ISSN:0975-9646

Saurabh Patodi et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (5) , 2015, 4667-4671

Data Mining: The Massive Data Set

[1] Saurabh Patodi, [2] Mini Jain, [3]. Teena Negi

[1] HOD, Department Of Computer Science,
S.J.H.S.Gujarati Innovative College Of Commerce And Science, Indore, Madhya Pradesh, India

[2] Software Developer, Computer Science Corporation ,
Indore, Madhya Pradesh, India

[3]Assistant Professor, Department Of Computer Science,
S.J.H.S.Gujarati Innovative College Of Commerce And Science Indore, Madhya Pradesh, India

Abstract-Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. It uses machine learning, statistical and visualization techniques to discovery and present knowledge in a form which is easily comprehensible to humans. Various popular data mining tools are available today. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. Data mining tools can answer business questions that traditionally were too time consuming to resolve. Data mining is the process of knowledge discovery where knowledge is gained by analyzing the data store in very large repositories, which are analyzed from various perspectives and the result is summarized it into useful information.

Keywords- Current and Future of Data Mining, Data Mining, Data Mining Trends, Heterogeneous Data, distributed data mining (DDM), Ubiquitous Data Mining (UDM), Statistical Relation Learning (SRL)

I. INTRODUCTION

In the process of *Data Mining* the input becomes the data and the output is the knowledge obtained from the same. One of the most important definitions from the past is provided by Fayyad, Piatetsky-Shapiro and Smyth (1996), who define it as "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data." In other words DATA mining can be said as the procedure of throwing questions and taking out required patterns, often in the past mysterious from huge capacities of data applying pattern matching or other way of thinking techniques.

Different types of data mining tools are available in the marketplace, each with their own strengths and weaknesses. Internal auditors need to be aware of the different kinds of data mining tools available and recommend the purchase of a tool that matches the organization's current detective needs.

Data mining, the new upcoming trend in development area for science, is a powerful technology with great potential to predict future trends and behaviors. It comes under a very new research area that aims at the discovery of useful information from large datasets. Data mining uses various tools for better results such as statistical analysis and inference to extract interesting trends and events, create useful reports, support decision making etc. The field of Data Mining is concerned with finding new patterns in large amounts of data.

II. DATA MINING PROCESS IN WEB MINING

Data mining is a multidisciplinary field with many techniques. With this technique you can create a mining model that described the data that you will use. Some elements in Data Mining Process are:

A. **Data Set:** It is a collection of data, usually presented in tabular form. Each column represents a particular variable. Each row corresponds to a given member of the data set in question. It lists values for each of the variables, such as height and weight of an object or values of random numbers. Each value is known as a datum. The data set may comprise data for one or more members, corresponding to the number of rows.

B. **Pre-processing:** Data mining requires substantial pre-processing of data. This was especially the case of the behavioral data. To make the data comparable, all data needs to be normalized.

C. **General Results:** This activity is related to overall assessment of the effort in order to find out whether some important issues might have been overlooked. This is the step where a decision upon further steps has to be made. If all previous steps were satisfactory and results fulfill problem objectives, the project can move to its conclusive phase.

D. **Decision Trees:** Decision trees are powerful and popular tools for classification and prediction. The attractiveness of decision trees is due to the fact that, these decision trees represent rules. Decision tree induction is a typical inductive approach to learn knowledge on classification. The key requirements to do mining with decision trees are: attribute-value description, predefined classes, discrete classes and sufficient data.

E. **Association Rules:** Association rules describe events that tend to occur together. They are formal statements in the form of X=>Y, where if X happens, Y is likely to happen.

