ISSN: 2348 - 2117

International Journal of Engineering Technology and Computer Research (IJETCR)
Available Online at www.ijetcr.org
Volume 2; Issue 6; Page No. 143-147

A Review: Comparison of Text Documents for feature selection on the use of side information with data miningtechniques

Ms. Sonal S. Deshmukh[1], Prof. R.N.Phursule[2]

[1]PG Research Scholar, [2]Associate Professor, Department of Computer Engineering

JSPM's, Imperial College of Engineering and Research,Pune, Maharashtra, India

sonal.deshmukh4@gmail.com

ABSTRACT

Nowadays, the usage of World Wide Web is growing enormously. The corpuses which are available on this world wide web is not structured as it is scattered data , due to thebroad availability of huge amounts of corpuses, there is a need to convert such data into useful information and knowledge. Therefore, there is need of data mining algorithm.However, the relative importance of this side-information or additional information may be difficult to estimate, especiallywhen part of the information is noisy. In such cases, it may be risky to assimilate side-information into the mining process, as it can either enhance the quality of the representation for mining the process, or can add noise to the process. Hence, we need a principled way to accomplish the mining process, so as to maximize the advantages by using this additional information. The study of this paper is immersed in effective clustering, classifying and mining approach with the help of side information or metadata.

Index Terms: Web mining, Data Mining, Parallel comparison, Clustering, Classification, Frequent Pattern, Visualization.

I. INTRODUCTION:

In recent years, mining has gradually become a new research topic. In many mining applications, side-information is available along with the documents like hyperlinks, non-textual attributes, and document provenance information. Due to this availability of side information, data mining has attracted a great deal of attention in the information industry and in society as a whole in recent year.

Data mining is the science of extracting data from large amount of datasets. Data or text mining i.e. "content mining" is a process used to discover knowledge out of data and presenting it in a form that is easily understand to humans. So, it is essential to do data mining that arranges document items into several groups so that similar items fall into the same group.

Mining finds a tiny set of valuable nuggets from a great deal of raw material. So working with both "data" and "mining" became a popular option. Text mining is the nontrivial extraction or the science of extracting useful information from large data sets.

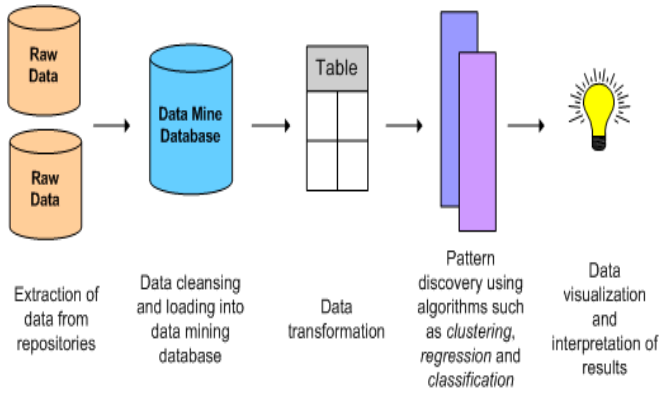

Figure 1: Data mining process

Alternatively, others view data mining as simply an essential step in the process of knowledge discovery. Knowledge discovery as the following sequence of steps [5], [7]:

1. **Data Cleaning:** To remove noise and inconsistent data.
2. **Data Integration:** Where multiple data sources may be combined
3. **Data Selection:** Where data relevant to the analysis task are retrieved from database.
4. **Data transformation:** Where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.

5. Data Mining: As essential process where intelligent methods are applied in order to extract data patterns.

6. Pattern Evaluation: To identify the truly interesting patterns representing knowledge based on some interesting measures.

7. Knowledge Presentation: Where visualization and knowledge representing techniques are used to present the mined knowledge to the user**.**

This survey consists of the following

- Text Mining
- Clustering
- Visualization

TEXT MINING

Text mining is the pertinent extraction, is the science for extracting useful information from large dataset.

Text mining = information retrieval + statistics + artificial intelligence.

The purpose of Text Mining is to process scattered (textual) information, elicit meaningful numeric indices from the text, thus, make the information contained in the text accessible to the various data miningalgorithms. Information can be extracted to obtain summaries for the words contained in the documents or to enumerate summaries for the documents based on the words contained in them.

