International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

COPY RIGHT

IJIEMR Transactions, online available on 23rd Dec 2019. Link :http://www.ijiemr.org/downloads.php?vol=Volume-08&issue=ISSUE-12

Title: DATA MINING AND ITS APPLICATIONS

Volume 08, Issue 12, Pages: 404–411.

Paper Authors

MR. POLASI SUDHAKAR , DR. V SURYANARAYANA

Ramachandra College of Engineering, Eluru, A.P, India

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

DATA MINING AND ITS APPLICATIONS

#1MR. POLASI SUDHAKAR *2DR. V SURYANARAYANA
[1]Assistant Professor, Department of CSE, Ramachandra College of Engineering, Eluru, A.P, India.
[2]Professor, Department of CSE, Ramachandra College of Engineering, Eluru, A.P, India.
sudhakar.forall@gmail.com, S_vadhri@yahoo.co.in,

ABSTRACT: Data Mining is the process of extracting or mining knowledge from large amount of data. Data mining is the technology that meets up to the challenge of solving our quest for knowledge from these vast data burdens. It provides us with a user oriented approach to novel hidden patterns in data. Important disciplines ranging from machine learning, information retrieval, statistics and artificial intelligence have had impacts on the development of data mining. This paper evaluates data mining in theory and in practice. An overview of database systems, data warehousing, data mining goals, applications and algorithms was carried out.

KEYWORDS: Data Mining, Data Warehouse Knowledge Discovery in Databases (KDD), Applications.

1. INTRODUCTION

In an information technology driven society, where knowledge is an invaluable asset to any individual, organization or government. Companies are supplied with huge amount of data in daily basis, and there is the need for them to focus on refining these data so as to get the most important and useful information in their data warehouses [1]. Data mining is a new technology which could be used in extracting valuable information from data warehouses and databases of companies and governments. It involves the extraction of hidden information from some huge dataset.It helps in detecting anomalies in data and predicting future patterns and attitude in a highly efficient way. Applying data mining makes it easier for companies and government, during quality decisions from available data, which would have taken longer time, based on human expertise [11, 12].

Data mining techniques could be applied in a wide range of organizations, so long as they deal with collecting data, and there are several data mining software been made available to the market today, to help companies tackle decision making problems and invariably overcome competition from other companies in the same business.Databases been the root technology that lead to data mining in form of evolution, then there is a brief literature on data warehousing and its relation to data mining, since all useful data collected by organizations are kept there, before they could be subjected to any further mining or analysis prior to decision making. There is an overview of data mining as a field, its evolution what motivated its coming into existence, data mining objective and the process of knowledge discovery in databases.

In this paper, an overview of database systems, its evolution, databases, data warehousing, and the relationship between data warehousing and data mining will be made. Database understanding would be incomplete without some knowledge of the major aspects which constitute the building and framework of database systems, and these fields include structured query language (SQL), extended mark-up language (XML), relational databases concepts, object-oriented concepts, client and servers, security, unified modelling language (UML), data warehousing, data mining and emerging applications.

Adding and retrieving information from databases is fundamentally achieved by the use of SQL, while interchanging data on the web was also possible and enhanced by publishing language like hyper text mark Up language (HTML) and XML.

Database systems could be classified as OLTP (on-line transaction process systems, and decision support systems, like warehouses, on-line analytical processing (OLAP) and mining [2,3]. Archive of data from OLTP form decision support systems which have the aim of learning from past instances. It involves many short, update-intensive commands and it is the main function of relational database management systems.

2. DATA WAREHOUSE AND DATA MINING

A database is a well structured aggregation of data that are associated in a meaningful way, which could be accessed in various logical ways by several users. Database systems are systems in which the translation and storage are of paramount value [9]. Database is a collection of organized data put in a way that a computer program could quickly and easily select required parts of the data. It can be presumed as an electronic filing system.

A traditional database is organized into fields, records and fields, where field implies single piece of information, record is a complete set of fields, and file a collection of records [9]. A database management system is needed to be able to access data or information from a database.

Data warehouse:

A data warehouse is an enabled relational database system designed to support very large databases at significantly higher level of performance and manageability. It is an environment and not a product [2]

.A data warehouse is also referred to as a subject-oriented, integrated, time variant and non-volatile collection of data which supports management decision making process. Subject-oriented depicts that all tangible relevant data pertaining to a subject are collected and stored as a single set in a useful format [7]

Integrated relates to the fact that data is being stored in a globally accepted style with consistent naming trends, measurement, encoding structure and physical features even when the underlying operational systems store the data differently.

Non-volatile simply implies that data in a data warehouse is in a read-only state, hence can be found and accessed in the warehouse.

Time-variant denotes the period the data has been available, because such data are usually of long term states.

