DOI 10.29042/2018-3731-3735

Helix Vol. 8(5): 3731- 3735

Efficient Techniques to Improve Clustering Accuracy on Web by Using Hybrid Approach

[1]Monika Sharma, [2]M A Rizvi
[1]M.Tech, Department of CEA, National Institute of Technical Teacher's Training and Research, Bhopal, M.P.
[2]Professor, Department of CEA, National Institute of Technical Teacher's Training and Research, Bhopal, M.P.
Email: [1]moni.sharma357@gmail.com , [2]marizvi@nitttrbpl.ac.in

Received: 09th July 2018, Accepted: 14th August 2018, Published: 31st August 2018

Abstract
Clustering in web mining has become a challenging task in the present scenario, which draws the attention of many researchers. Cluster analysis is mainly used to determine the clusters and algorithms for identifying the best cluster. Cluster analysis includes a number of unsupervised order techniques that are planned to discover a framework in datasets to form clusters where comparable information objects are gathered together into a well-organized groups. World Wide Web is a rich information source, along with a large number of the web users extracting information using web are increasing. The weblog records are increasing rapidly in a disorganized manner, it contains a lot of information and the major problems of internet users are to extract meaningful information and to discard irrelevant data. In this paper, various clustering methods are examined to recognize its significance with regards to the large dataset and observationally these have been tried on Wine Quality dataset taken from UCI repository to feature their qualities. In this paper, an attempt is made to develop a new hybrid technique is introduced to provide better clustering accuracy.

Keywords: Clustering, Web Mining, Hybrid Hierarchical Clustering.

Introduction

Recently, clustering in web mining gained a lot of attention and becomes a most popular technique for performing its main task to extract the interesting patterns from the internet. A cluster is characterize as a group of items that have a more advanced level of similarity one and all are correlated with objects that are not in the similar set. Still, there is an uncertainty respecting reasonable closeness metric for clustering. Various measures have been recommended to evaluating the closeness, i.e. Euclidean distance, density in data space and so on. The internet presents a lot of learning and data for users to access various types of information as per the user's choice.

But, the web is a collection of relevant as well as irrelevant information. Retrieving the significant data from the web turns out to be more complex to the users. To get rid of this problem, clustering technique is introduced.

In web mining, clustering provide an improved image of the information, since all items inside the cluster have less uncertainty in their attributes and they can be outline efficiently. Clustering has discovered operations in different areas like evaluating the missing qualities in information or recognizing exceptions in information. Various algorithms or calculations are exist for arranging information into clusters. Still, there is no universal answer for all the issues. The structure of web mining is shown in figure 1.

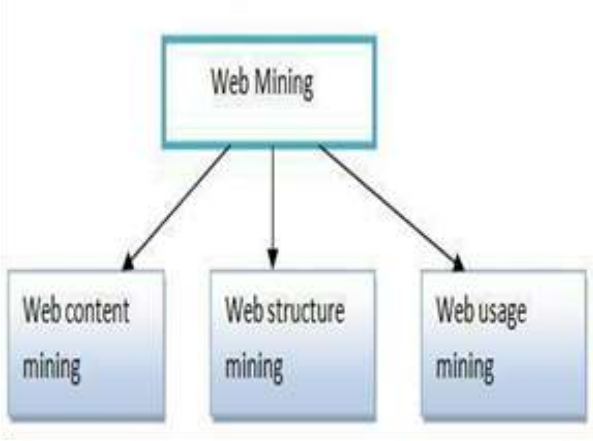

Figure 1: Taxonomy of Web Mining

There is no all-around target criteria for accuracy or clustering validity and every calculations or algorithms has its own disadvantages and accomplishments in taking care of the testing issue of unsupervised clustering. In this paper, Section 1 represents the procedure of web mining, section 2 presents the fundamental background on our concepts for researching the cluster structure, Section 3 describes the proposed methodology and Section 4 explains the experimental result and analysis for how to apply the algorithms to the whole dataset and identify the accuracy of clusters and at last Section 4 shows the final conclusion of the paper.

Procedure of Web Mining:

This process is usually identify three tasks in a particular manner, data pre-processing, pattern

discovery, and pattern analysis as demonstrates in the figure.

