IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 7, July 2015.
www.ijiset.com
ISSN 2348 – 7968

Preprocessing, Mining & CURE Hierarchical Clustering for Web Log Mining

Ms .Aashwini Thakare[1], Prof.M.S.Chaudhari[2]

[1] Computer Science & Engineering, PBCOE,
Nagpur, Maharashtra, India

[2] Computer Science & Engineering, PBCOE,
Nagpur, Maharashtra, India

Abstract

Now a day the World Wide Web becomes very popular and interactive for transferring of information. It is massive repository of web pages & links. It provides information about vast area for the internet user. The web is huge, diverse and active and thus increasing the scalability, multimedia data & temporal matters. The growth of the web has outcome in a huge amount of information that is now freely offered for user access. Since due to tremendous usage, the log files are growing at a faster rate and the size is becoming huge. Web page access and usage information, providing rich source for data mining. Natural Language Processing plays a vital role in efficient mining process as log data is normally noisy and indistinct. Data Mining is a process of analyzing large amounts of data and picking out the relevant information. It refers to extracting or mining knowledge from large amounts of data. Clustering, in data mining, is useful for discovering groups and identifying interesting distributions in the underlying data CURE (Clustering usage Representatives) method find clusters from a large database that is more robust to outliers, and identifies clusters having non-spherical shapes and wide variances in size.

Keywords: *WWW, Natural Language Processing, Data Mining, CURE Clustering*

1. Introduction

World Wide Web (WWW) is expanding tremendously everyday in the number of websites and also the population of users [7]. The basic purpose of website is to deliver useful information to its users efficiently and timely at the same time websites are competing to acquire their own shares of visitors. Websites are striving to improve themselves by offering personalized contents and services that supposedly is match best of the users' tastes or needs. The growth of the web has outcome in a huge amount of information that is now freely offered for user access. Since due to tremendous usage, the log files are growing at a faster rate and the size is becoming huge. Web page access and usage information, providing rich source for data mining [5].

1.1. Web Log Mining

Web usage mining also known as web log mining is the application of data mining techniques on large web log repositories to discover useful knowledge about user's behavioral patterns and website usage statistics that can be used for various website design tasks. The main source of data for web usage mining consists of textual logs collected by numerous web servers all around the world. There are four stages in web usage mining [7].

1.1.1. Data Collection

Data Collection is the first step in web usage mining process. It consists of gathering the relevant web data. Data source can be collected at the server-side, client-side, proxy servers, or obtain from an organization's database, which contains business data or consolidated Web data [7].

1.1.2. Data Preprocessing

The information available in the web is heterogeneous and unstructured. Therefore, the preprocessing phase is a prerequisite for discovering patterns. The goal of preprocessing is to transform the raw click stream data into a set of user profiles. Data preprocessing presents a number of unique challenges which led to a variety of algorithms and heuristic techniques for preprocessing tasks such as merging and cleaning, user and session identification etc [7].

1.1.3. Pattern Discovery

Once user transactions have been identified, a variety of data mining techniques are performed for pattern discovery in web usage mining. These methods represent

ISSN 2348 – 7968

the approaches that often appear in the data mining literature such as discovery of association rules and sequential patterns and clustering and classification etc. By using Apriori algorithm the biggest frequent access item sets from transaction databases that is the user access pattern are discovered [7].

1.1.4. Pattern Analysis

Pattern Analysis is the last stage of web usage mining. Mined patterns are not suitable for interpretations and judgments. So it is important to filter out uninteresting rules or patterns from the set found in the pattern discovery phase. In this stage tools are provided to facilitate the transformation of information into knowledge. The exact analysis methodology is usually governed by the application for which Web mining is done [7].

1.2. Application of Web Usage Mining

Users' behavior is used in different applications such as Personalization, e-commerce, to improve the system and to improve the system design as per their interest etc., Web personalization offers many functions such as simple user salutation to more complicate such as content delivery as per users interests. Content delivery is very important since non expert users are overwhelmed by the quantity of information available online. It is possible to anticipate the user behavior by analyzing the current navigation patterns with patterns which were extracted from past web log. Recommendation systems are the most common application. Personalized sites are example for recommendation systems. E-Commerce applications need customer details for Customer Relationship Management [7].

2. Literature Survey

B.Uma Maheswari & Dr.P.Sumathi presented preprocessing method for web log mining is the important method & preprocessing plays a vital role in efficient mining process as log data is normally noisy and indistinct [2].Reconstruction of sessions and paths are completed by appending missing pages in preprocessing. Additionally ,the transactions which illustrate the behaviors of n users are constructed exactly in preprocessing by calculating the reference length of user access by means of byte rate. Using Web clustering several types of objects can be clustered into different groups for various purposes.

Supinder Singh presented the review the existing web usage clustering techniques and proposed a swarm intelligence based PSO clustering algorithm for the clustering of web user sessions .The proposed algorithm works independently without hybridization with any other clustering algorithm. The results showed that the proposed approach performs better than the K-Means clustering .But as analyzed there is difference between the working styles of both the algorithm, because K-means works as a partitioning method and PSO works in hierarchical way [3].

