International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 5 Issue: 6 1125 – 1130

Fuzzy Clustering in Web Mining

Dr. A. B. Raut
Professor & Head,
Department of CSE,
H.V.P.M's, COET, Amravati,
Maharashtra,India.
anjali_dahake@rediffmail.com

Abstract- Web mining is the use of data mining techniques to automatically discover and extract information from web. Clustering is one of the possible techniques to improve the efficiency in information finding process. Conventional clustering classifies the given data objects into exclusive clusters. However such a partition is insufficient to represent many real situations. Hence a fuzzy clustering method is offered to construct clusters with uncertain boundaries and allows the object to belong to multiple clusters with degree of membership. Web data has fuzzy characteristics, so fuzzy clustering is better suitable for web mining in comparison with conventional clustering. In this paper, we have proposed two algorithms that are Fuzzy c-Means (FCM) and Clustering based on Fuzzy Equivalence Relations which can be used for web page mining and web usage mining. The results obtained from the proposed algorithm are more convincing. The experimental results are carried out on different algorithmic parameters on real data. The analysis is being done by comparing the proposed algorithm with conventional clustering algorithms.

*****

I. INTRODUCTION

The web is the largest information repository in the history of mankind. Today the size of data on the web is growing exponentially. It is very difficult for information user to find relevant information that one is looking for. "We are drowning in information but starving for Knowledge" said by John Nesbit long ago, is true even for today in the epoch of World Wide Web (WWW). They can encounter the Low Precision and Low Recall problem when interacting with the web. Antonino Gull`ı, identified the difficulties which modern web Information Retrieval (IR) tools need to handle that are discovering relevant documents, to handle huge quantity of information, to address subjective and time-varying search needs, finding fresh information and dealing with poor quality queries [1]. All these issues impose the creation of smart tools for "automatic knowledge extraction". Next generation of web IR tools should take advantage of new paradigm called as Clustering. Clustering is the process of forming groups (clusters) of similar objects from a given set of inputs. Modern web IR tools use clustering for web page content mining, search result mining and usage mining for classification of web documents and web users.

Author has proposed two algorithms for Fuzzy Clustering in Web Mining and the following diagram explains the different phases of proposed work.

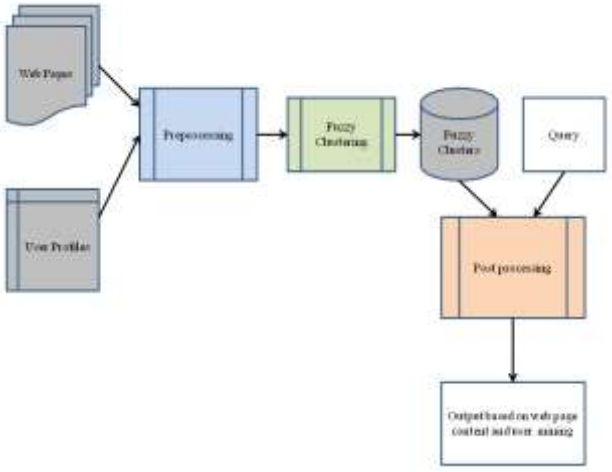

Fig. 1. Different phases of proposed work.

Web data is different from traditional data. Thus additional preprocessing is needed before Web Mining performs. Content Extraction, Automatic Extraction of Significant Terms, Creation of Term Document Matrix,Capturing Users' Access Behaviour Steps are performed in the preprocessing phase. This phase converts the web documents into the term document matrix or relation matrix which is then used by the proposed fuzzy clustering algorithms developed using Matlab. Fuzzy clusters were made by using two proposed algorithms. In postprocessing, the results of clustering have been saved in database which are then used for information retrieval.

 ISSN: 2321-8169
1125 – 1130

II. LITERATURE REVIEW

Web Mining is defined as the use of Data Mining techniques to automatically discover and extract information from web documents and services [2]. Initially two different approaches were taken for defining Web Mining. First was a "process-centric view", which defined Web Mining as a sequence of different processes as Resource Discovery, Information Extraction, Generalization and Analysis.