The process of constructing and using data warehouses is called data warehousing. Data warehouses comprise of consolidated data from several sources, augmented with summary information and covering a long time period. They are much larger than other kinds of databases, having sizes ranging from several gigabytes to terabytes

Typical workloads involving ad hoc, fairly complex queries and fast response times are important.

OLAP however is a basic function of a data warehouse system. It focuses on data analysis and decision making, based on the content of the data warehouse and it is subject oriented thus implying it is organized around a certain main subject [5]. It is built by integrating multiple, heterogeneous data sources like flat files, on-line transaction records and relational databases in a given format.

Data cleaning, integration and consolidation techniques are often employed to ensure consistency in nomenclature and could be viewed as an important pre-processing step for data mining, encoding structures, attribute measure and lots more among different data sources [2]. Data warehouses primarily provide information from a historical perspective.

Data Mining:

Data mining also termed as knowledge discovery is the computer-assisted process of digging through and analyzing enormous sets of data and then extracting the meaning of the data [1]. Knowledge discovery involves a process that yields new knowledge, it gives in details sequence of steps (data mining inclusive) that ought to be adhered to in order to discover knowledge in data, however each step is achieved using some software tools [4].It involves several steps, and each attempts to discover some knowledge using some knowledge discovery method.

Relation between data warehouse and data mining:

There has been an explosive growth in database technology and the amount of data collected. The huge size of data and the great computation involved in knowledge discovery hampers the ability to analyze the data readily available in order to extract more intelligent and useful information [13,14]. While data mining is all about to enhance decision making and predictions interestingly data warehousing provides online analytical processing tools for interactive analysis of multidimensional data of varied granularities which enhance data mining and mining functions such as prediction, classification and association could be integrated with OLAP operations thus enhancing mining of knowledge.

Knowledge discovery process models

The knowledge discovery process has been placed into two main models called the Fayyad et al (academic) model and the Anand and Buchner (industrial) model. The Fayyad model is represented in below figure

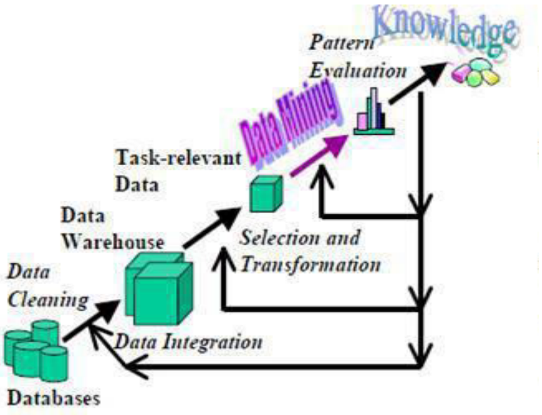

Figure 1. Knowledge Discovery Process Model Architecture

Developing and understanding the application domain entails, learning the useful afore-hand knowledge and aim of the end user for the discovered knowledge [6]. The next phase is creating a target data set, which involves querying the existing data to select the desired subset by selecting subsets of attributes and data points to be used for task [4].

Data cleaning and processing entails eradicating outliers, handling noise and missing values in data, and accounting for time sequence information and known changes. It leads to the data rejection and Projection. It consists of finding valuable attributes by utilizing dimension reduction and transformation methods, and discovering invariant representation of the data.

Consolidation of discovered knowledge involves incorporating discovered knowledge into the performance system, documenting and reporting. The Industrial model also tagged CRISP-DM knowledge discovery process is summarized in the graphs below

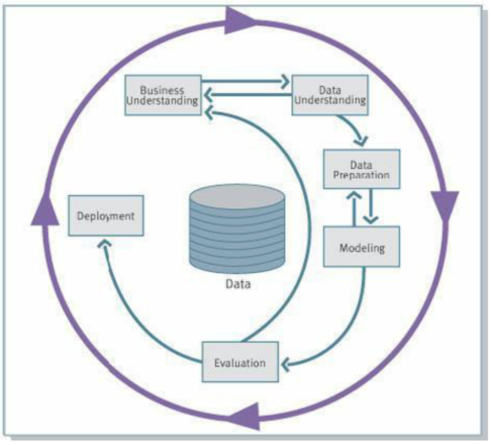

Figure 2. Knowledge Discovery Process Graph

Choosing the Data mining task and this involves matching the relevant prior knowledge and objective of a user with a specific data mining method [8]. Choosing the data mining algorithm basically involves selecting methods to search for patterns in the data and conclude on which models and yardsticks of the method is perfect on which models and yardsticks of the method is perfect. Data mining being the next phase involves pattern generation, in a particular representation form, such as classification, decision tree, etc.