V. Chitraa, Dr. Antony Selvdoss Davamani presented World Wide Web is a huge repository of web pages and links. It provides abundance of information for the Internet users. The growth of web is tremendous as approximately one million pages are added daily. Users' accesses are recorded in web logs. Because of the tremendous usage of web, the web log files are growing at a faster rate and the size is becoming huge. Web Usage Mining applies mining techniques in log data to extract the behavior of users which is used in various applications like personalized services, adaptive web sites, customer profiling, prefetching, creating attractive web sites etc., Web usage mining consists of three phases preprocessing, pattern discovery and pattern analysis [7].

Fabio Ciravegna, Sanda Harabagiu, presented The Recent Advances in Natural Language Processing Language is the most natural way of communication for humans. The vast majority of information is stored or passed in natural language (for example, in textual format or in dialogues). Natural language processing aims at defining methodologies, models, and systems able to cope with NL, to either understand or generate it. Opportunities for NLP applications range from querying archives (for example, the historical activity on querying databases), to accessing collections of texts and extracting information, to report generation, to machine translation [22].

Natural Language Processing (NLP) is the computerized approach to analyzing text that is based on both a set of theories and a set of technologies. And, being a very active area of research and development, there is not a single agreed-upon definition that would satisfy everyone, but there are some aspects, which would be part of any knowledgeable person's definition [48].

S.Veeramalai, N.Jaisankar, A.Kannan, presented Efficient Web Log Mining Using Enhanced Apriori Algorithm with Hash Tree and Fuzzy. The aim in web mining is to discover and retrieve useful and interesting patterns from a large dataset. In web mining, this dataset is the huge web data. Web data contains different kinds of information, including, web structure data, web log data, and user profiles data. Web mining is the application of data mining techniques to extract knowledge from web data, where at least one of structure or usage data is used in the mining process. Web usage mining has various application areas such as web pre-fetching, link prediction, site reorganization and web personalization. Most

ISSN 2348 – 7968

important phases of web usage mining are the reconstruction of user sessions by using heuristics techniques and discovering useful patterns from these sessions by using pattern discovery techniques like association rule mining, Apriori, etc[34].

Yogita Rani, Manju & Harish Rohil presented Comparative Analysis of BIRCH and CURE Hierarchical Clustering Algorithm using WEKA 3.6.9.The Hierarchical Clustering is the process of forming a maximal collection of subsets of objects (called clusters), with the property that any two clusters are either disjoint or nested. Hierarchical clustering combine data objects into clusters, those clusters into larger clusters, and so forth, creates a hierarchy of clusters, which may represent a tree structure called a dendrogram, in which the root of the tree consists of a single cluster containing all observations, and the leaves correspond to individual observations. BIRCH and CURE are two integrated hierarchical clustering algorithm [40].

CURE (Clustering Using Representatives) is a clustering algorithm that uses a variety of different techniques to create an approach which can handle large data sets, outliers, and clusters with non-spherical shapes and non-uniform sizes. CURE represents a cluster by using multiple "representative" points from the cluster. These points will, in theory, capture the geometry and shape of the cluster [38].

Seema Maitrey, C.K. Jhaa presented an integrated approach for CURE clustering using map-reduce technique. Data mining is a technique which allows to search precious information from huge collection of data. The explosive growth in databases has created a need to develop technologies that use information and knowledge intelligently. Therefore, Data Mining Technique has become an increasingly important research area [39].

One of the hierarchical clustering technique known as CURE clustering is used for handling large databases. In CURE, the concept of representative points is employed to find all the clusters of different shapes and sizes. Specifically, CURE represents a cluster by using multiple representative points from the cluster. These points capture the geometry and shape of the cluster and thus allow for non-globular clusters [44].

Seema Maitrey, C. K. Jha, Rajat Gupt , Jaiveer Singh presented Enhancement of CURE Clustering Technique in Data Mining. Clustering, in data mining, is useful for discovering groups and identifying interesting distributions in the underlying data. Among several clustering algorithms, we have considered CURE method from hierarchical clustering. CURE (Clustering usage Representatives) method find clusters from a large database that is more robust to outliers, and identifies clusters having non-spherical shapes and wide variances in size. CURE employs a combination of data collection, data reduction by using random sampling and partitioning [46].

3. Problem Definition

Now day, more amount of information is created weekly in websites. The advent of the Social Web has provided people with new content-sharing services that allow them to create and share their own contents, ideas, and opinions, in a time- and cost-efficient way, with virtually millions of other people connected to the World Wide Web. The first step is to define the time frame for which the log of data is collected and worked upon. This question is imperative to the size of the dataset. The size of the access log of 1 day is more than 250MB-500MB.It is difficult to obtain any useful pattern sequence by analyzing a single day's access log [2].

