Soft Computing Technique for Categorization of Unstructured Web Data

Ms. S. A. Deshpande
Muscat, Oman

Prof. Mahendra Patil
HOD Computer Department,
Atharva College of Engineering
Mumbai University, Mumbai

Prof. A. N. Boob
Assistant Professor,
School of C&IT,
Reva University,Banglore

Abstract- **The World Wide Web has huge amount of information that is retrieved using information retrieval tool like Search Engine. It becomes tedious for the user to manually extract real required information. The detection of common and distinctive topics within a document set, together with the generation of multi-document summaries, can greatly ease the burden of information management. In the present work, a technique is proposed called Soft Computing Technique for Categorization ofUnstructured Web Data that creates the clusters of web documents using fuzzy clustering which focuses on this problem of mining the useful information from the collected web document. We have used FCM (Fuzzy C mean) clustering algorithm and HCM (Hard C Mean) algorithm. FCM clustering is a clustering technique which is separated from hard C mean that employs hard partitioning. Such that a point can belong to all groups with different membership grades between 0 and 1. The evaluation of the performance is done by validation measures and it is evaluated by F-measure, entropy, and purity measure and time complexity. It is found that fuzzy clustering algorithms yields better results than hard clustering algorithms.**

Keywords: Search Engine, Web documents, Fuzzy C Means, Hard C Means, Entropy, Purity, F-Measure

I. INTRODUCTION

WWW is a huge repository of information consisting of hyperlinked documents spread over the internet. For a user, it is practically impossible to search through this extremely large database for the information needed by him. The search engine uses crawlers to gather information and stores it in database maintained at search engine side. For a given user's query the search engine searches in the local database and very quickly displays the results. The huge amount of information is retrieved using data mining tools. Classification, Clustering and Association tools etc. are used for data mining technique. Clustering plays a key role in searching for structures in data. As the number of available documents nowadays is large, hierarchical approaches are better suited because they permit categories to be defined at different pensiveness levels. The problem of clustering in finite set of data is to find several cluster canters that can properly characterize relevant classes of finite set of data such that degree of association is strong for data within blocks of the partition and weak for data in different blocks. When the weakness of a crisp partition of finite set of data is replaced with a fuzzy partition, this area is known as fuzzy clustering. Fuzzy clustering is a relevant technique for information retrieval. As a document might be relevant to multiple queries, this document should be given in the corresponding response sets, otherwise, the users would not be aware of it. Fuzzy clustering seems a natural technique for document categorization. There are two basic methods of fuzzy clustering, one which is based on fuzzy c-partitions, is called a fuzzy c-means clustering method and the other, based on the fuzzy equivalence relations, is called a fuzzy equivalence clustering method. The purpose of this paper is to propose a search methodology that consists of how to find relevant information from WWW [1]. In this paper, a method is being proposed of document clustering, which is based on fuzzy equivalence relation that helps information retrieval in the terms of time and relevant information.

II. LITERATURE REVIEW

Document clustering is the process of categorizing text document into a systematic cluster or group, such that the documents in the same cluster are similar whereas the documents in the other clusters are dissimilar. It is one of the vital processes in text mining. Due to growth and development in the field of internet and computational technologies, various clustering techniques have been proposed in the literature. Especially, text mining has gained lot of importance and it is demanding various tasks such as production of granular taxonomies, document summarization etc., for the scope of developing higher quality information from text [2]. Text mining is a knowledge concentrated technique where the user communicates with a document collection by using analysis tools. This is equivalent to data mining approach. It extracts the useful information from large volume of unstructured text. Text document used to identify simplified subset of document features that can be used to represent the particular document as the whole. This feature is said to be a representational model [3]. Each document in a collection is made up of large number of features, so that it affects the system approach, performance and design. The most widely used fuzzy clustering algorithm is Fuzzy c-means , a variation of the partitional k- means algorithm [4]. In fuzzy c-means each cluster is represented by a cluster prototype and the membership degree of a document to each cluster depends on the distance between the document and each cluster prototype. The closest the document is to a cluster prototype, the greater is the membership degree of the document in the cluster. In the year 1973 Dunn developed the Fuzzy C Means algorithm and later in 1981 Bezdek enhanced it. Fuzzy C Means algorithm is extensively used in pattern recognition. Fuzzy C Means algorithm uses the iterative process, which rejuvenates cluster centers for individual data point [5].

