UGC JOURNAL NUMBER 44557

International Journal of Allied Practice, Research and Review

Website: www.ijaprr.com (ISSN 2350-1294)

Clustering Analysis with Purity Calculation of Text and SQL Data using K-means Clustering Algorithm

Himanshi Jain and Reshu Grover
Computer Science & Engineering
LIET, Alwar, Rajasthan, India

Abstract - Today wide adoption of Internet has increased heterogeneous information, the data volume stored in database increases rapidly and in the large amounts of data much important information is hidden. Useful information can be extracted from the database they will create a lot of profit for the organization. The question they are asking is how to extract these values. Data mining is the key answer for it. Most objects and data in the real world are interconnected to form huge and sophisticated network. However, mostly people consider a database just as a warehouse that supports storing of data and retrieval rather than one or multiple heterogeneous information networks that contain rich, inter-related information.

The different types of data have to be managed and organized properly so that they can be accessed efficiently. Also, the web has grown exponentially in size and contains a large amount of publicly accessible web document distributed all over the world on thousands of servers. As document collection grows larger, they become more expensive to manage. The increasing rate of data sources on the web has brought great need of data mining techniques for extracting data from multiple heterogeneous sources. In this paper, we worked on heterogeneous data, data mining, we analyzed Text and SQL Data usingK-mean algorithm.

Keywords - *Unstructured Data, Heterogeneous Data, Data Mining, K-mean, unstructured data.*

I. Introduction

As heterogeneous has exploding growth data in every field of life. The size of database increases day by day. To extract the useful information from huge database is a challenging task. It has been noted that billions of database available for business record, university record, government record, social websites etc. Thousands of records are stored in one database. The increasing rate of heterogeneous data and database are required intelligent technique and tools, so that we can extract

useful knowledge. Clustering is the process of grouping data items or element in a way that makes the elements in a given group similar to each other in some aspect. Clustering has many applications such as data mining, statistical data analysis and bioinformatics [1]. It is also used for classifying large amount of data, which in turn is useful when analyzing data generated from search engine queries, papers and texts, images etc.

1.1 Problem Identification

In context with heterogeneous sources of data, we consider the challenging job of developing tentative analytical techniques to investigate different clustering techniques to partition heterogeneous datasets consist of heterogeneous domains such as categorical, numerical and binary or combination of all these data. The K-means method is a popular approach to clustering, the method is simple and allow for relatively efficient implementation that still produce good results. The exponential growth in the generation and collection of heterogeneous data gives us new term of data analysis and data extraction.

This paper proposed a framework for analyzing and mining the heterogeneous data from a multiple heterogeneous data sources. Clustering algorithms recognize only homogeneous attributes value. However, data in the every field occurs in heterogeneous forms, which if we convert data heterogeneous to homogeneous form can loss of information. In this research, we applied the K-mean clustering algorithm on real life heterogeneous datasets and analyses the result.

The Heterogeneous data are bulky, structured, semi structured or unstructured in nature, the performance level of storage, retrieval and analysis of heterogeneous data is a critical issue of research as indicated by literature review and its findings. This undertaken research has considered these issues to be taken care to evolve optimal solution. To achieve optimum levels of performance, the research is targeted to develop a suitable model and subsequent framework to achieve the targets of research by using data mining algorithm.

The organization of this paper is as follows; Section 2 Describes the related work, In Section 3 we have discussed our proposed approach. Section 4 shows the result analysis using K-means clustering approach, the final Section 5 concludes our research.

II. Related work

Data mining is distinguished by the fact that its main aim is to discovery of information, without a previously formulated assumption

AzraShaminet. al [2], proposed a framework for bio data analysis data mining technique on bio data as well as their proprietary data, Bio database is often distributed in nature. In this system take input from the user, preprocess the query and load it into local bio database. System will search the knowledge from Database and send it back to the user, if the data related to user query exists.

