we can confine the search to only the interesting patterns.

Graphical User Interface

This is an interface between users and the data mining system. This allows the user to interact with the system by specifying a data mining query or task, providing information to help focus the search, and performing exploratory data mining based on the intermediate data mining results. In addition, this component allows the user to browse database and data warehouse schemas or data structures, evaluate mined patterns, and visualize the patterns in different forms.

Architecture of a Typical Data Mining System.

Data mining involves an integration of techniques from multiple disciplines such as database technology, statistics, machine learning, high performance computing, pattern recognition, neural networks, data visualization, information retrieval, image and signal processing, and spatial data analysis. The discovered knowledge can be applied to decision making, process control, information management, query processing.

Data mining is performed on relational databases, data warehouses, transactional databases, advanced database systems, flat files, and the World-Wide Web.

Relational Databases

A relational database is a collection of tables. Each table is assigned a unique name and consists of a set of attributes (columns or fields). They store a large number of tuples (records or rows). Each tuple in a table represents an object identified by a unique key. These objects are described by a set of attribute values.

Relational data can be accessed by database queries. Queries are written in a relational query language, such as SQL, or with the assistance of graphical user interfaces. A given query is transformed into a set of relational operations, such as join, selection, and projection, and is then optimized for efficient processing. Querying allows retrieval of specified subsets of the data.

Through the use of relational queries, we can get a list of all items that were sold in the last quarter. Relational languages also include aggregate functions such as sum, avg (average), count, max (maximum), and min (minimum). By using these functions we can find out things like total sales of the last month, grouped by branch, or The number of sales transactions occurred in the month of December, or Which sales person had the highest amount of sales.

When data mining is applied to relational databases, one can go further by searching for trends or data patterns. For example, data mining systems may analyze customer data to predict the credit risk of new customers based on their income, age, and previous credit information.

Data Warehouses

A data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and which usually resides at a single site. Data warehouses are constructed via a process of data cleansing, data transformation, data integration, data loading, and periodic data refreshing.

A data warehouse is usually modeled by a multidimensional database structure, where each dimension corresponds to an attribute or a set of attributes in the schema, and each cell stores the value of some aggregate measure, such as count or sales amount. The actual physical structure of a data warehouse may be a relational data store or a multidimensional data cube. It provides a multidimensional view of data and allows the precomputation and fast accessing of summarized data. The difference between data warehouses and data marts is data warehouses are enterprise-wide, whereas **Data Marts** are department-wide.

Data warehouse tools help support data analysis, additional tools for data mining are required to allow more in-depth and automated analysis.

Transactional Databases

A transactional database consists of a file where each record represents a transaction. A transaction typically includes a unique transaction identity number (trans ID), and a list of the items making up the transaction (such as items purchased in a store). The transactional database may have additional tables associated with it, which contain other information regarding the sale, such as the date of the transaction, the customer ID number, the ID number of the sales person, and of the branch at which the sale occurred, and so on.

If we would like to dig deeper into the data, we could start by asking which items sold well together. This kind of market basket data analysis would enable you to bundle groups of items together as a strategy for maximizing sales. For example, given the knowledge that printers are commonly purchased together with computers, we can sell costly printers along with computers by giving a discount on printers.

Advanced Database Systems And Advanced Database Applications

The new database applications include handling spatial data (such as maps), engineering design data (such as the design of buildings, system components, or integrated circuits), hypertext and multimedia data (including text, image, video, and audio data), time-related data (such as historical records or stock exchange data), and the World-Wide Web (a huge, widely distributed information repository made available by the Internet). These applications require efficient data structures and scalable methods for handling complex object structures, variable length records, semi-structured or unstructured data, text and multimedia data, and database schemas with complex structures and dynamic changes. In response to these needs, advanced database systems and specific application-oriented database systems have been developed.

Apart from these Data Mining is performed on Object-Oriented Databases, Object-Relational Databases, Spatial Databases, Temporal Databases and Time-Series Databases, Text Databases and Multimedia Databases, Heterogeneous Databases and Legacy Databases and finally on The World Wide Web.

DATA MINING FUNCTIONALITIES

In general, data mining tasks can be classified into two categories: descriptive and predictive. Descriptive mining tasks characterize the general properties of the data in the database. Predictive mining tasks perform inference on the current data in order to make predictions.

Concept/Class Description: Characterization And Discrimination

Data can be associated with classes or concepts. For example, in an electronics store, classes of items for sale include computers and printers, and concepts of customers include big spenders and budget spenders. Such descriptions of a class or a concept are called class/concept descriptions. These descriptions can be derived via (1) **data characterization**, by summarizing the data of the class under study (often called the target class) in general terms, or (2) **data discrimination**, by comparison of the target class with one or a set of comparative classes (often called the contrasting classes), or (3) both data characterization and discrimination.