3. NEED AND GOALS OF DATA MINING

***Need of the Data Mining**:* The achievement of digital revolution and the escalation of the internet have brought about a great amount of multi-dimensional data in virtually all human endeavour, and the data type ranges from text, image, audio, speech, hypertext, graphics and video thus providing organizations with too many data, but the whole data might not be useful if it does not provide a tangible unique information that could be utilized in solving a problem [10]. The quest to generate information from existing data prompted the need for data mining.

***Data Mining Goals**:* Data mining is basically done with the aim of achieving certain objectives and it ranges from

Classification: This involves allocating data into classes or categories as a result of combining yardsticks. *Prediction:* Mining in this instance helps to single out features in a data, and their tendencies in the event of time.

***Identification**:* Trends or patterns in certain data could enhance in identifying the existence of items, events or action in a given scenario or case.

***Optimization*:** Mining also facilitates the optimization of the use of scarce resources in turn maximizing output variables within constraint conditions.

4. APPLICATIONS OF DATA MINING

The traditional approach to data analysis for decision making used to involve furnishing domain experts with statistical modelling techniques so as to develop hand crafted solutions for specific problems, but the influx of mega data having millions of rows and columns and the spontaneous constructions and deployment of data driven analytics coupled with demand by users for results easily readable and understandable has prompted the inevitable need for data mining [10].Data mining technologies are deployed in several decision-making scenarios in organizations. Its importance cannot be over emphasized, as it is applicable in several fields some of which as discussed below.

***Marketing** :* This involves analysis of customer behaviour in purchasing patterns, market strategies determination varying from advertising to location, targeted mailing, segmentation of customers, products, stores, catalos design and advertisement strategy

Supply chain visibility: Companies have automated portions of their supply chain enabling collection of significant data about inventory, supply performance and logistic of materials, and finished goods, material expenditures, accuracy of plans for order delivery. Data mining application also spans though price optimization and work force analysis in organizations.

Geospatial decision making: In climate data and earth ecosystem scenario, automatic extraction and analysis of interesting patterns involving modelling ecological data and designing efficient algorithm for finding spatiotemporal patterns in form of tele-connection patterns or recurring and persistent climate patterns. This operation is usually carried out using the clustering technique, which divides the data into meaningful groups, helping to automate the discovery of tele-connections

Biomedicine and science application : Biology used to be a field dominated by an attitude of formulate hypothesis, conduct experiment, evaluate results, but now upon the impact of data mining it has evolved into a field of big science attitude involving collecting and storing data, mine for new hypothesis, then confirm with data or supplemental experiment. It also includes discovery of patterns in radiological images, analysis of microarray (gene-chip) experimental data to cluster genes and to relate to symptoms or disease, analysis of side effects of drugs and effectiveness of certain drugs.

Manufacturing: The application in this aspect relates to optimizing the resources used in optimal design of manufacturing processes, and product design based on customers‟ feedback.

Telecommunications and control: It is applied to the vastly available high volume of data consisting of call records and other telecommunication related data, which in turn is applied in toll-fraud detection, consumer marketing and improving services. Data mining is also applied in security operations and services, information analysis and delivery, text and Web mining instances, banking and commercial applications as well as insurance.

5. DATA MINING ALGORITHMS

Data could only be useful when it is converted into information and it becomes paramount when some knowledge is gained from the generated information, as such is the most vital phase of data handling in any setup that deals with decision making, and this knowledge obtained could be inductive or deductive, where deductive knowledge deduces new information from applying pre specified logical rules on some data. The inductive knowledge is the form of knowledge referred to when data mining is concerned, as it discovers new rules and patterns from some given data. The knowledge acquired from data mining is classified in the forms below, though knowledge could be as a result of a combination of any of them:

Association rules: Simply involves correlating the presence of a set of items [3, 8] with another range of values for another set of variable [1].

Classification hierarchies: This aims at progressing from an existing set if transactions or actions to generate a hierarchy of classes.

Sequential patterns: *B*asically seeks some form of sequence from some events or activities.Patterns within time series: This involves detecting similarities within positions of time series of some data, implying sets of data obtained at regular intervals.

Clustering: This relates to segmentation of some given collection of items or actions into sets of similar elements

Data Mining Algorithms: Data mining algorithms are the mechanisms which create the data mining model, which is the main phase of the data mining process. In the subsequent sub headings the algorithms will be discussed.

1) Naïve Bayes algorithm

It is one of the important data mining algorithms which is used for the purpose of classification. It depends on bayes theorem**.**

2) Apriori algorithm

This algorithm applies a prior knowledge of an important attribute of frequent item-sets [8]. The Apriori property of any item-set declares that all non empty subsets of a frequent item-set has to be frequent, hence where a given item-set is not frequent (if it does not meet up to the minimum support threshold), then all superset of this item-set will also not be frequent, since it cannot occur more frequently than the original item-set.