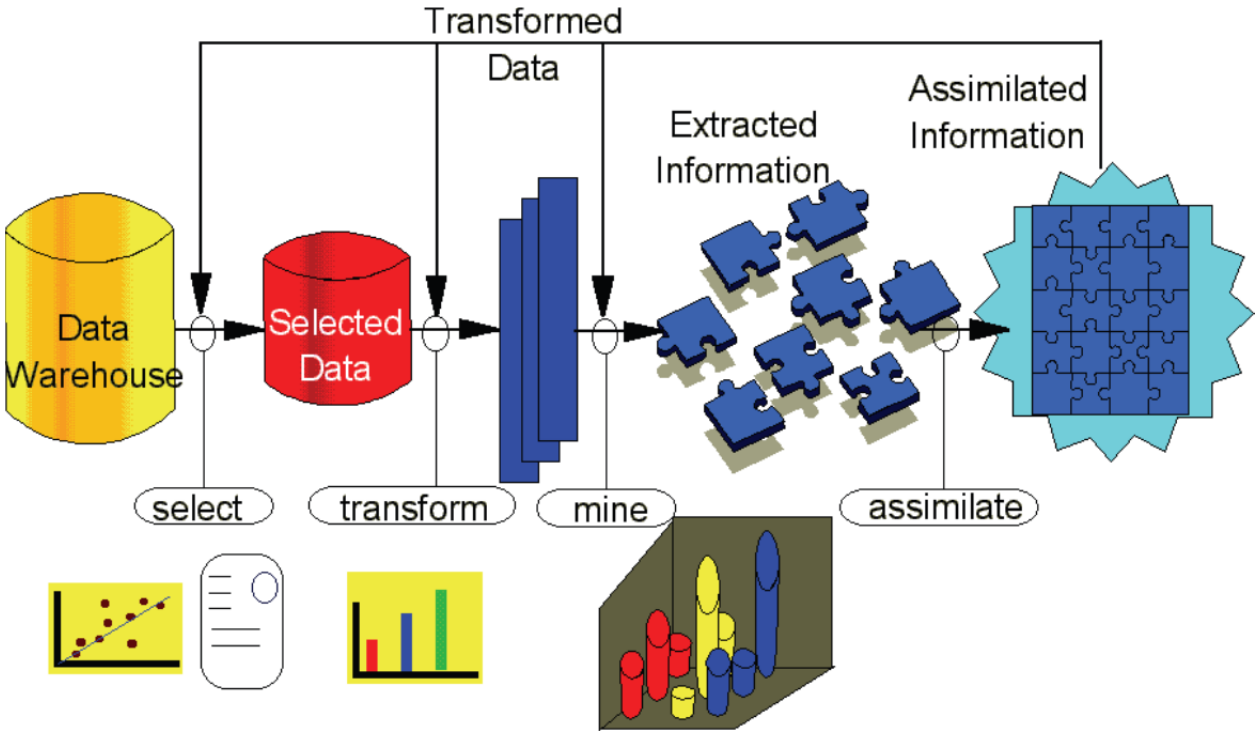

Fig 1. Data Mining Process

III. Roots Of Data Mining

Roots of Data Mining can be traced back along three lines

1. Statistics: The most important lines is statistics. Without statistics, there would be no data mining, as statistics are the foundation of most technologies on which data mining is built. Statistics embrace concepts such as regression analysis, standard distribution, standard deviation, standard variance, discriminant analysis, cluster analysis, and confidence intervals, all of which are used to study data and data relationships. These are the very building blocks with which more advanced statistical analyses are underpinned. Certainly, within the heart of today's data mining tools and techniques, classical statistical analysis plays a significant role.

2. Artificial Intelligence & Machine Learning: Data mining's second longest family line is artificial intelligence and machine learning. AI is built upon heuristics as opposed to statistics, and attempts to apply human-thought-like processing to statistical problems. Because this approach requires vast computer processing power, it was not practical until the early 1980s, when computers began to offer useful power at reasonable prices. AI found a few applications at the very high end scientific/government markets, but the required supercomputers of the era priced AI out of the reach of virtually everyone else. Machine Learning could be considered as an evolution of AI, because it blends AI heuristics with advanced statistical methods. It let computer programs learn about the data they study and then apply learned knowledge to data.

3. Databases: Third family is Databases. Huge amount of data needs to be stored in a repository, and that too needs to be managed. So, comes in light the databases. Earlier data was managed in records and fields, then in various models like hierarchical, network etc. Relational model served the needs of data storage for long while. Other advanced system that emerged are object relational databases. But in data mining, volume of data is too high, so we need specialized servers for it. We call the term as Data Warehousing. Data warehousing also supports OLAP operations to be applied on it, to support decision making .

4. Other Technologies: Apart from these, data mining inculcates various other areas, e.g. pattern discovery, visualization, business intelligence etc. The table summarizes the evolution data mining on the grounds of development in databases.

