IJCSMA
Anjali Saini, International Journal of Computer Science and Mobile Applications,
Vol.6 Issue. 1, January- 2018, pg. 31-37 ISSN: 2321-8363
Impact Factor: 5.515

AN APPROACH TO DATA MINING

Anjali Saini[1]
[1]Ph.D Scholar, Department of Computer Science and Engineering, Singhania University, Rajasthan, India
angel.anjali43@gmail.com

Abstract – Data Mining is the process of extracting patterns from data. Data Mining is seen as increasingly important tool by modern business to transform data into an informational advantage. The Knowledge Discovery in Databases (KDD) field of data mining is concerned with the development and enhancement of methods, algorithm and techniques which can make sense and evaluate of the available and required data. Knowledge Discovery in Database is useful in finding and searching different trends, patterns and anomalies in the databases which is useful to make accurate, effective and meaningful decisions for the future. A primary reason for using data mining is to assist in the analysis of collections of observations of behavior. This research paper gives an overview about data mining and its process which will provide great help to the researchers and scholars in their respective fields.

Keywords- Data Mining and its scope, Mining Components, Text Mining

1. INTRODUCTION

Data mining process includes understanding the business requirements and its needs. While understanding the business requirements both data and business requirements should be understand. Then, using this business requirement it identifies data source and data format in this the data is prepared and modeled for evaluation and then using these data source and data format, it builds data model. This data model is used to build data structure. Then, the mining operation is performed on this data structure. Data mining field comprises of four main disciplines- Statistics(defines tools for measuring significance in the data),Machine learning(provide algorithm to induce knowledge from the data),Artificial intelligence(involve knowledge for encoding and search techniques),Data management and databases(provides an efficient way of accessing and maintaining data). The amount of data especially in the field of web service-based applications, architectures of service-oriented and cloud computing, is increasing every day by day and to extract the valuable information from them different data mining technologies are being used. Data mining is defined as the technology to extract the valuable and useful information from the data and to analyze the patterns and association among these categorization .The categorization is a supervised form of machine learning. Machine learning comprises of supervised, unsupervised, semi- supervised and reinforced learning. In the supervised form of learning, the learning is from trained data available.

Data is the form of descriptive or the analytical information relative to some object, activity, domain or organization. A dataset is able to describe the complete characterization of a particular aspect. As the dataset having the larger information taken from random environment and from multiple sources, there is the requirement to apply some filtration and analysis to derive some decision.

A mining method is here proposed to process the information resources in an integrated form so that the new decisions and data processing will be carried out. The parallel processing and mining tool integration can be applied on massive datasets to find the answers of relative hypothesis. A conventional data processing is here applied to generate the data statistics and associated rule specific information derivation. The distinctive measures are here applied to provide attitude specific data analysis. The fact level extraction can be collaborated from previous and new information so that the predictive results will be optimized. The knowledge transition and its transformation to the valid data form for processing can be done to identify the associated relationship and behavior aspects. A tree specific wisdom is here defined for skill translated data processing.

Today we are drowning in information but starved for knowledge. The problem today is not that there is not enough data and information streaming in. We are in fact inundated with data in most fields. Rather, the problem is that there are not enough trained human analysts available who are skilled at translating all of this data into knowledge, and thence up the taxonomy tree into wisdom. The basic architecture of data mining processing behavior and characterization is shown in figure 1.