The huge amount of information, however, is mainly unstructured (because it is specifically produced for human consumption) and hence not directly machine process able. Because many citizens don't care about syntactic mistakes so that NLP approach is used to analyze a large corpus of the system to handle linguistic objects that fall outside the field of interest. The automatic analysis of text involves a deep understanding of natural language processing [19].

3.1. Aims & Objectives

1. The need of the data cleaning process will be a vital task and consumers the most amount of time. Unless you have the correct pre-processed data, it will be difficult to achieve good results.
2. The next important step will be feature extraction, where one needs to think out of the box keeping in mind the project goal. After the data will be process, the other important step will be to visualize the results obtained from the mining of the dataset.
3. In order to achieve all of this, the most important decision will be to choose the right set of tools that will support all the necessary requirements like handling huge amount of data, data pre-processing, mining algorithms and data visualization.

4. Research Methodology

4.1. System Model

The aim in web mining is to discover and retrieve useful and interesting patterns from a large dataset. In web mining, this dataset is the huge web data. Web data contains different kinds of information, including, web

ISSN 2348 – 7968

structure data, web log data, and user profiles data. Web mining is the application of data mining techniques to extract knowledge from web data, where at least one of structure or usage data is used in the mining process. Web usage mining has various application areas such as web pre-fetching, link prediction, site reorganization and web personalization.

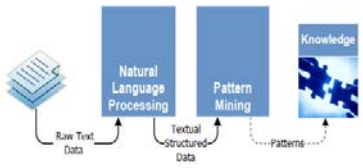

Figure 1.System Model

4.2. Natural Language Processing

Natural Language Processing can be a powerful technique to access structured and non structured information and to improve human-computer interaction so that authors aimed for discovering a NLP approach which improves e-democracy by increasing the citizens' participation in the decision-making process. NLP's goal is to test satisfiability of the requirements and to test robustness and the usability of the augmented phrase structure grammars in a highly sensitive environment. The NLP also aim at developing two tool sets to improve the communication of users in the context of urban planning. An important assumption of this approach is that the language fragment shows some (minimum) criteria of grammaticality [19].

Many citizens don't care about syntactic mistakes so that NLP approach is used to analyze a large corpus of the language and enable the system to handle linguistic objects that fall outside the field of interest. If we loosen certain rules of the grammars it may improve the performance for a system that deals with analyzing meaningful fragments in ungrammatical sentences. Some problems require skilled technical people who adopts on the system architecture but it is a difficult task so that the system may be designed with efficient NLP technique based on dirty NLP language that can manage an imperfect sentences. Natural Language Processing is also called as partial parsing & involves three important tasks such as Tokenization, Part of Speech Tagging & Chunking.

4.2.1. Tokenization

Tokenization is typically the first task in a pipeline of Natural Language Processing tools. It usually involves two sub-tasks which are often performed at the same time [27].

1. Separating punctuation symbol from words,
2. Detecting sentence boundaries.

4.2.2. Part of Speech Tagging

Part of speech (POS) Tagging is often performed as a preprocessing step in various speech & language processing tasks, form text chunking to dialog management to semantic role labeling .As POS Tag augment the information contained within words by explicitly indicating some of the structure inherent in language, their accuracy is especially critical to downstream natural language processing (NLP) applications [26].

Word can be used in variety of grammatical roles, for example nouns, adjectives, prepositions, verbs and so on. These categories are the basic grammatical units of language and are the basic parts of speech .Part of Speech tagging, or POS Tagging, is the task of automatically part labeling each taken in the sentence with its part of speech. This is crucial early step in part of speech understanding a sentence [22].

Part of speech tagging is the process of identifying the part of speech corresponding to each word in the text, based on both its definition, as well as its context i.e. relationship with adjacent and related words in a phrase or Sentence .POS Tagging aims at labeling each word with a unique tag that indicates its syntactic role, e.g. plural noun, adverb.

POS Tagging aims at labeling each word with a unique tag that indicates its syntactic role, e.g. plural noun, adverb [16].

For Example:

The function of sleep, according to one school of thought, is to consolidate memory.

The| DET function | NOUN of | PREP sleep | NOUN, |PUN according |V ERB to |PREP one |DET school | NOUN of | IN thought | NOUN, |PUN is | VERB to | PREP consolidate |VERB memory | NOUN, | PUN.

The| DET function | VERB of | PREP sleep | VERB, |PUN according |VERB to |PREP one |DET school | NOUN of | IN thought | VERB, | PUN is | VERB to | PREP consolidate |VERB memory | NOUN, | PUN.

The| DT function | NN of | IN sleep | NN, |, according |VBG to |TO one |CD school | NN of | IN thought | NN, |, is | VBZ to | TO consolidate |VB memory | NN.| [27].