Text analysis involves information retrieval and also contains lexical analysis to study word frequency distributions, pattern recognition, information extraction, data mining techniques including link and association analysis. It also includes visualization, and predictive analytics. The overall goal is, essentially, to turn text into data for analysis.

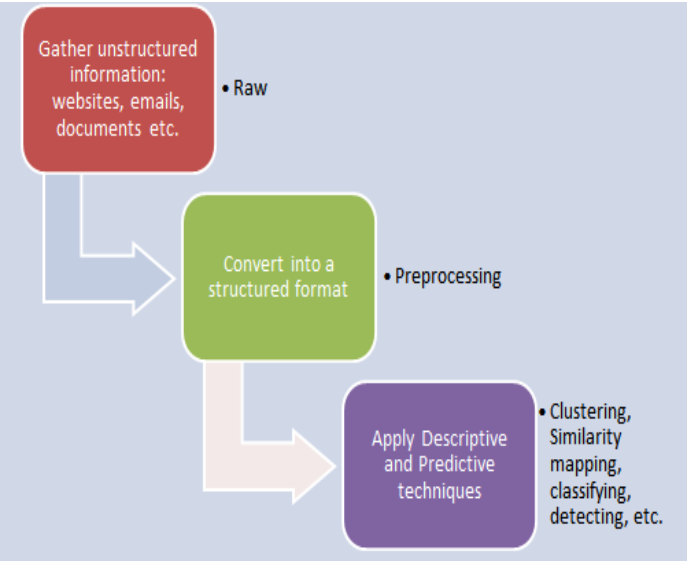

Figure 2: A high level- General Process in Text Mining

CLUSTERING

Clustering can be considered the most important *unsupervised learning* problem; so, as every other problem of this kind, it deals with finding a *structure* in a collection of unlabeled data.

A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way". A *cluster* is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters.

Clustering is the methods have been proposed, first to analyze these data by classifying them and then to organize documents into groups (or clusters), where each cluster represents a different topic. It is method of Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups it means objects within the cluster are more similar whereas objects of different cluster are less similar. It is based on information found in the data that describes the objects and their relationships.

Advantages:

- Reduce the size of large data sets.
- Reduce time i.e. workload get reduce.

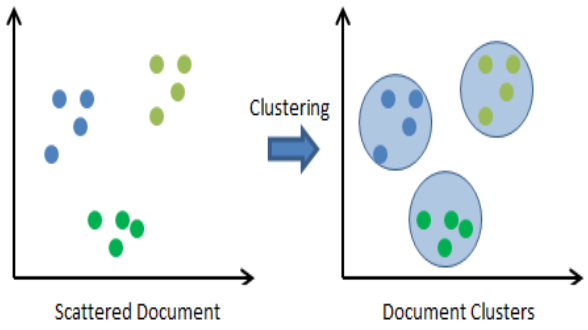

Figure 3:

DATA VISUALIZATION

Data visualization is the process of visually representing the data set in a meaningful and human understandable format. "Having the good information at the right time is good for making the good decisions". There is one type of data mining application i.e. Data Visualization which was considered as an information-modeling paradigm, Human beings understand graphics more easily than numbers and letters. Human brains can interpret graphs, charts, icons and models quicker than numbers in tables For example, a pie chart showing the classification of a university student will be understood quicker than the same data represented in a table.

II. LITERATURE REVIEW

Text clustering has widely become an important problem in recent years because of the large amount of unstructured data which is available in various forms. The problem of text clustering arises in the context of many application domains like the web, social networks, and other digital collections. [1] Showed that the use of side-information can greatly enhance the quality of text clustering, by maintaining a high level of efficiency.

Michael Steinbach presented the results of an experimental study of some common document clustering techniques that are agglomerative hierarchical clustering and K-means in [2] and measured as the bisecting K-means technique is good than the standard K-means approach and (somewhat surprisingly) as good than the hierarchical approaches. And also addressed that run time of bisecting K-means is better when compared to that of agglomerative hierarchical clustering techniques [2].

In traditional clustering, [3] studied the clustering problem in the presence of link structure information for the data set.

In [4], it is stated that clustering method BIRCH (Balanced Iterative Reducing and Clustering usingHierarchies) for very large datasets and made a large clustering problem tractable by concentrating on densely occupied portions, and using a compact summary.