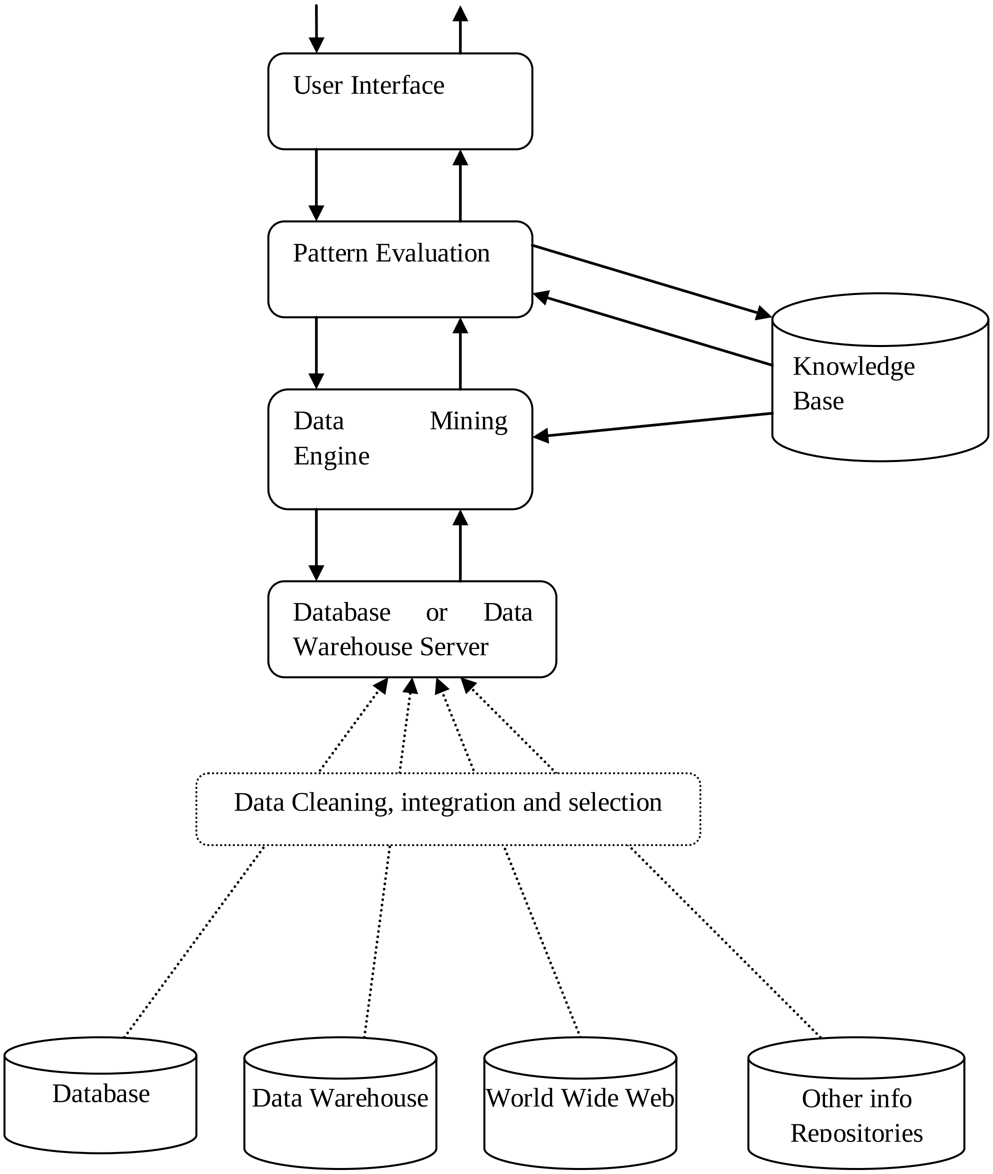

Figure1: Architecture Data Mining

The figure showing the mining model requires some interface in the form of tool where all the data processing will carried over the dataset. The first stage of this mining algorithm is to accept the dataset. The dataset extracted here is from the knowledge base in raw form. On which the filtration and transformation is applied to transform the data to the normalized form. This normalized data form is considered as knowledge base. While applying the application specific process, the first requirement is to generate the patterns over the datasets. The pattern generation and evaluation is here processed under mining engine for generating the decision rules. The data selection, integration and predictive analysis is here provided. Finally the user query is applied on it to take the decision specific work. The basic components of data mining and the process stages are here shown here under

Data Mining Scope

Data Mining is the scientific and structured process used to take the analytical and intelligent decisions over the dataset. The mining can be applied on any dataset at different stages. It is able to analyze the significance of individual attribute as well as to verify the significance of dataset. Various methods and aspects are integrated in the data mining to improve the productivity and significance of real time application. Data mining is having scope is different application areas including the health care system, business application, government system etc. The scope of data mining is increasing because of the following reasons:

- Can handle larger data space to generate more accurate decision
- The performance of data mining methods is better than traditional approaches or methods.
- Data mining offers number of predictive algorithms to generate the intelligent decisions
- It supports the vast number of applications and provide the constraints under the application or domain concern

The mining methods are having the increasing scope in commercial application because of the low cost and high performance implications. The survey based projects are analyzed by the mining algorithms to generate more predictive outcome which can be applied later to improve the productivity of the real systems. The support to the real time and larger dataset space also increases the requirements of data mining methods. There are number of industries, where these numbers can be much larger. The basic need of improved computational engines shows the importance of data mining to define in a cost-effective manner along with parallel computer system. In last few years, the scope and requirement of data mining applications and methods is growing very fast. The statistical decisions obtained from the mining methods helps to take more accurate decisions. The observation based analytical decision are more considerable than the experience oriented decision because of which the requirement of mining is increased in commercial applications.