ISSN 2348 – 7968

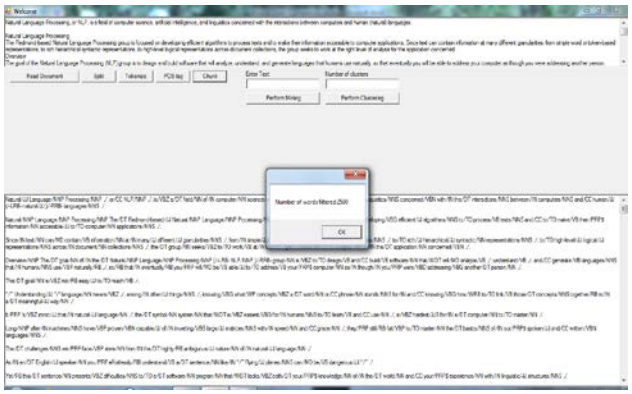

Figure 2 . Output of Tagging Process

4.2.3. Chunking

Chunking also called shallow parsing, aims at labeling segments of a sentence with syntactic constituents such as noun or verb phrase (NP or VP). Each word is assigned only one unique tag, often encoded as a begin-chunk (e.g. B-NP) or inside-chunk tag (e.g. INP)[20].Chunking is the process of dividing sentences into series of words that together constitutes a grammatical unit (mostly either noun or verb, or preposition phrases).The output is different from that of a fully parsed tree because it consists of series of words that do not overlap and that do not contain each other. This makes chunking an easier Natural Language Processing task than parsing [27].

Thus, chunking is a middle step between identifying the part of speech of individual words in a sentence, and providing a full parsed tree of it. Chunking can be useful for information retrieval, information extraction, and question answering since a complete chunk (Noun, Verb or Preposition Phrase) is likely to be semantically relevant for the requested information.

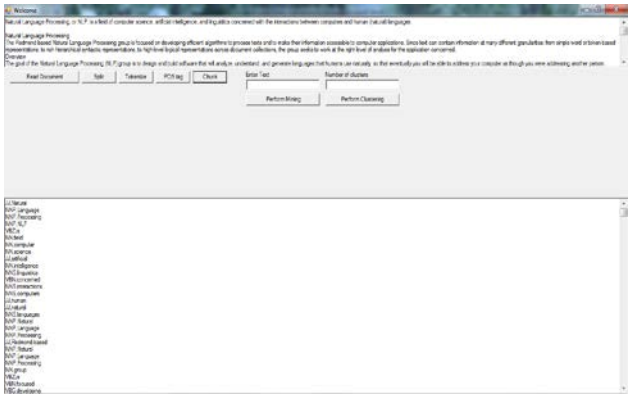

Figure 3. Output of Chunking Process

4.2.4. Application of Natural Language Processing

Natural language processing provides both theory and implementations for a range of applications. In fact, any application that utilizes text is a candidate for NLP [48]. The most frequent applications utilizing NLP include the following:

- **Information Retrieval** – given the significant presence of text in this application, it is surprising that so few implementations utilize NLP.
- **Information Extraction (IE)** – a more recent application area, IE focuses on the recognition, tagging, and extraction into a structured representation, certain key elements of information, e.g. persons, companies, locations, organizations, from large collections of text. These extractions can then be utilized for a range of applications including question-answering, visualization, and data mining.
- **Question-Answering** – in contrast to Information Retrieval, which provides a list of potentially relevant documents in response to a user's query, question-answering provides the user with either just the text of the answer itself or answer-providing passages.
- **Summarization** – the higher levels of NLP, particularly the discourse level, can empower an implementation that reduces a larger text into a shorter, yet richly constituted abbreviated narrative representation of the original document.

4.3. Data Mining

Data mining is an important method to increase efficiency, discover hidden, useful, valid and understandable knowledge from a massive database [32]. Data mining is the process of analyzing data from different perspective and summarizing the data into useful identical format of information that can be used to predict future trends or performances. The ultimate goal of data mining is to recognize pattern full information and predictions. Association mining is an important component of data mining. The Apriori Algorithm was proposed by Agrawal and Srikant in 1994 [31]. The algorithm finds the frequent set in the database. It makes use of the downward closure property. The algorithm is a bottom search, moving upward level-wise in the lattice. However, before reading the database at every level, it prunes many of the sets which are unlikely to be frequent sets, thus saving any extra efforts.

4.3.1. The Apriori Algorithm

Apriori employs an iterative approach, where k-item sets are used to explore (k+1)-item sets. First, the set of frequent 1-itemsets is found by scanning the database to

ISSN 2348 – 7968

count the occurrences for each item, and then collecting those items that satisfy minimum support defined. The result set obtained is denoted as L1. Now, L1 is used to find the set of frequent 2-itemsets, L2 and so on, until no more frequent k-item sets can be found. Thus the finding of each frequent k-item set requires one full scan of the database. To improve the efficiency of frequent item sets level wise generation, an important property called the Apriori property, is used. It reduces the search space. Apriori property states that "All nonempty subsets of a frequent item set must also be frequent". Thus it is divided into a two-step process which is used to find the frequent item sets: join and prune actions [49].

