Data Mining – A Review and Description

Nancy, Jasdeep Kaur, Ramneet Kaur, Nishu
pursuing m-tech in cse, Computer Science & Engineering Department,
Guru Nanak Dev Engg. College, Ludhiana, Punjab-India.
Email-ernancy.10@gamil.com

ABSRACT: Data mining is a powerful and new technique with great potential. It converts the raw data into the useful information. Data Mining is the process of extracting knowledge from data warehouses. To store databases, enterprises make data warehouses and data marts. Data warehouses and data marts contain large amounts of data. Due to extracting knowledge from large data warehouses or depositories, data mining plays great role in various fields of machine learning, advancements in static's, database system, pattern matching, and artificial intelligence. Various algorithms and programs are used for data mining approach.

*****

1. INTRODUCTION

Now the world of digital and information technology, all processes are become automated. Information technology is used in every field of human life such as business, engineering, medical.maathematical, scientific. All these fields of human life have lead to the large volume of data storage in various formats such as records, files, documents, images, sounds, recordings, videos and many new data. Collection of related data is also known as database. To extract hidden predictive information from data warehouses, the proper mechanism is used, that mechanism is also known as DATA MINING. It is a knowledge discovery technique (KDD). Its aims to extract useful and correct information from the large repository of data. Various functionalities and algorithms are used for data mining. The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining).

2. WHAT KIND OF DATA CAN BE MINED?

1. Flat Files
2. Relational Databases
3. Data Warehouses
4. Transaction Databases
5. Multimedia Databases
6. Spatial Databases
7. Time Series
8. World Wide Web

3. DATA PROCESSING

The Knowledge Discovery in Databases (KDD) process is commonly defined with the stages:

(1) Selection
(2) Pre-processing
(3) Transformation
(4) Data Mining
(5) Interpretation/Evaluation.

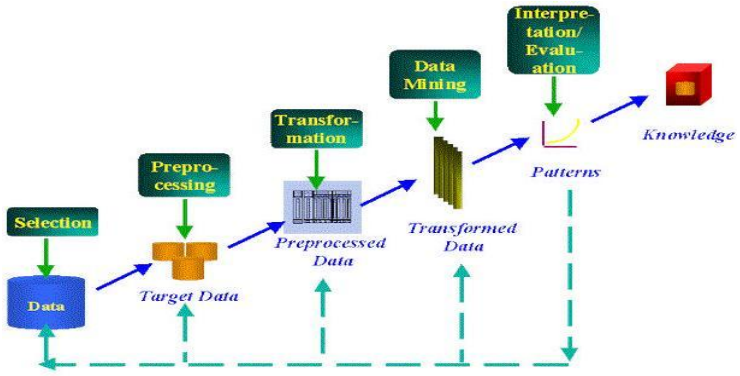

Figure.3.1

1. **Data cleaning:** It is also known as data cleansing; in this phase noise data and irrelevant data are removed from the collection.

2. **Data integration:** In this stage, multiple data sources, often heterogeneous, are combined in a common source.

3. **Data selection:** The data relevant to the analysis is decided on and retrieved from the data collection.

4. **Data transformation:** It is also known as data consolidation; in this phase the selected data is transformed into forms appropriate for the mining procedure.

5. **Data mining:** It is the crucial step in which clever techniques are applied to extract potentially useful patterns.
6. **Pattern evaluation:** In this step, interesting patterns representing knowledge are identified based on given measures.

7. **Knowledge representation:** It is the final phase in which the discovered knowledge is visually presented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

4. WHAT CAN BE DISCOVERED?

Two types of information can be discovered:-

1. **Descriptive Data Mining**: - a task that describes general properties of existing data.

2. **Predictive Data Mining**: - Task that attempts to do prediction based on inference on available data.

5. DATA MINING LIFE CYCLE

(1) Pre-processing
(2) Data mining and
(3) Results validation.

1. Pre-processing

Before data mining algorithms can be used, a target data set must be assembled. As data mining can only uncover patterns actually present in the data, the target dataset must be large enough to contain these patterns while remaining concise enough to be mined within an acceptable time limit. A common source for data is a data mart or data warehouse. Pre-processing is essential to analyze the multivariate datasets before data mining. The target set is then cleaned. Data cleaning removes the observations containing noise and those with missing data.