The scalability is another effective feature of mining method. It means, the new innovations or evolutions can be done based on the history mapping or the results observations of existing methods. The study is defined under the data access in a critical aspect for building the information drill so that the navigational applications can be implemented as well as criticality of the system will be resolved. [8]

2. MINING COMPONENTS

The mining can exist in different form including the data mining, textual mining, pattern mining etc. The mining also having the number of integrated algorithms and process stages. To perform any kind of mining using any algorithm some basic components exist in the system. These elements of components available in any mining application are listed as:

2.1 Knowledge Base

When the input data is processed and transformed the raw data in processing data form, it is called the knowledge base. The knowledge base is the actual repository on which the mining process is applied. During the preprocessing phase, the knowledge base is generated from the raw dataset. The dataset can be distributed for segmented respective to activity, object or the instance. The dataset is described with number of associated attributes and integrated instance specific observation. The user belief, constraints and metadata is defined to define the data in relational form.

2.2 Mining Processes

The data mining is not a single process but instead it is a group of associated processes. These are the result driven processes defined to provide the data characterization, its association and correlation with application, domain and other process stages. Some of the common mining processes include clustering, predictive algorithms, classification, rectification etc.

2.3 Pattern Evaluation

A component driven evaluation and measure is provided with module interactive mining model is provided. The pattern set generation and the filtration is the major challenge in this work area. The pattern discovery can be applied to control the process with associated filter specification. The feature evaluation and integrated mining model can be combined at different mining levels. The pattern correlation and the interest level observation is here defined to apply process mining. The stage is able to extract the effective data patterns which can be can define for decision making.

2.4 Visualization

Once the algorithmic implementation is defined and pattern rules are generated, the results are obtained from the system. These results are later on processed under some visualization algorithm to present them in effective way. The intermediate mining results, structures and component processing can be defined in different way to present the results effectively. The attribute level, test set level and the feature level visualization can be applied.

3. TEXT MINING

Text information processing is wide field as the complete document information; web data is available in the form of text data. This kind of data is having the various complexities including the unstructured data format. There is the requirement of some effective mining method to process this data and to acquire the data patterns. The pattern specific search can be applied to generate the relevant information. Mining of this textual information is required to generate the effective data search from web and the offline available documents. These documents are defined specific to some organization, activity or the object. These information data classes include government data, industry and the institutional data. Each of the information is available in the form of research papers, articles, emails, web pages and the digital libraries. The text data bases are generally in unstructured or semi structured form. The text information extraction and the interrelated information analysis is here provided. This kind of information can be applied in various text processing application including document clustering, document classification, sentiment analysis, recommender system etc. These online and offline textual documents based database system can be processed to generate the extracted textual information in semantic information form.

Text mining the process driven information applied to filter the unstructured data form and convert it to an organized and structural form. The level specific query processing and the knowledge access method is required to implement. The intelligent data processing and feature generation methods can be applied to generate this required feature data. This kind of information processing suffers various associated challenges. Some of the challenges relative to the textual processing is listed here

- Unstructured or the semi-structured data form
- Language specification
- Grammar problem
- Multiple meaning and verb forms
- Word sense observation
- Word relation mapping
- Contextual aspect specification
- Word categorization and contextual aspect identification
- Different meaning of same words
- Handling of abbreviations or the short forms of the words
- Symbolic data representation

To process this kind of complex data form, there is the requirement to apply a wider information processing and query processing methods so that the effective information extraction will be done. The data information extraction, intelligent methods can be applied using machine learning approaches. The search specific processing can be applied to generate the dataset information so that more sensitive data processing will be applied and the information can be converted to valid numerical form. This kind of transformed information is finally applied under some process methods. The textual information processing is shown here in figure 2.

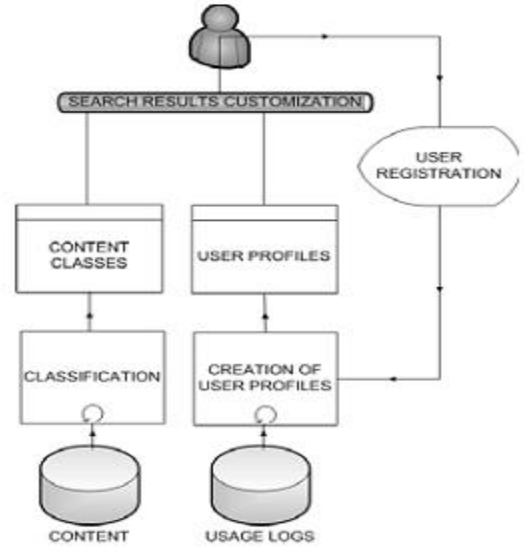

Figure 2 : Basic Textual Search Model [1]

Here figure 2 is showing the basic process model of information search apply on any content database. This database can be online or offline. To apply the search method, the user query is implied. This search is user specific for which the user registration is required. Now as the authenticated user apply the search, the user profile information is extracted. The profile map is applied on usage db to generate the usage pattern and the predictive user query and the content information results. The user query is also applied on the content database. The classification process is applied on this textual database. The content class analysis is applied to generate the topic specific classification. The usage information control is also applied to generate the effective search map.