IV. 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.

1. Traditional Data Mining Tools: Traditional data mining programs help companies establish data patterns and trends by using a number of complex algorithms and techniques. Some of these tools are installed on the desktop to monitor the data and highlight trends and others capture information residing outside a database. The majority are available in both Windows and UNIX versions, although some specialize in one operating system only. In addition, while some may concentrate on one database type, most will be able to handle any data using online analytical processing or a similar technology.

2. **Dashboards:** Installed in computers to monitor information in a database, dashboards reflect data changes and updates onscreen — often in the form of a chart or table — enabling the user to see how the business is performing. Historical data also can be referenced, enabling the user to see where things have changed (e.g., increase in sales from the same period last year). This functionality

makes dashboards easy to use and particularly appealing to managers who wish to have an overview of the company's performance.

C. Text-mining Tools : The third type of data mining tool sometimes is called a text-mining tool because of its ability to mine data from different kinds of text — from Microsoft Word and Acrobat PDF documents to simple text files, for example. These tools scan content and convert the selected data into a format that is compatible with the tool's database, thus providing users with an easy and convenient way of accessing data without the need to open different applications. Scanned content can be unstructured (i.e., information is scattered almost randomly across the document, including e-mails, Internet pages, audio and video data) or structured (i.e., the data's form and purpose is known, such as content found in a database). Capturing these inputs can provide organizations with a wealth of information that can be mined to discover trends, concepts, and attitudes.

V. DATA MINING: A NEW APPROACH TO INFORMATION OVERLOAD

Many vendors, consultants and analysts make data mining appear complex, difficult, mysterious and expensive. It may sometimes be complex (involving many parts), but it need not be mysterious or difficult. Data Mining simply means:

- **Finding patterns in your data which you can use to better conduct your business:** Some people, especially in the United States, use the term *knowledge discovery* instead of data mining. In this paper, the terms knowledge discovery and data mining are used interchangeably. Both describe the process of discovering a non-obvious pattern in data that can be used to for making better business decisions. It turns out that the vast majority of applications boil down to finding a relatively small number of types of data patterns.
- **Classification:** to which set of predefined *categories* does this case belong? In marketing, when planning a mail shot, the categories may simply be the people who will buy and the people who will not buy. In health care, they may be high-risk and low-risk patients.
- **Association:** which things occur together? For example, looking at shopping baskets you may find that people who buy beer tend also to buy nuts at the same time.
- **Sequence:** is essentially a time-ordered association, although the associated events may be spread far apart in time. For example, you may find that *after* marriage, people buy insurance.
- **Clustering** or **Segmentation:** is like classification except that the categories are not normally known beforehand. You might look at a collection of shopping baskets and discover that there are clusters corresponding to health food buyers, convenience food buyers, luxury food buyers and so on. Data mining is not mysterious; it is simply applied common or business sense.
- **Link Mining:** One spouseless property of the web are the links which are typically represent the structure of the web. Analyzing these links is often referred as Link Mining and is becoming very popular is the last years. Link Mining is mainly divided into three major tasks : the object related task like clustering based on links, prediction of (missing) links and a graph centered task like sub graph discovery.
- **Statistical Relation Learning:** Statistical Relation Learning (SRL) focuses on the combination of probabilistic and logic models with the goal to develop one combined approach which is better able to describe real world phenomena. Methods developed in this area are typically applied on richly structural data which are available for e.g. Hypertext classification, topic prediction of bibliographic entries, or on any kind of social networks.
- **Social Network Analysis:** SNA has a long-standing tradition, with important steps being the modeling of social relationships within 'sociograms'. SNA techniques analyze the network as a whole, or study properties of the individual nodes. Measures for the network as a whole comprise density (percentage of present edges among possible edges), diameter (length of the longest shortest path between any two nodes), clustering coefficient, etc.