3) Sampling algorithm

This algorithm is basically about taking a small sample of the main database of transactions, then establishing the frequent item sets from the sample. where such frequent item-sets form a superset of the frequent item-sets of the whole database, then one could affirm the real frequent item sets by scrutinizing the remainder if the database in order to determine the exact support values of the superset item-set.

4) Frequent-pattern tree algorithm

This is also an algorithm which came into been due to the fact that Apriori algorithm [3] involves creating and testing huge amount of item-sets. However, this algorithm eliminates the creation of such large candidate item-sets. A compressed sample of the database is first created, based on the frequent pattern tree; this tree keeps useful item-set information and gives an avenue for the efficient finding of frequent item-sets [11]. The main mining process is divided into smaller task and each functions on a conditional frequent pattern tree, which is a branch of the main tree.

5) Partition algorithm

Partitioning algorithm operates by splitting the database into non-overlapping subsets, which are taken for separate databases and all bulk item-sets for that partition are called Local Frequent item-sets, and they are created in one pass, after which the Apriori algorithm is then efficiently applied on each partition if it fits into the primary memory. Partitions are taken such that each every partition could be accommodated in the main memory, hence been checked only once.

6) Regression

Regression is an exclusive application of the classification rule. If a classification rule is regarded as a function over the variables that map these variables into target class variable, the rule is called a regression rule.

A common application of regression occurs when in place of mapping a tuple of data from some relation to a specific class, the value of variable is predicted based on the tuple itself. Regression involves smoothing data by fitting the data to a function. It could be linear or multiple, the linear involves finding the best line to fit two variables so that one could be used to predict the other, while the multiple one has to do with more than two variables.

7) Neural networks

This is a technique derived from artificial intelligence, using general regression and provides an iterative method to implement it. It operates using a curve fitting approach to infer a function from a given sample. It is a learning approach which uses a test sample for initial learning and inference. Neural networks are placed into two classes namely supervised and unsupervised networks.

8) Genetic algorithm

Genetic algorithms are a class of randomized search procedures capable of adaptive and large search over a large range of search space topologies [1]. It was developed by John Holland in the 1960s. The solutions generated by genetic algorithms are differentiated from that of other techniques because genetic algorithms use a set of solution during each generation instead of a single solution. Genetic algorithm is a randomized algorithm unlike other algorithms, and its ability to solve problems in parallel makes it powerful in data mining.

6. CONCLUSION

Data Mining is the process of extracting or mining knowledge from large amount of data. In this paper database system and techniques, data warehouse , data mining, goals, algorithms, relation between data mining and data warehouse are discussed. Data mining is n important topic of the computer science research in recent years, and it has a extensive applications in various fields. Data mining technology is an application oriented technology. It not only is a simple search, query and transfer on the particular database, but also analyzes, integrates and reasons these data to guide the solution of practical problems and find the relation between events, and even to predict future activities through using the existing data. Data mining brings a lot of benefits to businesses, society, governments as well as individual. However privacy, security and misuse of information are the big problem if it is not address correctly.

REFERENCES

1. Tan, P-N., Steinbach, M., and Vipin, K. (2006). Introduction to Data Mining.Addison-Wesley.
2. Jaiwei Han, Micheline Kamber,Data Mining Concepts and Techniques
3. Agrawal, R., Imielinski, T., and Swami, A. Mining Association Rules between Sets of Items in Large Databases. *Proceedings of ACM SIGMOD*, 2003.
4. Behnke, J., Dobinson, E., NASA Workshop on Issues in the Application of Data Mining to Scientific Data. *ACM SIGKDD Explorations* 2(1):70-79, 2000.
5. Chaudhuri, S., and Dayal, U. An overview of Data Warehousing and OLAP Technology. *SIGMOD Record* 26(1):65-74, 1997.
6. Fayad, U., Piatetsky-Shapiro, G., Smith, P., and Uthurusami, R. *Advances in Knowledge Discovery and Data Mining*, MIT Press, 1996.
7. Inmon, W. Building the Data Warehouse. John Wiley & Sons, 1996.
8. Klemettinen,M., et. al., “Finding Interesting Rules from Large Sets of Discovered Association Rules”, International Conference on Information and Knowledge Management, pp.401-407, Nov. 1994, Maryland
9. Abraham Silberschatz, Henry F. Korth, and Sudarshan (2002). Database System Concepts pp 445-489. 4th Edition, McGraw Hill
10. “Data mining tools”, by Ralf Mikut, Markus Reischl, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,2011

AUTHORS PROFILE

P.Sudhakar working as a Assistant. Professor, Department of Computer Sciences and Engineering at Ramachandra College of Engineering, Permanent Affiliated to JNTK, Kakinada, A.P., India. My researches Interests are data warehousing, Computer Network. He is life member of ISTE

Dr,SURYANARAYANA working as a Professor & HOD, Department of Computer Science and Engineering at Ramachandra College of Engineering, Eluru.