Various techniques for accurate clustering have been proposed [6], e.g. K-MEAN [7, 8], CURE [9], BIRCH [10], ROCK [11].K-MEAN clustering algorithm is used to partition objects into clusters while minimizing sum of distance between objects and their nearest center. In statistics and machine learning, *k*-means clustering is a method of cluster analysis which aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean. CURE (Clustering Using Representation) represents clusters by using multiple well scattered points called representatives. A constant number 'c' of well scattered points can be chosen from '2c' scattered points for merging two clusters. CURE can detect clusters with non-spherical shapes and works well with outliers. CURE achieves this by representing each cluster by a certain fixed number of points that are generated by selecting well scattered points from the cluster and then shrinking them toward the center of the cluster by a specified fraction. To handle large databases, CURE employs a combination of random sampling and partitioning.

BIRCH (Balance and Iterative Reducing and Clustering Hierarchies) is useful algorithm for data represented in vector space. It also works well with outliers like CURE. BIRCH incrementally and dynamically clusters incoming multi-dimensional metric data points to try to produce the best quality clustering with the available resources (i. e., available memory and time constraints). BIRCH can typically find a good clustering with a single scan of the data, and improve the quality further with a few additional scans. BIRCH is also the first clustering algorithm proposed in the database area to handle "noise)' (data points that are not part of the underlying pattern) effectively.

ROCK (Robust Clustering Algorithm for Categorical Attributes) gives better quality clusters involving categorical data as compared with other traditional algorithms.

III. ANALYSIS OF PROBLEM

With the recent explosive growth of the amount of content on the Internet, it has become increasingly difficult for users to find and utilize information and for content providers to classify and catalogue documents. Traditional web search engines often return hundreds or thousands of results for a search, which is time consuming for users to browse. The difficulties while using web for retrieval of information are:

- The Web is extremely large; there are more than 10 billion unique, publicly accessible pages on the Web.
- Web data changes rapidly. While the Web grows quickly in size, the information it contains is also updated constantly.
- The Web is poorly organized. Although small sections of the Web may be well structured and maintained, the Web as a whole is highly unstructured.
- The Web user community is very diverse. Users in different communities may have different backgrounds, interests, and preferences.

As a result of the above, users have increasing difficulty in locating the right information at the right time. Most Web users have had the experience of taking an hour or more to find a Web document that they can go through in five minutes. The amount of information vastly outstrips any individual's capability to survey it and how to find desired information efficiently and effectively has become an increasingly important and emergent issue

IV. IMPLEMENTATION

Web document mining helps users get the newest and worldwide information they are interested in which will be analysed and utilized further. Text clustering is unsupervised machine learning and all texts are unknown classification before being made in cluster. The similarity, among the texts in the same cluster, should be required as large as possible and the relation between clusters should be as minimum as possible to achieve this following two algorithms are implemented.

The HCM (K means) Clustering technique is simple, in this algorithm we decide centroids k, where K is user specified parameter namely number of cluster desired each point is then assigned to closest centroid and each collection of points assigned to a centroid is cluster. The centroid of each cluster is then updated based on the points assigned to the cluster. We repeat the assignment and update steps until no point's changes clusters or equivalently until the centroids remain the same.

In Fuzzy C means approach we first need to define the centroids and number of clusters, the main difference between HCM (K means)and Fuzzy C means is that an each point can belong to more than one cluster with some degree of membership.

