International Journal of Computer Science and Engineering Research and Development (IJCSERD), ISSN 2248-9363 (Print), ISSN 2248-9371 (Online)
Volume 1, Number 1, April- June (2011)
pp. 01- 09 © PRJ Publication
http://www.prjpublication.com/IJCSERD.asp

IJCSERD

© PRJ PUBLICATION

FUZZY CLUSTERING TECHNIQUE

Ms. Anjali B. Raut
Department of Computer Science & Engg, HVPM's COET, Amravati, India
anjali_dahake@rediffmail.com

Dr. G. R. Bamnot, Department of Computer Science & Engg,
PRMITR ,Badner, Amravati ,India.
grbamnote@rediffmail.com

ABSTRACT

Fuzzy clustering techniques are used to construct clusters with uncertain boundaries and allows that one object belongs to overlapping clusters with some membership degree. In other words, the fuzzy clustering is to consider not only the belonging status of object to the clusters, but also to consider to what degree do the object belong to the cluster. In this paper, a technique called "fuzzy hierarchical clustering" is being proposed that creates the clusters of web documents using fuzzy equivalence relation

Keywords-- Web Mining, Clustering, Search Engine, Fuzzy clustering

1. INTRODUCTION

World Wide Web has become a major source of information. Over the last decade there is tremendous growth of information on World Wide Web (WWW). According to a 2010 study [1], there are around 156 million websites and it continues to grow at roughly a million pages per day. Web creates the new challenges of information retrieval as the amount of information on the web and number of users using web growing rapidly. At present most users commonly use searching engines to find their information. Each searching engines having its own characteristics and algorithms to index, rank, and present web documents. But the limitations like Narrowly Searching Scope ,Low Precision ,Unable to Search Multimedia cannot overcome by search engines[2].

The ability to form meaningful groups of objects is one of the most fundamental modes of intelligence. Human perform this task with remarkable ease. Cluster analysis is a tool for exploring the structure of data. The core of cluster analysis is clustering; the process of grouping objects into clusters such that the objects from the same cluster are similar and objects from different cluster are dissimilar. The need to structure and learn vigorously growing amount of data has been a driving force for making clustering a highly active research area.

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. It is a Data Mining tool to use for grouping objects into clusters such that the objects from the same cluster are similar and objects from different cluster are dissimilar.

Web Mining has fuzzy characteristics, so fuzzy clustering is sometimes better suitable for Web Mining in comparison with conventional clustering.. 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 research is to propose a search methodology that consists of how to find relevant information from WWW. 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.

The paper is structured as follows: section 2 describes some related work about Web Mining and clustering algorithms. Section 3 shows the proposed method and section 4 presents an example, how to retrieve the relevant information from WWW. Section 5 shows the results. In section 6, conclusion and future work are presented.

2. RELATED WORK

Data Mining has emerged as a new discipline in world of increasingly massive datasets. Data Mining is the process of extracting or mining knowledge from data. Data Mining is becoming an increasingly important tool to transform data into information. Knowledge Discovery from Data i.e. KDD is synonym for Data Mining.

2.1 *Web Mining*

Over the last decade World Wide Web is a major source of information and it creates new challenges of information retrieval as the amount of information on the web increasing exponentially. Web Mining is use of Data Mining techniques to automatically discover and extract information from web documents and services [3]. Oren Etzioni was the person who coined the term Web Mining first time. 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 [3] 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 [4]. The second definition has become more acceptable, as is evident from the approach adopted in most research papers[5][6]. Web Mining is also a cross point of database, information retrieval and artificial intelligence [1].

2.2 Web Mining Process

Web mining may be decomposed into the following subtasks

1. Resource Discovery: process of retrieving the web resources.
2. Information Pre-processing : is the transform process of the result of resource discovery
3. Information Extraction: automatically extracting specific information from newly discovered Web resources.
4. Generalization: uncovering general patterns at individual Web sites and across multiple sites[1].

