THE EFFICIENT CLUSTERING OF WEB DATA MINING
BASED ON FUZZY LOGIC APPROACH

S. Senthil kumar , M.Phil.,
Research scholar
Research Department of Computer Science
INDOAMERICANCOLLEGE, Cheyyar.

Prof. L. THIYAGARAJAN
Head of the Department,
Research Department of Computer Science
INDOAMERICANCOLLEGE, Cheyyar

ABSTRACT:
Web is the huge collection of data repository in that finding the correct information is not easy. Searching and web mining techniques are used to retrieve the information from the web. Web document clustering is the most useful technique to improve the efficiency of information searching problem. The traditional web mining techniques has various difficulties in handling the data which are not clear. Web usage mining is a data mining technology to mining the data of the web server. It can find the searching patterns of the user and some kind of correlations between the web pages. Web usage mining gives the support for the website design, providing personalization server and other business making the decision, etc. Web mining applies the data mining, the artificial intelligence, and the charting technology .The web data and traces users' visiting characteristics, and then extracts the user's using pattern Fuzzy clustering methods have the potential to manage such type of situations efficiently. This project proposed a new approach named Fuzzy latent semantic clustering that provides clusters that can capture the uncertainty among Web user’s navigation behavior. This FLSC algorithm provides Effective web documents retrieval in search engines.
1. INTRODUCTION

1.1 INTRODUCTION TO WEB DOCUMENT RETRIEVAL

Modern days, Web document retrieval systems like search engines play vital role of people who search seeking related web document in the over the Internet. In the sense lot of new algorithms are proposed for due to the complexity retrieval web documents in clusters. Because it’s very challenging while retrieval web documents cluster based data. When the same web document or related web document may found in different web clusters in the sense search engine take more complex retrieve exact query user input. In addition the queries placed by immature users also make retrieval even more difficult job. To wrap all those concerns, a general search engine carries millions of documents in an answer to queries that may be specific or vague. This huge set of search results cannot be clutched by a user. And more importantly, the preferred information may not be there in the first display of the search engine window. Moreover the first few results are conquered by commonly search pages or documents.

Suppose user search the word crocodile for the instance web search that query a conventional search engines return the pages are related “crocodile Brand” irrespective to the user query. Crocodile query may imply multiple thinks animal or brands some others.

Hence a user has to traverse the search result, which that users seldom traverse all the pages of search result. Their searching is limited to only first few pages of the search engine. Therefore a method is needed to organize this gigantic search result. A way to organize this result set is to cluster them into thematic groups. 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. There are some clustering engines too which acquire search result by actually employing their own searching method. 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. cluster formation and labeling may go hand in hand. The objective of search result clustering is to provide users easiness in locating their information need rather than improving the ranking of documents in the result. A most common approach to cluster web search result is to cluster document snippets returned with each URL in the result.

A document is usually represented as a feature vector, which can be viewed as a point in the multi-dimensional space, to match users’ queries. 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. When the documents provide imprecise information, the use of fuzzy set theory is advisable. Fuzzy c-means and fuzzy hierarchical clustering algorithms were deployed for document clustering. 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.

Moreover, the similarity measures and bag of words were the main limitation of those methods to capture the semantics in the collection of documents. Based on Vector Space Model the similarity between two documents is measured with vector distance, such as Euclidean distance, Manhattan distance, and so on. These
methods do not take contextual meaning into consideration. Ontology-based fuzzy document clustering schemes were used to cluster documents with a limited subsets of selected terms based on the defined ontology. Those methods restrict the application domains, which makes them difficult to be generalized if the domain does not have a proper ontology. Some approaches consider the co-occurrence of terms, but neglect whether terms co-occur in the same context. For example, two terms ‘Wall’ and ‘Street’ do not represent a meaningful theme, say ‘Wall Street’ if these are located at different places in a document. Moreover, polysemy and other important issues are also ignored. The geometric aspect is latent semantics from the given set of documents. Basically, features (keywords or noun phrases) are taken as vertices and the associations (frequent co-occurring features) among those features are taken as links to represent semantic clusters, which are simplicial complexes in geometry. 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. Advocated using part-of-speech and Word Net lexicon to keep the terms in accordance to document subjects or titles, and then perform document clustering using Word Net hierarchical structures.