NiranjanLal et.al [3] proposed a framework for heterogeneous data being generated from various sources like digital audio, video, images data from different domains including healthcare, retail etc, and day to day life utilities. For instance, 30 billion web pages are accessed or the World

Wide Web (WWW). With billions of pages of that exist today, search engines plays an important role in the current Internet of Thing (IOT). So with billions of web pages accessible on the web today, a query entered by the user on the search engine will returns thousands of web pages in a second, and thus it becomes extremely important to rank these results in such a way that the most "related" or "important" or "authorized" pages are displayed first. This job of displaying the results is performed by ranking algorithms, and various search engines use different algorithms for ranking the results.

Rumi Ghosh et. al [4], proposed a framework to aggregate multiple heterogeneous documents to extract data from them. Therefore, in this paper, they propose a novel topic modeling framework, Probabilistic Source LDA which is planned to handle heterogeneous sources. Probabilistic Source LDA can compute latent topics for each source maintain topic-topic correspondence between sources and yet retain the distinct identity of each individual source. Therefore, it helps to mine and organize correlated information from many different sources.

PrakashR.Andhaleet. al[5], In this paper author represent the characteristics of HACE theorem which provides the description of heterogeneous data and proposes a model for processing of heterogeneous data from the view of data mining. This information extraction model involves the information extraction, data analysis and provides the security and privacy mechanism to the data.

Amir Ahmad et.al [6],performsK-mean clustering algorithm for both numerical and categorical data. In this paper, they present a modified approach of cluster center to overcome the numeric data only restriction of K-mean algorithm and provide a better categorization of clusters. The performance of the K-mean algorithm has been studied on real world data sets.

Dr. Goutam Chakra borty et.al [7], in this paper they represent a way at how to organize and analysis of textual data for getting useful data from huge collection of documents which improve the performance and business operations.

S. Geetha et.al [8], in this paper they represent a view for extraction of knowledge from huge amount of unstructured data by converting the data into structured format in the form of data relations. Set of rules are used for conversion of unstructured data into structured data which combine the data into relations and also creating

Yuanming Huang et.al.[9],in this paper they represent the theory and methods for audio, video information process for extraction of information in emergency system mechanism. The result of this research can form an intelligent information service platform, splitting the volume of information intelligent services in a various technical bottleneck.

Dr.MuhammadShahbaz et.al[10], in this paper they represent a solution of there work by development of a System which will provide number of features to process and determine text files for opinion mining at sentence level.

Ming-Syan Chen et.al [11], in this paper author describes about the basic definition of Data mining. Data Mining is defined as a process of extraction of useful knowledge from huge data warehouse. It has become valuable from the past decade in E-commerce to increase more information, to have a better perceptive of running a big business, and to discover new traditions to boost the big business.

Padhyet. al [12] and **Nicholas J Belkin et. al** [13], Describes about the survey of various data mining applications and how these are used in various fields in real life. They have focused on variety of techniques, approaches and different areas of the research which are helpful and marked as the important field of data mining Technologies. As we are aware that many MNC's and large organizations are operated in different places of the different countries. Each place of operation may generate large volumes of data. Corporate decision makers require access from all such sources and take strategic decisions .The data warehouse is used in the significant business value by improving the effectiveness of managerial decision-making. In an uncertain and highly competitive business environment, the value of strategic information systems such as these are easily recognized however in today's business environment, efficiency or speed is not the only key for competitiveness. These types of huge amount of data are available in the form of Tera to Peta bytes which has drastically changed in the areas of science and engineering. To analyze, manage and make a decision of such type of huge amount of data we need techniques called the data mining which will transforming in many fields.

1) Problem analysis: In this step we will determined whether the problem is suitable for data mining and what data and technologies are available. Also at this stage it will be important to determine what will be done with the results of the data mining to put the problem in context.

2) Data preparation: Should be part of the methodology and data processing. Processing involves pre-processing or cleansing of the data, data integration, variable transformation and splitting or sampling from the database.

3) Data exploration: This allows the analyst or data miner to discover the unexpected in the data as well as to confirm the expected.