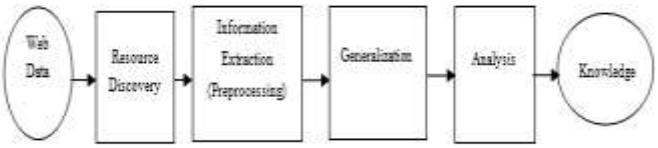

Fig. 2. Web mining Processes.

- Resource Discovery: It is the task of retrieving documents and services on the web.
- Information Extraction: It is the process of automatically extracting and preprocessing specific information from web resources.
- Generalization: It is a process of automatically discovering general patterns at individual web site and across multiple sites.
- Analysis: It is the process of validation and/or interpretation of the mined patterns.

Whereas, second was a "data-centric view", which defined Web Mining in terms of the type of data that was being used in the mining process[3].

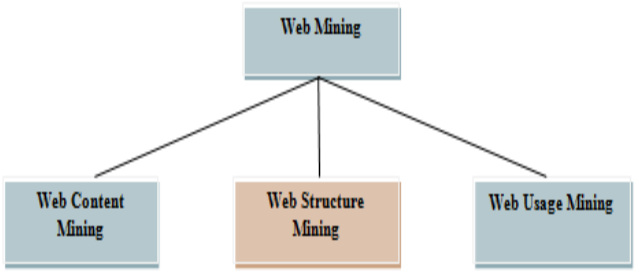

Fig.3. Web Mining Taxonomy.

Web Content Mining (WCM) describes the discovery of useful information from the web contents/data/documents/services [4][5]. Web Structure Mining (WSM) is defined as the process of discovering structure information from the web. Web Usage Mining (WUM) is the application of data mining techniques to discover interesting usage patterns from web usage data **[**5]**.** Web content mining and structure mining utilize the real or primary data on the web; usage mining mines secondary data generated by the users' interaction with the web.

A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. (i.e. there is high intra-cluster similarity and low inter-cluster similarity) **[**6]. The major clustering methods are Partitioning Methods, Hierarchical Methods, Density-based Methods, Grid-based Methods and Model-based Methods. Aside from the above categories of clustering method, there are two classes of clustering tasks that require special attention. One is clustering high dimensional data, and other is constraint based clustering. In crisp clustering, data is divided into distinct clusters, where each data element belongs to exactly one cluster. In fuzzy clustering (also referred to as soft clustering), data elements can belong to more than one cluster, and associated with each element is a set of membership degree. These indicate the strength of the association between that data element and a particular cluster. Joshi and colleagues used fuzzy techniques for web page clustering and usage mining. [7] and [8] used the mined knowledge to create adaptive web sites They argue that given the inherent ambiguity and complexity of the underlying data, clustering results should not be clearly demarcated sets but rather fuzzy sets - that is, overlapping clusters. For instance, a user can belong to multiple user interest groups because at different times he or she accesses the web for different information or merchandise. Insisting that each user fit only a single group is clearly inconsistent with this reality.

As web data has fuzzy characteristics, conventional clustering is not suitable for mining the web data. Hence a fuzzy clustering method is offered to construct clusters with uncertain boundaries and allows the object to belong to multiple clusters with degree of membership. Considering the immense potential of application of soft computing to Web Mining, new methodologies are being developed.

III. PROPOSED ALGORITHMS

Author has proposed two algorithms for Fuzzy Clustering in Web Mining. Following algorithm used Fuzzy c-Means algorithm suggested by Bezdek [23].