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. Initially, the dataset is partitioned into windows of fixed size based on the HMM log likelihood values. This provides a framework for identifying the most appropriate number of clusters (windows of varying sizes). After determining the number of clusters, the data values are then labelled and allocated to clusters. The algorithm is tested using a number of benchmark datasets. The proposed algorithm for both small and large datasets (KDD 1999 Intrusion Detection dataset) performed significantly better.
compared to other commonly used clustering algorithms.

In general the search results returned using any searching paradigm are not clustered automatically. But as the case is documents returned for a keyword may be of different nature depending upon the different meanings of the keyword. That is to say that the set of documents returned for a given keyword may further be subdivided into subsets of documents conveying similar sense of the keyword. Clustering the set of results will do this further subdivision and will present the results in a better way. It organizes the documents in such a way that the documents belonging to a group (cluster) are more similar to each other than to the ones which are a part of a different subgroup. Web mining has fuzzy characteristics, so fuzzy clustering is sometimes better suitable in comparison with conventional clustering.

There are two basic methods of fuzzy clustering, one which is based on fuzzy c partitions, is called a Fuzzy C-Means (FCM) clustering and another based on the fuzzy equivalence relations, is called a Fuzzy Equivalence Clustering. Data mining technique called association analysis, which is useful for discovering interesting relationship hidden in large data set also useful for clustering. There are two broad principles use for association analysis. One is Apriori and another is Frequent Pattern (FP) growth principle. FP-growth is a divide and conquer strategy that mines a complete set of frequent itemsets without candidate generation. FP-growth outperformance Apriori because Apriori incurs considerable I/O overhead since it requires making several passes over the transaction data set. In this paper a method is being proposed of web document clustering based on FP-growth and FCM that helps the search engine to retrieve relevant web documents needed for any user. Documents in the FCM are strongly correlated; however traditional FCM clusters are sensitive to the initialization of membership matrix and center. It also needs the number of clusters to be formed as initial parameter. This approach handles all this by using FLSC approach.
2. BACKGROUND AND LITERATURE SURVEY

2.1 LITERATURE SURVEY

The exploratory research and systematic literature review were used to analyze and apply the results of other research. Various methods of statistical analysis, operation research, data mining, and image processing techniques were applied. Experimental research was used to evaluate the proposed methods and compare them with alternative approaches.

Clustering can harvest either disjoint or covering partitions. In a covering partition, it is possible for a document to perform in multiple clusters whereas in disjoint clustering, each document appears in precisely one cluster. Document Clustering is dissimilar than document classification. In document classification, the classes (and their properties) are known a priori, and documents are allocated to these classes; whereas, in document clustering, the number, properties, or attachment (composition) of classes is not known in advance. Thus, classification is an illustration of supervised machine learning and clustering that of unsupervised machine learning. According to, document clustering is separated into two major subsections, hard clustering and soft clustering. Soft clustering also known as overlapping clustering is again distributed into partitioning, hierarchical, and frequent item set-based clustering. Hierarchical methods are categorized into agglomerative methods and disruptive methods. In an agglomerative method, each object forms a cluster. Two most parallel clusters are joined iteratively until some termination measure is fulfilled. Thus, it follows bottom up approach. Whereas, in a divisive method top-down approach is there; i.e., from a cluster containing of all the objects, one cluster is selected and divided into smaller clusters recursively until some termination criterion is fulfilled. The major decision criteria, at every step, are to discover which cluster to split and how to accomplish the split.

