**Available Online at www.ijcsmc.com**

**International Journal of Computer Science and Mobile Computing**

**A Monthly Journal of Computer Science and Information Technology**

**ISSN 2320–088X**

*IJCSMC, Vol. 3, Issue. 4, April 2014, pg.572 – 579*

**RESEARCH ARTICLE**

# Data Mining Technique its Needs and Using Applications

**Anup Arvind Lahoti**
Student of Master of Engineering in (CSE)
HVPM's college of Engineering and Technology
Amravati, India
anuplahotiom@gmail.com

**Prof. P. L. Ramteke**
Associate Professor and Head of the Department of (IT)
HVPM's College of Engineering and Technology
Amravati, India
pl_ramteke@rediffmail.com

***Abstract - Data mining is a process of using different algorithms to find useful patterns or models from data. It is a process of selecting, exploring and modeling large amount of data. Mostly used in technology and also used in different areas of real life like finance, marketing, business. It can also be used in different social science methodologies, such as psychology, cognitive science and human behavior etc. The ability to continually change and acquire new understanding is a driving force for the application of DM. This allows many new future applications of data mining [1] as in today's world the need of data in every field is growing very vast. So, for satisfying our need there should be proper Data Mining techniques available. In this paper we are presenting valuable information about data mining.***

***Keywords – Data Mining, Knowledge Discovery in Databases, Data Warehouses, Clusters, Educational data mining, Oracle Data Mining***

## I. Introduction

The term "Data Mining" appeared around 1990 in the database community. Data mining is the analysis step of the "Knowledge Discovery in Databases (KDD)" process [2]. This is like an interdisciplinary subfield of computer science,[3][4][5] is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable

 

structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, complexity considerations, post-processing of discovered structures, visualization, and online updating[3].

The actual data mining task is like an 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). This usually involves using database techniques such as spatial indices. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics. For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system. The data integration, data transformation, result interpretation and reporting are not part of the data mining step, but do belong to the overall KDD process as additional steps. These steps and its working are shown in the following figure.

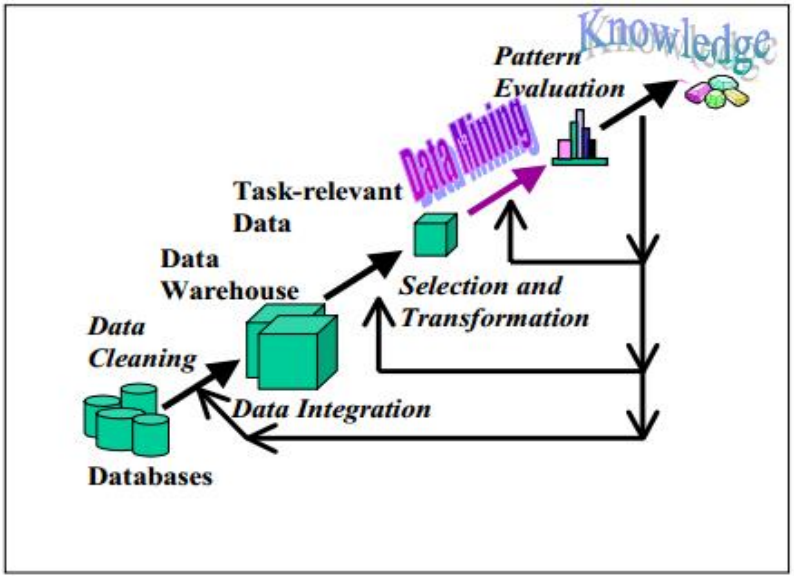


Figure 1: Data Mining is the core of Knowledge Discovery process

In principle, data mining is not specific to one type of media or data. Data mining should be applicable to any kind of information repository. However, algorithms and approaches may differ when applied to different types of data. Indeed, the challenges presented by different types of data vary significantly. Data mining is being put into use and studied for databases, data warehouses, transactional databases, the World Wide Web, etc. In this paper we are giving its brief introduction in section 1. In section 2 we introduce its background concept towards the requirement of data mining. The basics of data mining are given in section 3. How Data Mining works is given in next section. In section5 we are giving some application where data mining is used. And finally in last section we conclude the paper.