Tao Liu, Shengping Liu, Zheng Chen demonstrated that feature selection can improve the text clustering efficiency and performance, in that features are selected based on class information[5]. Document clustering techniques have been received more and more attentions as a fundamental and enabling tool for efficient organization, retrieval, and summarization of huge volumes of text documents [6]. The general goal of clustering is to group similar objects into one cluster while partitioning dissimilar objects into different clusters. Clustering has broad applications including the analysis of business and financial data, biological data, also like time series data and spatial data.

A tremendous number of clustering algorithms have been proposed in the past to learn with multiple views of the data. Some of them first extract a set of shared features from the multiple views and then apply any off-the-shelf clustering algorithm such as k-means on these features.

Document has recently gained explicit text mining support with the tm package enabling statisticians to answer many interesting research questions via statistical analysis or modeling of

(Text) corpora. However, we typically face two challenges when analyzing large corpora: (1) The amount of data to be processed in a single machine is usually limited by the available main memory (i.e., RAM), and (2) An increase of the amount of data to beanalyzed leads to increasing computational workload. Algorithms for clustering and cluster validity have proliferated due to their promise for sorting out complex interactions between variables in high dimensional data. Excellent surveys of many popular methods for conventional clustering using deterministic and statistical clustering criteria are available; for example, consult the books by Duda and Hart (1973), Tou and Gonzalez (1974), or Hartigan (1975) [7].

S. Alouane, M. S. Hidri, and K. Barkaoui have proposed a new similarity measure which iteratively computes similarity while combining fuzzy sets in a three-partite graph to cluster documents in [8].The Documents can be provided from multiple sites and make similarity computation an expensive processing. This problem motivated to use parallel computing [8].

Cluster analysis is often one of the first steps in the analysis of data; it is an effort at unsupervised learning usually in the context of very little a priori knowledge. Chen pointed out, "a common drawback to all clustering algorithms is that the performance is highly dependent on the user setting various parameters.

By utilizing efficient algorithms such as Apriori [9] to analyze very large transactional data -frequently from transactional sales data- will result in a large set of association rules. Normally, finding the association rules from very large data sets is considered and emphasized as the most challenging step in association mining in [10].

First, our work is motivated by research in the field of clustering. Hierarchical method can be divided into agglomerative and divisive variants. Hierarchical clustering is used to build a tree of clusters, also known as a dendrogram.Every node contains child clusters. In hierarchical clustering we assign each item to a cluster such that if we have N items then we have N clusters. Find closest pair of clusters and merge them into single cluster.So, that one cluster get reduce from the whole structure. Compute distances (similarities) between the new cluster and each of the old clusters. And it work until all k cluster get merge.

Therefore, we use two different approaches while clustering the data. In the first agglomerative approach, a bottom-up clustering method get commonly used. It works from bottom to up. This method starts with a single cluster which containsall objects, and then it splits resulting clusters until only clusters of individual objects

remain. It terminate when individual cluster contain a single object.

Second, our work is motivated by research in the field of mining and partitioning method is well technique for this. Partitioning method divide the data into different subset or group such that some criterion that evaluate the clustering quality is optimize. Partitioning algorithm is of different type such as: probabilistic clustering using the Naive Bayes or Gaussian mixture model, EM, K-Means, Aproiri, FP-growth, Fuzzy-C-Means etc.

Analyzing huge textual corpuses is the source of two challenges. Firstly, the amount of data to be processed in a single machine is usually limited by the main memory available. Secondly, the increase of the amount of data to be analyzed leads to an increasing computational workload. Thus, it is imperative to find a solution which overcomes the memory limitation (by splitting the data into several pieces) and to markedly reduce the runtime by distributing the workload across available computing resources (CPU cores or cloud instances).Recently there has been an increasing interest in parallel implementations of data clustering algorithms. In, a distributed programming model and a corresponding implementation called PFT-sim has been proposed. It allows efficient parallel processing of data in a functional programming. It take into account three parallel abstraction level (Document-Sentence-Word).Another method also proposed called Map reduce which is also a good technique for parallel comparison [8].

III. AN OVERVIEW OF PROPOSED SYSTEM

Nowadays, lots of corpuses are collected from the search engine like Google, yahoo and data warehouses. The main objective of search engine is to provide the most applicable and required documents for a user's query. The search engine might give the relevant documents related the information of the user's query.