2. Data mining

Data mining involves six common classes of tasks:

1. **Anomaly detection (Outlier/change/deviation detection)** – The identification of unusual data records

2. **Association rule learning (Dependency modeling)** – Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.

3.**Clustering** – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.

4. **Classification** – is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".

5**. Regression** – Attempts to find a function which models the data with the least error.

6. **Summarization** – providing a more compact representation of the data set, including visualization and report generation.

3. Results validation

The final step of knowledge discovery from data is to verify that the patterns produced by the data mining algorithms occur in the wider data set. Not all patterns found by the data mining algorithms are necessarily valid. It is common for the data mining algorithms to find patterns in the training set which are not present in the general data set. This is called over fitting. To overcome this, the evaluation uses a test set of data on which the data mining algorithm was not trained. The learned patterns are applied to this test set and the resulting output is compared to the desired output. For example, a data mining algorithm trying to distinguish "spam" from "legitimate" emails would be trained on a training set of sample e-mails. Once trained, the learned patterns would be applied to the test set of e-mails on which it had not been trained. The accuracy of the patterns can then be measured from how many e-mails they correctly classify. A number of statistical methods may be used to evaluate the algorithm, such as ROC curves.

If the learned patterns do not meet the desired standards, then it is necessary to re-evaluate and change the pre-processing and data mining steps. If the learned patterns do meet the desired standards, then the final step is to interpret the learned patterns and turn them into knowledge.

6. DATA MINING METHODS

The data mining methods are broadly categories as:

On-Line Analytical Processing (OLAP)
Classification
Association Rule Mining
Temporal Data Mining
Time Series Analysis
Spatial Mining,
Anomaly/outlier/change detection
Association rule learning
Cluster analysis
Decision trees
Factor analysis
Neural Networks
Regression analysis
Structured data analysis
Sequence mining
Text mining
Application domains
Analytics
Bioinformatics
Business intelligence
Data analysis
Data warehouse
Decision support system
Drug Discovery
Exploratory data analysis
Predictive analytics
Web Mining etc.

7. ISSUES IN DATA MINIG

1. Social issues
2. Security issues
3. User interface issues
4. Performance issues
5. Mining methodology issues
6. Data sources issues

8. DATA MINING APPLICATIONS

1. Games

Since the early 1960s, with the availability of oracles for certain combinatorial games, also called table bases (e.g. for 3x3-chess) with any beginning configuration, small-board dots-and-boxes, small-board-hex, and certain endgames in chess, dots-and-boxes, and hex; a new area for data mining has been opened. This is the extraction of human-usable strategies from these oracles. Current pattern recognition approaches do not seem to fully acquire the high level of abstraction required to be applied successfully. Instead, extensive experimentation with the table bases – combined with an intensive study of table base-answers to well designed problems, and with knowledge of prior art (i.e. pre-table base knowledge) – is used to yield insightful patterns. Berlekamp (in dots-and-boxes, etc.) and John Nunn (in chess endgames) are notable examples of researchers doing this work, though they were not – and are not – involved in table base generation.

2. Business

Data mining in customer relationship management applications can contribute significantly to the bottom line Rather than randomly contacting a prospect or customer through a call center or sending mail, a company can concentrate its efforts on prospects that are predicted to have a high likelihood of responding to an offer. More sophisticated methods may be used to optimize resources across campaigns so that one may predict to which channel and to which offer an individual is most likely to respond (across all potential offers). Additionally, sophisticated applications could be used to automate mailing. Once the results from data mining (potential prospect/customer and channel/offer) are determined, this "sophisticated application" can either automatically send an e-mail or a regular mail. Finally, in cases where many people will take an action without an offer, "uplift modeling" can be used to determine which people have the greatest increase in response if given an offer. Data clustering can also be used to automatically discover the segments or groups within a customer data set.

Data mining can also be helpful to human resources (HR) departments in identifying the characteristics of their most successful employees. Information obtained – such as universities attended by highly successful employees – can help HR focus recruiting efforts accordingly. Additionally, Strategic Enterprise Management applications help a company translate corporate-level goals, such as profit and margin share targets, into operational decisions, such as production plans and workforce levels

3. Science and engineering

In recent years, data mining has been used widely in the areas of science and engineering, such as bioinformatics, genetics, medicine, education and electrical power engineering.

