International Journal of Engineering Research ISSN: 2348-4039
& Management Technology
March-2016 Volume 3, Issue-2
Email: editor@ijermt.org www.ijermt.org

WEB DOCUMENTS RETRIEVAL USING FUZZY CLUSTERING

S.Kalaiselvi
Research Scholar
Department of Computer Science
and Applications
Adhiparasakthi College of Arts and Science
G.B.Nagar, Kalavai
Vellore, Tamilnadu

P.Hariharan
Assistant Professor
Department of Computer Science
and Applications
Adhiparasakthi College of Arts and Science
G.B.Nagar, Kalavai
Vellore, Tamilnadu

1. ABSTRACT:

The World Wide Web is biggest information warehouse in recent years. Growth of web is dramatically enlarged with new technologies. The search engines are ineffective when the number of document on the web have propagated. The same way query retrieval, most of which no relation with what the user was looking for. The web documented varied and multifaceted there is exists difficult relations with in one of web document and linking to others. This research work have focused about clustering algorithm discover the latent semantics in a text corpus from a fuzzy linguistic viewpoint. In addition the applicability in text fields, it can be extensive to the applications, such as data mining, bioinformatics, content-based or collaborative information sifting, and so forth. Secondly, retrieval document belongs to search topic that should different other topics the difference between the other topics. Web contents are able to be clustered into topics in the hierarchy depending on their fuzzy linguistic measures. This research is taking this problem to give optimistic solution. In future finding good algorithm that can be effective to retrieval web documents

Main theme:
Effective web documents retrieval in search engines.
How form web document clusters based on fuzzy clustering.

2. INTRODUCTION:

Clustering is a technique of grouping objects where objects of one group are similar to each other and dissimilar to objects of other groups. Clustering is an unverified machine erudition technique. The unsupervised feature makes it more proper for clustering search result as it is not possible to determine as to how many categories are there in search result. Clustering of web search involves four basic steps: a) search result acquisition, b) result preprocessing, c) cluster formation and d) labeling of clusters. Some clustering engines acquire search results from one or more search engines and then merge them into one unified result set. In preprocessing, each and every document of page of the search result is transformed into streams of words or phrases or sentences depending upon the attributes of the clustering method. Other tasks performed during pre-processing are stop word removal, stemming, filtering etc.

Many methods, including k- means, hierarchical clustering , and nearest-neighbor clustering , select a set of key terms or phrases to organize the feature vectors depending on the differences between documents to capture semantics in order to fit users' intents. Suffix-tree clustering is a phrase-based approach, which carries out document clustering depending on the similarities between documents. Fuzzy c-means and fuzzy hierarchical clustering need prior knowledge about 'number of clusters' and 'initial cluster centroids,' which are considered as serious drawbacks of these approaches. To address these drawbacks, ant-based fuzzy clustering algorithms and fuzzy k- means clustering algorithms were proposed that can deal with unknown number of clusters.

Based on Vector Space Model the similarity between two documents is measured with vector distance, such as Euclidean distance, Manhattan distance, and so on. A fuzzy hierarchical clustering approach to discover a set of highly-related fuzzy frequent item sets to represent the candidate clusters. Decomposition technique by breaking the network organized by co-occurrence keywords with the centroids, that is, the nodes with the maximum degree, and a cutoff threshold into several clusters.

Data Clustering Algorithm Based on Single Hidden Markov Model, which identifies a suitable number of clusters in a given dataset without using prior knowledge about the number of clusters.

3. EXISTING PROBLEMS:

3.1. EXISTING METHODS DRAWBACKS:

- Fuzzy c-means and fuzzy hierarchical clustering need prior knowledge about 'number of clusters' and 'initial cluster centroids,' which are considered as serious drawbacks of these approaches.
- In Vector Space Model restrict the application domains, which make them difficult to be generalized if the domain does not have a proper ontology.
- Most of the subsequent methods have tried to resolve the semantic clustering problem without much consideration to the semantic hierarchy in documents.
- Most of the subsequent methods have tried to resolve the semantic clustering problem without much consideration to the semantic hierarchy in documents.

