IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

"An integral part of existence: Orientation of Data mining as a unique digital world."

Dr. J. Arockia Venice

Professor and Registrar

DMI - St.Eugene University

Chibombo, Zambia.

Introduction:

'Data Mining' is the latest field of enriched study of the computer technology. In accordance to Britannica Encyclopedia, 'Data Mining', is also called knowledge discovery in databases (KDD), in computer science, the process of discovering interesting and useful patterns and relationships in large volumes of data. It can also be defined as 'the non trivial processof extracting interesting previously unknown, which brings about the discovery part andpotentially useful pattern or knowledge from huge amount of data. The definitions reflect the significance of data discovery and the manner in which it is patterned.

However, it has to be potentially useful and signifies to be an useful knowledge depending on theapplication. So, the knowledge often takes the form of patterns in data with some regularityor some kind of structure in the data taken from huge amount of data which is animportant aspect of comprehending the text content.

Literature review:

Jusoh and Alfawareh, 2012, state the concept of text data mining which is concerned with getting all the desired information out of 'mountains' of textual data is nearly as old as IR itself. However, text mining possesses essential features that distinguish it from IR as well as other related fields. Text mining aims at obtaining useful information from textual data that are inherently unstructured, unorganized, and erratic

Hearst (1999), one of the pioneers in the field of text data mining, provided a complete definition of TDM, wherein she made a clear distinction between TDM and the traditional IR. According to Hearst, traditional IR is concerned with the retrieval of documents that are relevant to a user's information needs (not the retrieval of the information itself), and then, selecting the desired information is left up to the user. On the other hand, TDM

does not only deal with the direct retrieval of information from documents, but also it attempts to discover new patterns of information from documents, such information are useful, non-trivial and unknown previously.

Many Literature reviews give us options to understand the concept of 'Data Mining', emphasizing the durability and reliability of this novel new experience of computer digitalization.

It is a decisively a significant reality due to continuous advancement in technology that is creating an increased digital space in terms with scientific and technical field of study. Here comes the challenge of tracing information or tracking the needed details by officials of computer technology, people of information communications technology and scientists. Hence, came the initiation of 'Data Mining', to help scientists, researchers and system analysts to gather useful information from the available large information spaces.

In fact, nearly 80% to 85% of multinational companies and corporate circles stores huge information in special web text spaces which share common features and digital characteristics.They are not structured in a perfect form, unruffled and fuzzy which make usage difficult. Henceforth, the concealed information is designed for clarity through modes and means of digital techniques and algorithm patterns.

Work compliance of Data Mining:

The specificity of 'Data Mining' lies in intercalation of techniques with combined tools from statistics and artificial intelligence along with data base management in order to analyze data sets, which forms the part of large digital collections. According to Oxford dictionary, a data set is a form of computing. It can be stated as a collection of related sets of information that is made of separate elements that can be manipulated as a unit by a computer. Many such data sets make a pattern of consolidated information made available in data mining.

It is a breakthrough of the mid 19th century and of the recent decade whereby, data mining is largely used in multi-dimensional outlook of the scenario of business. Concerns related to insurance, banking, retail, scientific research, governance and security make use of the concept of 'Data Mining'. Many such established systems from governmental to non-governmental make use of this transmission net works to store and gather data in a cumulative basis.

A breakthrough in 'Data Mining':

It is well known and a established fact that 'data Mining' is the most recent venture of the computer world. An interesting fact to know is that the first International Conference on 'Knowledge Discovery and Data Mining'(KDD), was held in Montreal, Canada in 1995, which was a welcome scenario for many computer analysts and system operators and an experience of novice for scientists and researchers to gather more information to make their projects, progressive work into a perfect form.

This conference led to the initiation and formulation of a journal for 'Data Mining' that made easy access to many diversified text content of eminent and budding scholars on Data Mining. It was in 1997 that the journal named 'Data Mining and Knowledge Discovery' came to form for the audience. Thenceforth, many issues on 'Data Mining' was published and at the same time many early data mining companies and their products were brought to an establishment. The seed sown in mid-19th century grew to a well formed, structured working unit

this 21st century. This has helped the macrocosm to derive and extract useful and novel patterns of knowledge acquisition in the form of data. More precise to state, that the data acquired is from historical data and not from the present existing concerns. Undoubtedly' Data Mining' is an innovative, new process that holds and transforms the information as and when needed.

