Review on Data Mining and Knowledge Discovery Process

Siddhant Sahare[1],Padmani Mahanand[1],Ronald David[1], Asst.Prof. Vandana Yadav[2]
Department of Computer Science & Engineering , Krishna Engineering College Khamharia, Junwani, Bhilai
Email: ronald777.david@gmail.com[1],vandanayadavbit10@gmail.com[2]

Abstract

This paper provides an introduction to the basic concept of data mining. Which gives overview of Data mining is used to extract momentous information and to develop significant relationships among variables stored in large-data-set (LDS) /data warehouse. In the case study reported in this paper, a data mining approach is applied to extract knowledge from a data set. Data mining is the process of discovering potentially useful, interesting, and previously unknown patterns from a large collection of data. Data mining is a multidisciplinary field, illustration work from areas including database technology, machine learning, statistics, pattern recognition, information retrieval, neural network, knowledge-based systems, artificial intelligence, high-performance computing, and data visualization. We present techniques for the innovation of patterns concealed in large-data-set (LDS), focusing on issues relating to their likelihood, efficacy and effectiveness. The automated, prospective analyses obtainable by data mining progress beyond the analyses of past events provided by retrospective-tools typical of decision-support-systems (DSS).

Keywords- *Data Warehouse, Data Mining, Clustering, Data Integration, Pattern evaluation, Knowledge representation, Retrospective tool.*

1. INTRODUCTION

Data mining is a course of action to extract the inherent information and knowledge which is potentially useful and people do not know in advance, and this extraction is from the mass, incomplete, noisy, fuzzy and random data.

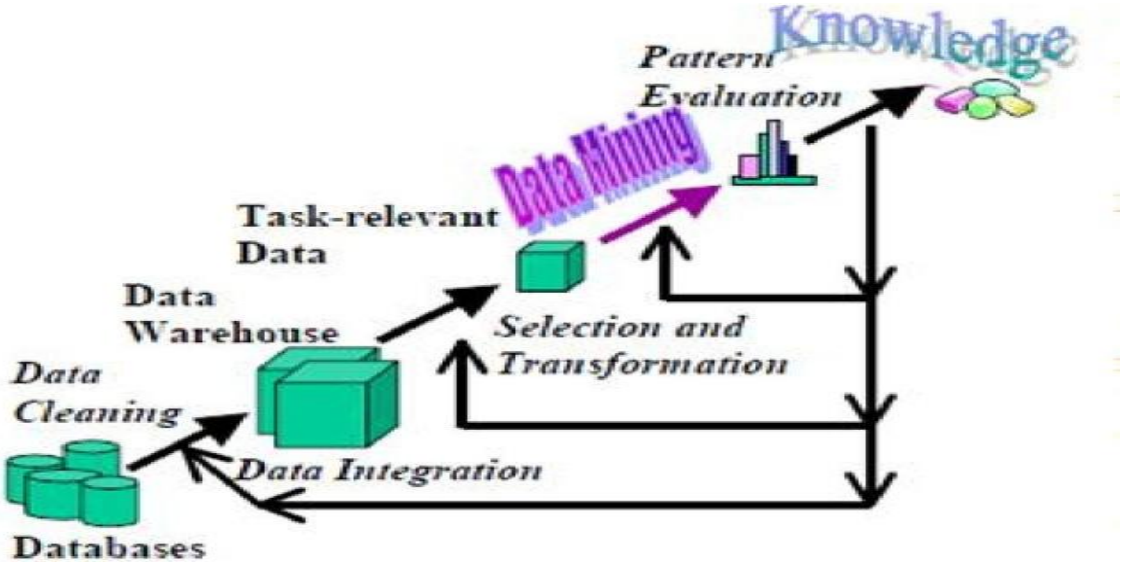

Fig.1: Data mining is the core of Knowledge Discovery Process

The essential difference between the data mining and the conventional data analysis (such as query, reporting and on-line application of analysis) is that the data mining is to mine information and discover knowledge on the basis of no clear assumption.[1] In addition to industry driven demand for standards and interoperability, professional and academic activity have also made considerable contributions to the evolution of the methods and models; an article published in a 2012 issue of the *International Journal of Computer Technology and Electronics Engineering (IJCTEE)* the results of a literature survey which traces and analyzes this evolution.

Data mining is the use of computerized data analysis techniques to uncover previously undetected relationships among data items. Data mining often involves the analysis of data stored in a data warehouse. Three of the major data-mining-techniques (DMT) are regression, classification and clustering. *Data Mining*, also popularly known as *Knowledge Discovery in Databases* (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. The following figure (Figure 1) shows data mining as a step in an iterative knowledge discovery process

The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge.