Data characterization is a summarization of the general characteristics or features of a target class of data. For example, to study the characteristics of software products whose sales increased by 10% in the last year, one can collect the data related to such products by executing an SQL query. **Data discrimination** is a comparison of the general features of target class data objects with the general features of objects from one or a set of contrasting classes. The target and contrasting classes can be specified by the user, and the corresponding data objects retrieved through data base queries. For example, one may like to compare the general features of software products whose sales increased by 10% in the last year with those whose sales decreased by at least 30% during the same period.

Association Analysis

Association analysis is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data. Association analysis is widely used for market basket or transaction data analysis.

Example

Given an electronics relational database, a data mining system may find association rules like $\text{age}(X; "20 \text{ -- } 29") \wedge \text{income}(X; "20 \text{ -- } 30K") \rightarrow \text{buys}(X; "CD \text{ player} ")$ [support = 2%; confidence = 60%] meaning that customers under study, 2% (support) are 20-29 years of age with an income of 20-30K and have purchased a CD player at electronics shop. There is a 60% probability (confidence, or certainty) that a customer in this age and income group will purchase a CD player.

Note that this is an association between more than one attribute, or predicate (i.e., age, income, and buys).

Classification And Prediction

Classification is the processing of finding a set of models (or functions) which describe and distinguish data classes or concepts, for the purposes of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data (i.e., data objects whose class label is known). The derived model may be represented in various forms, such as classification (IF-THEN) rules, decision trees, mathematical formulae, or neural networks.

Classification can be used for predicting the class label of data objects. However, in many applications, one may like to predict some missing or unavailable data values rather than class labels. This is usually the case when the predicted values are numerical data, and is often specifically referred to as **prediction**.

Example

Suppose, as sales manager of an electronics company, you would like to classify a large set of items in the store, based on three kinds of responses to a sales campaign: good response, mild response, and no response. You would like to derive a model for each of these three classes based on the descriptive features of the items, such as price, brand, place made, type, and category. The resulting classification should maximally distinguish each class from the others, presenting an organized picture of the data set. Suppose that the resulting classification is expressed in the form of a decision tree. The decision tree, for instance, may identify price as being the single factor which best distinguishes the three classes. The tree may reveal that, after price, other features which help further distinguish objects of each class from another include brand and place made.

Cluster Analysis

Clustering analyzes data objects without consulting a known class label. In general, the class labels are not present in the training data simply because they are not known to begin with. Clustering can be used to generate such labels. The objects are clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity. That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters.

Outlier Analysis

A database may contain data objects that do not comply with the general behavior or model of the data. These data objects are outliers. Most data mining methods discard them as noise or exceptions. Deviation-based methods identify outliers by examining differences in the main characteristics of objects in a group.

A data mining generates patterns. A pattern is interesting if 1) it is easily understood by us, 2) valid on new or test data with some degree of certainty, 3) potentially useful and 4) novel. An interesting pattern represents **knowledge**.

CLASSIFICATION OF DATA MINING SYSTEMS

They are classified based on the following criteria

Classification According To The Kinds Of Databases Mined

If classifying according to data models, we may have a relational, transactional, object-oriented, object-relational, or data warehouse mining system. If classifying according to the special types of data handled, we may have a spatial, time-series, text, or multimedia data mining system, or a World-Wide Web mining system. Other system types include heterogeneous data mining systems, and legacy data mining systems.

Classification According To The Kinds Of Knowledge Mined.

Data mining systems can be categorized according to the kinds of knowledge they mine, i.e., based on data mining functionalities, such as characterization, discrimination, association, classification, clustering, trend and evolution analysis, deviation analysis, similarity analysis, etc.

Classification According To The Kinds Of Techniques Utilized.

Data mining systems can also be categorized according to the underlying data mining techniques employed. These techniques can be described according to the degree of user interaction involved (e.g., autonomous systems, interactive exploratory systems, query-driven systems), or the methods of data analysis employed (e.g., database-oriented or data warehouse-oriented techniques, machine learning, statistics, visualization, pattern recognition, neural networks, and so on). A sophisticated data mining system will often adopt multiple data mining techniques or work out an effective, integrated technique which combines the merits of a few individual approaches.

Classifications According To Applications Adapted.

There could be a data mining system tailored specifically for finance, telecommunications, DNA, stock markets and so on.