4) Pattern generation: Should follow which involves applying the algorithms and validating and interpreting the patterns that result.

5) Model validation: This is required in order to confirm the usability of the developed model. Validation can be conducted using a validation data set and assesses the quality of the model fit to the data as well as protecting the model from over- or under-fitting.

6) Interpretation and decision making: Conclude the methodology and attempt to transform the patterns discovered during data mining into knowledge. There are a number of initiatives for the development of a formal/documented data mining process in the world. It is reassuring to the data mining community that the processes emerging from all of these initiatives reveal a large degree of similarity.

E. Rahm, P.A. Bernsteinet. al [14] and **Piatetsky-Shapiro et. al** [15], Describes about the process of data mining with business perspective. There is increased interest in a process or methodology for data mining. This process is another important aspect that needs to be examined as it layouts clear steps that can be followed in the data mining process. It is argued that such a formalized process will widen the exploitation of data mining as an enabling technology for solving business problems. It will allow people with varying expertise in data mining and from different business sectors to carry out successful data mining projects with a high degree of consistency

believes that to be effective in data mining, successful data analysts generally following a four step process.

Romeroet. al [16], Describes about the process of knowledge discovery in database. Data mining and KDD are the synonym of each other. There is widespread agreement on the main steps or stages involved in such a process.

1) Goal definition: This involves defining the goal or objective for the data mining project. This should be a business goal or objective which normally relates to a business event such as arrears in mortgage repayment, customer attrition (churn), energy consumption in a process, etc. This stage also involves the design of how the discovered patterns would be utilized as part of the overall business solution.

2) Data selection: This is the process of identifying the data needed for the data mining project and the sources of this data.

3) Data preparation: This involves cleansing the data, joining/merging data sources and the derivation of new columns (fields) in the data through aggregation, calculations or text manipulation of existing data fields. The end result is normally a flat table ready for the application of the data mining itself (i.e. the discovery algorithms to generate patterns). Such a table is normally split into two data sets; one set for pattern discovery and one set for pattern verification.

4) Data exploration: This involves the exploration of the prepared data to get a better feel prior to pattern discovery and also to validate the results of the data preparation. Typically, this involves examining the statistics (minimum, maximum, average, etc.) and the frequency distribution of individual data fields. It also involves field versus field graphs to understand the dependency between fields.

5) Pattern Discovery: This is the stage of applying the pattern discovery algorithm to generate patterns. The process of pattern discovery is most effective when applied as an exploration process assisted by the discovery algorithm. This allows business users to interact with and to impart their business knowledge to the discovery process.

S. R. Dhamankaret. al[17], in this author describes about ontology and also state that more complex will be the application, the bigger the space comes into existence between application and users. The data mining applications illustrate the concepts and characters, and later than planned a selection model to match these business requirements to data mining categories to connect complex data mining concepts with business problems and assists users to choose the best data mining solution. Knowledge discovery in databases, interesting Knowledge, regularities, or high level information can be extracted from the relevant sets of data in databases and be investigated.

III. Proposed Methodology

In this research, as shown in Figure 1 Firstly, we convert the heterogeneous data into same format for analyzing data, then we apply k- mean clustering algorithm using Euclidean distance on the heterogeneous data. As we can also apply Euclidean distance, for heterogeneous data. Before we studied Euclidean distance is possible only for numerical values. It is also possible for heterogeneous data. It only depends how we store data and convert in a format which is suitable for clustering of heterogeneous data. According to survey almost 91% of the data is heterogeneous. So, we require a technique or tool which can handle heterogeneity of data.
The different stages involved in our algorithm are shown in Figure 1.