4. PROPOSED ALGORITHM:

The proposed algorithm is **Fuzzy Latent Semantic Clustering (FLSC)** that covers the latent semantics of web documents that can applicable in text domains, it can be extended to the applications such as Data mining Bio informatics, Content based or collaborative information filtering.

Latent Semantic Clustering (LSC) is a technique in overlapping the cluster processing, in particular distributional semantics, of analyzing relationships between a set of cluster and the terms they contain by producing a set of concepts related to the results and terms. A matrix containing query counts per cluster is constructed from a large piece of text and a mathematical technique called singular value decomposition (SVD) is used to reduce the number of clusters while preserving the similarity structure among clusters. Queries are then compared by taking the cosine of the angle between the two vectors (or the dot product between the normalizations of the two vectors) formed by any two rows. Values close to 1 represent very similar search query while values close to 0 represent very dissimilar search query.

Fuzzy Latent Semantic Clustering (FLSC) approach first Fuzzy c-means to find the cluster based on the user queries.

F (U,V) =$\sum_{i=1}^{C} \sum u^{m}_{i\,j}$ *distance (vi,xi)* subject to $\sum$ **u_{ij} = 1 for all j.**

Where distance pattern x from the i_{th} cluster U_{ik} the membership function (in the interval [0,1]) of point x_k in the i_{th} cluster such that 0¡ _nj =1 U_{ij} ¡n < U is the fuzzy c-partition of the data set, V is a set of c-prototypes and m¿1 is the fuzzier. However, even a few outliers or inherent noise in real data can affect the result of this algorithm. Fuzzy c-means has problems ending correct clusters in the presence of noise or outliers, because of its assumption that any point in a dataset must essentially belong to a cluster.

Fuzzy logic is based on the theory of fuzzy sets, a theory which relates to classes of objects with un-sharp boundaries in which membership is a matter of degree. Documents, queries and their characteristics could easily be viewed as fuzzy granular classes of objects with un-sharp boundaries and fuzzy memberships in many concept areas .Since the concept of fuzzy logic is quite intuitive, the fuzzy logic model provides a framework that is easy to understand for a common user of IR system.

4.1. ALGORITHM:

```
Initialize number of clusters
Initialize Cj (cluster centers)
Initialize $\alpha$ (threshold value)
Repeat
For i=1 to n: update $\mu j$ (Xi)
For k=1 to p ;
Sum=0
Count=0
For i=1 to n:
If $\mu$ (Xi) is maximum in Ck
then
If $\mu$ (Xi)>=$\alpha$
Sum=sum+Xi
Count= count+1
Ck=sum/count
        Until Cj estimate stabilize.
```

The clustering framing as follows

A set clusters C= {C1, C2, C3 ….Ck}

Maximum precision values:

$$\textbf{Precision (Cj, Li)} = \frac{(|Ci\cap Lj|)}{|Ci|}$$

Where the maximum is taken over all clusters F (Cj, Li) is defined as

$$F(Cj,Li)= \frac{2x\ Recall\ (Cj,Li)\ X\ Precision\ (Cj,Li)}{Recall\ (Cj,\ Li)\ X\ Precision\ (Cj,\ Li)}$$