MAJOR ISSUES IN DATA MINING

- Mining different kinds of knowledge in databases.
- Interactive mining of knowledge at multiple levels of abstraction.
- Incorporation of background knowledge.
- Data mining Query languages and ad hoc data mining
- Presentation and visualization of data mining results
- Handling noisy or incomplete data
- Pattern evaluation
- Efficiency and scalability of data mining algorithms
- Parallel, distributed, and incremental mining algorithms.
- Handling of relational and complex types of data.
- Mining information from heterogeneous databases and global information systems.

The Scope Of Data Mining

Data mining derives its name from the similarities between searching for valuable business information in a large database. Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing these capabilities:

Automatic Prediction Of Trends And Behaviors

Data mining automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data - quickly. A typical example of a predictive problem is targeted marketing.

Automatic Discovery Of Previously Unknown Patterns

Data mining tools sweep through databases and identify previously hidden patterns in one step. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying *anomalous data* that could represent data entry keying errors.

A Comparison Of Leading Data Mining Tools

Distinctives	Strengths	Weaknesses
<i>Clementine</i>	visual interface; algorithm breadth	scalability
<i>Darwin</i>	efficient client-server; intuitive interface options	no unsupervised; limited visualization
<i>DataCruncher</i>	ease of use	single algorithm
<i>Enterprise Miner</i>	depth of algorithms; visual interface	harder to use; new product issues
<i>GainSmarts</i>	data transformations, built on SAS; algorithm option depth	no unsupervised; limited visualization
<i>Intelligent Miner</i>	algorithm breadth; graphical tree/cluster output	few algorithm options; no automation
<i>MineSet</i>	data visualization	few algorithms; no model export
<i>Model I</i>	ease of use; automated model discovery	really a vertical tool
<i>ModelQuest</i>	breadth of algorithms	some non-intuitive interface options
<i>PRW</i>	extensive algorithms; automated model selection	limited visualization
<i>CART</i>	depth of tree options	difficult file I/O; limited visualization
<i>Scenario</i>	ease of use	narrow analysis path
<i>NeuroShell</i>	multiple neural network architectures	unorthodox interface; only neural networks
<i>OLPARS</i>	multiple statistical algorithms; class-based visualization	dated interface; difficult file I/O
<i>See5</i>	depth of tree options	limited visualization; few data options
<i>S-Plus</i>	depth of algorithms; visualization; programable/extendable	limited inductive methods; steep learning curve
<i>WizWhy</i>	ease of use; ease of model understanding	limited visualization

REFERENCES

1. www.theartling.com
2. <http://www.eco.utexas.edu/~norman/BUS.FOR/course.mat/Alex/>
3. <http://www.data-mining-guide.net/Data-Mining-Overview.html>
4. <http://www.kdnuggets.com/websites/data-mining.html>
5. http://www.pcc.qub.ac.uk/tec/courses/datamining/stu_notes/dm_book_1.html
6. <http://infolab.stanford.edu/~ullman/mining/mining.html>
7. <http://home.planet.nl/~maane093/framehol.html>
8. Book: *Data Mining Concepts and Techniques* by Jiawei Han and Micheline Kamber.
9. Technology Impact in Retail, Oct 19 2005 <http://www.microsoft.com/net/business/industry/retail.msp>
10. The Impact of Business Performance Management, David A. Kelly 15 Jan 2006
http://www.ebizq.net/topics/tech_in_biz/features/6617.html
11. New Hope for Old Risks, Robert J.Shiller, winter 2003 <http://www.cio.com/archive/092203/shiller.html>
12. The impact of technology on organizational transformations, Verner M.Kiernan, March 3 1995
<http://www.mindspring.com/~kiernan/mgt6107.html#Impact>
13. Information Technology Impacts, Roger Clarke, 18 May 1996
<http://www.anu.edu.au/people/Roger.Clarke/SOS/ITImpacts.html#Ec>
14. How does E-Commerce Work, Linda Wokosin <http://www.stc.org/confproceed/2001/PDFs/STC48-000128.PDF#search=%22how%20does%20e%20commerce%20work%22>
15. Advertising on the Internet, Steven Armstrong, 2nd Edition, Jan 1 2001, Korgan page publication

BIOGRAPHIES

Ihssan Alkadi is on the faculty at University of Louisiana at Lafayette (UL Lafayette). He received his B.S. Degree in Computer Science at SLU, May 1985. In May 1992 he earned his MS. In Systems Science from Louisiana State University (LSU). He earned his Doctoral degree in Computer Science at LSU. His areas of expertise include software engineering in general, testing in particular, Internet, HTML, and operating systems. His research interests include testing in object oriented systems, systems validation, and system verification.