Survey on Parallel Comparison of Text Document with Input Data Mining and VizSFP

Priyanka P. Palsaniya[1], D. C. Dhanwani[2]

[1]Student, Computer Science & Engineering, P.R.Pote College / Sant. Gadge Baba Amravati University, India

[2]Assistant Professor, Computer Science & Engineering, P.R.Pote College / Sant. Gadge Baba Amravati University, India

Abstract: ***Nowadays, the usage of world wide web is growing tremendously. The corpuses available on this world wide web is not structured or scattered data , due to this wide availability of huge amounts of corpuses there is a need for turning such data into useful information and knowledge. Therefore, data mining algorithm get used. In this work, various data is taken as input and mining algorithm get apply on this data for getting frequent pattern. After this, Parallel comparator gets used to detect the similar pattern from the corpuses. Then the data get arrange into different groups of similar type by using clustering algorithm. To overcome the problem of representing this large amount of data, data mining and clustering results we are using information visualization techniques for visualizing similar frequent patterns.***

Keyword: Web mining, Data Mining, Parallel comparison, Clustering, Classification, Frequent Pattern

1. Introduction

In recent years, mining has gradually become a new research topic. The Web today contains a lot of information such as information related to organizations, products, industry and many useful knowledge etc. that may be of wide interest. Web Mining is the application of data mining techniques to discover patterns from the Web. Web mining can be divided in 3 categories. Web usage mining is the application that uses data mining to analyze and discover interesting patterns on the web. Web content mining is the combination of both data and text mining process to discover useful information in the web. Web structure mining means using graph theory to analyze the node and connection structure of a web site [1], [4].

In many mining applications, side-information is available along with the documents. Example of side-information such as the hyperlinks or non-textual attributes, document provenance information, user-access behavior from web logs. Due the availability of this side-information, Data mining has attracted a great deal of attention in the information industry and in society as a whole in recent years. Because of this, there is a need of using a principled way to perform the mining process, so that system are able to maximize the advantages from using this side information which led to an interest in creating scalable and effective mining algorithms [2],[3] .

Simply stated, data mining refers to extracting or “mining” knowledge from large amounts of data means it is the process that finds a small set of precious nuggets from a great deal of raw material. Remember that the mining of gold from rocks or sand is referred to as gold mining rather than rock or sand mining. Thus, data mining should have been more appropriately named “knowledge mining from data,” which is unfortunately somewhat long. “Knowledge mining,” a shorter term may not reflect the emphasis on mining from large amounts of data. Nevertheless, mining is a term characterizing the process that finds a small set of precious nuggets from a great deal of raw material.

Many people treat data mining as a synonym for another popularly used term, Knowledge Discovery in database or KDD. 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 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.

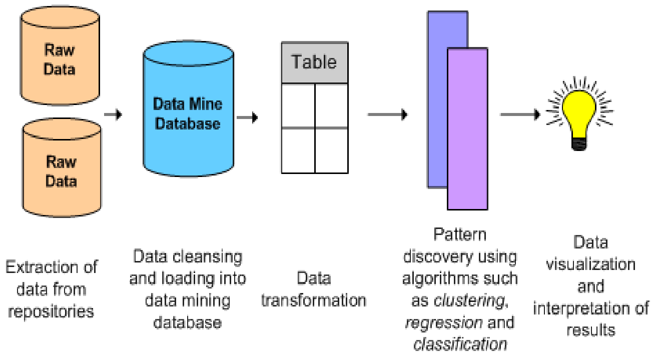
Raw Data
Raw Data
Data Mine Database
Table
Extraction of data from repositories
Data cleansing and loading into data mining database
Data transformation
Pattern discovery using algorithms such as *clustering, regression* and *classification*
Data visualization and interpretation of results

Figure 1: Data Mining

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

Text mining = information retrieval + statistics + artificial intelligence (natural language processing, machine learning / pattern recognition).

Clustering is the method 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 where as objects of different cluster are less similar. Clustering has several advantages which are as follows : Reduce the size of large data sets, Reduce time i.e. workload get reduce.

Data visualization is the process of visually representing the data set in a meaningful and human understandable format. “It is good to have good information at the right time 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 [16].