4.2. PROPOSED SYSTEM ARCHITECTURE:

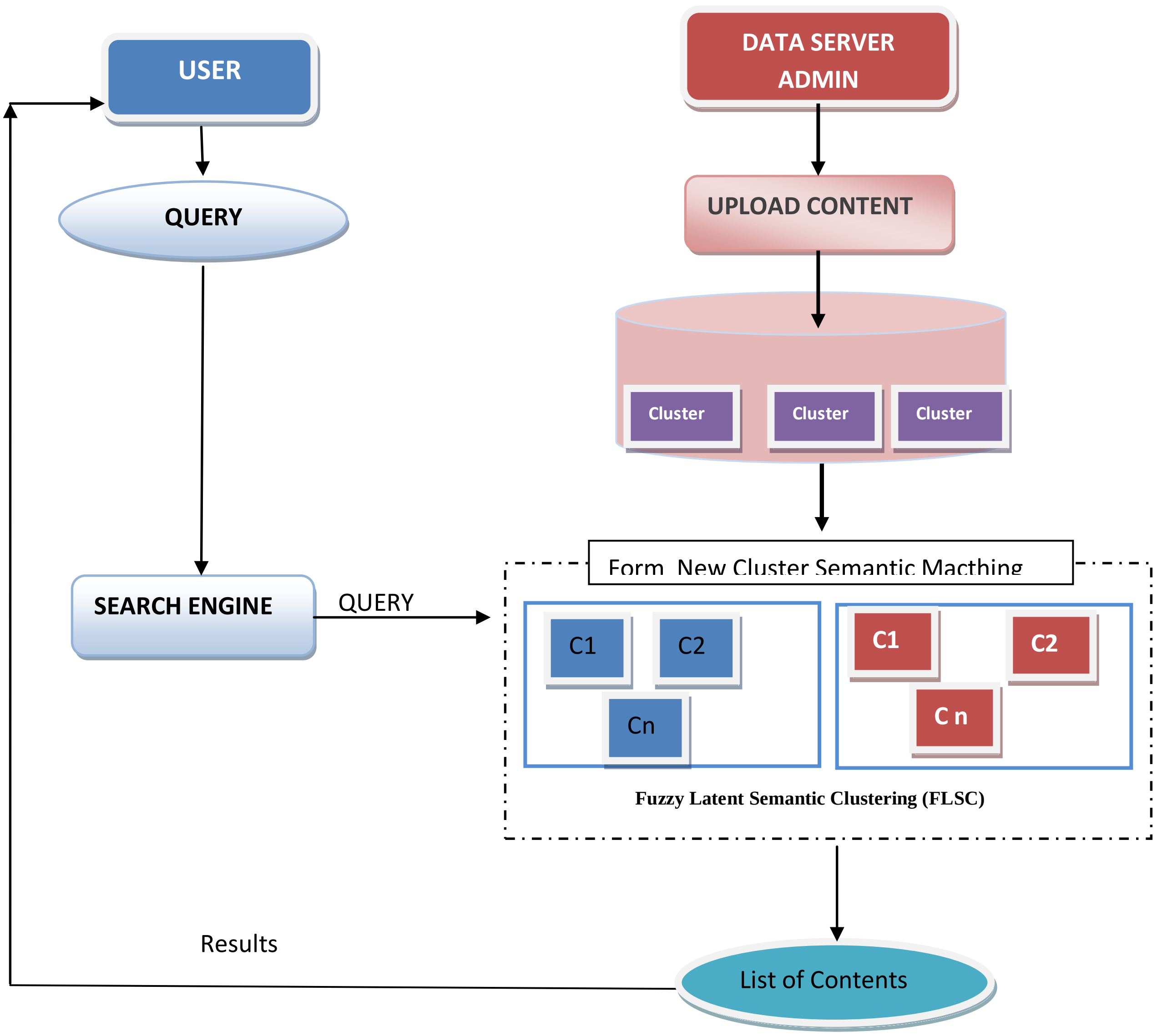

Fig 4.1 Proposed System Architecture

4.3. ADVANTAGES OF PROPOSED ALGORITHM:

- This algorithm can effectively discover such as maximal fuzzy simplexes and use them to cluster the collection of web documents.
- FLSC is a very good way to organize the unstructured and semi structured data into several semantic topics. It also illustrates that geometric complexes are an effective model for automatic web documents clustering.
- FLSC algorithm effectively retrieves the web documents and it's filtered to retrieve unnecessary web documents using this approach.
- It provides us to create new applications in web.

COMPARISON BETWEEN SOME FUZZY BASED CLUSTERING ALGORITHMS:

Comparison Metric	Hard C-Means	Fuzzy C-Means	Fuzzy C- Mediods	Fuzzy K-Means	FLSC
Speed	Average	Slow	Faster than C-Means And Hard C-Means	Average	Much Faster
Belongingness of each data	Only in Single Cluster	More than One Cluster	More than One Cluster	More than One Clusters	Form new cluster from more than one cluster
Combination of Clusters	Very Slow	Average	More than Average	Average	Very Quickly
Computational Time	Short	Long	Long	Long	Long
Overlapping The Clusters	No	Overlapping	Possible to Overlapping	No	Overlapping
Susceptible	High	Less	Very Less	High	Very Less Noise
Data Set Type	Crisp Set	Fuzzy	Fuzzy	Fuzzy	Fuzzy
Time Complexity	O(ncdi)	O(ndc^{2}i)	O(n^2)	O(nmkT)	O(n^2)
Accuracy	65	76	78	80.5	85

Table 4.1 Comparison between Different Fuzzy based Clustering Algorithms.