VI. CHALLENGES IN WEB MINING

The web poses great challenges for resource and knowledge discovery based on the following observations –

- **The web is too huge** − The size of the web is very huge and rapidly increasing. This seems that the web is too huge for data warehousing and data mining.
- **Complexity of Web pages** − The web pages do not have unifying structure. They are very complex as compared to traditional text document. There are huge amount of documents in digital library of web. These libraries are not arranged according to any particular sorted order.
- **Web is dynamic information source** − The information on the web is rapidly updated. The data such as news, stock markets, weather, sports, shopping, etc., are regularly updated.
- **Diversity of user communities** − The user community on the web is rapidly expanding. These users have different backgrounds, interests, and usage purposes. There are more than 100 million workstations that are connected to the Internet and still rapidly increasing.

VII. THE SCOPE OF DATA MINING

Data mining derives its name from the similarities between searching for valuable business information in a large database — for example, finding linked products in gigabytes of store scanner data — and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find exactly where the value resides. Given databases of sufficient size and quality, data

mining technology can generate new business opportunities by providing these capabilities:

- **Automated 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. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.
- **Automated 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.

VIII. FUTURE TRENDS AND APPLICATIONS

1. DISTRIBUTED/COLLECTIVE DATA MINING

One area of data mining which is attracting a good amount of attention is that of distributed and collective data mining. Much of the data mining which is being done currently focuses on a database or data warehouse of information which is physically located in one place. However, the situation arises where information may be located in different places, in different physical locations. This is known generally as distributed data mining (DDM). Therefore, the goal is to effectively mine distributed data which is located in heterogeneous sites.

2. UBIQUITOUS DATA MINING (UDM)

The advent of laptops, palmtops, cell phones, and wearable computers is making ubiquitous access to large quantity of data possible. Advanced analysis of data for extracting useful knowledge is the next natural step in the world of ubiquitous computing. Accessing and analyzing data from a ubiquitous computing device offer many challenges.

3. HYPERTEXT AND HYPERMEDIA DATA MINING

Hypertext and hypermedia data mining can be characterized as mining data which includes text, hyperlinks, text mark-ups, and various other forms of hypermedia information. As such, it is closely related to both web mining, and multimedia mining, which are covered separately in this section, but in reality are quite close in terms of content and applications. While the World Wide Web is substantially composed of hypertext and hypermedia elements, there are other kinds of hypertext/hypermedia data sources which are not found on the web. Examples of these include the information found in online catalogues, digital libraries, online information databases, and the like.. Some of the important data mining techniques used for hypertext and hypermedia data mining include classification (supervised learning), clustering (unsupervised learning), semi-structured learning, and social network analysis. In the case of classification, or supervised learning, the process starts off by reviewing training data in which items are marked as being part of a certain class or group.

4. MULTIMEDIA DATA MINING

Multimedia Data Mining is the mining and analysis of various types of data, including images, video, audio, and animation. The idea of mining data which contains different kinds of information is the main objective of multimedia data mining. As multimedia data mining incorporates the areas of text mining, as well as hypertext/hypermedia mining, these fields are closely related. Much of the information describing these other areas also applies to multimedia data mining. This field is also rather new, but holds much promise for the future.

5. SPATIAL AND GEOGRAPHIC DATA MINING

The data types which come to mind when the term data mining is mentioned involves data as we know it—statistical, generally numerical data of varying kinds. However, it is also important to consider information which is of an entirely different kind—spatial and geographic data which could contain information about astronomical data, natural resources, or even orbiting satellites and spacecraft which transmit images of earth from out in space. Much of this data is image-oriented, and can represent a great deal of information if properly analyzed and mined. A definition of spatial data mining is as follows: ―the extraction of implicit knowledge, spatial relationships, or other patterns not explicitly stored in spatial databases. Some of the components of spatial data which differentiate it from other kinds include distance and topological information, which can be indexed using multidimensional structures, and required special spatial data access methods, together with spatial knowledge representation and data access methods, along with the ability to handle geometric calculations.

6. CONSTRAINT- BASED DATA MINING

Many of the data mining techniques which currently exist are very useful but lack the benefit of any guidance or user control. One method of implementing some form of human involvement into data mining is in the form of constraint-based data mining. This form of data mining incorporates the use of constraints which guides the process. Frequently this is combined with the benefits of multidimensional mining to add greater power to the process. There are several categories of constraints which can be used, each of which has its own characteristics and purpose. These are:

- **Knowledge-type constraints.** This type of constraint specifies the ―type of knowledge‖ which is to be mined, and is typically specified at the beginning of any data mining query. Some of the types of constraints which can be used include clustering, association, and classification.
- **Data constraints**. This constraint identifies the data which is to be used in the specific data mining query. Since constraint-based mining is ideally conducted within the framework of an ad-hoc, query driven system, data constraints can be specified in a form similar to that of a SQL query.