2. Literature Survey

Literature Survey is the most important step in software development process. Generally, clustering methods can be categorized as Hierarchical and Partitional (Non Hierarchical). Therefore, document clustering can be done either by using Hierarchical and Partitonal technique. It is possible to use different types of algorithms to extract most important information from a database [14].

In traditional clustering, [20] studied the clustering problem in the presence of link structure information for the data set. [21] stated clustering method BIRCH(Balanced Iterative Reducing and Clustering using Hierarchies) for very large datasets and made a large clustering problem tractable by concentrating on densely occupied portions, and using a compact summary. 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 contains all objects, and then it splits resulting clusters until only clusters of individual objects remain. It terminate when individual cluster contain a single object. In the second divisive approach a top-down clustering method and is less commonly used. It works in the opposite direction to that of agglomerative technique. It work from top to down and other steps is same as that of agglomerative approach [15].The major focus of this work has been on scalable clustering of multi-dimensional data of different types. There are various types of text clustering methods; one of the most well-known techniques for text-clustering is the scatter-gather technique, which uses a combination of agglomerative and partitional clustering [12].

In a similar fashion, Kshitij Bhagwat, Dhanshri More, Sayali Shinde, Akshay Daga, Assistant Prof. Rupali Tornekar describe a Comparative Study of Brain Tumor Detection Using K Means , Fuzzy C Means and Hierarchical Clustering Algorithms[19].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 K-Means, Aproiri, FP-growth, Fuzzy-C-Means etc. These algorithms are among the most important data mining algorithms in the research community.

Michael Steinbach presented the results of an experimental study of some common document clustering techniques that are agglomerative hierarchical clustering and K-means in [13] 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 [13].The most popular partitioning algorithm is the k-Means algorithm. All this approach is define as a Partition objects into k nonempty subsets. If the data point or a mean point (object closest to the centroid of a cluster) is closest to its own cluster, leave it where it is. If the data point is not closest to its own cluster, move it into the closest cluster. For each data point: Calculate the distance from the data point to each cluster then Assign each object to the closest cluster and if we get results then no data point moving from one cluster to another. So, we say that clusters are stable and the clustering process terminate. K-means clustering is Simple, understandable and items automatically assigned to clusters.

Many algorithms are developed and they are divided into two classes: candidate generation or pattern growth. Apriori algorithm first generates an association rule for frequent pattern matching. Apriori is a representative the candidate generation approach. It generates length (k+1) candidate item sets based on length (k) frequent item sets. Since Apriori algorithm was first introduced and as experience was accumulated, there have been many attempts to devise more efficient algorithms of frequent item set mining [6]. The most outstanding improvement over Apriori would be a method called FP-growth (frequent pattern growth) that succeeded in eliminating candidate generation. The frequency of item sets is defined by counting their occurrence in transactions. FP-growth, is proposed by Han in 2000, represents pattern growth approach, it used specific data structure (FP-tree), FP-growth discover the frequent item sets by finding all frequent in 1-itemsets into condition pattern base , the condition pattern base is constructed efficiently based on the link of node structure that association with FP-tree. FP-growth does not generate candidate item sets explicitly.

S. Alouane, M. S. Hidri, and K. Barkaoui describe a Fuzzy triadic similarity for text categorization Analyzing huge textual corpuses is the source of two challenges. Which iteratively computes similarity while combining fuzzy sets in a three-partite graph to cluster documents in.The Documents can be provided from multiple sites and make similarity computation an expensive processing. This problem motivated to use parallel computing. 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 [10]. Saleh Abdel-Hafeez, Ann Gordon-Ross and Behrooz Parhami describe an Scalable Digital CMOS Comparator Using a Parallel Prefix Tree [24].

There are a number of visualization techniques which can be used for visualizing the. Such as x-y plots, bar charts, line graphs, etc., with that a number of sophisticated visualization techniques going to be used in data visualization. Data usually consist of number of dimension and variables depending on data different visualization are possible. According variables and dimension data divided into one dimensional data, two dimensional data, multidimensional data and more complex form text and hypertext data, hierarchy of graph, data type from the field of algorithm and software .To represent one dimensional data histogram or pie chart method is used, for two dimensional data scatter plot and line graph is used, and for multidimensional data icon based method, pixel based method, dynamic parallel coordinate system. No single algorithm or method best at all time. Performance of visualization is highly data dependent. For visualize data, data must be preprocess and classify according to dimensions and for recovery data mining techniques such as preprocessing, classification are used [23].