 ISSN: 2321-8169
1125 – 1130

Algorithm 1: Fuzzy c-Means for Web Mining
Input : Set of HTML pages
Output: Fuzzy Clusters
1: D ← Input Documents (web pages)
/* **Content Extraction** */
2: for all $d \in D$ do
3: Identify and extract the main content of block of document
4: end for
/* **Feature Extraction and Document Vector Preparation** */
5: for all $d \in D$ do
6: Extract the features
7: end for
8: F ← Dominant Feature set /* Discover the dominant feature set */
9: M ← term-document matrix /* prepare term-document matrix of documents and terms as dominant features using term weighting scheme */
/* **Clustering with FCM** */
10: U ← FCM(M) /* Find fuzzy partions by applying Fuzzy c-Means algorithm
/* **Cluster Profile Discovery** */
11: Crisp assignment of documents to cluster
12: For each cluster
12: Find out most frequent terms in the clusters
13: end for
14: Assign each cluster with the significant terms

stering Algorithm based on Fuzzy Equivalence Relations for Web Mining
Input: Page visiting table for usage mining
Term Document matrix for page mining
Output: Fuzzy Equivalence relation or Similarity Relation
Step 1: **Determine fuzzy compatibility relation.**
A fuzzy compatibility relation R (reflexive and symmetric) on X is determined in terms of an appropriate distance function of the Minkowski class.

$$R(x_i, x_k) = 1 - \delta\left(\sum_{j=1}^{p} |x_{ij} - x_{kj}|^q\right)^{\frac{1}{q}} \qquad (2.1)$$

For all pairs $\langle x_i, x_k \rangle \in X$
Where $q \in \mathbb{R}^+$,
and δ is a constant that ensures that $R(x_i, x_k) \in [0,1]$,
δ is inverse value of largest distance in X.
Step 2: Determine fuzzy equivalence relation
Relation R obtained by equation (2.1) is a fuzzy compatibility relation, but not necessarily a fuzzy equivalence relation. Hence determine transitive closure R_T of R .

Step 3: Find Fuzzy partitions
Find fuzzy partitions of relation of its α cuts. Use fuzzy partitions to find out the similarity between users or pages.

IV. EXPERIMENTAL RESULTS

The proposed algorithm is applied on the data set which is a collection of about 1500 web pages obtained from the IEEE Xplore. The abstracts correspond to topics Artificial Intelligence, Association Analysis, Classification, Clustering, Data Mining, Data Preprocessing, Data Structures, Neural Networks, Web Mining, Cloud Computing, Image Processing etc. Clusters quality is

ISSN: 2321-8169
1125 – 1130

measured by using Silhouette Index and Cophenet Coefficient metrics. The Following figure shows the silhouette index for FCM using Euclidean, Cityblock and Cosine distance for weighting exponent i.e. m=2.

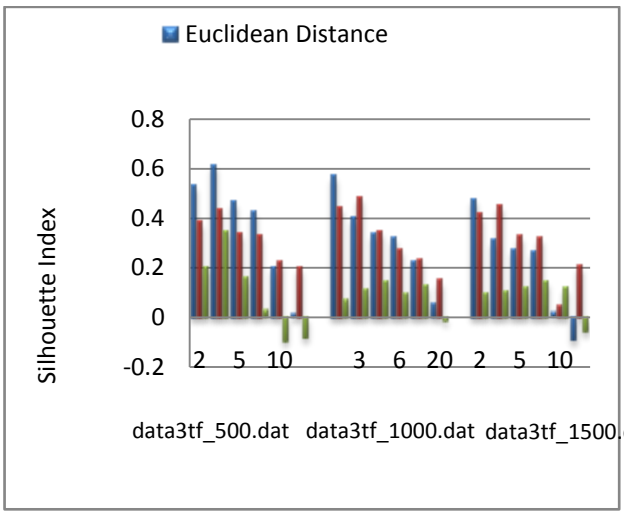

Fig. 4. Silhouette Index for Fuzzy c-Means

Following figures gives the degree of membership plot produces by Fuzzy c-Means for 5 clusters of dataset containing 1000 web documents.

