Fuzzy Clustering of Web Documents Using Equivalence Relations and Fuzzy Hierarchical Clustering

Neha Arora[#1], Devendra Kumar[#2]
Department of Computer Science and Engineering
IFTM University Moradabad
[1]apexneha2009@gmail.com
[2]dev625@yahoo.com

Abstract—***WWW is a fertile area for data mining research,[1] as huge amount of information is available in the form of unstructured and semi structured text databases[2] .It becomes typical to mine the relevant content or information from the web. So method of document clustering has been introduced as a methodology for improving document retrieval process. Clustering is a useful method for the textual data mining. Traditional clustering technique uses hard clustering algorithm in which each document use to belong to only one and exactly one cluster which creates problem to detect multiple themes of the documents. Clustering can be considered the most important unsupervised learning process which deals with finding the clusters according to logical relationship or consumer preferences. A cluster can be a structure in a collection of unlabeled data. The analysis of clusters deals with organizing the data objects into various clusters which has least inter cluster similarity and more intra cluster similarity [4]. Many clustering algorithms have been proposed by researchers. Partitioning clustering and hierarchical clustering are two main approaches to clustering. This paper summarizes the agglomerative hierarchical clustering method and presenting the clusters in the form of a dendrogram. Then Birch multiphase hierarchical clustering is applied in which clustering features are measured using clustering feature tree.***

Keywords— **Search Engine, Web Mining, Agglomerative Hierarchical Clustering, Fuzzy clustering.**

I. INTRODUCTION

The explosive growth of information sources on the www has made necessary for users to utilize automated tools to retrieve the desired information resources and to scan and analyse their usage pattern .Web creates the challenge for information retrieval on the web because it is typical to search the large database for the information required by the users that raises the need of search engines.

Web mining is the process that makes use of the data mining techniques to extract information and automatically discover the documents and services of web. According to analysis targets, web mining process can be broadly divided into three different types, which are Web content mining, Web usage mining and Web structure mining. Web content mining is the mining, integration and extraction of useful data, information and knowledge from Web page . Web usage mining is the process of extracting useful information from server logs e.g. use Web usage mining is the process of finding out what users are looking for on the Internet. Web structure mining is the process of using graph theory to analyse the node and connection structure of a web site [6]. The www is dynamic, huge, diverse which raises problems when interacting with the web like low precision which arises due to the irrelevance of the many search results and problem of low recall which arises due to the inability to index all the information available on the web[1].

Clustering is a process of partitioning data objects into meaningful subclasses, called clusters. Clustering can be said as automated process to group the related records together .Records that are related are grouped together on the basis of having similar values for attributes. This approach of partitioning the database through clustering analysis is often used as an exploratory technique. Clustering is an unsupervised learning technique because there are no predefined classes in it .Cluster analyses itself is not a specific algorithm. It can be achieved by various algorithms that differ significantly in their approach of what constitutes a cluster and how to efficiently identify them. Clustering techniques have been widely used in many areas such as data mining, artificial intelligence, pattern recognition, bioinformatics, segmentation and machine learning [3].

Web Mining makes use of the data mining methods to extract the text from the web. The possible techniques of clustering are applied which can help to improve the efficiency of the searching process of the information. As the number of available documents nowadays is large, hierarchical approaches are better suited because they permit categories to be defined at different levels.

There are various orthogonal aspects with which clustering methods could be compared: The partitioning criteria, separation of clusters similarity measure, clustering space. There are various clustering methods: Partitioning methods, Hierarchical method, Density based method and Grid based method.

Fuzzy clustering is a class of algorithms for the cluster analysis in which the allocation of data points to clusters is "soft" means "fuzzy" in the sense that an object belongs to a clusters with some membership level called as fuzzy logic. Fuzzy clustering is a process of assigning these membership levels, and then using them to assign data elements to one or more clusters [7]. Membership level indicates the strength of the association between a particular cluster and the data

element. Fuzzy clustering is a process of assigning these membership levels, and then using membership level to assign data elements to one or more clusters. In many situations, fuzzy clustering is more appropriate than hard clustering. Objects on the boundaries between several classes are not forced to fully belong to one of the classes, but rather are assigned membership degrees between 0 and 1 indicating their partial membership [8].