2. METHODOLOGY

The iterative method consists of the following steps:

- **Data cleaning**: also known as data cleansing, it is a phase in which noise data and extraneous data are removed from the collection.
- **Data integration**: at this stage, multiple data sources, often assorted, may be combined in a common source.
- **Data selection**: at this step, the data appropriate to the analysis is decided on and retrieved from the data collection.
- **Data transformation**: also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.
- **Data mining**: it is the crucial step in which clever techniques are applied to extract patterns potentially useful.
- **Pattern evaluation**: in this step, strictly interesting patterns representing knowledge are identified based on given measures.
- **Knowledge representation**: is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

Data mining commonly involves four classes of tasks: [1]

i) 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.

Clustering is a data mining (machine learning) technique used to place data elements into related groups without advance knowledge of the group definitions. Popular clustering techniques include k-means clustering and expectation-maximization (EM) clustering.

ii)Classification - is the task of generalizing known structure to apply to new data. For example, an email program might attempt to classify an email as legitimate or spam. Common algorithms include decision tree learning, nearest neighbor, naive Bayesian classification, neural networks and support vector machines.

Working with categorical data or a mixture of continuous numeric and categorical data? Classification analysis might suit your needs well. This technique is capable of processing a wider variety of data than regression and is growing in popularity.

iii)Regression - Attempts to find a function which models the data with the least error.

Regression is the oldest and most well-known statistical technique that the data mining community utilizes. Basically, regression takes a numerical dataset and develops a mathematical formula that fits the data. When you're ready to use the results to predict future behavior, you simply take your new data, plug it into the developed formula and you've got a prediction! The major limitation of this technique is that it only works well with continuous quantitative data (like weight, speed or age). If you're working with categorical data where order is not significant (like color, name or gender) you're better off choosing another technique.

Regression is a data mining (machine learning) technique used to fit an equation to a dataset. The simplest form of regression, linear regression, uses the formula of a straight line ($y = mx + b$) and determines the appropriate values for m and b to predict the value of y based upon a given value of x. Advanced techniques, such as multiple regression, allow the

use of more than one input variable and allow for the fitting of more complex models, such as a quadratic equation.

iv)Association-rule-learning (ARL) - 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.

- **Data preparation**

Data preparation generally consists of two processes: data collection and data collation. Data collection is the first step of data mining, and the data can come from the existing transaction processing systems, also can be obtained from the data warehouse; data collation is to eliminate noise or inconsistent data, it is the necessary link of data mining. The data obtained from the phase of the data collection may have a certain degree of "pollution", which refers to that in the data may be its own inconsistency, or some missing data, so the collation of the data is essential. At the same time, through data collation the data can be done on a simple generalization processing, thus on the basis of the original data more rich data information will be obtained, which will facilitate the next data mining step.

- **Data mining**

Data mining is the core stage of the entire process, it mainly uses the collected mining tools and techniques to deal with the data, thus the rules, patterns and trends will be found.

- **Information expression**

Information expression is to use visualization and knowledge information expression technology to provide the mined knowledge information for users, is an important means to show the data mining results. Clear and effective mining result information expression will greatly facilitate the accuracy and efficiency of the decision-making.

- **Analysis and decision-making**

The ultimate goal of data mining is to assist the decision making. Decision-makers can analyze the results of data mining and adjust the decision-making strategies combining with the actual situation.

3.Data-mining-architecture (DMA):

There are three tiers in the tight-coupling data mining architecture:

1. Data layer: as mentioned above, data layer can be database and/or data warehouse systems. This layer is an interface for all data sources. Data mining results are stored in data layer so it can be presented to end-user in form of reports or other kind of visualization.
2. Data mining application layer is used to retrieve data from database. Some transformation routine can be performed here to transform data into desired format. Then data is processed using various data mining algorithms.
3. Front-end layer provides intuitive and friendly user interface for end-user to interact with data mining system. Data mining result presented in visualization form to the user in the front-end layer.