- **Dimension/level constraints**. Because much of the information being mined is in the form of a database or multidimensional data warehouse, it is possible to specify constraints which specify the levels or dimensions to be included in the current query.
- **Interestingness constraints**. It would also be useful to determine what ranges of a particular variable or measures are considered to be particularly interesting and should be included in the query.

CONCLUSION

Data mining initially generated a great deal of excitement and press coverage, and, as is common with new "technologies", overblown expectations. 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] J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2000.

[2] G. Piatetsky-Shapiro, U. M. Fayyad, and P. Smyth. From data mining to knowledge discovery: An overview. In U.M. Fayyad, et al. (eds.), Advances in Knowledge Discovery and Data Mining, 1-35. AAAI/MIT Press, 1996.

[3] The WEKA data mining sodtware: An update, Mark Hall, Eibe Frank, G. Holmes, B. Pfahringer, P. Reutemann, IH Witten, ACM SIGKDD Explorations, Newsletter, Pages 10-18, volume 11 issue 1, june 2009.

[4] DBMiner: A System for Data Mining in Relational Databases and Data Warehouses, Data Mining Research Group, Intelligent Database Systems Research Laboratory School of Computing Science, Simon Fraser University, British Columbia, Canada, http: db.cs.sfu.ca DBMiner.

[5] www.uea.ac.uk/polopoly_fs/1.3589!introductionkdd.pdf

[6] Heikki, Mannila, ―Data mining: machine learning, statistics, and databases‖, Statistics and Scientific Data Management, pp. 2-9. 1996.

[7] Knowledge Discovery in Databases, AAAI Press / the MIT Press, Massachusetts Institute of Technology. ISBN 0– 26256097–6. MIT1996.

[8] Chakrabarti, van den Berg, and Dom. ―Distributed Hypertext Resource Discovery through Examples, ―Proceedings of the 25thVLDB (International Conference on Very Large Data Bases), Edinburgh Scotland, 1999.

[9] Han, J. and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2001.

[10] Han, J., M. Kamber, and A. K. H. Tung, "Spatial Clustering Methods in Data Mining: A Survey", H. Miller and J. Han (eds.), Geographic Data Mining and Knowledge Discovery, Taylor and Francis, 2001.

[11] Miller and J. Han (eds.), Geographic Data Mining and Knowledge Discovery, Taylor and Francis, 2001.

[12] Kotsiantis, S., Kanellopoulos, D., Pintelas, P., ―Multimedia mining‖, SEAS Transactions on Systems, No 3, s. 3263-3268, 2005.

[13] Huysmans, Baesens, Martens, Denys and Vanthienen, ―New Trends in Data Mining‖, Tijdschrift voor Economie en Management, vol. L, 4, 2005.

[14] Olfa Nasraoui and Maha Soliman, ―Market-Based Profile Infrastructure: Giving Back to the User‖, Next Generation of Data Mining, Taylor and Francis, 2008.

[15] Salmin, Sultana et al., ―Ubiquitous Secretary: A Ubiquitous Computing Application Based on Web Services Architecture‖, InternationalJournal of Multimedia and Ubiquitous Engineering Vol. 4, No. 4, October, 2009.

[16] Jing He, ―Advances in Data Mining: History and Future‖, Third international Symposium on Information Technology Application, 978-0-7695-3859-4/09 IEEE 2009 DOI 10.1109/IITA.2009.204.

[17] M.S. Chen, J. Han, and P.S. Yu. ―Data mining: An overview from database perspective‖, IEEE Transactions on Knowledge and Data Eng., 8(6):866-883, December 1999

[18] Venkatadari M., Dr. Lokanataha C. Reddy, ―A Review on Data Mining From Past to Future‖, International Journal of Computer Applications, pp.19-22, vol. 15, No. 7, Feb 2011.

[19]Agrawal, R., Srikant, R.: Privacy-Preserving Data Mining. In: SIGMOD Conference.