The bisecting K-means, a variant of K-means, is a divisive hierarchical clustering algorithm. The algorithm recursively chooses the largest cluster and uses the simple K-means algorithm to divide it into two sub-clusters until the preferred number
of clusters is formed. Hierarchical clustering gives superior quality clustering, but is inadequate because of its quadratic time complexity.

Whereas, partitioning methods like K-means and its variants have a linear time complexity, making it more appropriate for clustering huge datasets, but are thought to create inferior clusters. Also, the major problem with K-means is that it is sensitive to the assortment of the initial partition and may converge to local targets. Regular item set-based: These methods use common item sets generated by the connotation rule mining to cluster the documents. Also, these approaches reduce the dimensionality of term features proficiently for very large datasets, thus advances the accuracy and scalability of the clustering algorithms. Another benefit of frequent item set based clustering method is that each cluster can be considered by the attained frequent item sets shared by the documents in the identical cluster.

These methods comprise Hierarchical Frequent Term-based Clustering (HFTC), Hierarchical Document Clustering Using Frequent Item sets (FIHC), and Fuzzy Frequent Item set-based Document Clustering (F2IDC). HFTC method reduces the overlap of clusters in terms of shared documents. But the experimentalations of Fung et al. showed that HFTC is not scalable. For a huge datasets in FIHC algorithm is given where recurrent item sets derived from the association rule mining are used to build a hierarchical topic tree for clusters. FIHC uses only the global frequent items in document directions, which significantly reduces the dimensionality of the document set. Thus, FIHC is not only scalable, but also perfect. In F2IDC fuzzy association rule mining is united with WordNet. A term hierarchy created from WordNet is useful to discover generalized recurrent item sets as candidate cluster labels for grouping documents. The produced clusters with conceptual labels are easier to understand than clusters marked by isolated terms for classifying the content of individual clusters.

2.2 SOME RECENT WOKS
Algorithm Proposed: Prediction at Prefetch.
Author Name: Narendra S.Chaudhri and Avisheek Chosh.
**Paper Name:** Feature Extraction Using Fuzzy Rule Based System.  
**Year:** 2011.

This approach authors proposed two-dimensional scatter plots of the projected data which helps in easier visual assessment. The scatter plots for the IRIS data from each algorithm as can be easily seen, Sammon’s algorithm works effectively for Iris as the size of the data set is relatively small. Structure preservation is achieved as there is slight overlap between the two classes and the other is distinctly separate. For the fuzzy rule based model the results are similar to that of Sammon’s algorithm and the projection is good. Fig. 3 displays the two dimensional scatter plots for the New Thyroid data set. The results from Sammon’s algorithm and fuzzy logic approach are nearly similar. Thus it is evident that the fuzzy rules extracted using the MA model provide good generalization capability. However for the diabetes data set the results from the fuzzy model are comparatively better than the Sammon’s algorithm as can be seen from Fig. 4. Though there is significant overlap among the two classes the fuzzy model does a better job at classifying them. Hence for larger data sets, not only does the fuzzy model give a better. Data projection is an important tool in exploratory data analysis. Sammon’s non linear projection method lacks predictability and is ineffective for large data sets. To introduce predictability we implement an extension of Sammon’s algorithm using fuzzy logic approach. The implemented algorithm is tested with a few datasets and is found to have good predictability and works well with large datasets.

In this method achieve a low cost data projection algorithm with prediction ability has been successfully implemented in the .NET framework. The algorithm integrates the structure preserving ability of the Sammon’s algorithm and the generalization capability of rule based fuzzy systems. Different data sets from the UCI Repository were used and the comparison of the results makes it evident that the fuzzy model is significantly better than the Sammon’s method as it has additional features of predictability and reduced computational overhead.
Algorithm Proposed: Sammon’s algorithm for Fuzzy Clustering.

Author Name: Sachin Ashok Shinde and Seema Singh Solanki.