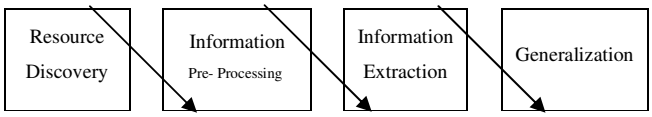

Figure 2.1: Web mining Process

2.3 Web Mining Taxonomy

Web mining can be categorized into three different classes based on which part of the web is to be mined i.e. Web content mining, Web structure mining and Web usage mining [2][6][7][8]. Following Figure 2.2 shows the Web Mining Taxonomy.

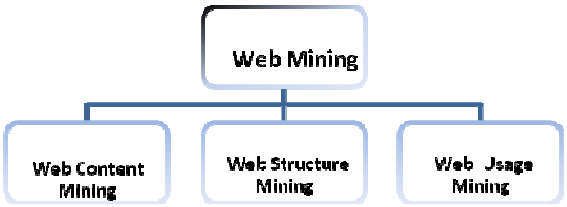

Figure 2.2: Web mining Taxonomy

Web content mining is the process of extracting useful information from the contents of web documents. Web structure mining is the process of discovering structure information from the web. Web usage mining is the application of data mining techniques to discover interesting usage patterns from web usage data. WCM and WSM uses real or primary data on the web whereas WUM mines the secondary data derived from the interaction of the users while interacting with the web.

2.4 Document Clustering

The Web is the largest information repository in the history of mankind. Finding the relevant information on www is not an easy task. The information user can encounter the following problems when interacting with the web[2].

- Low precision: Today's search tools have the low precision problem , which is due to the irrelevance of many search results. This results in a difficulty finding the relevant information.
- Low recall: It is due to the inability to index all the information available on the web . This results in a difficulty finding the unindexed information that is relevant.

Clustering is one of the Data Mining techniques to improve the efficiency in information finding process. Document clustering is widely applicable in areas such as search engines, web mining and information retrieval. Most document clustering methods perform several pre-processing steps including stop words removal and stemming on the document set [9]. Most of the document clustering algorithm worked on BOW (Bag Of Words)model[6]. Each document is represented by a vector of frequencies (TF) of remaining terms within the document. Some document clustering algorithms employ an extra pre-processing step that divides the actual term frequency by the overall frequency of the term in the entire document set(TF-IDF).It has great potentials in applications like object recognition, image segmentation and information filtering and retrieval [10].

Most of the clustering techniques fall into two major categories, and these are the hierarchical clustering and the partitional clustering [10]. Hierarchical clustering can further be categorized as agglomerative and divisive, based on how hierarchyis formed. Hierarchical method produce a nested sequence of partitions, with a single all inclusive cluster at the top and singleton clusters of individual objects at the bottom.

These algorithms start with the set of objects as individual clusters, then, at each step merges the two most similar clusters. This process is repeated until a minimal number of clusters have been reached, or if a complete hierarchy is required then the process continues until only one cluster is left. These algorithms are slow when applied to large document collections; single link and group-average can be implemented in O+ (n2) time (where n is the number of items) [11], while complete link requires O (n3) time and therefore tends to be too slow to meet the speed requirements when clustering several items. In terms of quality, complete links tend to produce "tight" clusters, in which all documents are similar to one another, while single link have the tendency to create elongated clusters which is a disadvantage in noisy domains (such as the web), because it results in one or two large clusters, and many extremely small ones[11]. This method is simple but needs to specify how to compute the distance between two clusters. The three commonly used methods for

computing distance are the single linkage, complete linkage and the average linkage method respectively. Divisive hierarchical clustering methods work from top to bottom, starting with the whole data set as one cluster, and at each step split a cluster until only singleton clusters of individual objects remain. They basically differ in two things, (i) Which cluster to split next (ii) How to perform the split. A divisive method begins with all patterns in a single cluster and performs the split until a stopping criterion is met [11].