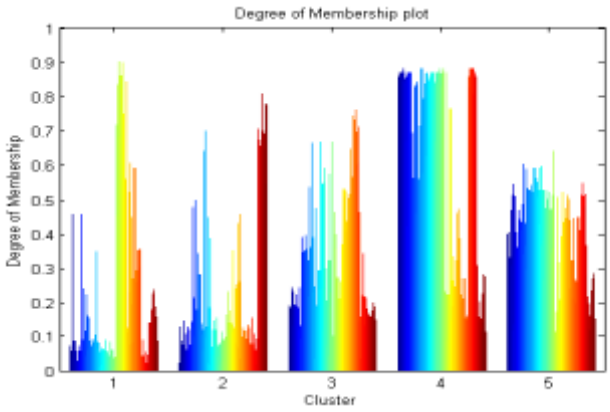

Fig.5. Degree of membership plot for Fuzzy c-Means for 5 clusters

Following figures describes the silhouette index and running time for Hard and Fuzzy c-Means Algorithm

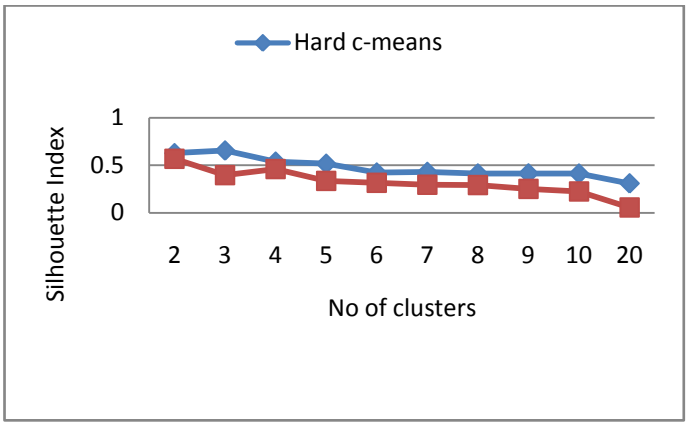

Fig. 6. Silhouette Index for Hard and Fuzzy c-Means Algorithm

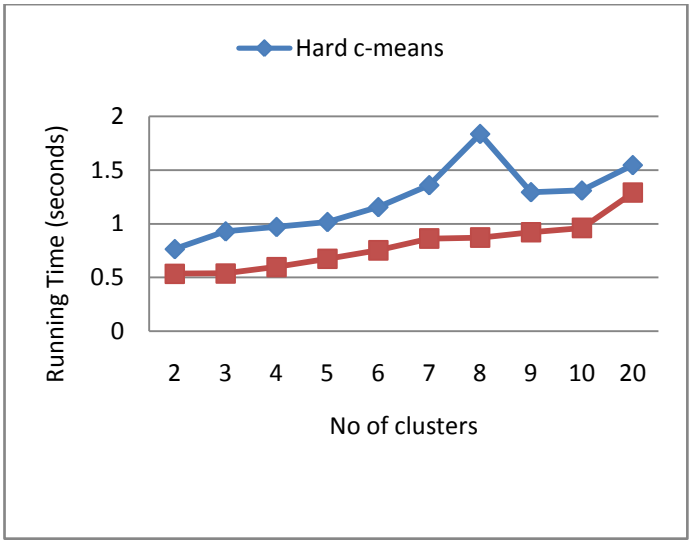

Fig. 7. Graph showing No. of Clusters vs. Running Time for Hard and Fuzzy C-Means Algorithm

The following figure shows the analysis of cophenet coefficient and running time of hierarchical clustering and clustering based on fuzzy equivalence relation using Euclidean and City block distance measure.