The modeling of imprecise and qualitative knowledge, as well as handling of uncertainty at various stages is possible through the use of fuzzy sets.

Fuzzy logic is capable of supporting, to a reasonable extent, human type reasoning in natural form by allowing partial membership for data items in fuzzy subsets. Fuzzy clustering is a method of document clustering There are various method of fuzzy clustering like fuzzy c-partition Fuzzy C-Mean (FCM), Possibility C-Mean (PCM)[9] and the other is based fuzzy equivalence relation of discrete mathematics .

BIRCH: Multiphase Hierarchical clustering using Clustering Feature Tree [15].

Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) is designed for clustering a large amount data by integrating hierarchical clustering and other clustering methods. It uses clustering feature to summarize a cluster and clustering feature tree to represent a cluster hierarchy. Birch is a integrated hierarchical clustering algorithm which are not considered as pure clustering algorithm, so Other clustering algorithm are merged into hierarchical clustering to improve the quality of cluster and also to perform multiple phase clustering [16]. Hierarchical algorithm suffer from the problem that once we have performed either merge or split step, it can never be undone. It leads to smaller computation costs by not having to worry about a combinatorial number of different choices. However, such type of techniques cannot correct mistaken decisions that once have been taken. There are two approaches that could help in improving the quality of hierarchical clustering, the first approach is to perform careful analysis of object linkages at each hierarchical partitioning and second approach is by integrating hierarchical agglomeration and other approaches by first using a hierarchical agglomerative method and group objects into micro-clusters, and then performing macro-clustering on the micro-clusters using another clustering method such as iterative relocation [Jiawei Han & Micheline Kamber, 2006] .This approach is similar to B+-Tree or R-Tree.

The objective of this paper is to search or retrieve the relevant content from the web in less time and efficiently and tries to improve the quality of the clusters formed.

This paper consists of following sections:

II. RELATED WORK

Data Mining is a process of extracting interesting patterns and knowledge from large amount of data. Many people treat data mining as a synonym of Knowledge discovery from data or KDD while others view data mining an essential step in the process of knowledge discovery.

The Web Mining research is at cross road of research from several research communities such as database, information retrieval and within Ai, especially the sub areas of machine learning and natural language processing [1].With the tremendous growth of information on WWW, it creates a challenge to retrieve the information efficiently. Web mining is the use of data mining technique to automatically discover and extract document from the web [2].

The term web mining was first coined by the Oren Etzoini made the hypothesis that the information on the web is structured. Oren Etzioni suggested to decompose web mining into following sub tasks namely:

A. Web Mining Process

Web mining may be decomposed into the following subtasks:

- Resource Discovery: process of retrieving the web resources.
- Information Pre-processing: is the transform process of the result of resource discovery
- Information Extraction: automatically extracting specific information from newly discovered Web resources.
- Generalization: uncovering general patterns at individual Web sites and across multiple sites. [10]

B. Web Mining Taxonomy [11]

Web Content Mining (WCM): Web Content Mining is the process of extracting useful information from the contents of web documents. A web page contains the collections of facts or content data. However, Web contents are not only text, but encompass a very good range of data such as audio, video, symbolic, metadata and hyperlinked data.