Structural entity of Data Mining:

More to refer on the importance of 'Data Mining' and in lieu, with the above graphic representation it can be well said that it is possibly one of the most effective kinds of data.However, it is framed and structured synchronizing with relational database, data warehouse or transactional data base. Precisely, it is taken to concern with more advanced applications such as sensor streams. To bring to focus at this juncture, the application of industrial sensors that is used in measuring temperature and pressure in an area. The sensors as used in tracing the sequences of a stock markets. Many more instances of the framework of 'Data Mining' is used in networks like Facebook or as Object Oriented Database. Explication of 'Data Mining' is commonly applied in spatial data geographical information system as a special temporal data. Very interesting to know is the usage of pattern analysis of 'Data Mining' which can be used to discern inconsistent, abnormal behavior of an individual involved in a fraudulent action or any other criminal activity.

Research Methodology:

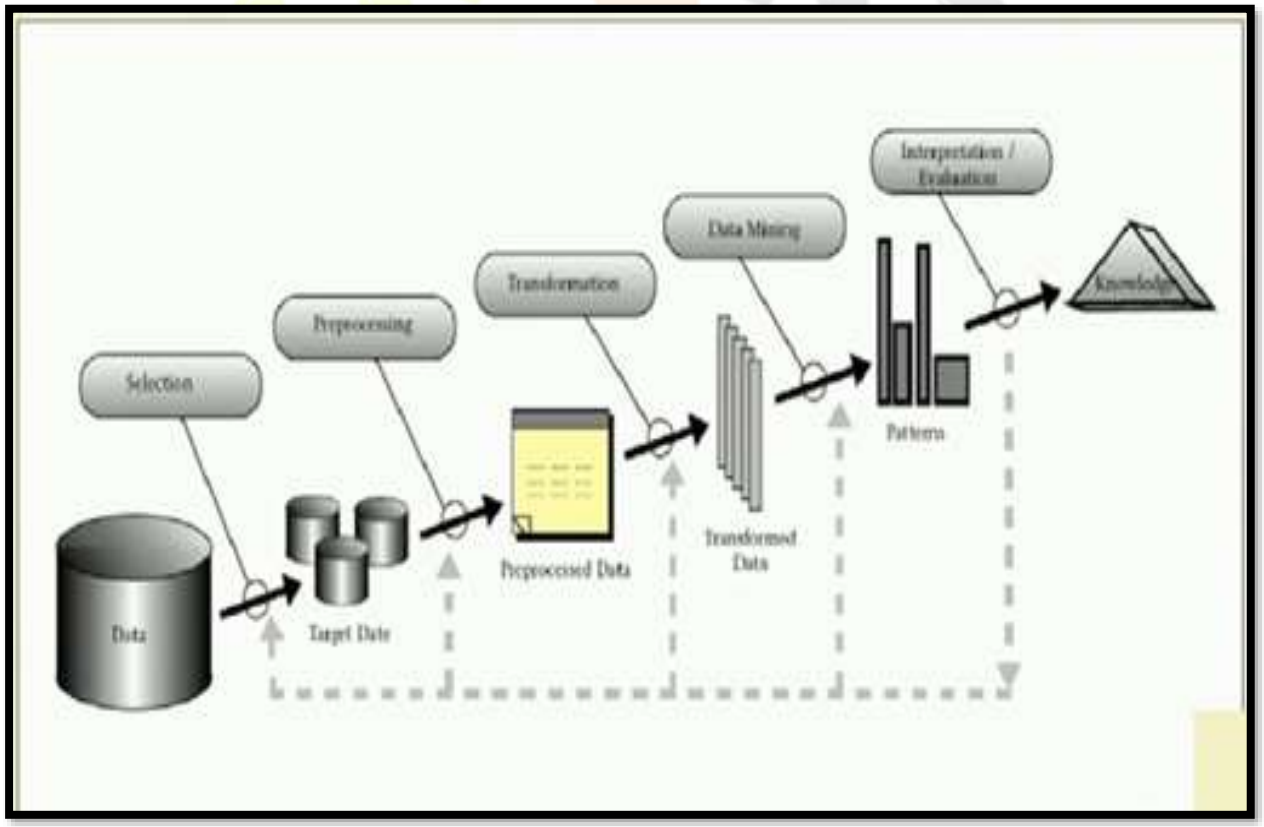
There has been an explicit growth in thevolume of data, because of ease of storage, ease of data collection and because of more computerization of details in various companies intheir day to day applications. There are so many options, tasks, techniques, tools, formats, and approaches to data mining that can support industrial engineers when they find it very difficult to design and implement projects. Although methodologies of many

types of varied functional attributes exist, most probably in computer analysis they are designed for specific software packages. Most of these methodologies use a traditional statistical approach.