## II. Background towards Requirement

Now in today's increasing computational environment, nearly all of our activities from birth until death Stored or leave as digital traces. Health records, schools attended, college record, office record, wages earned, — these and countless other data capturing the details of our daily lives serve as our digital social footprint. Collectively, these digital records—across a group, town, county, state, or nation—form a population's *social genome*, it act as the footprints of our society in general. If data is properly integrated, analyzed, and interpreted, social genome data could offer crucial insights to best serve our greatest societal priorities like healthcare, economics, education, and employment. Population information is the burgeoning field at the intersection of social sciences, health sciences, computer science, and statistics that applies quantitative methods and computational tools

 

for large amount of data. Personal data stored in the Cloud may contain account numbers, passwords, notes, and other important information that could be used at any time.

Today, cloud computing generates a lot of hype, it's both promising and scary. Businesses see its potential but also have many concerns. It's based on the concept of time-shared remote services that isn't new one; cloud computing infrastructures use new technologies and services, some of which haven't been fully evaluated with respect to security. Now a days, Personal health record (PHR) is an emerging patient-centric model of health information exchange, which is often outsourced to be stored at a third party, such as cloud providers. However, there have been wide privacy concerns as personal health information could be exposed to those third party servers and to unauthorized parties. There should be some assurance to the patients' control over access to their own PHRs.

It is a promising method to encrypt the PHRs before outsourcing. In the course of doing business, sometimes sensitive data must be handed over to supposedly trusted third parties. For example, a hospital may give patient records to researchers who will devise new treatments. Similarly, a company may have partnerships with other companies that require sharing customer data. Another enterprise may outsource its data processing, so data must be given to various other companies. Privacy protection is one of the fundamental security requirements for database outsourcing. A major threat is information leakage from database access patterns generated by query executions while mining data from such different data sets. The standard private information retrieval (PIR) schemes, which are widely regarded as theoretical solutions, entails computational overhead per query for a database with items.

The manual extraction of patterns from data has occurred for centuries. Early methods of identifying patterns in data include Bayes' theorem (1700s) and regression analysis (1800s). The proliferation, ubiquity and increasing power of computer technology has dramatically increased data collection, storage, and manipulation ability. As data sets have grown in size and complexity, direct "hands-on" data analysis has increasingly been augmented with indirect, automated data processing, aided by other discoveries in computer science, such as neural networks, cluster analysis, genetic algorithms (1950s), decision trees (1960s), and support vector machines (1990s). Data mining is the process of applying these methods with the intention of uncovering hidden patterns [5] in large data sets. It bridges the gap from applied statistics and artificial intelligence that usually provide the mathematical background to database management by exploiting the way data is stored and indexed in databases to execute the actual learning and discovery algorithms more efficiently, allowing such methods to be applied to ever larger data sets. 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 sets.

## III. Data Mining Basics

Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Although data mining is a relatively new term, the technology is not. There is continuous innovation in this field companies have used powerful computers to sift through volumes of supermarket scanner data and analyze market research reports for years.

### *A. Data*

Data are any facts, numbers, or text that can be processed by a computer. Today, organizations are accumulating vast and growing amounts of data in different formats and different databases. This includes:

- operational or transactional data such as, sales, cost, inventory, payroll, and accounting
- nonoperational data, such as industry sales, forecast data, and macro-economic data
- meta data - data about the data itself, such as logical database design or data dictionary definitions

### *B. Information*

The patterns, associations, or relationships among all this *data* can provide *information*. For example, analysis of retail point of sale transaction data can yield information on which products are selling and when.

### *C. Knowledge*

Information can be converted into *knowledge* about historical patterns and future trends. For example, summary information on retail supermarket sales can be analyzed in light of promotional efforts to provide knowledge of consumer buying behavior. Thus, a manufacturer or retailer could determine which items are most susceptible to promotional efforts.

 

*D. Data Warehouses*

Dramatic advances in data capture, processing power, data transmission, and storage capabilities are enabling organizations to integrate their various databases into *data warehouses*. Data warehousing is defined as a process of centralized data management and retrieval.. Data warehousing represents an ideal vision of maintaining a central repository of all organizational data. Centralization of data is needed to maximize user access and analysis of their data. Dramatic technological advances are making this vision a reality for many companies. And, equally dramatic advances in data analysis software are allowing users to access this data freely.