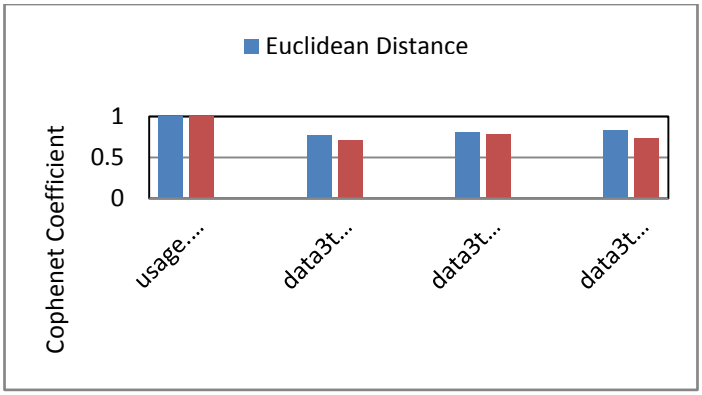

Fig.7. Cophenet Coefficient for clustering based on Fuzzy equivalence relation

V. CONCLUSION

Experimental results demonstrate that proposed algorithms produce fuzzy clusters of web data. The proposed algorithms are implemented for web page and web user mining successfully. During the preprocessing, it has been observed that the term list is extremely narrowed i.e. from 2844 words to 264 words after stemming for 1300 web pages. It is observed that Euclidean distance gives higher performance than Cityblock and Cosine distance measures when silhouette index is evaluated for Fuzzy c-Means. The running time required for Fuzzy c-Means algorithm is low as compared with hard c-Means algorithm. While comparing cophenet coefficient for Fuzzy equivalence relation algorithm Euclidean distance gives higher performance than Cityblock measure. Comparing proposed approach with other crisp and fuzzy approaches shows that fuzzy c-Means algorithm outperforms for Web Mining.

Web mining can be viewed as the extraction of structure from an unlabeled, semi-structured data set containing the characteristics of web data which has fuzzy characteristics i.e. data with uncertain boundaries. So fuzzy clustering is a better solution to classify web data properly so that it can be used to mine the web data effectively. Finally it can be concluded that fuzzy clustering methods offer rewards in the area of Web Mining.

REFERENCES

[1]. [Antonino Gull`I, 2006] Antonino Gull`I,, "On Two Web IR Boosting Tools: Clustering and Ranking", Doctoral Dissertation Report, University of Pisa, 2006.

[2]. [Etzioni, 1996] O. Etzioni, "The World Wide Web: Quagmire or Gold Mine?", *Communications of ACM*, pp. 1-6, 1996.

[3]. [Cooley et al.,1997] R. Cooley, B. Mobasher, and J. Srivastava, "Web Mining: Information and Pattern Discovery on the World Wide Web", *in Proceeding of 9th IEEE International Conference on Tools with Artificial Intelligence*, pp. 558–567, 1997.

[4]. [Kosala & Blockeel, 2000] R. Kosala and H. Blockeel, "Web Mining Research: A Survey", *Journal of ACM SIGKDD*, Vol. 2, pp. 1-15, 2000.

[5]. [Srivastava et al., 2000] J. Srivastava, R. Cooley, M. Deshpande and P. Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", *ACM SIGKDD Explorations*, Vol. 1, No. 2, pp. 12–23, 2000.

[6]. [Zaïane, 1999] O. R. Zaïane, "Principles of Knowledge Discovery in Databases ,1999- Chapter 8: Data Clustering" http://www.cs.ualberta.ca/~zaiane/courses/cmput690/slides/Chapter8/index.html

[7]. [Joshi and Krishnapuram, 1998]A. Joshi and R. Krishnapuram, "Robust Fuzzy Clustering Methods To Support Web Mining", *in Proceedings of Workshop in Data Mining And Knoweldge Discovery SIGMOD*, pp. 15-1-15-8, 1998.

[8]. [Kamdar, 2001] T. Kamdar, "*Creating Adaptive Web Servers Using Incremental Weblog Mining", Masters thesis, Computer Science Dept., University of Maryland, Baltimore*, 2001.

[9]. [Ansari et al., 2011] Z. Ansari, A. Babuy, W. Ahmed and M. Azeemz , "A Fuzzy Set Theoretic Approach to Discover User Sessions From Web Navigational Data" , *in Proceedings of IEEE Conference on Recent Advances in Intelligent Computational Systems (RAICS)*, pp. 879-884, 2011.