In the study of human genetics, sequence mining helps address the important goal of understanding the mapping relationship between the inter-individual variations in human DNA sequence and the variability in disease susceptibility. In simple terms, it aims to find out how the changes in an individual's DNA sequence affects the risks of developing common diseases such as cancer, which is of great importance to improving methods of diagnosing, preventing, and treating these diseases. The data mining method that is used to perform this task is known as multifactor dimensionality reduction.

4. Human rights

Data mining of government records – particularly records of the justice system (i.e. courts, prisons) – enables the discovery of systemic human rights violations in connection to generation and publication of invalid or fraudulent legal records by various government agencies.

5. Spatial data mining

Spatial data mining is the application of data mining methods to spatial data. The end objective of spatial data mining is to find patterns in data with respect to geography. So far, data mining and Geographic Information Systems (GIS) have existed as two separate technologies, each with its own methods, traditions, and approaches to visualization and data analysis. Particularly, most contemporary GIS have only very basic spatial analysis functionality. The immense explosion in geographically referenced data occasioned by developments in IT, digital mapping, remote sensing, and the global diffusion of GIS emphasize the importance of developing data-driven inductive approaches to geographical analysis and modeling.

Challenges

Geospatial data repositories tend to be very large. Moreover, existing GIS datasets are often splintered into feature and attribute components that are conventionally archived in hybrid data management systems. Algorithmic requirements differ substantially for relational (attribute) data management and for topological (feature) data management. Related to this is the range and diversity of geographic data formats, which present unique challenges. The digital geographic data revolution is creating new types of data formats beyond the traditional "vector" and "raster" formats. Geographic data

repositories increasingly include ill-structured data, such as imagery and geo-referenced multi-media.

Current geographic knowledge discovery (GKD) methods generally use very simple representations of geographic objects and spatial relationships Geographic data mining methods should recognize more complex geographic objects (i.e. lines and polygons) and relationships (i.e. non-Euclidean distances, direction, connectivity, and interaction through attributed geographic space such as terrain). Furthermore, the time dimension needs to be more fully integrated into these geographic representations and relationships.
Geographic knowledge discovery using diverse data types: GKD methods should be developed that can handle diverse data types beyond the traditional raster and vector models, including imagery and geo-referenced multimedia, as well as dynamic data types (video streams, animation).

In four annual surveys of data miners, data mining practitioners consistently identify three key challenges that they face more than any others, specifically (a) dirty data, (b) explaining data mining to others, and (c) unavailability of data/difficult access to data. In the 2010 survey data miners also shared their experiences in overcoming these particular challenges.

Sensor data mining

Wireless sensor networks can be used for facilitating the collection of data for spatial data mining for a variety of applications such as air pollution monitoring. A characteristic of such networks is that nearby sensor nodes monitoring an environmental feature typically register similar values. This kind of data redundancy due to the spatial correlation between sensor observations inspires the techniques for in-network data aggregation and mining. By measuring the spatial correlation between data sampled by different sensors, a wide class of specialized algorithms can be developed to develop more efficient spatial data mining algorithms.

Visual data mining

In the process of turning from analogical into digital, large data sets have been generated, collected, and stored discovering statistical patterns, trends and information which is hidden in data, in order to build predictive patterns. Studies suggest visual data mining is faster and much more intuitive than is traditional data mining.

Music data mining

Data mining techniques, and in particular co-occurrence analysis, has been used to discover relevant similarities among music corpora (radio lists, CD databases) for the purpose of classifying music into genres in a more objective manner.

Surveillance

Data mining has been used to stop terrorist programs under the U.S. Data mining has been used to stop terrorist programs under the U.S. government, including the Total Information Awareness (TIA) program, Secure Flight (formerly known as Computer-Assisted Passenger Prescreening System (CAPPS II)), Analysis, Dissemination, Visualization, Insight, Semantic Enhancement (ADVISE), and the Multi-state Anti-Terrorism Information Exchange (MATRIX).These programs have been discontinued due to controversy over whether they violate the 4th Amendment to the United States Constitution, although many programs that were formed under them continue to be funded by different organizations or under different names.

In the context of combating terrorism, two particularly plausible methods of data mining are "pattern mining" and "subject-based data mining".