Daniel A. Keim, Hans-Peter Kriegel describes various Visualization Techniques for Mining Large Databases [22]**.** Michael Hahsler and Sudheer Chelluboina describe a Visualizing association rules in hierarchical group [17]. Interesting association rules and frequent patterns can be discovered by using 3 types of measures such as i) Objective Measures (ii) Subjective Measures and (iii) Semantic Measures. Objective measure is a data-driven approach for evaluating the quality of association patterns. Since, these measures calculate the frequent patterns based on the statistical parameters, different measures produce different results. Hence they are not sufficient for determining the interestingness of a discovered rule. Subjective measures generally operate by comparing the beliefs of a user against the patterns discovered by the data mining algorithm. But these measures are also not sufficient, as the rules generated are user biased. Ertek G. and Demiriz A. also present a Framework for Visualizing Association Mining Results Semantic measures uses natural language processing (NLP) techniques such as domain ontologies and web ontologies to identify relationships amongst the patterns. SSFPOA extracts and clusters semantically similar frequent patterns. It uses both domain dependent and domain independent ontologies, and considers the entire path to conform the semantic similarity between elements along with their structural information such as the number of the children for each node, the number of subclasses for each class within the ontology. These patterns can be best displayed using visualization methods [18].

3. Overview of Proposed Methodology

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 met the information need of the user’s query.

These large amounts of corpuses led to an interest in creating scalable and effective mining algorithm. After this KDD process get perform on these collected corpuses .So, that the data will be in human readable format and data will be efficient that means it does not contain any noisy information apart from the original information. 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, give us 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 [10]. 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 with input data mining and visualization.

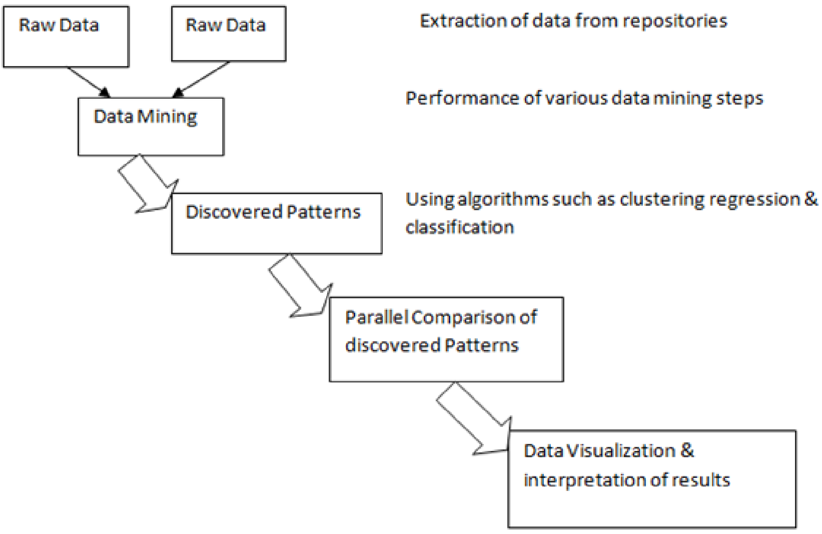
Raw Data
Raw Data
Extraction of data from repositories
Data Mining
Performance of various data mining steps
Discovered Patterns
Using algorithms such as clustering regression & classification
Parallel Comparison of discovered Patterns
Data Visualization & interpretation of results

Figure 2: An Overview of Proposed System

4. 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. 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.

References

[1] C. C. Aggarwal and C.-X. Zhai, *Mining Text Data*, Springer, 2012.

[2] I. S. Dhillon, J. Fan, and Y. Guan. Efficient clustering of very large document collections. In R. Grossman, C. Kamath, P. Kegelmeyer, V. Kumar, and R. Namburu, editors, Data Mining for Scientific and Engineering Applications, pages 357–381. Kluwer Academic Publishers, 2001.