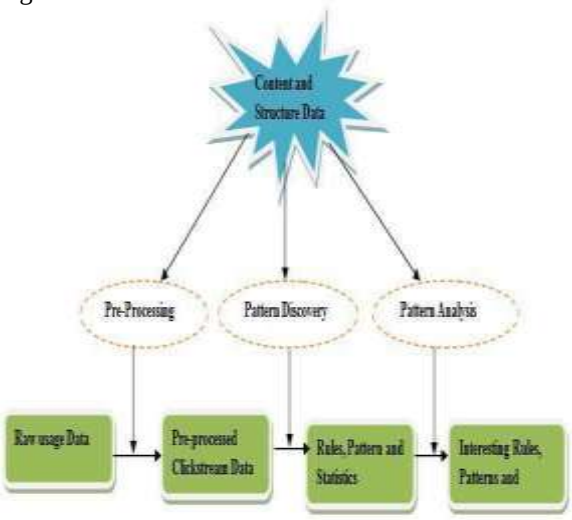

Figure 2: Web Mining Procedure

Web usage mining determines the recurrence of page access by the users and after that, it finds the common traversal ways of the users. The primary task is gathering the information from web server log records.

Data Gathering and Pre-Processing

Web information can be gathered and utilized as a part of the web personalization. Data gathering or data collection is a process of accumulation of web log records from the server by a procedure of confirmation is acknowledged as information gathering. And during the process of data pre-processing relevant attributes are maintained to decrease the content of weblog records [2]. Preprocessing provides accurate, concise data for data mining. The unnecessary data can be reduced by data pre-processing process. [1] After the data cleaning process, the particular number of users and sessions distinguished. And at the final phase of data preprocessing we get the successive user access pattern from the web server get to log file information.

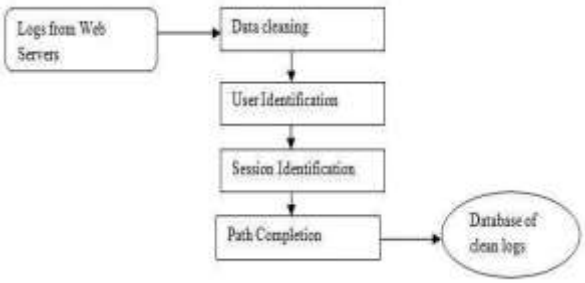

Figure 3: Process of Data Pre-processing

Pattern discovery

To identify a suitable pattern by the preprocessing data, some methods are tested such as path analysis, Association rules, Sequential pattern, Clustering, classification etc. The fundamental strategies are classification analysis, association rule discovery, dependency modeling, and clustering analysis.

Pattern Analysis

The investigation of the pre-processed information is actually valuable to all the organizations performing various businesses over the web. The administrators of websites are interested in frequent usage pattern of websites [1]. In this paper, our main purpose is to provide better clustering accuracy as compared to existing systems.

Related Study

Frank Klawonn proposed a method known as dynamic data assigning assessment clustering that was expected both to characteristic cluster structure in as well as to discover the clusters. This new approach can recognize clusters.[15]

Pranav Nerurkar describes the different clustering methods and these have been tested on artificial benchmarks to discuss their strengths and weaknesses [16].

D.S. Anupama et al. present a method which tested on two datasets specifically www.enggresources.com and NASA. They present algorithm of hierarchal agglomerative which reduces the inputs of prediction through determining the session representatives and it deal with the nearness of a grouping in the user navigation pattern that reduce limitations of partition based algorithms [2].

V.Sujatha et.al. proposed the method of prediction of user navigation patterns by using clustering and classification (PUCC) by the web log data. The results can present the better quality of clustering for navigation pattern in web usage mining [7].

Archana Patel et.al. present a novel approach for identifying user navigation patterns for predicting user's request based on clustering users searching behavior and knowledge[8].

Leet.Ramesh et.al. present the concept of web mining where MDA is defined for describing the nature of agents in the proposed system[14].

B. Nigam presents different models and their variations for predicting the next one web page accessed by the web user. They use Markov model for identify variations for web prediction [10].

Meera Narvekar et al. proposed a hybrid model which combines the benefits of the models such as Markov model and Hidden Markov Model. This paper shows various prediction challenges i.e. long preparing time, more prediction time, less prediction accuracy etc. They present a new way by which system focus to decrease the complexity of prediction and increase the accuracy of prediction [6].