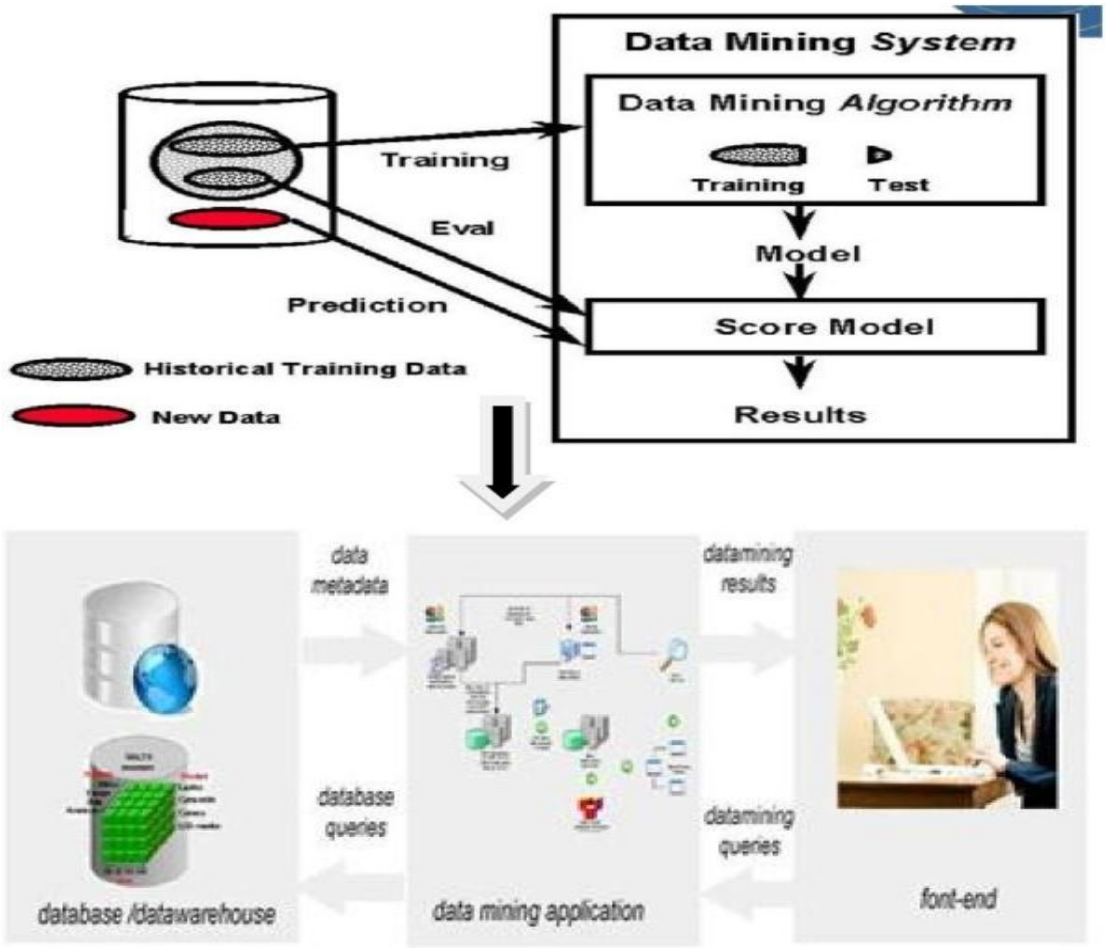

Fig.2: Data-Mining-Architecture(DMA)

In this article, we've discussed various *data mining architectures*, its advantages and disadvantages. And then we looked into a tight-couple **data mining architecture** – the most desired, high performance, high scalable data mining architecture.

Data mining based on decision tree

Decision-tree-learning(DTL), used in statistics, data mining and machine learning, uses a decision-tree(DT) as a predictive model which maps observations about an item to conclusions about the item's target value. More descriptive names for such tree models are **classification-trees(CT)** or **regression-trees(RT)**. In these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels.

In decision analysis, a decision-tree(DT) can be used to visually and explicitly represent decisions and decision making. In data mining, a decision-tree(DT) describes data but not decisions; rather the resulting classification tree can be an input for decision making

A **decision-support-system** (**DSS**) is a computer-based information system that supports business or organizational decision-making activities. DSSs serve the management, operations, and planning levels of an organization and help to make decisions, which may be rapidly changing and not easily specified in advance.

DSSs include knowledge-based systems. A properly designed DSS is an interactive software-based system intended to help decision makers compile useful information from a combination of raw data, documents, personal knowledge, or business models to identify and solve problems and make decisions.

Data mining requires data preparation which can uncover information or patterns which may compromise confidentiality and privacy obligations. A common way for this to occur is through data aggregation. Data aggregation be when the data are accrued, possibly from various sources, and put together so that they can be analyzed. This is not data mining per se, but a result of the preparation of data before and for the purposes of the analysis. The threat to an individual's privacy comes into play when the data, once compiled, cause the data miner, or anyone who has access to the newly compiled data set, to be able to identify specific individuals, especially when originally the data were anonymous.