1. The Join Step:

A set of candidate k-item set is generated by joining Lk-1 with itself. This set of candidates is denoted as Ck.

2. The Prune Step:

The members of Ck may or may not be frequent, but all of the frequent k-item sets are members of Ck. A scan of the database is required to determine the count of each candidate in Ck. The scan of the database result in the formation of Lk i.e. it contains all candidates having a count no less than the minimum support count are frequent by definition, and therefore belong to Lk. The Apriori property is used to reduce the size of Ck, as follows. Any (K-1)-item set that is not frequent cannot be a subset of a frequent k-item set. Hence, if any (K-1)-subset of the candidate k-item set is not in Lk-1, then the candidate item cannot be frequent either and so can be deleted from Ck [49].

Figure 4 . Output of Apriori Algorithm

4.4. Clustering

Clustering is the process of grouping the data into classes or clusters, so that objects within a cluster have high similarity in comparison to one another but these objects are very dissimilar to the objects that are in other clusters. Clustering methods are mainly divided into two groups: hierarchical and partitioning methods. Hierarchical clustering combine data objects into clusters, those clusters into larger clusters, and so forth, creating a hierarchy of clusters. In partitioning clustering methods various partitions are constructed and then evaluations of these partitions are performed by some criterion [38].

4.3.2. Goals of Clustering

The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. But how to decide what constitutes a good clustering. It can be shown that there is no absolute "best" criterion which would be independent of the final aim of the clustering. Consequently, it is the user which must supply this criterion, in such a way that the result of the clustering will suit their needs. For instance, we could be interested in finding representatives for homogeneous groups (data reduction), in finding "natural clusters" and describe their unknown properties ("natural" data types), in finding useful and suitable groupings ("useful" data classes) or in finding unusual data objects (outlier detection) [46].

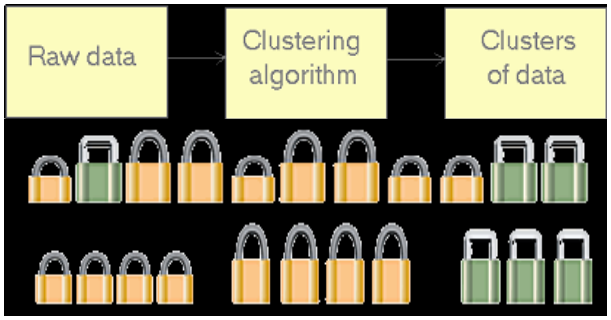

Figure 5. Stages of Clustering

4.4.2. Hierarchical Clustering Algorithm

Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. The quality of a pure hierarchical clustering method suffers from its inability to perform adjustment, once a merge or split decision has been executed. Then it will neither undo what was done previously, nor perform object swapping between clusters. Thus merge or split decision, if not well chosen at some step, may lead to some-what low-quality clusters. One promising direction for improving the

ISSN 2348 – 7968

clustering quality of hierarchical methods is to integrate hierarchical clustering with other techniques for multiple phase clustering [46].

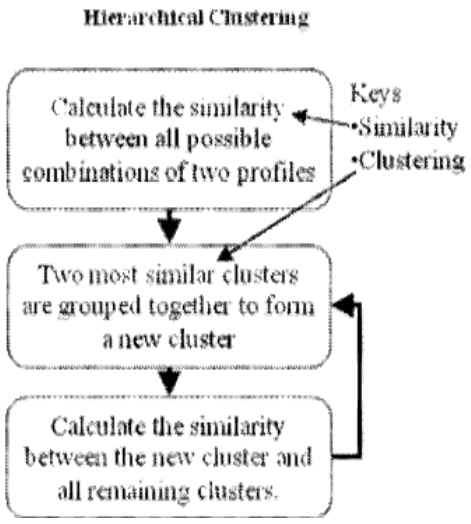

Figure 6.Hierarchial Clustering

4.4.3. CURE (Clustering Using RE presentatives) Clustering Algorithm

CURE is an agglomerative hierarchical clustering algorithm that creates a balance between centroid and all point approaches. Basically CURE is a hierarchical clustering algorithm that uses partitioning of dataset. A combination of random sampling and partitioning is used here so that large database can be handled. In this process a random sample drawn from the dataset is first partitioned and then each partition is partially clustered. The partial clusters are then again clustered in a second pass to yield the desired clusters. It is confirmed by the experiments that the quality of clusters produced by CURE is much better than those found by other existing algorithm[40].