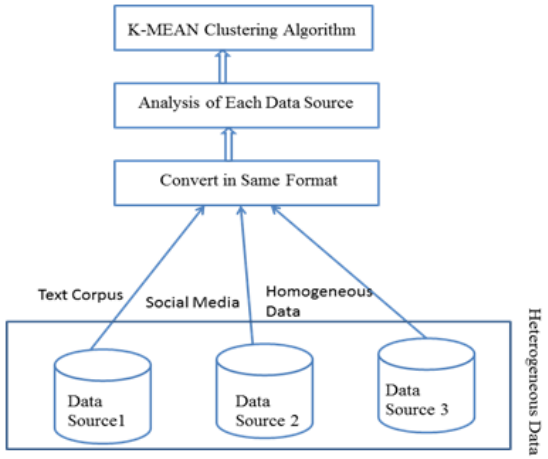

Figure1. A view of proposed Framework

To summarize the contribution of the research as follows:

1) Retrieve the result individually from all the data sources into one format, 2) Analysis of all the heterogeneous sources including text corpus, social media, image and homogeneous data., 3) Applying the K-mean clustering algorithm individually on each heterogeneous data source for extracting the hidden knowledge, 4) Applying the clustering algorithm K-mean on heterogeneous large dataset.

3.1 K-mean on Heterogeneous Data

Clustering is the process of aggregation of a set of items in a way such that items in the same group are more identical to each other than in other groups. In other words, main objective of clustering is to divide the items into uniform and different group for an output. Mostly, clustering methods can only handle either numerical data or non-numerical data [9]. We also manipulate or change the operations on our data according to our requirement. Clustering algorithms depends on multiple purposes as the type of data available for clustering. There exists several numbers of techniques for clustering such as hierarchical, center based partitioning, density and graph based clustering.

Similarity Based Cluster: This type of cluster contains data items which are similar or related to each other in some way. Generally, partitioning clustering methods are K-mediods and K-mean. Here we are using only K-mean for clustering. The basic requirement of K-mean clustering is that taking the no. of cluster as 'k' from the user initially. Representation of cluster as mean value which is based on similarity of the data items in a cluster[27]. The mean or center point of cluster is known as 'centroid'. Centroid is a value which can be find out through the mean of related points. K-mean algorithms have $\sqrt{a^2 + b^2}$ several advantages such as simplicity, high speed to access database on very large scale.

3.2 Purity Calculation for Data Analysis

The purity measure is a measure for evaluating the quality of cluster on the basis of labeled samples. A cluster will be considered pure if the value of purity is one; it means the entire labeled are correctly classified. Similarly, a cluster will be considered impure if its value is close to zero or close to zero. It means all the labeled value occurs in different classes. Purity is defined as below:

Purity=$1/N \sum_{i=1}^{k} max_j$ $|C_i \cap t_j|$

N = No. of data points

K = no. of cluster

C_i= Cluster in C

$t_j$is specify the max. count for each cluster C_i

IV. Analysis of Heterogeneous Data

In This section we have analyzed the Text and SQL data by K-means clustering algorithm on using R Programming on RStudio.For experimental analysis, we have considered two real dataset. One dataset is purely categorical and also in unstructured format, second dataset is purely numeric. We have done analysis by taking real life datasets collected from UCI [18].

4.1Analysis of Unstructured Data

The words cloud and Histogram of words is shown in Figure 2 and Figure 3 respectively.

Figure 2.Words Cloud of Text

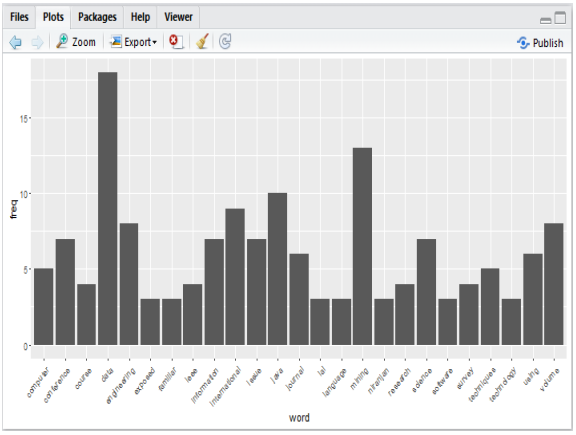

Figure 3. A Histogram of Words

4.2 SQL Data Analysis

SQL data analysis is analysis of homogeneous s data. SQL data is in the form of specific format. It is store in the form of tables. We analyze the data in histogram format as shown in Figure 4.

