DOCUMENT SIMILARITY EVALUATION USING A FUZZY CLUSTERING APPROACH

M. Umamaheswari,

Assistant Professor in CSE,

KSR College of Engineering, Tiruchengode-637 215.

umadeena@gmail.com

A. Viswanathan,

Associate Professor in CSE,

KSR College of Engineering, Tiruchengode-637 215.

professorvichu@gmail.com

S. Anguraj

Assistant Professor in IT

K.S.R. College of Engineering, Tiruchengode-637 215

anguangu@gmail.com

M. Sathya,

Assistant Professor in CSE

K.S.R College of Engineering, Tiruchengode -637215

sathimanogaran@gmail.com

ABSTRACT

The Vector Space Model and other techniques to document clustering rely on single term analysis of the document data set. In these circumstances, it is particularly important to use more useful criteria to categorise documents more accurately, such as phrases and their weights. A taxonomy of documents may be constructed, automated document categorization, grouping search engine results, and other uses for document clustering that are particularly advantageous. Because of this, the Fuzzy Clustering method is better at producing the intended results. Our research presents the key idea behind efficient Fuzzy document clustering. The first element, the Document Index Graph, is a document index design that enables steady construction of the index for the document set while putting a focus on efficiency as opposed to relying solely on single-term indexes. It provides efficient phrase matching, that can be used to determine how similar two documents are. This model is adaptable in that it can go back to a compact version of the vector space model if we don't index phrases. Two computational models are applied in both phases: the Gaussian Mixture Model and Expectation Maximization. These two elements work together to create a robust and reliable document similarity computation model, which produces far better Web document clustering results than previous methods.

1. INTRODUCTION

A data collection seems to be divided into clusters during clustering such that each cluster contains data that, in theory, share some characteristic, such as closeness to other clusters as determined by a predetermined distance metric. Data mining, machine learning, pattern recognition, image processing, and bioinformatics all use data clustering as a statistical data processing technique. The computational task of grouping a data collection into k clusters is referred to as "k-clustering."

Data clustering (or simply clustering) has several names that have similar meanings, including clustering methods, automated categorization, numerical taxonomy, and typological evaluation.

Each data point can be a member of many clusters when using fuzzy clustering, also known as soft clustering or soft k-means. Assigning data points to clusters in a way that makes items in the same cluster as similar as feasible and items in separate clusters as dissimilar as possible is known as clustering or cluster analysis. Through the use of similarity metrics, clusters are found. Distance, connectedness, and intensity are some of these similarity metrics. Depending on the data or the application, several similarity metrics may be selected.

In order to make clusters out of a particular document set that are more similar to one another than clusters in other sets are clusters in other clusters, document clustering divides the documents into groups in an unsupervised manner. It is a tool for several information retrieval tasks, such as the effective browsing, organising, and summarising of enormous volumes of text. The goal of cluster analysis is to classify patterns into clusters depending on how similar they are. Clustering has been used to tackle issues in a variety of fields, including mathematics, computer science, statistics, biology, and economics.

2. RELATED WORKS

The vector space model and other document clustering techniques rely solely term examination of the document data set. In these situations, more information qualities, such phrases and their weights, are especially crucial to get more precise document classification. A taxonomy of documents may be constructed, automatic document classification, grouping search engine results, and other uses for document clustering are all extremely advantageous. In this article, we'll talk about two crucial components of document clustering (Hammouda et al. 2004). A cluster overlap rate may be calculated using Wang and Sun's (2004) novel theory for cluster overlap, which turns out to be a very accurate measure of cluster similarity.

Based on this metric, they create a new Fuzzy image segmentation method that addresses the problem of determining the optimal number of clusters in part. The experimental results show the utility of the novel algorithm. A novel cluster overlap theory developed by Wang and Sun (2004) enables the computation of a cluster overlap rate, which turns out to be a reliable indicator of cluster similarity. They developed a new Fuzzy image segmentation method based on this measure that partially solves the issue of choosing the right number of clusters. Experimental findings support the proposed algorithm's usefulness.

3. METHODOLOGY

CHALLENGES IN DOCUMENT CLUSTERING

- High dimensionality: Each unique word in the document collection is a dimension. There might be 1520,000 dimensions as a consequence. Due of this high dimensionality, many existing clustering algorithms lack scalability and efficiency. This has been explained in the sentences that follow. substantial data: Processing tens of thousands to hundreds of thousands of documents is required for text mining.
- Why High accuracy on a consistent basis: Current strategies could work well for some document sets but not others.A useful cluster description is essential for the end user. Navigating should be simpler thanks to the resulting hierarchy.