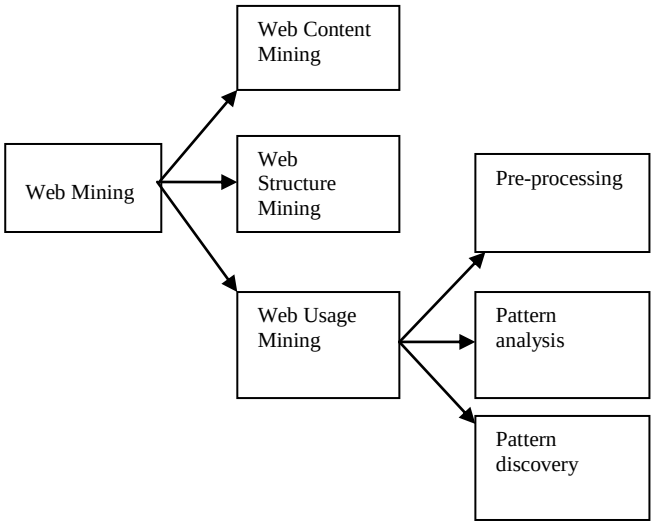

Fig. 1 Web Mining Taxonomy

Web Structure Mining (WSM): WSM deals with mining the structure of hyperlinks within the web itself. WSM reveals more information than just the information contained in documents .Web structure utilizes the hyperlinks structure of the web to apply social network analysis to model the underlying links.

Web Usage Mining (WUM): Content and structure mining utilizes the primary data or real data on the web, usage mining mines the secondary data generated by the users interaction with the web. Web usage data includes data from, proxy server logs, web server access logs ,browser logs ,user profiles ,user sessions or transactions, user queries , mouse clicks and scrolls and any others data generated by the interaction of users and the web.

C. Clustering

Document clustering is widely applicable in areas such as search engines, web mining, information retrieval, and topological analysis. Raymond Kosala and Hedrick Blockeel [1] surveyed that today's search tool have the following problems .The first problem is the problem of low recall and other is problem low precision .There are various clustering which could be implemented. Basically, Clustering algorithm are divided into following categories: Partitioning Technique ,Hierarchical Technique, Density based and grid based technique A.K. Jain , M.N. Murty [3] has surveyed various data clustering algorithm Hierarchical clustering is better quality clustering algorithm approach in terms of its time complexity [13].

Hierarchical clustering requires pre computed documents similarity matrix. Hierarchical document clustering organizes clusters into a tree or a hierarchy that facilitates browsing .The parent-child relationship among the nodes in the tree can be viewed as a topic-subtopic relationship in a subject hierarchy such as the Yahoo! directory. Hierarchical clustering can be divided into two categories i,e agglomerative hierarchical and divisive hierarchical clustering.

Agglomerative algorithms (bottom up approach) begin with each element as a separate cluster and merge them in successively larger clusters [14]. .Hierarchical Clustering algorithm include Single linkage ,Complete linkage, Group average and Ward method .In the agglomerative approach, the entities that are most similar and are the least far apart are grouped together to form clusters.

Partitioning algorithms are based on specifying an initial number of groups, and iteratively reallocating objects among groups to convergence [14].

Michael Steinbach, George Karypis and Vipin Kumar [13] compared the document clustering algorithm and presented their results. They focussed on the two main approaches of the document clustering that are agglomerative hierarchical clustering and k means method. They portrayed that hierarchical clustering is better quality clustering approach.

Agglomerative hierarchical clustering has different variants like Single link and complete link in which single link takes O(n2) time and complete link takes O(n3) time.

Menahen Friedman [4] had proposed fuzzy based document clustering which clusters the document mined from the web. A. K. Jain, M.N. Murty and P.J Flynn surveyed the various clustering algorithm and had broadly classified the clustering algorithms into hierarchical and partition techniques .He proposed a fuzzy based document that are represented by variable length vector which consist of two fields . The first field is the identification of keywords and the second field denotes the frequency associated with the keyword. Keole Ranjit [2] presented various clustering algorithm such as K-means, Fuzzy c-means and Hypershperical. Oren Etzoini has transformed the web into massive layered database to facilitate data mining. Etzoini was the first person who coined the term web mining. There were two different approaches for web mining, first is the process centric view which detailed that web mining is a sequence of different process and the other is data centric view which detailed web mining in terms of the type of the data that was being used for mining process.

Traditional clustering is different from the fuzzy hierarchical clustering in the way that in conventional clustering each document belong to only one cluster which made it difficult to the retrieve the documents from the web. Conventional clustering is also called as hard clustering .On the other hand fuzzy clustering is a different concept which is termed as soft clustering. Dealing with uncertainty and vagueness fuzzy clustering works, in which the documents can belong to various clusters on the basis of similarity. This methodology allows data object to belong to various clusters with some membership degree. Jursic Matjaz *et al.* [15] have been presented the fuzzy clustering of two dimensional points and documents. For the need of documents clustering they implemented fuzzy c-means.