Due to these large amounts of corpuses, force to an interest in creating scalable and effective mining algorithm. Once the corpuses get collected, KDD process get perform on the corpuses. Therefore, the data will be in human readable format and now data will be efficient for further use, that means it does not contain any kind of noise. The data mining is one of the applications of Knowledge Discovery in Database process. So in this KDD process there are various steps on the information are Cleaning, Integration, Selection, Transformation, data mining etc.

This data mining, gives the relevant data. So, that the data will be divided into various groups for comparing with each other. These parallel comparisons are considered as a preprocessing step to clustering [8]. After this we are using clustering technique, to form the various group of data means to form the cluster. To summarize all this gathered information or cluster, visualization of data with the help of various graphs is good technique e.g. Line graph, histogram, etc. So, that we can measure the performance.

The following diagram states the overview of parallel comparison of documents for feature selection with data mining techniques.

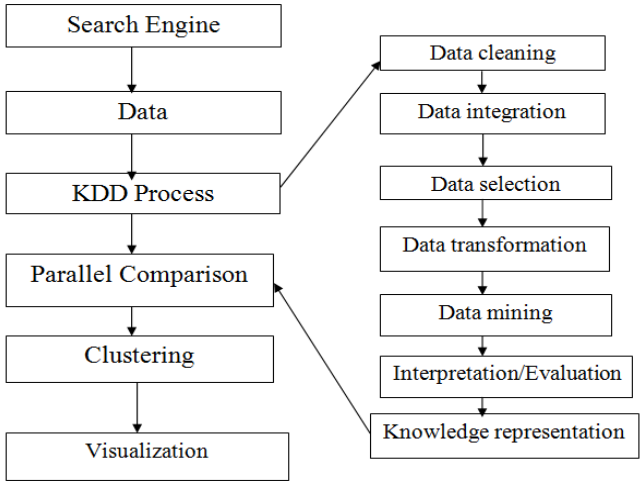

Figure 4: An Overview of Proposed System

IV. CONCLUDING REMARK

The objective of this paper was to give an introduction to basic issues of clustering text data with side information for large data sets. We are using clustering and classification techniques to improve the text data mining with other available information. The paper therefore aimed to provide early stage researchers with an introductory guide to this complex matter, raising awareness of the fact that very simple rules and guidelines might have a lasting impact on the text data with side information. And also the use of parallel comparator to compare text documents. So, that the efficiency will be improved also proposed an Visualization technique for visualizing cluster data so, that the data will be in human readable manner which reduce the workload and time.

V. REFERENCES:

1. C.C.Aggarwal and P.S.Yu, "On text clustering with side information," in Proc. IEEE ICDE Conf., Washington, DC, USA, 2012.
2. M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," inProc. Text Mining Workshop KDD, 2000, pp. 109–110.

3. R. Angelova and S. Siersdorfer, "A neighborhood-based approach for clustering of linkeddocument collections," inProc. CIKM Conf., New York, NY, USA, 2006, pp. 778–779.
4. T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An efficient data clustering method for very large databases," inProc. ACM SIGMOD Conf., New York, NY, USA, 1996, pp. 103–114.
5. T. Liu, S. Liu, Z. Chen, and W.-Y. Ma, "An evaluation of feature selection for text clustering," in Proc. ICML Conf., Washington, DC, USA, 2003, pp. 488–495.
6. W. Xu, X. Liu, and Y. Gong, "Document clustering based on nonnegative matrix factorization," inProc. ACM SIGIR Conf.,New York, NY, USA, 2003, pp. 267–273.
7. J. C. Bezdek, "Fcm: The fuzzy c-means clustering algorithm,"Computers and Geosciences, vol. 10(2-3), pp. 191–203, 1984.
8. S. Alouane, M. S. Hidri, and K. Barkaoui, "Fuzzy triadic similarity for text categorization: Towards parallel computing," in the 5th International Conference on Web and Information Technologies, 2013, pp.265–274.
9. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. ProceedingsOf the 20th VLDB Conference, Santiago, Chile. (1994) 487–499.
10. Ertek G. and Demiriz A, "A Framework for Visualizing Association Mining Results," in ISCIS, pp. 593-602, 2006.