Partitional clustering algorithms work by identifying potential clusters while updating the clusters iteratively, guided by the minimization of some objective function. The most common class are the K-means and its variants, Kmeans, according to [12], is a linear time clustering algorithm. It is a representative of the partition based algorithms where the number of clusters needs to be known apriori, it uses a minimum "within class sum of squares from the centers" criterion to select the clusters. K-means, according to [13], is a partitional algorithm that derives clusters based upon longest distance calculations of the elements in a dataset, and then it assigns each element to the closest centroid (the data points; that is the mean of the value in each dimension of asset of multidimensional data points). However, according to [14], this method has since been refined and can deal with ellipse shaped data clusters as well as ball shaped ones and does not suffer from the dead unit problem plague of the earlier Kmeans algorithm. Also, this new K-means algorithm performs proper clustering without pre-determining the exact cluster number and it is proven to be efficient and accurate[14].

Experimental results of K-means algorithm have been shown in [13]. In detail, it randomly selects K of the instances to represent the clusters based on the selected attributes; all remaining instances are assigned to their cluster center. Kmeans then computes the new centers by taking the mean of all data points belonging to the same cluster. The operation is iterated until there is no change in the gravity centers. If K cannot be known ahead of time, various values of K can be evaluated until the most suitable one is found. The effectiveness of this method, as well as of others, relies heavily in the objective function used in measuring the distance between instances. Several variants of K-mean algorithm have been reported in the literature, such as the K-median. The K-mode algorithm [15] is a recent partitioning algorithm that uses the simple matching coefficient measure to deal with categorical attributes. The K-prototype algorithm by [15] integrated the K-means and the K-modes algorithm to allow for clustering instance described by mixed attributes. Some of them attempt to select a good initial partition so that the algorithm is more likely to find the global minimum value. Another variation is to permit splitting and merging of the resulting clusters, i.e. a cluster is split when its variance is above a specified threshold, and the two clusters are merged when the distance between their centroids is below another pre-specified threshold [15]. Using this variant, it is possible to obtain the optimal partition starting from any arbitrary initial partition, provided proper threshold values are specified. Another variation of the Kmeans algorithm involves selecting a different criterion function altogether [15]. The dynamic clustering algorithm (which permits representation other than the centroids for each was proposed in [16] and [17] and describes a dynamic clustering approach obtained by formulating the clustering problem in the framework of maximum likelihood estimation [10]. It is less sensitive to outliers than traditional K-means due to the characteristics of the norm.

Oren Zamir and Oren Etzioni[12] in their research listed the key requirements of web document clustering methods as relevance, browsable summaries, overlap, snippet tolerance, speed and accuracy. They have given STC (Suffix Tree Clustering) algorithm which creates clusters based on phrase shared between documents.It is a linear time clustering algorithm. A phrase in this context is an ordered sequence of one or more words and a base cluster to be a set of documents that share a common phrase. Suffix tree, as defined by [19], is a concept representation of a tie (retrieval) corresponding to the suffixes of a given string where all the nodes with one 'child' are merged with their 'parents'. It is a divisive method which begins with the dataset as a whole and divides it into progressively smaller clusters, each composed of a node with suffixes branching like leaves as given by [12].

Buckshot algorithm is an hybrid clustering method that combines the partitioning and hierarchical clustering methods. More precisely, it is a K-means algorithm where the initial cluster centroids are created by applying agglomerative hierarchical clustering (AHC) to a sample of the document collection [12]. Single pass algorithm attempts to find spherical clusters of equal size [11]. It is an incremental algorithm that uses a greedy agglomerative clustering algorithm, assigning each document to a cluster only once. The first processed document is used to start the first cluster. Every additional document is compared to all existing clusters and the most similar cluster is found. If its similarity to this cluster is above a predefined threshold, the document will be added to that cluster; otherwise it will be used to create a new cluster. Fractionation Clustering Algorithm is an approximation of the AHC where the search for the two closest clusters is not performed globally, but locally, and in bound regions [12]. Fractionation algorithm finds centers by initially breaking the corpus of documents into a set number of buckets of predetermined size [13]. The cluster subroutine is then applied to each bucket individually, breaking the contents of a bucket into yet smaller groups within the bucket. This process is repeated until a set number of groups is found, and this end up as the K centers [13].