[10]. [Chen et al., 2009] H. Chen , X. Li , M. Chau , Y. Ho and C. Tseng , "Using Open Web APIs in Teaching Web Mining", *IEEE Transactions on Education*, Vol.52, No.4, pp.482-490, 2009.

[11]. [Darouich et al., 2013] A. Darouich, F. Khoukhi and K. Douzi , "Mining Fuzzy Motivation Indicator in Learning Environment through Human Computer Interaction", *in Proceedings of Science and Information Conference (SAI)*,pp. 712-720, 2013.

[12]. [Dey et al., 2010] L. Dey, S. Haque and N. Raj , "Mining Customer Feedbacks for Actionable Intelligence", *in Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, Vol. 3, pp. 239-242, 2010.

[13]. [Gholamzadeh & Taghiyareh, 2010] N. Gholamzadeh and F. Taghiyareh , "Ontology-based Fuzzy Web Services Clustering" , *in Proceedings of 5th* International Symposium on Telecommunications (IST), pp. 721-725, 2010.

[14]. [Joshi et al., 2010] M.Joshi, P. Lingras, Y. Yiyu and C. Virendrakumar , "Rough, Fuzzy, Interval Clustering for Web Usage Mining", *in Proceedings of 10th International Conference on Intelligent Systems Design and Applications (ISDA)*, pp. 397-402, 2010.

[15]. [Nadi et al., 2010] S. Nadi, M. Saraee and M. Davarpanah-Jazi , "A Fuzzy Recommender System for Dynamic Prediction of User's Behavior", *in Proceedings of International Conference on Internet Technology and Secured Transactions (ICITST)*, pp. 1-5, 2010.

[16]. [Ramudu & Murty, 2012] B. Ramudu and M. Murty , "Topic Based Semantic Clustering using Wikipedia Knowledge", *in Proceedings of International Conference on Data Science & Engineering (ICDSE)*, pp. 1-7, 2012.

[17]. [Wu, 2009] J. Wu ," Web News Summarization via Soft Clustering Algorithm", *in Proceedings of 6th International Conference on Fuzzy Systems and Knowledge Discovery*, Vol. 7, pp. 18-21, 2009.

[18]. [Yang & Li, 2009] Ming Yang and Hong Li , "User Analysis Based on Fuzzy Clustering", *in Proceedings of International Conference on Business Intelligence and Financial Engineering*, pp. 194-196, 2009.

[19]. [Yaxiu & Xin-Wei, 2009] Y. Yaxiu and W. Xin-Wei , "Web Usage Mining Based on Fuzzy Clustering" , *in Proceedings of . International Forum on Information*

Technology and Applications, IFITA '09, Vol. 2, pp. 268-271, 2009.

[20]. [Wang et al., 2009]A. Wang , Y. Li and W. Wang , "Text Clustering Based on Key Phrases", *in Proceedings of 1st International Conference on Information Science and Engineering (ICISE)*, pp. 986-989, 2009.

[21]. [Zhang et al., 2009] J. Zhang ,P. Zhao, L. Shang and L. Wang , " Web Usage Mining Based on Fuzzy Clustering in Identifying Target Group*" in Proceedings of ISECS International Colloquium on Computing, Communication, Control, and Management*, Vol. 4, pp. 209-212, 2009.

[22]. [Zhao et al., 2010] J. Zhao , S. Gu and L. He, "A Novel Approach to Clustering Access Patterns in E-Learning Environment", *in Proceedings of 2nd International Conference on Education Technology and Computer (ICETC)*, pp. V1-393 - V1-397, 2010.

[23]. [Bezdek, 1981] J. Bezdek, "*Pattern Recognition with Fuzzy Objective Function Algorithms*", Plenum Press, New York, ISBN: 0306406713, 1981.