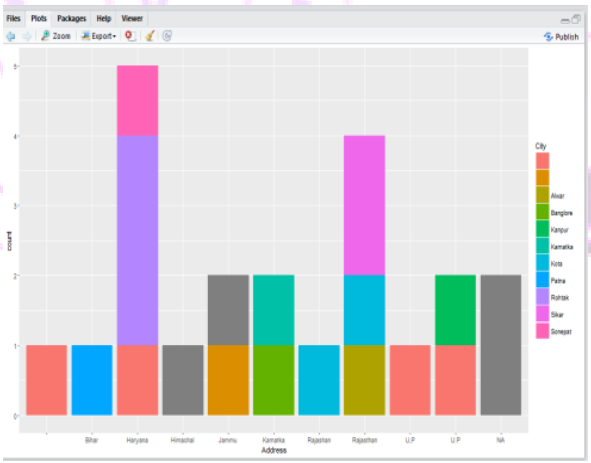

Figure 4 A view of SQL data Analysis

4.3K-mean on Text Data

Large document corpus may afford a lot of useful information to people. Corpus consist of no. of documents like pdf, ppt, text files. If the document is in the form of ppt or pdf or any other form first we have to convert it into text format. Then we perform pre-processing step to remove noise from the data. After that we convert the data in csv format. After converting the whole data in csv format we analyze the data by word cloud or histogram. After analyzing the data we apply the clustering algorithm on heterogeneous data as shown in Figure 5.

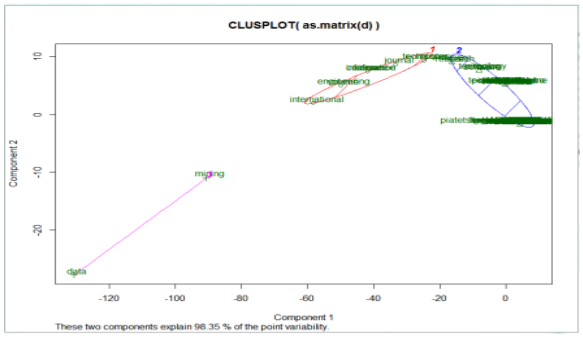

Figure 5 A view of Text clustering using K-mean

4.4 K-mean clustering on the William Shakespeare book

Here we perform the clustering of book using K-mean. The Shakespeare book has no pre-defined format. We apply the clustering for large scale dataset and analyse the cluster. We get the solution of our goal we can apply the K-mean clustering algorithm using distance metric for heterogeneous data and extract the useful data from huge dataset as shown in Figure 6.