*E. What can data mining do?*

Data mining is primarily used today by companies with a strong consumer focus - retail, financial, communication, and marketing organizations. It enables these companies to determine relationships among "internal as well as external". Internal factors such as price, product positioning, or staff skills, and external factors such as economic indicators, competition, and customer demographics. And, it enables them to determine the impact on sales, customer satisfaction, and corporate profits. Finally, it enables them to "drill down" into summary information to view detail transactional data.

With data mining, a retailer could use point-of-sale records of customer purchases to send targeted promotions based on an individual's purchase history. By mining demographic data from comment or warranty cards, the retailer could develop products and promotions to appeal to specific customer segments.

For example, WalMart is pioneering massive data mining to transform its supplier relationships. WalMart captures point-of-sale transactions from over 2,900 stores in 6 countries and continuously transmits this data to its massive 7.5 terabyte Teradata data warehouse. WalMart allows more than 3,500 suppliers, to access data on their products and perform data analyses.

## IV. Working of data mining

While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks. Generally, any of four types of relationships are sought:

- Classes: Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.
- Clusters: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.
- Associations: Data can be mined to identify associations. The beer-diaper example is an example of associative mining.
- Sequential patterns: Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

*A. Data mining involves six common classes of tasks[2]*

- Anomaly detection – The identification of unusual data records, that might be interesting or data errors that require further investigation.
- Association rule learning (Dependency modeling) – Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits. For example- 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.
- 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.
- 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".
- Regression – attempts to find a function which models the data with the least error.

 

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

### *B. Different levels of analysis*

- Artificial neural networks: Non-linear predictive models that learn through training and resemble biological neural networks in structure.
- Genetic algorithms: Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.
- Decision trees: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID) . CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. CART segments a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID.
- Nearest neighbor method: A technique that classifies each record in a dataset based on a combination of the classes of the *k* record(s) most similar to it in a historical dataset (where *k* 1). Sometimes called the *k*-nearest neighbor technique.
- Rule induction: The extraction of useful if-then rules from data based on statistical significance.
- Data visualization: The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

### *C. Technological infrastructure which is required*

Today, data mining applications are available on all size systems for mainframe, client/server, and PC platforms. System prices range from several thousand dollars for the smallest applications up to $1 million a terabyte for the largest. Enterprise-wide applications generally range in size from 10 gigabytes to over 11 terabytes. NCR has the capacity to deliver applications exceeding 100 terabytes. Relational database storage and management technology is adequate for many data mining applications less than 50 gigabytes.

There are two critical technological drivers:

- Size of the database: the more data being processed and maintained, the more powerful the system required.
- Query complexity: the more complex the queries and the greater the number of queries being processed, the more powerful the system required.

## V. Application Examples of Data Mining

### *A. Educational Data Mining*

This work aims at the concept of enhancing the Technical Higher Education to address sustainability and globalization, by doing the research in a novel merged area "Educational data mining" (EDM). Sustainability in education can be achieved by improving the existing education system and by restructuring and reorienting the same. As per the research done in Shandong University, to attain sustainability, more focus should be given in creating student's opportunities and conducting sustainability related researches in Higher Education. A novel merged area called Educational data mining (EDM) can be used in creating students opportunities by using different DM algorithms.

Educational Data Mining is an emerging interdisciplinary filed, concerned with developing data mining methods for exploring the massive data of the educational institutions , and using those methods to better understand students, and the environment in which they learn. The educational data is taken from interactive learning environments, computer-supported collaborative learning environment, or administrative data from the universities, It allows its researchers to use the data mining models to attempt the diversity of research in educational field, working in all the environment like online teaching, classroom or the mixture of both i.e. the blended mode. Researchers study a variety of areas, including individual learning factors, factors that are associated with student's failure or non-retention in courses. In particular, the EDM focuses on the prediction, innovative work done on

 

existing models to understand new knowledge of students learning pattern, their emotional and intelligent quotients, the teaching learning process. [7], [8].