Pattern mining

"Pattern mining" is a data mining method that involves finding existing patterns in data. In this context patterns often means association rules. The original motivation for searching association rules came from the desire to analyze supermarket transaction data, that is, to examine customer behavior in terms of the purchased products. For example, an association rules "beer ⇒ potato chips (80%)" states that four out of five customers that bought beer also bought potato chips.

In the context of pattern mining as a tool to identify terrorist activity, the National Research Council provides the following definition: "Pattern-based data mining looks for patterns (including anomalous data patterns) that might be associated with terrorist activity — these patterns might be regarded as small signals in a large ocean of noise.” Pattern Mining includes new areas such a Music Information Retrieval (MIR) where patterns seen both in the temporal and non temporal domains are imported to classical knowledge discovery search methods.

Subject-based data mining

"Subject-based data mining" is a data mining method involving the search for associations between individuals in data. In the context of combating terrorism, the National Research Council provides the following definition: "Subject-based data mining uses an initiating individual or other datum that is considered, based on other information, to be of high interest, and the goal is to determine what other persons or financial transactions or movements, etc.,."

Knowledge grid

Knowledge discovery "On the Grid" generally refers to conducting knowledge discovery in an open environment using grid computing concepts, allowing users to integrate

data from various online data sources, as well make use of remote resources, for executing their data mining tasks. The earliest example was the Discovery Net, developed at Imperial College London, which won the "Most Innovative Data-Intensive Application Award" at the ACM SC02 (Supercomputing 2002) conference and exhibition, based on a demonstration of a fully interactive distributed knowledge discovery application for a bioinformatics application. Other examples include work conducted by researchers at the University of Calabria, who developed Knowledge Grid architecture for distributed knowledge discovery, based on grid computing.

6. **The Digital Library retrieves**, collects, stores and preserves the digital data. The advent of electronic resources and their increased use in libraries has brought about significant changes in Library. The data and information are available in the different formats. These formats include Text, Images, Video, Audio, Picture, Maps, etc. therefore digital library is a suitable domain for application of data mining.

8. CONCLUSION

In this paper I briefly reviewed data mining, concepts of data mining, processing, methods, life cycle model, issues, and types of data mining, application of it.

This review will be helpful to you to easily understand what data mining is, previously used data mining techniques, now days use data mining techniques and also process of data mining.

In this I completely, explain applications and types of data mining.

9. FUTURE SCOPE

Data mining is a very wide concept. It contains many concepts such as various algorithms to extract knowledge from large databases.
Soft Computing techniques like Fuzzy logic, Neural Networks and Genetic Programming which contains Complex data objects Includes high dimensional, high speed data streams, sequence, noise in the time series, graph, Multi instance objects, Multi represented objects and temporal data etc. these are used in Business, Web, Medical diagnosis, Scientific and Research analysis fields (bio, remote sensing etc…), Social networking etc.

10. REFERENCES

1. WIKIPEDIA.Com

2. Mr. S. P. Deshpande1 and Dr. V. M. Thakar International Journal of Distributed and Parallel systems (IJDPS) Vol.1, No.1, September 2010 DOI.

3. A Review on Data mining from Past to the Future.

Venkatadri.M Research Scholar, Dept. of Computer Science, Dravidian University, India. Lokanatha C. Reddy Professor, Dept. of Computer Science, International Journal of Computer Applications (0975 – 8887) *Volume 15– No.7, February 2011*

4. Data Mining for High Performance Data Cloud using Association Rule Mining.
1 T.V.Mahendra 2N.Deepika 3N.Keasava Rao Professor & HOD, IT, Narayana Engg. College, Nellore, AP, India.
2 Sr.Assistant Professor, Dept. of ISE, New Horizon College of Engineering, Bangalore, India.
3 Associate Professor, IT, Narayana Engg. College, Gudur, AP, India.

5. Data Mining and KDD: A Shifting Mosaic
By Joseph M. Firestone, Ph.D. White Paper No. Two March 12, 1997 the Idea.

6. Data mining and static's**:** what's the connection? Jerome h. Friedman.

7. Google.Com.

8. "Data mining" slidesahre.net/ernancy/CLAHE. Nancy.GNDEC Ludhiana Punjab India.

9. IEEEXPLORE.Com.

10.www.cs.ualberta.ca/~zaiane/courses/cmput690/notes/Chapter1.