It is still not clear that this approach to data mining is sufficient for obtaining the vast array of data needed for many multitasking industrial engineering applications. Thus, a data mining methodology is designed and patterned to meet the specific requirements of industrial engineering and other discovery references in established concern. Such a methodology is applied in selecting appropriate data mining tools and implementing data mining projects from a systems perspective on a large scale.

In most latest and modern perspectives, Data Mining structural details are commonly used business that work on the principles of e-commerce such as Amazon, Flipkart and web-marketing sequences. On a larger perspective, the implications of 'Data Mining' is of foremost importance in search engines of 'Google' wherein the bank transactions, stock market exchanges, remote sensing applications, storage of biological data, scientific stimulations, social media arenas that use digital photography and You Tube applications are put to usage and made secure. At this stance, a huge volume of data is stored and may not readily provide enough knowledge for any first user but on refined strategically enforced modulation is sure to become more adaptable for anyone's use. Henceforth, a method of automated analysis of massive data together with meaningful knowledge is done which has given birth to the concept of 'Data Mining'. The alternate name to this conceptualized idea is 'Knowledge Discovery in Database' (KDD).

A Pictorial Reflection of the process of "Knowledge Discovery in Database'



Working of process of KDD or 'Data Mining':

The initiation step of the process is that the data base is created to hold and store the data. It should be able to store multiple databases of all contributing agents. The next step leads to selection of data. It works on the data that an individual is interested. Segregation of data is made possible. This data of interest for an individual that he has selected is put into data warehouse. Data warehouse is a place where preprocessing techniques are applied. Nevertheless, data warehouse can be related as a place where a normal relational data is modified and stored with further integration into the required form is done that is adequate enough for further data mining that can be retrieved or discovered from the storage.

The preprocessed data can be used to do a data transformation in various methodologies in order to change the data into a suitable form. For a fulfilling perfection of such a stage of application, we have to redo the transformations many times using data analysis tasks. The data transformed from whichever form it has been derived will turn into more insightful information with proper exact information as required. This data so assimilated will be the final selected, preprocessed and transformed data ready for use as core collective data in 'Data Mining'.

Typical Data Mining System: Reflected in a narrative form

This would be a structure of a typical data mining system; you wouldhave a data base in the bottom most layer. We have integration selection cleaning done, then you have to move the data to data warehouse. Further, using a data mining engine, you would find out the patterns of the model and finally to evaluate the patterns and visualize them through an user interface and generate moreknowledge and store it in a knowledge base for further action. The illustration given below will enhance the steps of the narration for clarity.