III. PROPOSED WORK

A Fuzzy based web documents clustering is being proposed is being proposed which is based on fuzzy concept of equivalence relations of discrete mathematics. The database stores the downloaded documents which contains the respective keywords. The keywords are fetched from the documents after the elimination of stop words.

TABLE 1
DOCUMENTS AND KEYWORDS

Documents	Keywords	Sites
0	Hierarchical	www.altavista.com
1	Partitioning	www.ask.com
2	Hierarchical	www.bing.com
3	Density	www.google.com
4	Agglomerative	www.gopher.com
5	Partitioning	www.monster.com
6	Hierarchical	www.yahoo.com

TABLE 2
MAPPING OF KEYWORDS AND THEIR ID'S

Keywords	Keyword id	Sites
Hierarchical	0	www.altavista.com
Partitioning	1	www.ask.com
Density	2	www.google.com
Agglomerative	3	www.gopher.com

A Fuzzy binary relation that is reflexive, symmetric and transitive is known as fuzzy equivalence relation or similarity relation .A binary relation that is reflexive and symmetric is known as compatibility or tolerance relation. Compatablity relation is applied to form clusters in terms of using an appropriate distance function like Euclidean distance, Manhattan distance and Minkowski distance to measure the distance between two points, here objects.

TABLE 3
KEYWORD ID AND DOCUMENT ID

K	**1**	**2**	**3**	**4**	**5**	**6**	**7**
X_{k1}	0	1	2	3	4	5	6
X_{k2}	0	1	0	2	3	1	0

Distance measures used for computing the dissimilarity of objects described by numeric attributes.These measures include the Euclidean ,Manhattan and Minkowski distance. The compatibility relation is defined is defined in the form of an appropriate distance function defined by Minkowski distance function.

$$d(i,k)=1-\delta(\sum | x_{ik} - x_{jk}|^{q})^{1/q}$$

δ is a constant which ensures that relation lies between [0,1]
After applying transitive closure algorithm a hierarchical cluster tree is generated in which similar documents are kept in the same cluster while different documents are kept in different cluster in terms of least time and good efficiency.

IV. RESULTS

Generating the values for q , where q is a real no.

x_1=(0,0) x_2=(1,1) x_3=(2,0) x_4 =(3,2) x_5=(4,3) x_6=(5,1) x_7=(6,0)

Applying the distance function on the above mentioned values as defined above, we obtain the following relation matrix of the data points.

1.0	0 .76	0 .67	0 .4	0 .17	0.15	0.0
0.76	1.0	0 .76	0 .63	0 .4	0.33	0.15
0.67	0.76	1.0	0 .63	0.4	0.47	0.33
0.4	0 .63	0 .63	1.0	0 .76	0.63	0.4
0.17	0 .4	0 .4	0 .76	1.0	0 .63	0.4
0.15	0.33	0.47	0 .63	0.63	1.0	0.76
0.0	0.15	0.33	0 .4	0.4	0.76	1.0

Fig. 2 Relation Matrix

A fuzzy equivalence relation is defined in terms of transitive closure of the relation matrix.

1.0	0.76	0.76	0.63	0.63	0.63	0.63
0.76	1.0	0.76	0.63	0.63	0.63	0.63
0.76	0 .76	1.0	0.63	0.63	0.63	0.63
0.63	0 .63	0 .63	1.0	0.76	0.63	0.63
0.63	0.63	0.63	0.76	1.0	0.63	0.63
0.63	0.63	0.63	0.63	0.63	1.0	0.76
0.63	0.63	0.63	0.63	0.63	0.76	1.0

Fig. 3 Transitive closure

Alpha Cut

Alpha cut is a property of a fuzzy set.An α-cut or α-level set of a fuzzy set $A \subseteq X$ is an ORDINARY SET $A_\alpha \subseteq X$, such that:

$$A_\alpha=\{\mu_A(x)\geq\alpha, \forall x\in X\}$$

$\alpha \in [0.0, 0.63] : \{x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$
$\alpha \in [0.63,0.76]: \{\{x_1,x_2,x_3\}\{x_4, x_5\}\{x_6,x_7\}$
$\alpha \in [0.76,1.0]: \{\{x_1\}\{x_2\}\{x_3\}\{x_4\}\{x_5\}\{x_6\}\{x_7\}...\}$

The result is the dendrogram after the analysis of the alpha cut which is the property of fuzzy clustering. Dendrogram is commonly used to represent the process of hierarchical clustering .It shows how are the objects grouped together (agglomerative) or partitioned (partitioning) step by step.