[3] Kraft, M. R., Desouza, K. C., Androwich, I., "Data Mining in Healthcare Information Systems: Case Study of a Veterans' Administration Spinal Cord Injury Population". *IEEE, Proceedings of the 36th Hawaii International Conference on System Sciences, 0-7695-1874-5/03, 2002.*

[4] Han J. and Kamber M., Data Mining: Concepts and Techniques, 2nd ed., San Francisco, Morgan Kauffmann Publishers,2001.

[5] Silvia Rissino and Germano Lambert-Torres, Rough Set Theory – Fundamental Concepts, Principals, Data Extraction, and Applications, Data Mining and Knowledge Discovery in Real Life Applications, pp. 438, February 2009.

[6] R. Agrawal and R. Srikant. "Fast algorithms for mining association rules in large databases". In Proc. VLDB 1994, pp. 487–499.

[7] J. Han, J. Pei, Y. Yin, and R. Mao. "Mining frequent patterns without candidate generation: a frequent-pattern tree approach". Data Mining and Knowledge Discovery, 8(1), pp. 53–87, 2004 .

[8] F. de Carvalho, Y. Lechevallier, and F. M. de Melo, "Partitioning hard clustering algorithms based on multiple dissimilarity matrices," *Pattern Recognition*, vol. 45, pp. 447–464, 2012.

[9] J. C. Bezdek, "Fcm: The fuzzy c-means clustering algorithm," *Computers and Geosciences*, vol. 10(2-3), pp. 191–203, 1984.

[10] S. Alouane, M. S. Hidri, and K. Barkaoui, "Fuzzy triadic similarity for text categorization: Towards parallel computing," in *the 5th Inter-national Conference on Web and Information Technologies*, 2013, pp. 265–274.

[11] L. A. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, pp. 338–353, 1996. R. Benosman, Y. Albrieux, and K. Barkaoui, "Performance evaluation of a massively parallel esb-oriented architecture," in *Service-Oriented Computing and Applications*, 2012, pp. 1–4.

[12] D. Cutting, D. Karger, J. Pederson, and J. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In *Proceedings of ACM SIGIR*, 1992.

[13]]M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," inProc. Text Mining Workshop KDD, 2000, pp. 109–110.

[14] H. Schutze and C. Silverstein, "Projections for Efficient Document Clustering," in *ACM SIGIR Conf.*, pp. 74–81, 1997.

[15] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," in *Text Mining Workshop, KDD*, pp. 109–110, 2000.

[16] S. Vasavi, S. Jayaprada, V. Srinivasa Rao, "Extracting Semantically Similar Frequent Patterns Using Ontologies", Proceedings of the Second international conference on Swarm, Evolutionary, and Memetic Computing - Volume Part II, Pages 157-165, 2011.

[17] Michael Hahsler and Sudheer Chelluboina, "Visualizing association rules in hierarchical groups", In Computing Science and Statistics, Vol. 4242nd Symposium on the Interface: Statistical, Machine Learning, and Visualization Algorithms" Interface 2011.

[18] Ertek G. and Demiriz A, "A Framework for Visualizing Association Mining Results," in ISCIS,pp.593-60,2006.

[19] Kshitij Bhagwat, Dhanshri More, Sayali Shinde, Akshay Daga, Assistant Prof. Rupali Tornekar, " Comparative Study of Brain Tumor Detection Using K Means , Fuzzy C Means and Hierarchical Clustering Algorithms" International Journal of Scientific & Engineering Research, Volume 4, Issue 6, June-2013 626 ISSN 2229-5518.

[20] R. Angelova and S. Siersdorfer, "A neighborhood-based approach for clustering of linked document collections," in Proc. CIKM Conf., New York, NY, USA, 2006, pp. 778–779.

[21] 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.

[22] 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.

[23] Mr. Sushilkumar Chavhan1, Ms.S.M.Nirkhi," Visualization Techniques for Digital forensics: A Survey," in International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)Volume-2 Number-4 Issue-6 December-2012.

[24] Saleh Abdel-Hafeez, Ann Gordon-Ross and Behrooz Parhami ,"Scalable Digital CMOS Comparator Using a Parallel Prefix Tree," in IEEE transactions on very large scale integration (vlsi) systems, vol. 21, no. 11, november 2013