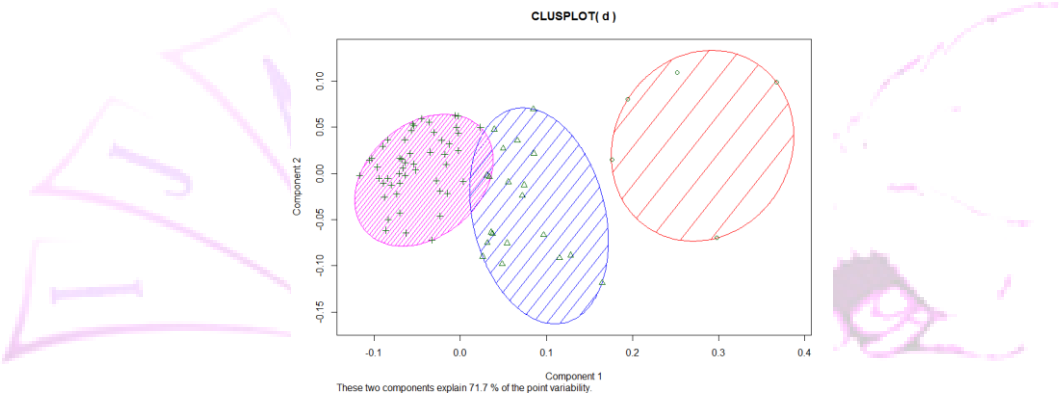

Figure 6 A view of William Shakespeare book clustering using K-mean

4.5 Cluster Evaluation using Zoo dataset

It is purely mixed data set consist of 101 data points with 18 attribute. One attribute is categorical 15 Boolean attribute and 2 numerical attribute. This dataset is collected from UCI Machine learning dataset available online. The cluster evaluation of K=3 for wine dataset are represented below in Figure 8. The quality of cluster is defined using purity and rand index. Purity value of 1; which indicates the best cluster quality. Purity value 1 shows there is no misclassification of data points. Using the K-mean algorithm the purity value is 0.94. The purity value is close to 1 which shows K-mean is good model for heterogeneous dataset.

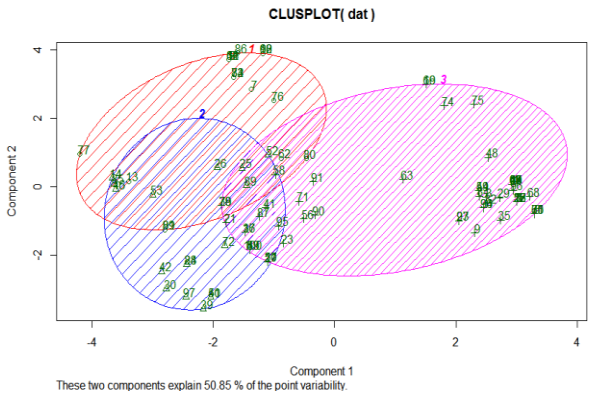

Figure 7 A view of Zoo dataset clustering using K-mean

4.6 Purity Comparison

Here, we perform the comparison between the purity values for different value of K using K-mean clustering algorithm. We are using two dataset explained above Wine dataset and Zoo dataset. We have noticed that purity value become constant after reaching at some point. Description of purity for different clusters is shown in Table 1.

Table 1 A Description of Purity for different cluster

Cluster no.(K)	Purity for Wine Dataset	Purity for Zoo dataset
3	0.96	0.94
4	0.97	0.96
5	0.97	0.96

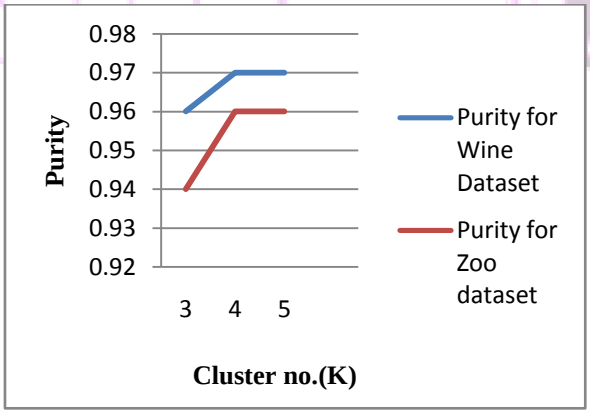

Figure 8 Purity comparisons for numerical and mixed dataset

Here we analyze heterogeneous data purity increased rapidly as compared to other dataset and also increases at some extent and then become constant as shown in Figure 8. So, K-mean algorithm is best for heterogeneous data.

V. Conclusion

In this paper, K-mean clustering is used to act on heterogeneous dataset using Euclidean distance. K-mean clustering algorithm for heterogeneous dataset has relevance in almost every field commercial, education and also in medical sector. We analyze K-mean clustering using distance measure is suitable for heterogeneous type of data.

Our main contribution in this dissertation, we have designed the framework which can work on heterogeneous data using existing clustering technique which is efficient for analysis of heterogeneous data.