Proposed Method:

Web log information contains numerous unessential information. Three types of unnecessary information are images requests, wrong demands, and bug navigation requests. By using a logical data cleaning process unnecessary information are eliminated from the web log files. In the proposed work we use a dataset of Wine Quality, taken from the UCI repository. We have use Agglomerative hierarchical clustering algorithm and hybrid hierarchical clustering algorithms to compare the difference of results between them on the basis of correlation coefficient, i.e agglomerative coefficient and Cophenetic coefficient.

Agglomerative hierarchical clustering algorithm is performing as a bottom-up technique. That is, every item is at first considered as a single cluster. At each progression of the calculation, the two clusters which more comparative are consolidated into another bigger cluster (hubs). This technique is focuses until all the point are individual from only one single huge group i.e. root.

Divisive hierarchical clustering algorithm performs as a top-down approach. It starts by including the root, where entire items are combined in a single cluster. At every point of progression, the better composite cluster is split into two. This procedure is frequently works until the point when all items are in their own groups. Agglomerative Hierarchical Clustering is better in determining the small clusters, While Divisive Hierarchical Clustering is better in determining the large cluster.

Hybrid Hierarchical Clustering Algorithm (HHCA) combines the best characteristics of both hierarchical techniques i.e. Agglomerative and Divisive hierarchical clustering. It acquires the benefits of hierarchical clustering. Currently, a large number of researchers has concentrated its investigation on implementing data mining methods to web logs for consequently produce a prediction for user navigation pattern. [3]

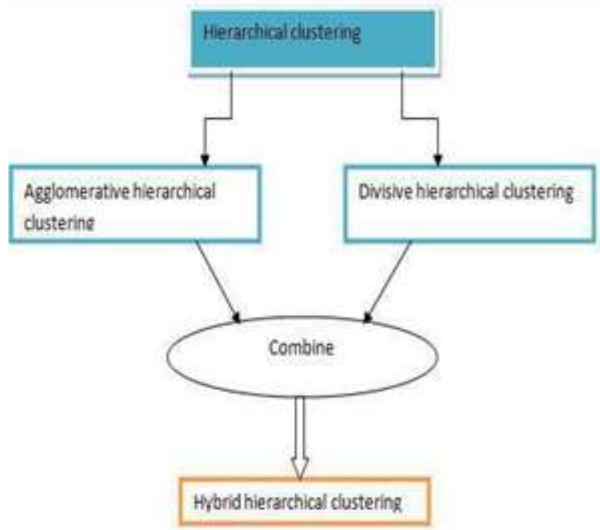

Figure 4: Basic Structure of Hybrid Hierarchical Clustering

Experimental Result and Analysis:

The proposed method is implemented by using the R software using the dataset of Wine Quality. The dataset of wine quality are taken from UCI repository, it is a multivariate dataset which contain 4898 number of instances and 12 input variables. The dataset first has to be clean and pre-processed to removing the unessential information from the dataset. We are using agglomerative hierarchical clustering and hybrid hierarchical clustering algorithms to calculate the accuracy measure by using coefficient of algorithms i.e. agnes and cophenetic coefficient. The following figure shows the pre-processed dataset.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
fixed.acidity	0.4904126	-0.0340705	0.2348647	-0.2067868	-0.0737283	0.0213518	-0.1902478	-0.1679815	0.2392506
volatile.acidity	-0.2500425	-0.4175907	0.0379531	0.0600582	-0.3896780	0.2059167	-0.5032296	-0.2361562	0.301986
citric.acid	0.4612757	0.1880421	0.0232969	-0.0177646	0.1530568	0.1677997	0.3236680	-0.0686908	0.380574
residual.sugar	0.1770457	-0.1075760	0.4414404	0.6912022	-0.2522720	0.1403373	0.1890659	0.1338685	-0.407711
chlorides	0.1980726	-0.1840579	-0.5868053	0.1043259	-0.2792350	0.3284687	0.2045207	0.4260818	0.196171
total.sulfur.dioxide	0.0290360	-0.2421770	-0.1541016	0.5139024	0.7896102	0.1742588	-0.2226109	-0.0114009	0.183965
density	0.3816467	-0.3470780	0.2353204	0.0996077	-0.1295264	-0.4326026	0.0629676	0.0188373	0.264585
ph	-0.4338348	0.0431409	0.0473490	0.2659020	-0.0708446	-0.4529256	0.3017625	0.0407125	0.508141
sulphates	0.2470789	0.1577817	-0.5397019	0.3079646	-0.1547071	-0.4191368	-0.1069901	-0.4934258	-0.232467
alcohol	-0.0640454	0.5289369	0.1239474	0.2512682	-0.2062121	0.4007244	0.0340426	-0.3653451	0.259865
quality	0.0958453	0.5620588	0.0384785	0.1655456	-0.0723728	-0.2077524	-0.5585160	0.5685601	0.136147