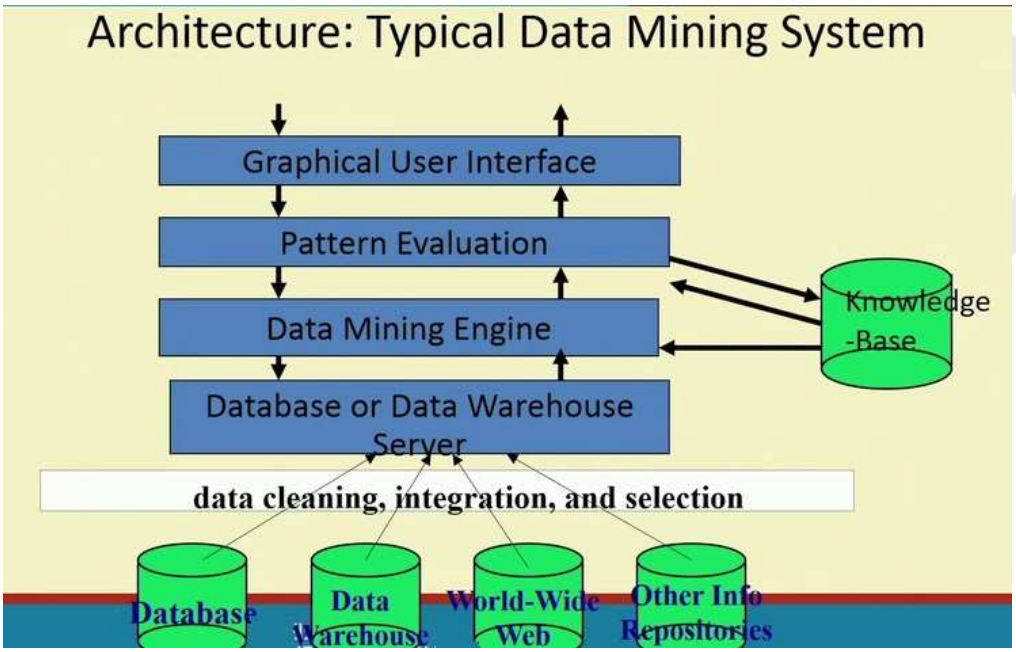


Fig (ii) Data Mining System

The Major concerns of 'Data Mining'

The following are the major issues in Data Mining:

1. Methodology:
   1.1.Mining different types of knowledge from diverse data types (eg. Bio, stream and web)
   1.2.Performance : Efficiency, effectiveness and scalability
   1.3.Pattern Evaluation : the interestingness problem
   1.4.Incorporation of background knowledge
   1.5.Handling noise and incomplete data
   1.6.Parallel,distributed and incremental mining methods
   1.7.Integration of the discovered knowledge with existing one
2. User Interaction
   2.1.Data Mining query languages and ad-hoc mining
   2.2.Expression and visualization of data mining results
   2.3.Interactive mining of knowledge at multiple levels of abstraction
3. Applications and Social impacts
   3.1.Domain specific data mining and invisible data mining
   3.2.Protection of data security, integrity and Privacy

Data Preprocessing :