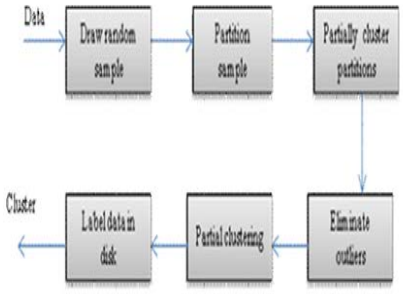

Figure 7.CURE Clustering Algorithm

4.4.4. Implementation of CURE Clustering Algorithm

CURE (no. of points k)
Input: A set of points S

Output: *k* clusters
1. r.mean and r.rep store the mean of the points in the cluster for every r, d a set of *c* representative points of the cluster (initially *c* = 1) since each cluster has one data point.
2. r.closest stores the cluster closest to r.
3. A k-d tree T is used to insert the input points.
4. Treat each input point as separate cluster.
5. Compute r.closest for each r and then insert each cluster into the heap Q.
6. Clusters are arranged in increasing order of Distances between r and r.closest.
7. While size(Q) > *k*
8. Remove the top element of Q (say r) and merge it with its closest cluster r.closest (say v)
9. Compute the new representative points for the merged cluster w.
10. Remove r and v from T and Q.
11. For all the clusters r in Q, update r.closest and displace r
12. introduce w into Q
13. Repeat [46].

CURE (Clustering usage Representatives) method find clusters from a large database that is more robust to outliers, and identifies clusters having non-spherical shapes and wide variances in size. CURE employs a combination of data collection, data reduction by using random sampling and partitioning. With the availability of large data sets in application areas like bioinformatics, medical informatics, scientific data analysis, financial analysis, telecommunications, retailing, and marketing, it is becoming increasingly important to execute data mining tasks in parallel [43].

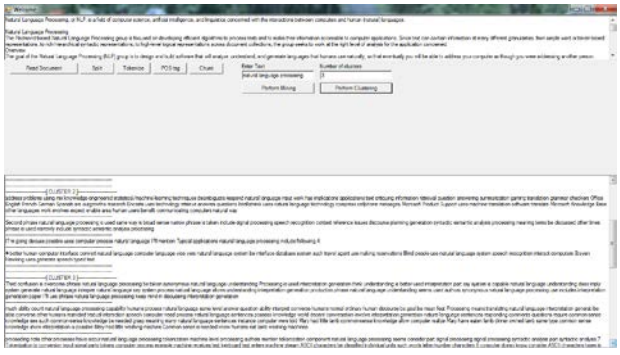

Figure 8. Output of CURE Clustering Algorithm

5. Conclusion

ISSN 2348 – 7968

The information era has brought us vast amounts of digitized text that are generated, propagated, exchanged, stored, and accessed through the internet each day across the world. The accumulation of this data is making information acquisition increasingly difficult, with language becoming a critical obstacle to growth .To overcome these difficulties, the Natural Language Processing (NLP) focuses on to transform unstructured text into structured document that can be searched & browsed in flexible ways. The goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. The concept aims to find frequent patterns, interesting correlations, and associations among sets of items in the transaction databases or other data repositories. Clustering, in data mining, is useful for discovering groups and identifying interesting distributions in the underlying data CURE (Clustering usage Representatives) method find clusters from a large database that is more robust to outliers, and identifies clusters having non-spherical shapes and wide variances in size. Thus it can be concluded as the time taken to form the clusters increases as the number of cluster increases.

Acknowledgments

My sincere thanks to all the people who have contributed in any way in carrying out this work.

References

[1] M.Rathamani, Dr. P.Sivaprakasam," Cloud Mining: Web usage mining and user behavior analysis using fuzzy C-means clustering" ,in IOSR Journal of Computer Engineering (IOSRJCE) ISSN: 2278-0661, ISBN: 2278-8727 Volume 7, Issue 2 (Nov-Dec. 2012), PP 09-15.// L.S.1

[2] B.Uma Maheshwari ,Dr. P. Sumathi,"A New Clustering and Preprocessing for Web Log Mining ",in 2014 World Congress on Computing and Communication Technologies.//L.S.2S

[3] Supinder Singh,"Web Log File Data Clustering Using – Means and Decision Tree",IJARCSSE All Rights Reserved Page|379 Volume3,Issue 8,August 2013, ISSN:2277 128.//3

[4] Bhupinder Singh,Usvir Kaur ,Dr.Dhreerendra Singh,"Web Usage Clustering Algorithms:A Review",International Journal of Latest Scientific Research and Technology 1(2),July -2014,pp 1-7,ISSN:2348-9464.// ls 4

[5] J Vellingiri, S.Chenthur Pandian," A Survey on Web Usage Mining" , in Global Journal of Computer Science and Technology, Volume 11- Issue 4 Version 1.0 March 2011,ISSN : 0975-4172,PP 67-70.

[6] Hemanshu Rana, Mayank Patel ," A Study of Web Log Analysis Using Clustering Techniques", in International Journal of Innovative Research in Computer and Communication Engineering ,Vol. 1, Issue 4, June 2013, ISSN:2320-9798, PP 925-929.

[7] V.Chitraa, Dr. Antony Selvdoss Davamani," A Survey on Preprocessing Methods for Web Usage Data", in International Journal of Computer Science and Information Security (IJCSIS), Vol. 7, No. 3, 2010,ISSN:1947-5500,PP 78-83.