3. PROPOSED WORK

The crawler collects the web pages from the web and stored it in database . The indexer extracts all words from the entire set of documents and eliminates words i.e. stop words such as “a”, “and”, “the” etc from each documents. The documents are stored in indexed database based on keywords Now, the proposed fuzzy clustering method based upon fuzzy equivalence relations is applied on the indexed database. A list of common words called keywords is generated in table 3.1.

Table 3.1: Document No. and Keywords

Document No	Keywords
0	Web
1	Fuzzy
2	Cluster
3	Fuzzy
4	Web

Each keyword is assigned a Keyword ID as shown in table 3.2

Table 3.2: Keywords and Keyword ID

Keywords	Keyword ID
Web	0
Fuzzy	1
Database	2
Cluster	3

The information contained in table 3.1 and table 3.2 is used to generate the required document clustering data for applying fuzzy equivalence relation. As it is not directly possible; so first determine a fuzzy compatibility relation (reflexive and symmetric) in terms of an appropriate distance function applied on given data. Then, a meaningful fuzzy equivalence relation is defined as the transitive closure of this fuzzy compatibility relation.
A set of data X is consisting of the following points in R2 (p-tuples of R_p) as shown in figure 3.1.

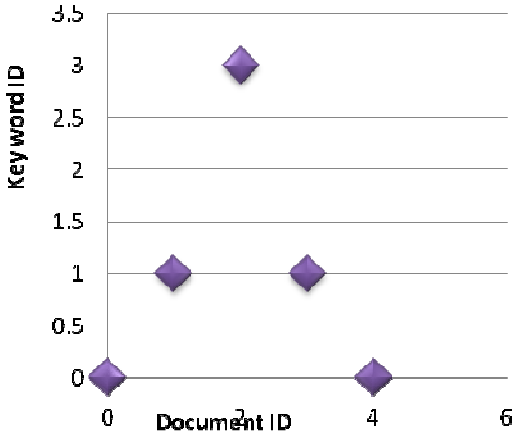

Figure 3.1: Web Document Data

The data X is shown in table 3.3

Table 3.3

K	1	2	3	4	5
X_{K1}	0	1	2	3	4
X_{K2}	0	1	3	1	0

Let a fuzzy compatibility relation, R, on X be defined in terms of an appropriate distance function of the Minkowski class by the following formula

$$R(x_i,x_k) = 1- \delta\left(\sum_{J=1}^{p} | x_{ij} - x_{kj} |^{q} \right)^{1/q} \text{.....(i)}$$

For all pairs $(x_i, x_k) \in X$, where $q \in RT$, and δ is a constant that ensures that $R(x_i,x_k) \in [0,1]$, Clearly , δ is the inverse value of the largest distance in X. In general, R defined by equation (i) is a fuzzy compatibility relation, but not necessarily a fuzzy equivalence relation. Hence, there is need to determine the transitive closure of R.
Given a relation R(X,X), its transitive closure R_T (X,X) can be determined by simple algorithm that consists of the following three steps:

1. R' = R U (R o R)
2. If R' ≠ R, make R = R' and go to step 1
3. Stop R' = R_T

This algorithm is applicable to both crisp and fuzzy relations. However, the type of

composition and set union in step 1 must be compatibile with the definition of transitivity employed. After applying this algorithm a hierarchical cluster tree will be generated. Each cluster has similar documents which help to find the related documents in the terms of time and relevancy.

4. EXAMPLE

To illustrate the method based on fuzzy equivalence relation, let us take a example. In this example there are five web documents and four keywords as shown in figure 3.1. By applying above algorithm, analyze the data for q= 1, 2.

As the First step,we analysis it for q=2, which is corresponds to the Euclidean distance .