The fuzzy clustering is nothing but set of theory and is suitable to handle problems with the vague and boundaries of the clusters. In the fuzzy clustering each object is assigned to belong to handle all of the clusters with certain degree of membership. The fuzzy clustering is superior to crisp clustering when the boundaries among the cluster are vague and ambiguous. The fuzzy clustering is synthesis between clustering and fuzzy set theory. Our proposed approach attempts to resemble to unsupervised issues of supervised clustering. The FRBC is fuzzy rule clustering based algorithms starts to repeatedly extract the entire cluster in problems pattern. To extract each cluster the FRBC considered all the unlabeled data patterns of problems as main data have to make labeled to that data. Cluster membership is a matter of degree of the relevant cluster. The Data have divided into two parts, This tells about the computing process of data, Which shows the adjacency matrix A and form the diagonal matrix D. then it makes to representation of Eigen values of A’ form of matrix. Then it converts cluster to initial rules by using the optimize rule parameter and generate rule base. Finally it builds the fuzzy rules based classifier. There is number of basic types of fuzzy clustering which is shown as below.

A cluster is collection of data objects that are similar to one another with the same cluster and not similar with other cluster. Also it is study on fuzzy rule based clustering development novel. So this can say about the clustering, it is nothing but grouping of set of physical objects into the classes of similar objects. Up to yet the cluster never gets identified by the human directly but it was possible for the machines or system to identify cluster easily as per the requirements of dataset. The cluster which is fuzzy in nature is difficult to understand. The most limitations of fuzzy and crisp clustering algorithm are there sensitivity to number of potential cluster and their initial position. The clustering is not easy to understand for
the human up to yet. These will be the ideas behind concept of this fuzzy clustering to make it possible understand to the human, And also to make the crisp and boundaries easy for the cluster. The idea behind this is developments of rule based algorithm for human to understand of the cluster. The accuracy of the finding cluster should be maintained. This will be another attempt to make it possible.

**Algorithm Proposed:** K-means Clustering and C-means clustering methods.

**Author Name:** Faraz Zaidi.

**Paper Name:** “Fuzzy Clustering and Visualization of Information for Web Search Results”.

**Year:** 2012.

Author presents a decomposition method to identify the important keywords in the co-occurrence network. The method exploits the fact that nodes having high degree are responsible for keeping large size networks as a single connected component. This fact can be used to identify, what we call themes or subjects around which the different web pages are organized & have calculated the cluster centroids, the remaining algorithm to generate a fuzzy clustering is quite simple. First we associate all the remaining nodes in the co-occurrence network by assigning them to one of the centroids. The member that the edges between any two nodes are weighted by the number of documents that appears together. This weight can be calculated for a node and a centroid by adding the weight of all the edges between a node and the nodes of a centroid. As a result of this association, we generate a hard and partitioned clustering for the network. We then calculate for each document in the collection its degree or relevance to these centroids giving us a fuzzy clustering where a document can belong to more than one centroid. To generate a hierarchical clustering, we run the algorithm for different values of cutoff where at each level, only the nodes belonging to a cluster are considered rather than the whole network.

Searching for information on the web is a common task. Often information on the web is distributed, semi-structured, overlapping and heterogeneous. Organization and Visualization of this information is an active area of research where the goal is to help users locate required information in web pages efficiently.
The most widely used data organization technique is clustering. This paper introduces a new clustering algorithm to organize web pages, and a visualization method which facilitates users to search information efficiently from the web. The algorithm presented is a hierarchical fuzzy clustering algorithm which uses domain knowledge to determine input parameters as opposed to other existing algorithms in the literature. The comparative results show that the algorithm performs as well as existing algorithms. Next, we present a methodology to visualize the clustered collection of documents and their contents such that users can visually explore data and extract information. A detailed example is presented to demonstrate various views to visualize clusters, documents and the keywords present in the web pages.