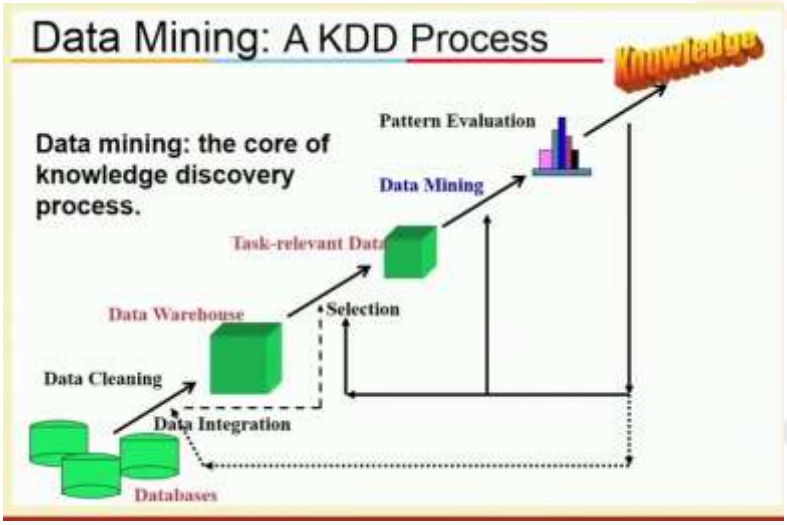


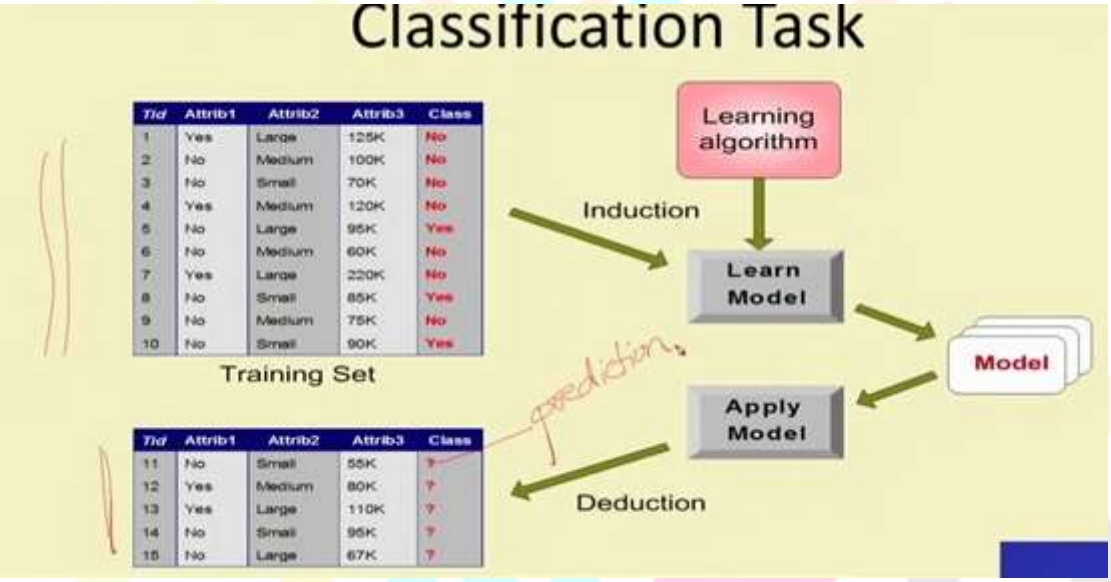
Fig (iii) how data is preprocessed: a methodical application.

As I had discussed before the data mining process consist of the following steps. So, youhave your data in a database, on which you do integration among all the sources of thedata. So, it is not just one database, but

multiple databases and you integrate the data,you clean the data and this integrated and cleaned data is in a form where it is stored insomething called a data warehouse.

So, relevant data means not only selecting the records which are relevant, but also theattributes,and the characteristics of the data of various aspects of the data, in order to formthe relevant data. So, this step of getting information from the data warehouse, where you have integrated data as a relevant data is known as the datapreprocessing step. And finally, what we do is that on this preprocessed data, we fit amathematical model, which describes the patterns present in the data either associationrules or classification rules, we have a mathematical description of this clean data andthen we evaluate the patterns, evaluate the models and those evaluated and visualizedmodel are usable byhuman into actionable form of data, that is what we call aknowledge.

Classification in 'Data Mining': It defines the idea of a given set ofattributes and a set of problems belonging to a class.



Ex 1: Classification task table

Suppose I have a table like this

Let me focus on the table. Look at this table. It is a tableconsisting of some "Attributes'put in form as attribute one, attribute two, attribute three, whose values arelike yes, no, large, medium, small, 125,000 k, 100,000 k values. And the final columnmarked in red is something called a 'Class'. So, each of the rows are object.

So, you can imagine that each row is a person - record ofa person, whose attribute is that he is employed or unemployed, then addressed as yes or no. Attribute2 maybe to his place of habitation,whether he stays in a small home or a large home or a medium home. Attribute 3 is on his annual income, 100,000 k or 125,000k.

And then maybe I want to I know that each of these persons as referred in each of these rows belong toone of the two categories taking to reference, either you give them a loan or if they apply for a loan. Further the thought initiation will be if they will repaytheir loan or if you should not give them a loan, then the options would be 'yes' or 'no'. I should give them a loan or Ishould not give them a loan. So, what I will do is that I have to refer something called an experience or a training set where I will have ten such persons whose referential attribute details andthe class would be in line with the loan to be given or not. It may be a known component.

So, I have this ten persons for whom, it is a known attribute and my job is to work upon five more persons whose attribute values that I know. Right from the fact that they are employed or not, if their house is small or big, if their income is below or more than 55,000 k and if for them whether they have actually been given a loan ornot which I may not know. Then, I want to predict whether they should be given a loan or not. So, among these persons if I know, they are given loan or not given a loan. The transformation of information makes a complete storage of information in 'Data Mining'.

Practical References Enriched Knowledge:

Classification Techniques:

1. Decision Tree based methods
2. Rule-based methods
3. Memory based reasoning
4. Neural networks
5. Naïve Bayes and Bayesian Networks
6. Support Vector machines

Let us see the decision tree based methods and Bayes classification

Decision Trees:

Decision tree to represent learned target functions

- Each internal node tests an attribute
- Each branch corresponds to attribute value
- Each leaf node assigns a classification
- Can be represented as logical formulas