[8] Tasawar Hussain, Dr. Sohail Asghar, Simon Fong," A Hierarchical Cluster Based Preprocessing Methodology for Web Usage Mining,PP 472-477.

[9] Vijayashri Losarwar, Dr. Madhuri Joshi," Data Preprocessing in Web Usage Mining", in International Conference on Artificial Intelligence and Embedded Systems (ICAIES'2012) July 15-16, 2012 Singapore,PP 1-5.

[10] Vishal Gupta, Gurpreet S.Lehal ,"A Survey of Text Mining Techniques and Applications", in Journal of Emerging Technologies in Wen Ingelligence,vol 1,Issue 1, August 2009,ISSN : 60-76.

[11] Ranan Collobert ,Janson Weston ,Leon Botton ,Michael Karlen,Koray Kavukcuoglu,Pavel Kuska ,"Natural Language Processing (Almost) from Scratch ",in Jornal of machine Learning Research 12(2011) 2493-2537 ,PP 2493 -2537.

[12] Richa Chaurasia ,Prof.Preeti Chaudhary,"A Survey On Web Log Pre-Processing and Evidence Preservation for Web Mining ", in International Journal of Innovative Research in Technology & Science(IJIRTS),ISSN :2321-1156,PP 47-50.

[13] B. Madasamy, Dr. J. Jebmalar Tamilselvi "General Web Knowledge Mining Framework", International Journal on Computer Science and Engineering (IJCSE) ISSN : 0975-3397, Vol. 4 No. 10 Oct 2012, PP 1744-1750.

[14] Dhanasekaran.K, Rajeswari.R," A Research-oriented Survey and Current Status on Feature Extraction, Ontology Construction towords Natural Language Processing",in International Journal of Computer Science Issues ,Vol 9 Issue 3 ,May 2012, ISSN : 1694-0814,PP 239 -248.

[15] Enik Cambria ,Bebo White , A Review Article on "Jumping NLP Curves :A Review of Natural Language Processing Research" in IEEE Computational Intelligence Magazine ,May 2014 ,PP 48-57.

[16] Delia Rusu ,Lorand Dali,Blaz Fortuna ,Marko Grobelnik ,Dunja Mladenic ,"Triplet Extraction From Senteces" .

[17] Christopher D. Manning ,Mihai Sarudeanu,John Bauer,Jenny Finkel,Steven J. Bethard,David McClosky,"The Stanford Core NLP Natural Language Processing Toolkit".

[18] Adam L. Berger ,Stephen A. Della Pietra,Vincent J. Della Pietra,"A Maximum Entropy Apprach to Natural Language Processing ".

[19] Ralph Weischedel ,Jaime Carbonell ,Barbara Grosz,Wendy Lehnert,Mitchell Marcus,Raymond Perrault,Robert Wilensky ," Natural Laguage Processing",PP 481-493.

[20] Ronan Collobert,Jason Wetson ,"A Unified Architecture for Natural Language Processing :Deep Neural Networks with Multitask Learning" .

[21] Enik Cambria, Bebo White, A Review Article on "Jumping NLP Curves: A Review of Natural Language Processing Research" in IEEE Computational Intelligence Magazine, May 2014, pp 48-57.

[22] Fabio Ciravegna, Sanda Harabagiu," Recent Advances in Natural Language Processing", 1094-7167/03/$17.00 © 2003 IEEE ,IEEE Intelligent Systems *Published by the IEEE Computer Society.*

[23] Mohamed Attia, Mohsen A.A.Roshwan & Mohamed A.S.A.A.A1-Badrashiny,"A Semi –Automatic Visual Interactive Tool For Morphological, POS-Tags, Phonetic and Semantic Annotation of Arabic Text Corpora", in IEEE Transaction on Audio, Speech &Language Processing , Vol.-17, No. 5, July -2009.

[24] Francisco Tacoa ,Hirashi Uchida,Mitsuru Ishizuka ,"A Word Sense Disambiguation Approach For Converting Natural Language Text into A Common Semantic Desciption ", in IEEE Fourth International Conference on Semantic Computing.

[25] Zhimao Lu Dong Mei, Fun Rubo Zhang,"A Word Sense Disambiguation Based on Vicarious Words", in Fourth Internatiional Conference on Natural Computation.

[26] Yahang Gao, Wanxiang Che, Ting Liu & Sheng Li,"Semi – supervised Domain Adaptation For WSD: using a Word by Word Selection Approach ",in International Conference on Congnitive Informatics (ICCI)10.

[27] Jerome R.Bellegardu ,"Part-of Speech Tagging By Latent Analogy", in IEEE Journal of Selected Topic Signal Processing Vol.-4,No.6,December 2014.

[28] John Bos, Malvina Nissim,"From Shallow to Deep Natural Language Processing A Hands on Tutorial ".

[29] Antonio LaTorre, Jose M.Peria, Victor Robles, Maria S.Perez, " A Survey in Web Page Clustering Technique ".