Data mining based on neural network:

The data mining based on neural network is composed by data preparation, rules extracting and rules assessment three phases, as shown in Fig. 3.

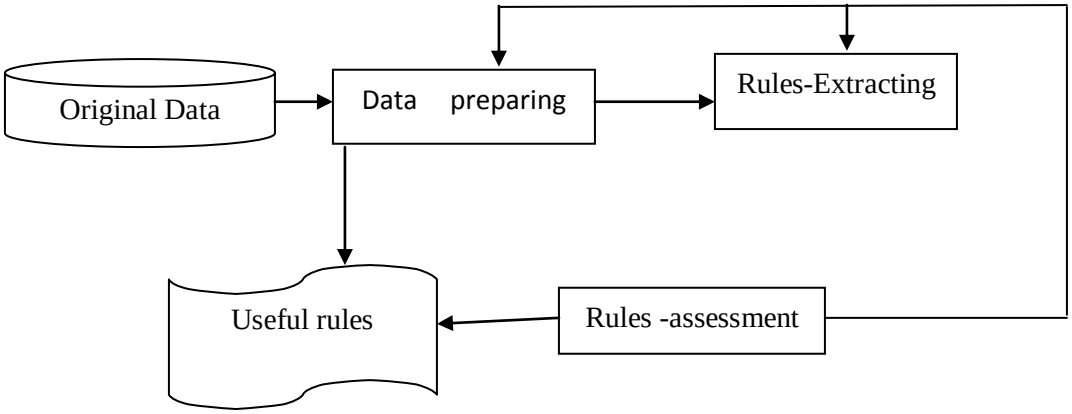

Fig.3: Data mining process on neural network

4. CONCLUSION

From the word mining it is clear that something is to be extracted and that is data. The extraction of information or data from large-data-sets (LDS) is termed as Data mining. This is an effective and influential technology to help IT companies focus on the most important part of data or information or in short the most important data. Data mining is supported by or based on three technologies. First is Massive Data Collection, Second is Powerful Multiprocessor Computer, and third is Data mining Algorithm. Lastly Data mining help us to focus on the important data or information in our data warehouses & data mining is also known as Knowledge-Discovery (KDD) in DBMS (Database Management System).

REFERENCES

[1]: Hemlata Sahu, Shalini Shrma, Seema Gondhalakar " A Brief Overview on Data Mining Survey", International Journal of Computer Technology and Electronics Engineering (IJCTEE) 2012 Volume 1, Issue 3.

[2] Ming-Syan Chen, Jiawei Han, Philip S yu. "Data Mining: An Overview from a Database Perspective[J]". IEEE Transactions on Knowledge and Data Engineering, l996, 8(6):866-883.

[3] R Agrawal ,T 1 mielinski, A Swami." Database Mining: A Performance Perspective[J]"· IEEE Transactions on Knowledge and Data Engineering, 1993,12:914-925.

[4] Fayyad, Usama; Gregory Piatetsky-Shapiro, and Padhraic Smyth (1996). "From Data Mining to Knowledge Discovery in Databases". http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996- Fayyad.pdf Retrieved 2008-12-17..

[5] Y. Peng, G. Kou, Y. Shi, Z. Chen (2008). "A Descriptive Framework for the Field of Data Mining and Knowledge Discovery" International Journal of Information Technology and Decision Making, Volume 7, Issue 4 7: 639 – 682. doi:10.1142/S0219622008003204.

[6] Data mining:Ford, C.W.; Chia-Chu Chiang; Hao Wu; Chilka, R.R.; Talburt,J.R.;" Information Technology: Coding and Computing", 2005. ITCC 2005 InternationalConference Volume: Digital Object Identifier: 10.1109/ITCC.2005.270 Publication Year: 2005 , Page(s): 122 - 127 Vol. 1

[7] Han, J. & M. Kamber, "Data mining: concepts and techniques", San Francisco: Morgan Kaufman (2001).

[8] "Data mining tools", by Ralf Mikut, Markus Reischl, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2011

[9] "Data mining and ware housing". Electronics Computer Technology (ICECT), 2011 3rd International Conference on Volume:1, Publication Year: 2011 , Page(s): 1 – 5

[10] "The applied research on data mining in the financial analysis of university with the analysis of college students „arrears as an example" Chen Hongfei; Wang Xiaoyan; Business Management and Electronic Information (BMEI), 2011 International Conference on Volume:2 Digital Object Identifier: 10.1109/ICBMEI.2011.5917992 Publication Year: 2011 , Page(s): 633 - 636

.