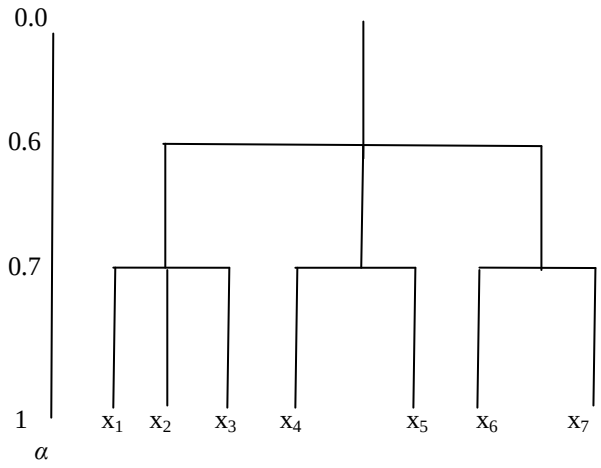

Fig. 4 Dendrogram

Applying the BIRCH algorithm on the generated data points. [Tian Zhang et al.,1996].

Consider a cluster of n dimensional data objects or points. The clustering feature of cluster is a 3d vector summarizing information about clusters of objects .It is defined as

$$CF= \{N, LS, SS\}$$

Where LS is the linear sum of the n points and SS is the square sum of the data points. A CF tree is height balanced tree. A CF tree has two parameters: branching factor B, and threshold T. The branching factor specifies the maximum no. of children per non leaf node .The threshold specifies the maximum diameter of the sub clusters stored at leaf node of the tree.

Phase 1: Birch scans the database to build an initial in memory cf tree, which can be viewed as a multilevel compression of the data that tries to preserve the data inherent clustering structure.

Phase 2: It applies a clustering algorithm to cluster the leaf nodes of the CF tree which removes sparse clusters as outliers and group dense clusters into larger ones.

CF1+CF2={n1+n2,LS1+LS2,SS1+SS2}

CF_1= <3,(0+1+2,0+1+0),($0^2+1^2+2^2$,,$0^2+1^2+0^2$)>

=<3(3,1),(5,1)>

$CF_{2=}$<2(3+4, 2+3),(3^2+4^2,2^2+3^2)>

=<2(7,5)(25,13)>

CF_3=<2(5+6,1+0),(5^2+6^2,$1^2+0^{2)}$>

=<2(11,1),(61,1)

The clustering feature of a new cluster after merging all the individual clusters C_1, C_2 and C_3 is

CF_4=<3+2+2(3+7+11,1+5+1),(5+25+61,1+13+1>

=<7(21,7), (91,15)>

An important consideration in Birch is to minimize the time required for input/ output. Birch is a multiphase clustering technique in which a single scan of the data set yields a basic, good clustering and one or more additional scans can optionally be used to further improve quality.

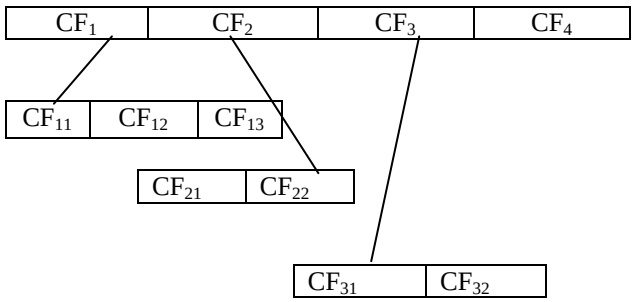

Fig. 5 CF- tree structure

All the entries in a leaf node must satisfy the threshold requirements with respect to the threshold value T, i.e the diameter of the sub cluster must be less than T. If the diameter of the sub cluster stored in the leaf node after insertion is larger than the threshold value, then the leaf node and other nodes are split. After the insertion of the new object, information about it is passed toward the root of the tree. The size of the CF tree can be changed by modifying the threshold.