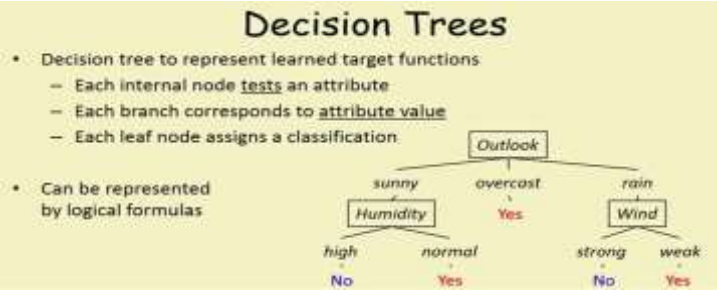


Fig (IV) representation of decision tree

A decision tree is nothing but salient attributes that branch asnodes and further branching on them which finally, get to the formation of a mind map resembling the branches, leaves and a tree.

Clustering:

Clustering means organizing data into classes such that there is

- High intra-class similarity
- Low intra-class similarity
- Finding the class labels and number of classes directly from the data

Clustering is to organize the data the points and the facts into groups which are homogeneous. Homogeneous means if you refer to points from thesame group they are close together, whereas, if you take the points from differentgroups they are apart. So, you need to maximize intra class similarity and minimize interclass similarity among the points. So, this actually amounts to finding the natural groups in the data.

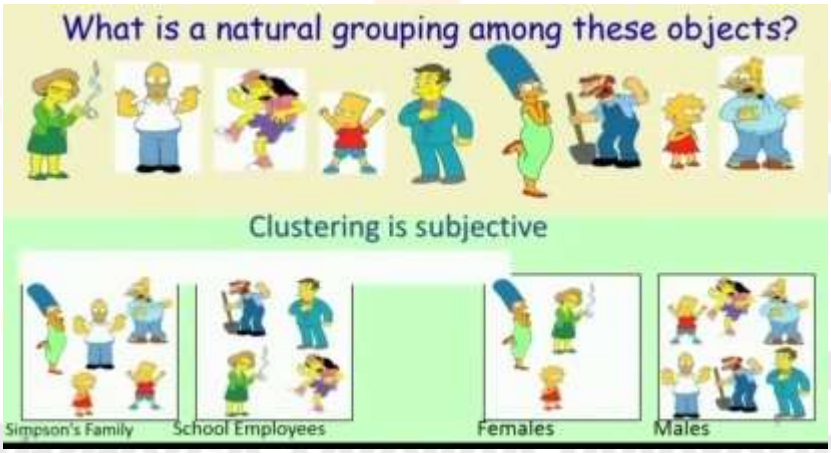


Fig (V) Clustering

So, what do natural groups mean? Taking this image, in accordance to the previous definition, it is said if data is from similar objects, then it is similar. So, this similarity measure defines what a natural group is.

For example, what is the natural grouping among these objects that is shown in thepicture several possible groupings can be made. For example, two groupscan be made with 9 images as seen in the first row. Accordingly, from this assortment one can identify those who are employed or just the members of a family.

Then another clustering can be done in groups of male and female living forms. This is a reflection of a data which is identifiable and knowledge discovered.

Regression:

Regression is predictive data mining for attribute values of an example, for which one have to predict the output. It is not a class. It is a real value.

Now, we will look at another problem, Regression as we know is a predictive data mining; thatit represents values of an example. If we predict an output, but we need to know that output is not a class. On the other had if it is taken to be a class, we can probably represent them by replacing them in 'tiers'. Sayspam is 0; non-spam is 1 ok, fraud is 0; non-fraud is 1, fraud is minus 1, non fraud can be represented but here it is not a class. So, output is no longer an integer,output is a real value ok.

So, regression is predictive mining where output is real and you still have a training set.

So, it is supervised and learning becomes accessible and is ok.

An example stated for relevance:

- Maybe if you have visited the website 'weather dot com' and you wish to predict the rainfall.
- Let us say rainfall received in centimeters for the next one month.
- Take a real value say rainfall received is 230 centimeters,
- Or for example, you want to predict the value of a stock.
- In the next step that is also a real value whereas, if you predict the stock goes up or goes down, then that is a 2 class classification problem of the stated scenario.
- Another example, many people refer the number of users who will or the fraction of users who will click on certain internet advertisement.
- In reliance to the stated examples, an user of data mining is sure to analyze the situation as required as the data is predicative and is fed in the system as a derivative that could be used prior before the actual observations. The significant usage factor of such a practice is the most valuable asset of 'Data mining'

This type of predictions being done is a mathematical process, where future events and outcomes are predicted. The details given as input are analyzed for which historical data and current data to forecast information is patterned.