3. PROBLEM IDENTIFICATION AND DEFINITION

3.1 EXISTING PROBLEMS

Information retrieval process deals with information access, representation, storage and organization of information. The objective of an information retrieval system is to enable users to find relevant information from an organized collection of documents. The criteria for complete success are the retrieval of all relevant information items stored in a given system and the rejection of all the non-relevant ones.

David Grossman et al. (2004) discussed the importance of document-clustering to group the documents based on the contents reducing the search space for the given query. The keyword based methodologies to cluster the documents are not convenient since they do not capture the semantic structure of documents. Moreover, the keyword based methodologies for document-clustering are not effective. To overcome the problems faced by the keyword based methodologies document-clustering is performed by combining ontology with optimization techniques like PSO. The document-clustering approaches use vector space model (VSM) for document representation. But using VSM ignores the semantic relatedness among documents. For example, having “Fruits” in one document and “Apple” in another document does not contribute to similarity measurement unless semantic relatedness is
considered. Semantic relationship is not included in most of the clustering approaches.

According to Maedche et al (2003) use of ontology provides good background knowledge and improves document-clustering. Recent works have shown that ontology is useful to improve the performance of text clustering in these situations. Currently, a challenge when querying information using semantics offered by ontology is how to extract information from ontology more efficiently (Aleman-Meza et al 2003). Semantic annotation is about assigning to the entities in the text links to their semantic description (Maedche et al 2003). Annotation provides additional information about web contents so that better decision on content can be made. Annotation of ontology tells us what kind of property and value types should be used in describing a resource. The usage of domain ontologies is used for annotation. Information retrieval system does not actually retrieve information but rather documents from which the information can be obtained if they are read and understood. Figure 3.1 shows the basic concepts of an information retrieval system.

3.2 Existing system

Fuzzy c-means and fuzzy hierarchical clustering need prior knowledge about ‘number of clusters’ and ‘initial cluster centroids’.

In conventional clustering all objects either belongs to a category or not at all. But this is not always true because exception is there. With the use of fuzzy clustering, the documents can belong to more than one domain topic, represented by one group, with varying degrees of relevance. The relevance of documents with relation to groups can be represented by means of linguistic terms, which resembles in a more appropriate way the indication of importance given by human beings. For example, a document can belong “very much”, or a “bit” to a particular group/topic or a topic can be “very important”, or “minor” for the user’s query.

3.3 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.

4.1 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. LSC assumes that words that are close in meaning will occur in similar pieces of query. 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.

The use of fuzzy logic has been widely studied in clustering problems, and the results have also been extended to handle high dimensional data. The importance of clustering to Web mining, especially in the domains of Web Content and Web Usage mining, make Web clustering an interesting topic of research. This includes
clustering of Web documents, snippets and access logs. Usually the Web involves overlapping clusters. So a crisp usage of metrics is better replaced by fuzzy sets which can reflect, in a more natural manner, the degree of belongingness/membership to a cluster.

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

\[ F(U,V) = \sum_{i=1}^{C} \sum_{j} u_{ij}^m \text{distance}(v_i,x_i) \text{ subject to } \sum_{j} u_{ij} = 1 \text{ for all } j. \]

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 \( 0 \leq nj = 1 \) \( U_{ij} \in [0,1] \) 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.

The Information Retrieval system retrieves documents based on a given user query. Both the documents and in most cases, the queries, are instances of natural language. Natural language is often vague and uncertain. It is difficult to judge something that is vague and uncertain with deterministic crisp formulas and/or crisp logical rules.

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.

Web Document clustering methods create classes so that documents in one class are more similar to each other than to documents in other clusters. The theoretical foundation that document clustering may improve retrieval performance was formulated approach to document classification is relatively new, additional research is needed in this
area. Therefore this work focuses on the evaluation of fuzzy-clustering methods in web document retrieval.

4.1. PROPOSED SYSTEM ARCHITECTURE.

4.2. 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.