As an example the student's enrollment database has all the attributes of the students from their past data from which we can predict the student's caliber, intelligence, and interest in subject opting for. Sometimes students select the subject under external pressure, which can be monitored at the entry level and can guide the students to select the subject of his / her interest by their own. This leads the university to improve the learning and hence the research. Also Factors like previous year results, attendance, financial status of family, parent educational qualification plays an important role in students' education. By applying various DM algorithm to these factors we can predict the student's outcome. The faculty database can be helpful to faculties for their research, their appraisal, and overall talent management to the HR.

*B. Computer based employee performance management data mining system*

Majority of the organizations in the NPS use manual paper technologies in the form of time-books, logbooks, among others, as process monitoring and tracking mechanisms. Because of these complexities, the APER is designed to award subjective ratings, since evidential justification is not a requirement. Consequently, the supervisors are at liberty to use their whims and caprices in rewarding and penalizing perceived cronies and foes. This trend is dangerous since it leads to inequity in the evaluation standard.

However, [13] and [14] warn that the practice of isolating employees' performance management process from organizational business processes is inappropriate, as it will introduce many kinds of errors in the data used for staff performance evaluations, which could lead to inaccurate compensations and missed career opportunities, and consequently, create employee dissatisfaction. The feedback reports can contain such messages as: commendations, opportunities to achieve, the scope to develop skills, and guidance on career paths. All these are nonfinancial rewards, which can make a longer-lasting and more powerful impact than financial rewards. And [10] also condemns this trend and states that a culture that allows a once-a-year feedback in the form of employee performance evaluations is a culture that encourages management malfeasance. To reverse these negative trends, the right kind of information has to be collected on a continuous basis, and used to inform learning and decisions-making, which in turn should lead to performance improvements. For this purpose, we present the structures for the development of a computer based Employee Performance Management Data Mining System.

The weaknesses identified in the previous management system could be strengthened with a computer based system that integrates the employee's performance BPI in such a way that it is promptly available, directly accessible to management, and independently verifiable, with a view to removing biases from supervisors. Such design will also provide regular monitoring and reporting, as well as reduce the burden of computational complexities on the human beings. To achieve this, we adopt the PMS process recommended in [12], who affirm for a PMS to be effective, it has to be a network of activities in the stages. As stated, the PMS process is an on-going communications process. The outcome of one stage provides the input for the next one in a dynamic manner. Consequently, it should be represented as a cycle of inter-related and interdependent activities that contribute to the attainment of organizational goals.

*C. Oracle Data Mining*

Oracle Data Mining (ODM) provides powerful data mining functionality as that of native SQL functions within the Oracle Database. It enables users to discover new insights hidden in data and to leverage investments in Oracle Database technology. With Oracle Data Mining, you can build and apply predictive models that help you target your best customers, develop detailed customer profiles, and find and prevent fraud. Oracle Data Mining, a component of the Oracle Advanced Analytics Option, helps companies better "compete on analytics." The Oracle Data Miner "work flow" based GUI, an extension to SQL Developer, allows data analysts to explore their data, build and evaluate models, apply them to new data and save and share their analytical methodologies. Data analysts and application developers can use the SQL APIs to build next-generation applications that automatically mine star schema data to build and deploy predictive models that deliver real-time results and predictions throughout the enterprise. Because the data, models and results remain in the Oracle Database, data movement is eliminated, information latency is minimized and security is maintained.

Data analysts can quickly access their Oracle data using Oracle Data Miner work flow based graphical user interface and explore their data to find patterns, relationships, and hidden insights. Oracle Data Mining provides a collection of in-database data mining algorithms that solve a wide range of business problems. Anyone who can access data stored in an Oracle Database can access Oracle Data Mining results-predictions, recommendations, and

 

discoveries using Solutions. It is an extension to Oracle SQL Developer that enables data analysts to work directly with data inside the database, explore the data graphically, build and evaluate multiple data mining models, apply Oracle Data Mining models to new data and deploy Oracle Data Mining's predictions and insights throughout the enterprise. Oracle Data Miner work flows capture and document the user's analytical methodology and can be saved and shared with others to automate advanced analytical methodologies.