[30] Jayshree Jha, Leena Ragha," Educational Data Mining using Improved Apriori Algorithm**" ,**in International Journal of Information and Computation Technology. ISSN 0974-2239 Volume 3, Number 5 (2013), pp. 411-418.

[31] Jyoti Arora, Nidhi Bhalla, Sanjeev Rao," A Review On Association Rule Mining Algorithms",in International Journal of Innovative Research in Computer and Communication Engineering Vol. 1, Issue 5, July 2013,ISSN NO. ISSN (Print) : 2320 – 9798 ISSN (Online**):** 2320 – 9801.

[32] Pratibha Mandave , Megha Mane,Prof.Sharda Patil ,"Data mining using Association rule based on APRIORI algorithm and improved approach" in International Journal of Latest Trends in Engineering and Technology (IJLTET) , Vol. 3 Issue2 November 2013, ISSN: 2278-621X,pp 107-113.

[33] Jiao Yabing," Research of an Improved Apriori Algorithm in Data Mining Association Rules" , in International Journal of Computer and Communication Engineering, Vol. 2, No. 1, January 2013,pp 1-3.

[34] S.Veeramalai, N.Jaisankar, A.Kannan," Efficient Web Log Mining Using Enhanced Apriori Algorithm with Hash Tree and Fuzzy" , in International journal of computer science & information Technology (IJCSIT) Vol.2, No.4, August 2010 ,pp 1-15.

[35] S. M. Kamruzzaman , Farhana Haider , Ahmed Ryadh Hasan,"Text Classification Using Data Mining " in ICTM 2005 ,pp 1-19.

[36] Anupriya,Ashok Kumar," Analysis on Parallelization of Apriori Algorithm in Data Mining" in International Journal of Computer Applications (IJCA) (0975 – 8887),pp 1-3.

[37] Abhang Swati Ashok, Joresandeep S ," The Apriori Algorithm:Data Mining Appraches Is To Find Frequent Item Sets From A Transaction Dataset " ,In International Journal Of Innovative Research In Science, Engineering And Technology , Volume 3, Special Issue 4, April 2014, ISSN (Online) : 2319 – 8753 ,ISSN (Print) : 2347 – 6710,pp 1-5.

[38] Yogita Rani and Dr. Harish Rohil , "A Study of Hierarchical Clustering Algorithm",in .International Journal of Information and Computation Technology. ISSN 0974-2239 Volume 3, Number 11 (2013), pp. 1225-1232.

[39] Seema Maitrey , C.K. Jhaa,"An integrated approach for CURE clustering using map-reduce technique " pp 631-637.

[40] Yogita Rani, Manju & Harish Rohil, "Comparative Analysis of BIRCH and CURE Hierarchical Clustering Algorithm using WEKA 3.6.9",in The SIJ Transactions on Computer Science Engineering & its Applications (CSEA), Vol. 2, No. 1, January-February 2014 ISSN: 2321 – 2381 © 2014 | Published by The Standard International Journals (The SIJ) .

[41] Periklis Andritsos ,"Data Clustering Techniques".

[42] "An Introduction to Cluster Analysis for Data Mining".

[43] .Sudipto Guha,Rajeev Rastogi & Kyuseok Shim " CURE:An Efficient Clustering Algorithm for large Database".

[44] S.Ravthi,Dr.T.Nalini,"Performance Comparision of Various Clustering Algorithm ", in International Journal of Advanced Research in Computer Science & Software Engineering , ISSN : 2277-128X ,Volume 3,Issue 2,February 2013 ,pp 67-72.

[45] George Karypsis ,Eui-Hong (Sam) Han,Vipin Kumar," Chemeleon :Hierarchial Clustering Using Dynamic Modelling",ISSN No. 0018-9162/99/$10.00 © 1999 IEEE,pp 68-75.

[46] Seema Maitrey, C. K. Jha,Rajat Gupta ,Jaiveer Singh ,"Enhancement of CURE Clustering Technique in Data Mining", presented in National Conference on Development of Reliable Information Systems, Techniques and Related Issues (DRISTI) 2012 Proceedings published in International Journal of Computer Applications® (IJCA).

[47] Joakim NIVRE , "On Statistical Methods in Natural Language Processing".

[48] "Natural Language Processing".

[49] Gurneet Kaur," Improving The Efficiency Of Apriori Algorithm In Data Mining ",In International Journal Of Science, Engineering And Technology, ISSN: 2348-4098

ISSN 2348 – 7968

Volume 02 Issue 05 June- 2014,pp 1-12.

Ms. Aashwini Thakare has completed her B.E. in Information Technology from RTMNU,Nagpur. She is pursuing M.Tech in Computer Science & Engineering from same University. Her Research interest including Data Mining.

Prof. M.S.Chaudhari is Head of Department in Computer Science & Engineering Department in Priyadarshni Bhagwati College of Engineering, Nagpur.