5. RESULTS AND DISCUSSIONS

Here this approach is taken three fuzzy based cluster algorithms to retrieve the web documents and compare with FLSC algorithm. Our FLSC algorithm used to consider new clusters based on the similar web documents found in the different clusters. Once Data admin upload the
web document into the data server its created a new cluster and storing into the data server whether user given query search all the cluster and retrieve the web documents but it take time to retrieve the web documents and also it is not retrieve effectively to the user queries.

5.1 DISCUSSION

FLSC is an iterative algorithm. The aim of FLSC is to find clusters which very similar web documents found in the different clusters that minimize a dissimilarity function. In Fuzzy clustering each member is associated some membership value, that indicate the strength of association between a data element and a particular cluster. FLSC find clusters centers that minimize a dissimilarity function. FLSC iteratively moves the cluster centers to the “right” location within a dataset. Fuzzy set allows for degree of membership A single point can have partial membership in more than one class. There can be no empty classes and no class that contains no data points.

FLSC iteratively moves the cluster centers to the “right” location within a dataset. FLSC is an algorithm that has objective function that is robust in nature. In other words, a single outlier object could lead to a very unintuitive clustering result. To overcome this FCMedd is being there. This algorithm is robust in nature because the performance is not affected by the presence outliers.

Taking into account the enormous latent of application of soft computing to web mining, this Research is timely and suitable. In our Research summarized the dissimilar types of web mining and its basic components, along with their current states of art. The limitations of the existing web mining methods/tools are explained. The significance of soft computing and consequence of fuzzy logic is already discussed. The possible prospect directions of using FL for
some of these tasks are given in detail. Last, the use of Hard C-means (HCM), Fuzzy C-means (FCM) and Fuzzy C-Mediods (FCMedd) clustering are discussed in detail in next. Fuzzy clustering, 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.

5.1 COMPARISON BETWEEN SOME FUZZY BASED CLUSTERING ALGORITHMS

Fig 5.1. Accuracy between Fuzzy based Clustering Algorithms.

The human generated summary as a gold summary standard for assessment that has turn out to be standards of automatic assessment of summaries. It compares the summaries generated by our method with the human generated (gold standard) summaries. For comparison, our valuation was done using accuracy percentage which was found to have the highest correlation with human judgments, namely, at a confidence level of 85%. It is claimed that our summary correlates highly with human assessments and has high recall and precision significance test with manual evaluation results. So we choose precision, recall as the measurement of our experiment results.

Fig 5.2 Combinations between Clusters.

In conventional clustering all objects either belongs to a category or not at all. But this is not always true because exception is there. With the use of fuzzy clustering, the documents can belong to more than one domain topic, represented by one group, with varying degrees of relevance. The relevance of documents with relation to groups can be represented by means of linguistic terms, which resembles in a more appropriate way the indication of importance given by human beings.
Above Graph shows ability of combined between the clusters on different clustering algorithm. In FLSC Based algorithm resembles new cluster based to combine of same domain name web documents presents in different clusters.

6. SCREENSHOTS

**Fig 6.1 Form of new clustering**

<table>
<thead>
<tr>
<th>Domain/Name</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>1</td>
</tr>
<tr>
<td>Artificial Intelligence</td>
<td>4</td>
</tr>
<tr>
<td>Data Mining</td>
<td>7</td>
</tr>
<tr>
<td>Data Structure</td>
<td>3</td>
</tr>
<tr>
<td>How to Study</td>
<td>1</td>
</tr>
<tr>
<td>Mobile Computing</td>
<td>2</td>
</tr>
<tr>
<td>Neural Networks</td>
<td>5</td>
</tr>
<tr>
<td>Ties</td>
<td>1</td>
</tr>
</tbody>
</table>

**Fig 6.5 User Search Page**

7. 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 overwhelmed 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.

8. FUTURE ENHANCEMENT

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. 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.

REFERENCES