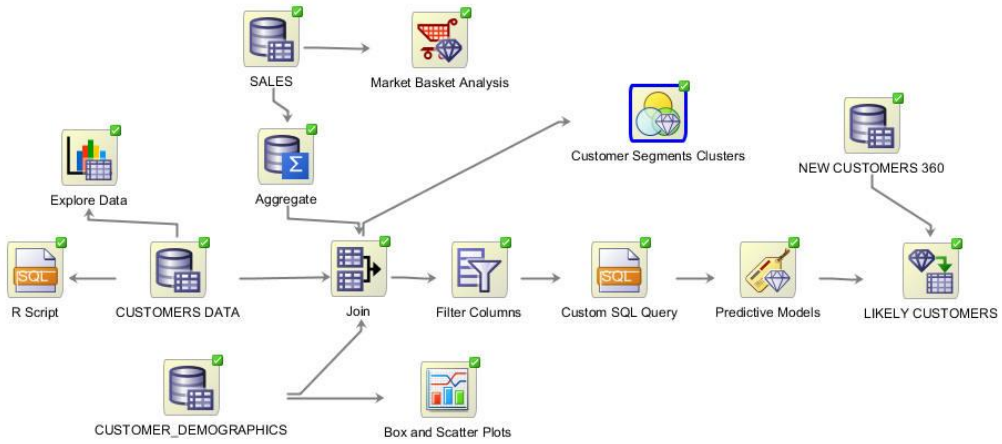


Figure 2. Oracle Data Mining Framework

## VI. Conclusion

With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, if not necessary, to develop powerful means for analysis and perhaps interpretation of such data and for the extraction of interesting knowledge that could help in decision-making. 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 data, information, knowledge and data warehouse as a storehouse, is a repository of data collected from multiple data sources all are important terms, there should be some proper mechanism to mining data from them. We also had seen some live applications where Data Mining is used in educational sector and management sector and performing well.

## References

[1] Shu-Hsien Liao, Pei-Hui Chu, Pei-Yuan Hsiao, "Data mining techniques and applications – A decade review from 2000 to 2011", Expert Systems with Applications 39 (2012) 11303–11311 .
[2] Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996). "From Data Mining to Knowledge Discovery in Databases". Retrieved 17 December 2008.
[3] [a b c d] "Data Mining Curriculum". ACM SIGKDD. 2006-04-30. Retrieved 2011-10-28.
[4] Clifton, Christopher (2010). "Encyclopædia Britannica: Definition of Data Mining". Retrieved 2010-12-09.
[5] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2009). "The Elements of Statistical Learning: Data Mining, Inference, and Prediction". Retrieved 2012-08-07.
[6] Kantardzic, Mehmed (2003). *Data Mining: Concepts, Models, Methods, and Algorithms*. John Wiley & Sons. ISBN 0-471-22852-4. OCLC 50055336.
[7] S. ELATIA, D. IPPERCIEL, & A. HAMMAD," Implications and Challenges to Using Data Mining in Educational" , Canadian Journal of Education 35, 2 (2012): 101-119

 

[8] Romero, C., Ventura, S. (2007). Educational Data Mining: A Survey from 1995 to 2005. Expert Systems with Applications, 33, 125-146
[9] Tang, T., McCalla, G. (2005) Smart recommendation for an evolving elearning system: architecture and experiment, International Journal onE-Learning, 4(1), 105–129.,
[10] API, "Performance Management in the Public Sector – What's the score?", 2011
[11] M. Armstrong, "Performance Management: Key Strategies and Practical Guidelines," London: Kogan Page, 2009
[12] A. S. DeNisi, and R. W. Griffin, "Human Resource Management," New York: Houghton Mifflin Company, 2001
[13] K. H. Han, S. H. Choi, J. G. Kang and G. Lee, "Performance-Centric Business Activity Monitoring Framework for Continuous Process Improvement," Recent Advances in Artificial Intelligence, Knowledge Engineering and Data Bases. Seoul: Autoever Systems Corp, 2011
[14] D. Otley, "Performance Management: A Framework for Analysis," In R. Thorpe & J. Holloway (Eds.), Performance Management: Multidisciplinary Perspectives (pp. 24-39). New York: PALGRAVE MACMILLAN, 2008
[15] http://webdocs.cs.ualberta.ca/~zaiane/courses/cmput690/notes/Chapter1/ch1.pdf
[16] http://pages.cs.wisc.edu/~dbbook/openAccess/thirdEdition/slides/slides3ed-english/Ch26-DataMining-95.pdf

 