5. CONCLUSION:

There is a growing consciousness that, in perform; it is easy to discover a huge amount of information from the Web, where most of these prototypes are actually obvious, outmoded, and useless or monotonous to the user. To prevent the user from being over whelmed by a large number of uninteresting patterns, techniques are needed to identify only the useful/interesting patterns and present them to the user. Fuzzy sets, which constitute the oldest component of soft computing, are suitable for handling the issues related to understandability of patterns, incomplete/noisy data, mixed media information and human interaction, and can provide approximate solutions faster. In this research present new algorithm Fuzzy Latent Semantic Clustering (FLSC) that performs well for semantic web documents and results shows its accuracy and speed combined user query and retrieve web documents effectively. In Future this will apply this algorithm for temporal dataset like images and videos in social networks.

6. REFERENCES:

1. Magne Setnes "*Supervised Fuzzy Clustering for Rule Extraction"* Fuzzy Systems Conference Proceedings, 1999. FUZZ-IEEE '99. 1999 IEEE International .2000.
2. Ajith Abraham "*Natural Computation for Business Intelligence from Web Usage Mining"* 3 IEEE Congress on Evolutionary Computation (CEC2003), Australia, IEEE Press, ISBN 0780378040, pp. 1384-1391,2005.
3. B. de la Ossa, J. A. Gil, J. Sahuquillo and A. Pont " *Improving Web Prefetching by Making Predictions at Prefetch" ,IEEE* Next Generation Internet Networks, 3rd EuroNGI Conference on 2007.
4. Narendra S.Chaudhri and Avisheek Chosh." *Feature Extraction Using Fuzzy Rule Based System*" International Journal of Computer Science and Applications, ©Technomathematics Research Foundation Vol. 5, No. 3, pp 1 – 2011.

5. Sachin Ashok Shinde and Seema Singh Solanki "*A Fuzzy Rule Based Clustering Development Novel"* International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 – 2012.
6. Faraz Zaidi. "*Fuzzy Clustering and Visualization of Information for Web Search Results",* Fuzzy Clustering and Visualization of Information for Web Search Results. Journal of Internet Technology, Taiwan Academic Network, 13 (6) (939-952) -2012.
7. R. R. Papalkar and G. Chandel "**"***fuzzy clustering in web text mining and its application in ieee abstract classification" - 2013.*
8. Yingdi Sara El Manar El Bouanani and Ismail Kassou ""*A Comparison of Fuzzy Clustering Algorithms to Cluster Web Messages*". World Academy of Science, Engineering and Technology International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol:7, No:7,-2013.
9. Yingdi Guo, Kunhong Liu, Qingqiang Wu, Qingqi Hong and Haiying Zhang "WSEAS TRANSACTIONS on COMPUTERS. Yingdi Guo, Kunhong Liu, Qingqiang Wu, Qingqi Hong, Haiying Zhang .-2014.
10. I-Jen Chiang, Charles Chih-Ho Liu, Yi-Hsin Tsai, and Ajit Kumar" *Discovering Latent Semantics in Web Documents using Fuzzy Clustering*" IEEE Transactions on Fuzzy Systems (Volume:23 , Issue: 6) -2015.
11. M. Thangamani and P. Thangaraj, "*Ontology based fuzzy document clustering scheme*," *Modern Applied Science*, vol. 4, no. 7, pp. 148– 156, 2010.
12. S. Song, Z. Guo, and P. Chen, "*Fuzzy document clustering using w ghted conceptual model*," *Information Technology Journal*, vol. 10, no. 6, pp. 1178–1185, 2011.