We compare the Fuzzy C means (FCM) clustering algorithm and Hard C mean (K means) algorithm. In non-fuzzy or hard clustering, data is divided into clusters, where each data point belongs to exactly one cluster.

- Used to classify data.
- Each data point will be assigned to only one cluster.
- Clusters are also known as partitions.
- U is a matrix with c rows and n columns.
- The cardinality gives number of unique c partitions for n data points.

In this clustering technique partial membership is not allowed. HCM (K means) is used to classify data. By this we mean that each data point will be assigned to one and only one data cluster. In this sense, these clusters are also called as partitions that are partitions of the data. In case of hard c mean each data element can be a member of one and only one cluster at a time.

V. RESULT

We compare the Fuzzy C means (FCM) clustering algorithm and Hard C mean (K means) algorithm. In Hard C mean (K means) algorithm, data is divided into clusters, where each data point belongs to exactly one cluster and in Fuzzy C means (FCM) data is divided into clusters but each data point can belong to each cluster with some degree of membership. To measure the performance of implemented algorithms the above mentioned parameters are considered

Table1: Comparative Analysis according to clustering measures

Clustering Measure	Dataset	HCM (K-means)	Fuzzy c-Means
F-Measures	20 Pages	0.676	0.893
	53 Pages	0.62	0.85
Entropy Measure	20 Pages	0.376	0.216
	53 Pages	0.25	0.143
Purity	20 Pages	0.14	0.175
	53 Pages	0.053	0.059

Table2: Comparative Analysis according to number of clusters

No. of Clusters	HCM (K-means)	Fuzzy c-Means
1	108	288
2	364	4608
3	864	12960
4	1152	25344

VI. CONCLUSION

The present work implements an algorithm for clustering of web documents according to hard and fuzzy approach. The limitation of HCM (k means) clustering is, overcome by fuzzy c means clustering in which overlapping clusters were formed.

To measure the performance of both the algorithms, entropy, F- measure, purity and time complexity parameters were considered. It is observed the greater the value of purity indicates good clustering. The entropy is negative measure, the lower the entropy the better clustering it is. The higher the F-measure indicates better clustering. The results show that the Fuzzy C Means has high value of purity, F-measure and low value of entropy. This indicates good clustering. The HCM (k means) has lower value of purity and high value of entropy compared to Fuzzy C Means. It shows that FCM performs better than HCM (k means) i.e one document can belong to more than one cluster.

REFERENCES

[1] R. Kosala and H.Blockeel *"Web Mining Research: A Survey"*, SIGKDD Explorations ACM SIGKDD, July 2000.

[2] Cooley, R.,Mobasher, B.,Srivastava, J. *"Web Mining: Information And Pattern Discovery On The World Wide Web"*University of Minnesota, Technical Report TR 97-207, 1999.

[3] Mendelzon A., Mihaila, G. Milo, T,*"Querying the World Wide Web"*, Journal of digital library,pp 68-288 1997.

[4] Cunning Ham, H,"Information Extraction, Automatic ,Encyclopaedia of Languages and linguitics",2005.

[5] Bezdek J. C, "Pattern Recognition with Fuzzy Objective Function Algorithms",Plenum Press, New York, 1988.

[6] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, Angela Y. Wu, "An Efficient k-Means Clustering Algorithm: Analysis and Implementation", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No. 7, July 2002..

[7] PawanLingras ,Rui Yan and Chad West, " Fuzzy C- Means Clustering of Web Users for Educational *Sites"*, Springer Publication ,2003.

[8] LinasBaltruns and JuozasGordevicius, *"Implementation of CURE Clustering Algorithm",* SIGMOD Seattle, WA, USA ACM February 1, 2005.

[9] Tian Zhang, Raghu Ramakrishna, MironLivny, "BIRCH: An Efficent Data Clustering Method for Very Large Databases" SIGMOD '96 6/96 Montreal, Canada IQ ACM1996.