Data Processing

Mining is a process to discover the hidden information or pattern from text by applying some descriptive and quantitative analysis. Text is the most generic input submitted by the user to present his opinion, thoughts or suggestions. A story, news, social media post, emails or the personal documents are generally in textual form. Text mining defines the automated process to analyze the text and obtain the meaningful information and decisions. Various features and aspect relative to the domain and topic can be analyzed to generate such meaningful patterns. Once the meaningful patterns or features are generated, the decisions can be concluded for the input text. The predictive algorithm can be applied over the textual input to generate the aggregative and concluded decisions. Various online and offline application uses the text mining to summarize the documents, categorize the documents and to identify the hidden sentiments from these documents[33][34][35]. In this section, different forms of text mining operations are discussed.

1) Summarization

Summarization is the text mining process applied on larger documents or paragraph to acquire the compact and meaningful information. Various documents such as news, research articles, web page text etc. have larger textual contents. Processing the complete text requires more efforts and space requirements. Summarization is a mining process that can analyze the text and transform it to more meaningful and smaller text. Summarization is also effective to assign the topic name, annotation to some article or document. In many applications, summarization is included in preprocessing stage to remove the irrelevant information.

2) Classification

Another aspect of textual information is to obtain the data class. Let, a large document set is available without specification of document type and only data classes are available. In such case, the classification algorithm is helpful to identify the most appropriate class for a textual element or string. The classification is the generalized process defined under the specification of training set and testing set. The training set is defined as the dataset with the specification of relative class. To perform the classification, the training set is process to generate the rules which are processed on testing set or new data to recognize the data class. The classification accuracy can be obtained in terms of identification of appropriate class for textual data.

3) Clustering

Clustering is a type of segmentation process which partitions the available data in smaller groups where each group contains the mutually similar data. Clustering is also considered as the classification process in which data elements are divided in certain groups. There are number of available clustering algorithms. These algorithms are defined around the specification of cancroids and the center to element distance analysis is performed to obtain the cluster members. Cluster is the unsupervised classification approach that is not defined with specification of class. The division of the data groups is done based on distance measures. Some other parameters can also be considered for data clustering .

4. CONCLUSION AND FUTURE SCOPE

Data Mining itself is defined the processing algorithm associated with different operational and application specific concepts. It is used to derive the information from the dataset and to perform the prediction respective to the data processing. These kinds of systems are able to provide the quality information processing to generate the dataset with technological specification and the data quality based processing. These kinds of systems are opportunity specific and provide the capability under the specification of business application. Some of the major scopes of data mining are explained here under

- Major scope of the mining is in predictive applications. In such application, the data processing is done to identify the data patterns that can be later used to predict the future aspects. These aspects can be used to derive more innovative future solution and the operational specification. Such as the strategy making based on the analysis on market trends can improve the company growth.

- Data mining perform the analysis on statistical information of firm to capture the targets effectively so that the maximum return from market will be achieved. Other predictive problems include the detection of bankruptcy and or the frauds.

REFERENCES

[1] S. Gaion, S. Mininel, F. Vatta and W. Ukovich, "*Design of a domain model for clinical engineering within the HL7 Reference Information Model*," 2010 IEEE Workshop on Health Care Management (WHCM), Venice, 2010, pp. 1-6
[2] Ghada Almodaifer," *Discovering Medical Association Rules from Medical Datasets"*, 978-1- 61284-704-7/11, 2011 IEEE
[3] Kapil Bakshi, "*Considerations for Big Data: Architecture and Approach*", IEEE, 2012
[4] Patrick C. H. Ma," *An Iterative Data Mining Approach for Mining Overlapping Coexpression Patterns in Noisy Gene Expression Data", IEEE* Transactions On Nanobioscience 1536-1241 , 2009.
[5] Bolin Ding," *Efficient Mining of Closed Repetitive Gapped Subsequences from a Sequence Database*", IEEE International Conference on Data Engineering 1084-4627/09 , 2009 IEEE
[6] Knuth D.,Morris.J Pratt, *V Fast pattern matching in strings*, SIAM Journal on computing, vol 6(1),323-350, 2005.
[7] Jiuyong Li," *Mining Causal Association Rules*", 978-0-7695-5109-8/13,2013 IEEE
[8] Antonia Azzini," *Consistent Process Mining Over Big Data Triple Stores*", 2013 IEEE International Congress on Big Data 978-0-7695-5006-0/13 , 2013 IEEE
[9] Surajit Chaudhuri," *An Overview of Data Warehousing and OLAP Technology".*
[10] Arvind Arasu, "*Towards a Domain Independent Platform for Data Cleaning*".
[11] Ranjit Singh, "A *Descriptive Classification of Causes of Data Quality Problems in Data Warehousing*", IJCSI International Journal of Computer Science Issues ISSN (Online): 1694-0784 ISSN (Print): 1694-0814