Dimensionality Reduction:

This means, you have a number of attributes in your data, imagine, often a large number ofattributes, but the question is if all these attributes are not really required or importantfor the data mining task. So, what we do in dimensionality reduction is done to reduce thenumber of attributes that we actually need.

Purpose:

- Avoid curse of dimensionality
- Reduce amount of time and memory required by data mining algorithms
- Allow data more to be more visualized
- May help to eliminate irrelevant features

Techniques:

- Principle component analysis
- Singular value decomposition
- Supervised and non-linear techniques

If there are more of dimensions and more training set, then it grows exponentially; that means, if thedimensionality goes up by 1, you need square of the number of training sets on it,that ascends up rapidly. This can be taken as the first motivation.

Second motivation of course, if you increase the dimensionality, your computation time willincrease and your storage will increase. It is then we need to reduce dimensionality to improve upon computation time, and storage which is considered as the second motivation.

The third motivation is that you can add more of the dimensions by which more difficult it is tointerpret the data and to visualize the data. In fact, actually we human beings cannotvisualize beyond 2 or 3 dimensions and herein, this dimension reduction will help us toeliminate the noisy features, the features which are not relevant.

Findings summarized:

'Data mining' is a perfect strategic tool which makes easy access to information and knowledge acquisition. Moreover, it is easy to implement, being cost-effective less time consuming. Most high level projects are designed under the framework of 'Data Mining'. In areas where careful planning, preparation and study analysis is to be done in order, 'Data Mining' is of foremost support to obtain significant results.

The methodology of 'Data Mining' is a successful application in industrial engineering. It provides complete information for decision-making unlike the methodology of traditional statistical point of view. 'Data mining' is more applicable as it looks into the role of the organization and the stakeholders in specific that the compatibility of resources is enhanced.

Challenges:

Foremost, industrial engineers find it hard to work as they are not sure of the organizational goals and strategies. The selected tools and techniques that they adapt may not be vividly suitable for their applications. The models that they design may not truly adhere to the organizational plan and entity which hampers their project evaluation.

The significant factor is that the organizational goals and their data requirements should be exactly focused in making use of 'Data Mining'. The need of informational identification, analyzing existing data sources and scheduling plans with the existing data stores should be done by accessing the existing data after appropriate clarification. Hence, the stakeholders' needs and organizations' requirements with strategies must be specifically analyzed before making use of 'Data Mining'. Success of a project through organizational plans need the support of stakeholders' and the necessary details gathered should be accurate and precise.

Further, effective documentation should be done for enhanced usage of 'Data Mining' that bring forth perfection to project development under the pretext of 'Data Mining'.

Conclusion:

The paper has been an evidential deliberation on the need and significance of the concept of 'Data Mining' which is of increasing usage in industrial, commercial and corporate sectors. The futurity of many scope of development lies in this novice impact of 'Data Mining'. Nevertheless, the paper has been taken for a discussion with the issues of software implications, network designing, implementation, process orientation and applications of 'Data Mining'.

Undoubtedly, the future of digital world is getting enriched with advanced developments being made possible through the well planned organized system of data mining. The related algorithms taken to focus are sure to advance the usage and practicality of data mining. Further, this paper has been bringing in relevance of different components of data mining that is sure to impart knowledge to the young researcher in this vibrant field of study. The striking feature of this paper is the reflection of mini thesis writing with a few literature reviews visualized to enhance the theoretical knowledge of the discourse. Well acknowledged is the fact that the importance of 'Data Mining' is sure to be a greater impetus to any user as the power of this concept will be in the hands of the end-users as the multidimensional utility of such a field of study will create interest to any digital savvy personnel. The attractive force of 'Data Mining' is the frequent use of visuals and this concept of visualization to feature data patterns, trends and relationships is an added specialty of 'Data Mining'.

Reference:

- Deck, S. (1999). “Data mining”, Computerworld, Vol. 33 No. 13
- Han, J. and Kamber, M. (2001), Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, Academic Press, California.
- Anonymous. (2002). “Data mining digs deep to improve on quality”, Professional Engineering, Vol. 15 No. 11