In future, As we know data mining of heterogeneous data is new area for latest research. So, we can apply k-mean algorithm for mining of streams of data, big data analysis with appropriate data mining techniques.

VI. References

[1] Anil Kumar Jain, "Data Clustering: 50 Years Beyond K-means" , Pattern Recognition Letters, Vol. 31, Issue no. 8, pp. 651-666, 2010.

[2] AzraShamim, VimalaBalakrishnan, MadihaKazmi, and ZunairaSattar, "Intelligent Data Mining in Autonomous Heterogeneous Distributed and Dynamic Data Sources", 2nd International Conference on Innovations in Engineering and Technology (ICCET'2014) Sept. 19-20, 2014.

[3] NiranjanLal, SamimulQamar, "Comparison of Ranking Algorithm with Dataspace", International Conference On Advances in Computer Engineering and Application(ICACEA),pp. 565-572, March 2015.

[4] Rumi Ghosh, SitaramAsur, "Mining Information from Heterogeneous Sources: A Topic Modeling Approach" ACM 978-1-4503-2321,2013.

[5] Prakash R.Andhale1 , S.M.Rokade2, "A Decisive Mining for Heterogeneous Data", International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 12,pp. 43-437, December 2015.

[6] Amir Ahmad, Lipika De, "A K-mean clustering algorithm for mixed numeric and categorical data" Data & Knowledge Engineering Elsevier,pp. 503-527,2007.

[7] Dr. Goutam Chakra Borty,Murali Krishna Pagolu, Analysis of Unstrucured Data: Application of Text Analytics and Sentiment Mining,2014.

[8] S.Geetha,Dr. G.S Anandha Mala, "Effectual Extraction of Data Relations from Unstructured Data",3rd International Conference on Sustainable Energy and Intelligent Systems (SEISCON 2012) ,pp 1-4, 27-29 December 2012.

[9] YuanMingHuang,YujieZheng, "Research on Theory and Methods on Massive Audio, Video Unstructured Information Intelligent Process in Emergency System",2nd IEEE International Conference on Information and Engineering (ICIFE),pp. 856-859,17-19 Sept. 2010.

[10] Dr. Muhammad Shahbaz, Dr. Aziz Guergachi, RanaTanzeelurRehman, "Sentiment Miner: A Prototype for Sentiment Analysis of Unstructured Data and Text", 27th IEEE Canadian Conference Electrical and Computer Engineering (CCECE), pp 1-7,4-7 May 2013.

[11] Ming-Syan Chen, Jiawei Han, and Philip S.Yu, "Data Mining – An Overview from Database Perspective",Knowledge and Data Engineering, IEEE Transactions on ,Volume 8 , No.6 , pp 866-883,Dec 1996.

[12] Neelamadhab Padhy, Dr. Pragnyaban Mishra , and Rasmita Panigrahi, "The survey of Data Mining Applications and Future Scope", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3, June 2012.

[13] Nicholas J Belkin and W Bruce Croft,"Retrieval techniques", Annual Review of Information Science and Technology,Volume 22,pp 109-45,Information Today,1987.

[14] E. Rahm, P.A. Bernstein. "A Survey of Approaches to Automatic Schema Matching". VLDB Journal, Volume 10, No. 4, pp 334-350,2001.

[15] Piatetsky-Shapiro, Gregory, "The Data-Mining Industry Coming of Age" ,IEEE Intelligent Systems, Volume 6, pp 32-34,2000.

[16] Romero, Cristobal, Sebastián Ventura, and Paul De Bra, "Knowledge discovery with genetic programming for providing feedback to courseware authors." Volume 14,Issue 5, pp 425-464, 2004.

[17] S. R. Dhamankar, Y. Lee, A. Doan, A. Halevy, and P. Domingos, "iMAP: Discovering Complex Semantic Matches between Database Schemas", International Conference on Management of Data,ACM SIGMOD, pp 383-394,2004.

[18] https://archive.ics.uci.edu/ml/datasets/