Figure: 5.1: Pre-Processed Wine Quality Dataset

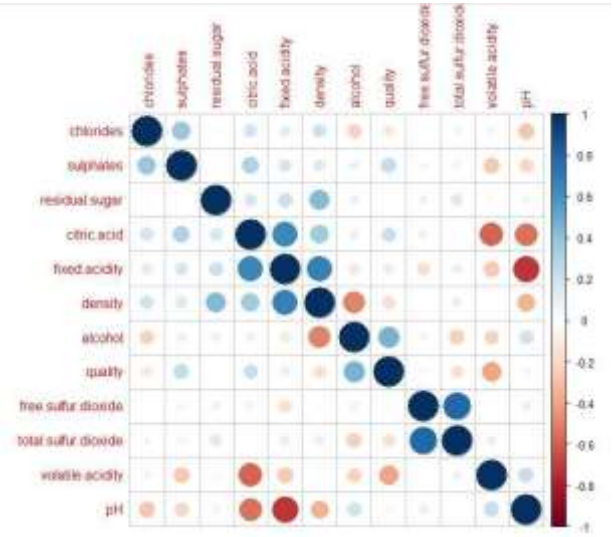

Figure 5.2: Plot of Pre-Processed Dataset

By applying the agglomerative hierarchical clustering algorithm, divisive hierarchical clustering algorithm and hybrid hierarchical clustering algorithm on the required pre-processed datasets, we calculate the correlation coefficient i.e. Agglomerative coefficient (agnes) and Cophenetic coefficient on the pre-processed dataset of wine quality. Where agglomerative coefficient determines the clustering structure of the dataset and Cophenetic Coefficient determines the cophenetic distances for the hierarchical clustering. The results of the coefficient on the pre-processed dataset is shown in following table:

Methods	Agglomerative Hierarchical Clustering (agnes)	Hybrid Hierarchical Clustering (Cophenetic Coefficient)
Single	0.1748079	0.7784353
Complete	0.2004841	0.5688289
Average	0.1604813	0.7947383
Ward	0.1749945	0.6999272

Table 1: Comparison of Clustering Algorithms

The following figure shows the cluster plot of the given dataset

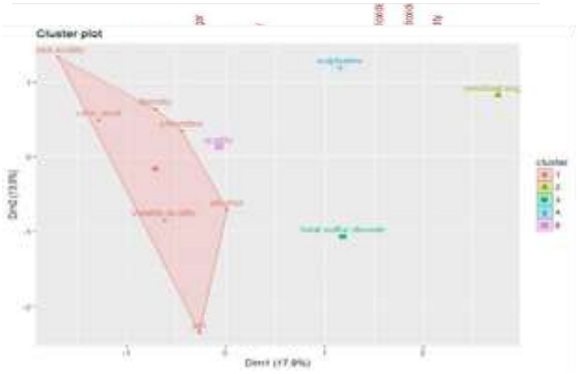

Figure.6: Cluster Plot

As result shows that the hybrid hierarchical clustering algorithm performs better on the dataset as compared to agglomerative hierarchical clustering algorithms. It forms the better clusters as compared to other algorithms. This algorithm is more efficient for improving the clusters accuracy.

The following table and correlation plot shows the similarity and difference between different clustering methods.

	ward.D	single	complete	average
ward.D	1.0000000	0.6755518	0.3117991	0.7556994
single	0.6755518	1.0000000	0.4811656	0.9794864
complete	0.3117991	0.4811656	1.0000000	0.5401331
average	0.7556994	0.9794864	0.5401331	1.0000000