An important consideration in Birch is to minimize the time required for input/ output. Birch is a multiphase clustering technique in which a single scan of the data set yields a basic, good clustering and one or more additional scans can optionally be used to further improve quality.

The advantages of hierarchical clustering are:

- Embedded flexibility regarding a level of granularity.
- Ease of handling of any forms of similarity or distance.
- Consequently applicability to any attributes types.
- Hierarchical clustering algorithms are more versatile

V. CONCLUSION AND PROBLEM STATEMENT

Agglomerative Hierarchal Clustering is more suitable for Web Mining as it is useful to detect the outlier data point or documents. This technique keeps the related documents in the same cluster so that searching of documents becomes more efficient in terms of time complexity. In future work we can also improve the relevancy factor of Hierarchal clustering and applying Birch algorithm structures help the clustering method achieve good speed and scalability in large databases.

.

REFERENCES

[1]. Raymond Kosala and Hendrik Blockeel, "Web Mining Research: A Survey" Volume 2, Issue 1, July 2000

[2]. Mr. Keole Ranjit R and Dr. Karde Pravin P,'Information Retrieval From Web Document Using Clustering Tech' ,Volume 2 Issue 3 March 2013 Page No. 759-76.

[3]. Jain A, Murty M and Flynn P. Data clustering: A review ACM Computing Surveys, 31(3), pp. 264–323, 1999.

[4]. Menahem Friedman and Abraham Kandel, 'A Fuzzy-Based Algorithm for Web Document', 2004 IEEE.

[5]. Han J and Kamber M. Data mining: concepts and techniques, Morgan Kaufmann, San Francisco, 2001.

[6]. http://en.wikipedia.org/wiki/Web_mining

[7]. http://en.wikipedia.org/wiki/Fuzzy_clustering

[8]. http://homes.di.unimi.it/~valenti/SlideCorsi/Bioinformatica05/Fuzzy-Clustering-lecture Babuska.pdf

[9]. International Journal off Industrial Engineering & Production Research,, December 2012,, Vol.. 23, No.4.

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

[11]. Sankar K. Pal and Pabitra Mitra, 'Web Mining in Soft Computing Framework: Relevance, State of the Art and Future Directions', IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 13, NO. 5, SEPTEMBER 2002.

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

[13]. Michael Steinbach,George Karypis and Vipin Kumar "A comparison of document clustering technique**".**

[14]. T. Soni Madhulatha -" AN OVERVIEW ON CLUSTERING METHODS" IOSR Journal of Engineering Apr. 2012, Vol. 2(4).

[15]. http://www.cse.hcmut.edu.vn/~chauvtn/data_mining/Texts/[1] Data Mining - Concepts and Techniques (3rd Ed).

[16]. Comparative Analysis of BIRCH and CURE Hierarchical Clustering Algorithm using WEKA 3.6.9.

[17]. Yogita Rani*, Manju** & Harish Rohil The SIJ Transactions on Computer Science Engineering & its Applications (CSEA), Vol. 2, No. 1, January-February 2014.

[18]. Matjaz Jursic and Nada Lavrac‖ *Fuzzy Clustering of Documents*‖ IMCIS‖, October 17, 2008, Ljubljana, Slovenia.

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

[20]. 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.

[21]. Matjaz Jursic et al. Journal of Universal Computer Science, vol. 16, no. 9 (*2010).*

[22]. A Random Indexing Approach for Web User Clustering and Web Prefetching Miao Wan1, Arne J¨onsson2, Cong Wang1, Lixiang Li1, and Yixian Yang1 L. Cao et al. (Eds.): Springer-Verlag Berlin Heidelberg 2012.

[23]. Manpreet kaur and Usvir Kaur "Comparison Between K-Mean and Hierarchical Algorithm Using Query Redirection", Volume 3, Issue 7, July 2013