Following are data points for q =1,2

x_1= (0,0) , x_2= (1,1) ,x_3=(2,3) , x_4= (3,1), x_5= (4,0)

There is need to determine the value of δ for equation (i). The largest Euclidean distance between any pair of given data points is 4 (between x_1 and x_5)

then we have $\delta = 1/4 = 0.25$

Now calculate membership grade of R for equation (i)

For example

$R(x_1, x_3) = 1 - 0.25(2^2 + 3^2)^{0.5} = 0.1$

When determined, relation R may conveniently be represented by the matrix for the following data points

$$R = \begin{pmatrix} 1 & .65 & .1 & .21 & 0 \\ .65 & \mathbf{1} & .44 & .5 & .21 \\ .1 & .44 & 1 & .44 & .1 \\ .21 & .5 & .44 & 1 & .65 \\ 0 & .21 & .1 & .65 & 1 \end{pmatrix}$$

This relation is not max-min transitive ; its transitive closure is

$$R_T = \begin{pmatrix} 1 & .65 & .44 & .5 & .5 \\ .65 & \mathbf{1} & .44 & .5 & .5 \\ .44 & .44 & 1 & .44 & .44 \\ .5 & .5 & .44 & 1 & .65 \\ .5 & .5 & .44 & .65 & 1 \end{pmatrix}$$

This relation includes four distinct partitions of its α – cuts:

$\alpha \in [0, 0.44]$: { { x_1 , x_2 ,x_3 , x_4, x_5, x}}

$\alpha \in (0.44, 0.5]$: { { x_1 , x_2 , x_4, x_5},{x_3}}

$\alpha \in (0.5, .65]$: { { x_1,x_2} , {x_3 }, {x_4, x_5}}

$\alpha \in (0.65, .1]$: { { x_1} , {x_2 },{x_3 }, {x_4},{ x_5}}

Now repeat the analysis for q = 1 in eq (i) which represents the Hamming distance. Since the largest hamming distance in the data is 5(between x_1 and x_3 and between x_3 and x_5) ,

we have δ=1/5=0.2.

The matrix form of relation R is given by eq.(i) is now

R=

$$\begin{pmatrix} 1 & .6 & 0 & .2 & .2 \\ .6 & \mathbf{1} & .4 & .6 & .2 \\ 0 & .4 & 1 & .4 & 0 \\ .2 & .6 & .4 & 1 & .6 \\ .2 & .2 & 0 & .6 & 1 \end{pmatrix}$$

And its transitive closure is

RT=

$$\begin{pmatrix} 1 & .6 & .4 & .6 & .6 \\ .6 & \mathbf{1} & .4 & .6 & .6 \\ .4 & .4 & 1 & .4 & .4 \\ .6 & .6 & .4 & 1 & .6 \\ .6 & .6 & .4 & .6 & 1 \end{pmatrix}$$

The relation gives the following partions in its α - cuts

$\alpha \in [0, 0.4]$: { { x_1 , x_2 ,x_3 , x_4, x_5}}

$\alpha \in (0.4, 0.6]$: { { x_1 , x_2 , x_4, x_5},{x_3}}

$\alpha \in (0.6, 1]$: { { x_1} , {x_{21} },{x_3 }, {x_4},{ x_5}}

5. RESULTS AND SNAPSHOTS

This result agrees with our visual perception of geometric clusters in the data.. The dendrogram is a graphical representation of the results of hierarchical cluster analysis. This is a tree-like plot where each step of hierarchical clustering is represented as a merging of two branches of the tree into a single one. The branches represent clusters obtained on each step of hierarchical clustering. The result of above example is described in the form of dendrogram, in snapshots shown in Figure 5.1 and figure 5.2