Table 2: Similarity/Difference Between Different Clustering Methods

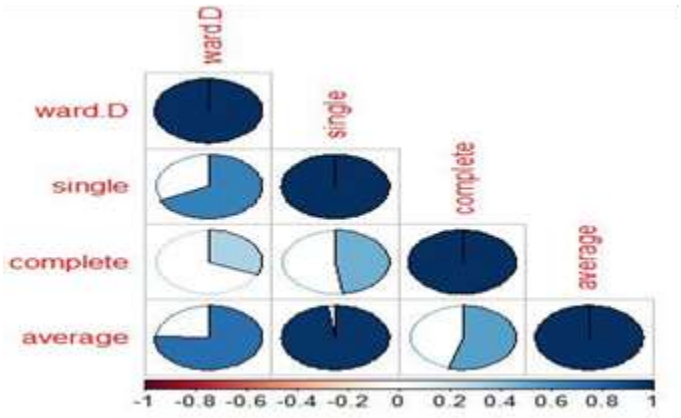

Figure 7: Correlation Plot of Comparison and Similarity between Different Clustering Methods

Conclusion

In this research paper, we target to produce a good quality of clusters by evaluating the comparison between various algorithms. As result shows that hybrid hierarchical clustering algorithm performs better as compared to other algorithm i.e. agglomerative hierarchical clustering. Results prove that accuracy is approximately 79% achieved by average method in hybrid hierarchical clustering algorithm. It is more efficient for clustering the web pages. The proposed techniques can also be evaluated on different datasets.

References

[1] Priyanka S. Panchal, Prof. Urmi D. Agravat " Hybrid technique for user's web page access prediction based on markov model", IEEE – 31661, IEEE Xplore, 2013.

[2] Anupama D.S, Sahana D. Gowda " Clustering of web user sessions to maintain occurrence of sequence in navigation pattern" Elsevier, Procedia Computer Science 58 (2015) 558 – 564.

[3] Yue Xu " Hybrid clustering with application to web mining ", IEEE Xplore , 2005.

[4] Data mining algorithms in R, "https ://en.wikibooks.org/wiki/Data_Mining_Algorit hms_In_R/Clustering" .

[5] Vinh-Trung Luu, Germain Forestier, Mathis Ripken,Frederic Fondement, Pierre-Alain Muller " Web usage prediction and recommendation using web session clustering".

[6] Meera Narvekar, Shaikh Sakina Banu, "Predicting user's web navigation behavior using hybrid approach" , Elsevier, procedia computer science 45(2015) p.no.- 3-12.

[7] V. Sujatha, Punithavalli, "Improved user navigation pattern prediction technique from web log data", Elsevier, procedia engineering 30 (2012) p.no. 92-99.

[8] Ms. Archana Patel, Ms. Prachi Joshi, "A survey on new techniques in web path recommendation systems in data mining", International Journal For Technological Research In Engineering, Volume 1, Issue 4, December – 2013, ISSN (Online): 2347 – 4718.

[9]Prajyoti Lopes, Bidisha Roy, "Dynamic recommendation system using web usage mining for e-commerce users", Elsevier, procedia computer science 45(2015) p.no.- 60-69.

[10] B. Nigam, S. Tokekar, and S. Jain, "Evaluation of models for predicting user's next request in web usage mining," international Journal on Cybernetics & informatics (UCI), vol. 4, pp. 1–13.

[11] Hugh Chipman, Robert Tibshirani, "Hybrid hierarchical clustering with applications to microarray data", Biostatistics (2006), 7, 2, pp. 286– 301.

[12] Priyanka Makkar, Payal Gulati, Dr. A.K. Sharma, "A Novel Approach for Predicting User Behavior for Improving Web Performance ", International Journal of Computer Science and Engineering,Vol. 02, No. 04, 2010, 1233-1236.

[13] B. Mobasher, H. Dai, T. Luo and M. Nakagawa, "Effective personalization based on association rule discovery from Web usage data," in Proc. ACM Workshop WIDM, Atlanta, GA, Nov. 2001.

[14] Leet. Ramesh A. Medar, A.H. Kulkarni, "Design of an information intelligent system based on web data mining", International journal of computer science engineering and information technology research, Vol.1, Issue 2. Dec 2011 , 9-21.

[15] Frank Klawonn "Exploring data sets for clusters and validating single clusters" Elsevier Procedia Computer Science 96 (2016) 1381 – 1390.

[16] Pranan Nerurkar, Archana Shirke, Madhav Chandane, Sunil Bhirud "Empirical Analysis of Data Clustering Algorithms" Elsevier Procedia Computer Science 125(2018) 770-779.