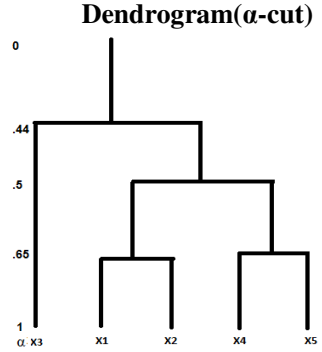

Figure 5.1 : Snapshot of Dendogram for Euclidean distance

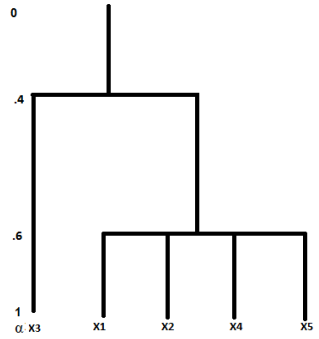

Figure 5.2 : Snapshot of Dendogram for hamming distance

6. CONCLUSION AND FUTURE RESEARCH

Web data has fuzzy characteristics, so fuzzy clustering is sometimes better suitable for Web Mining in comparison with conventional clustering. This proposed technique for clustering, based on fuzzy approach improves relevancy factor. This technique keeps the related documents in the same cluster so that searching of documents becomes more efficient in terms of time complexity.

REFERENCES

[1] http://news.netcraft.com

[2] WangBin and LiuZhijing , *Web Mining Research* , In Proceeding of the 5th International Conference on Computational Intelligence and Multimedia Applications(ICCIMA'03) 2003.

[3] Oren Etzioni, *The World Wide Web: quagmire or gold mine?* ,Communications of ACM", Nov 96.

[4] R. Cooley,B. Mobasher and J. Srivastava ,*Web Mining: Information and Pattern Discovery on the World Wide Web*, In the Proceeding of ninth IEEE International Conference on Tools with Artificial Intelligence(ICTAI'97),1997.

[5] Hillol Kargupta, Anupam Joshi, Krishnamoorthy Sivakumar and Yelena Yesha,*Data Mining: Next Generation Challenges and Future Directions*, MIT Press,USA , 2004.

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

[7] Sankar K. Pal,Varun Talwar and Pabitra Mitra , *Web Mining in Soft Computing Framework : Relevance, State of the Art and Future Directions* , IEEE Transactions on Neural Network , Vol 13,No 5,Sept 2002 .

[8] Andreas Hotho and Gerd Stumme, *Mining the World Wide Web- Methods, Application and Perceptivities*, in Künstliche Intelligenz, July 2007.

[9] C.M. Benjamin, K.W. Fung, and E. Martin, *Encyclopaedia of DataWarehousing and Mining*. Montclair State University, USA. 2006.

[10] A. Jain, and M. Murty, *Data Clustering: A Review* ACM Computing Surveys, vol. 31, pp. 264-323. 1999.

[11] E.Z. Oren, *Clustering Web Documents: A Phrase-Based Method for Grouping Search Engine Results*, Ph.D. Thesis, University of Washington.1999.

[12] O. Zamir, O. Etzioni, *Web Document Clustering*, Department of Computer Science and Engineering, University of Washington, Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 46-54.1998.

[13] S. Sambasivam, and N. Theodosopoulos, *Advanced Data Clustering Methods of Mining Web Documents*. Issues in Informing Science and Information Technology. Vol. 3, pp. 563-579.

[14] Y.M. Cheung, *K*-means: A New Generalized k-means ClusteringAlgorithm*. Pattern Recognition Letters, vol. 24, pp. 2883-2893.2003.

[15] Z. Huang, *Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values*.Data Mining and Knowledge Discovery, vol. 2, pp. 283-304.1998

[16] E. Diday, *The Dynamic Cluster Method in Non-Hierarchical Clustering*. Journal of Computer Information Science. Vol. 2, pp. 61-88. 1973.

[17] M.J. Symon, *Clustering Criterion and Multi-Variate Normal Mixture*.Biometrics, vol. 77, pp. 35-46. 1977.

[18] A. K. Jain,M. N. Murty and P. J. Flynn, *Data clustering: A review*, ACM computing surveys 31(3):264-323,Sept 1999.

[19] King-Ip Lin and Ravikumar Kondadadi, *A Similarity Based Soft Clustering Algorithm for Documents*, in Proceeding of the 7th International Conference on Database Systems for Advanced Applications DASFAA-2001), April 2001.