EXISTING SYSTEM

The next frequent item set that best describes the next cluster is chosen by HFTC greedy in order to minimise overlap between documents that include both the item set and some other item sets. In other words, the order in which the item sets are selected, which is controlled by the greedy heuristic, determines the clustering outcome. This method doesn't choose clusters in a particular order. Instead, we group documents into the most pertinent cluster.

PROPOSED SYSTEM

This section describes about proposed work in detail.

Flow of the proposed work

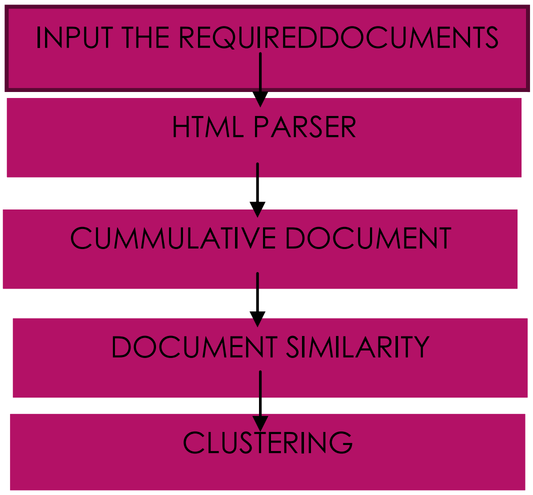


Figure 1: Design Layout

.

(1) The specifications for the weights are being set up.

(2) Using the EM approach to estimate their means and covariance.

(3) Dividing the data into classes based on the weights assigned to each class and the probability density values for each class.

(4) Continue doing step 1 until either the cluster number achieves the required value or the highest OLR drops below the predetermined threshold value. Print the output before moving on to step 3. This method's distinguishing characteristic is that it measures cluster similarity using the overlap rate.

The process of turning a theoretical concept into a functional system is known as project implementation. It may be viewed as the most important step in assuring the success of a new system and providing the user confidence that the system will operate and be successful as a consequence. The implementation process requires careful planning, study into the current system and its limitations on implementation, the development of changeover techniques, and evaluation of those methods.

Pre-processing stages for documents

- Tokenization: The process of treating a document as a string (or collection of words) and then dividing it into a list of tokens.

- Eliminating stop words: Stop words are inconsequential, often recurring words. The stop words are removed in this phase.

• Stemming word: In this phase, tokens are combined into their fundamental forms (connection -> connect).

Depiction of documents

- Tokenization: The process of first processing a document as a string (or as a group of words) and then breaking it into tokens.
- Removing stop words: Repetitive, meaningless words are known as stop words. In this stage, the stop words are eliminated.

The method of stemming words involves conflating tokens to their root form (connection -> connect).

Making use of term weights

- Inverse Document Frequency; Term Frequency.
- Determine the weighting for the TF-IDF.

Comparing the resemblance of two papers

To quantify how similar two publications are, the cosine similarity measurement is employed. Calculating the cosine of the angle between two document vectors yields their cosine similarity.

HTML Parser

- The first step when a document enters the process state is parsing, which is the separation or identification of meta tags in an HTML page.
- Here, the whole tree structure's nodes are parsed from the raw HTML file that was read.

Cumulative Document

- The cumulative document is made up of all of the documents' meta-tags and contains them all.
- In the input base document, we look for references (to other pages), then we read other documents to look for references, and so on.
- As a consequence, all of the documents' meta-tags are found beginning with the base document.

6.1.4 Document Similarity

- To assess the degree of similarity between two publications, the cosine-similarity measure approach is applied.
- The TF-IDF measure of the words (meta-tags) in the two publications is used to calculate the cosine-similarity weights.
- The phrase weights are calculated to achieve this.TF = C / T
- IDF = DO / DF.

  DO→ quotient of the total number of documents

DF → number of times each word is found in the entire corpus

C → quotient of no of times a word appears in each document

T → total number of words in the document

TFIDF = TF * IDF

Clustering

- Clustering is the division of data into sets of related elements.
- Although the procedure is made simpler, representing the data with fewer clusters leads in the loss of certain fine features.
- The documents are grouped together in a cluster if their cosine similarity measure falls below a predetermined threshold.

WORKING

Divisive Fuzzy clustering's fundamental method is as follows: If there are N items in the collection, and there is a N*N distance (or similarity) matrix:

STEP 1: Assign each item to a cluster such that, in the case of N items, there are N clusters, each containing a single item. Allow the distances (similarities) between the items in the clusters to match those between the clusters.

STEP 2: Finding the closest (most similar) set of clusters using oh tf - itf and merging them into one cluster results in one less cluster.

STEP 3: Calculate the separations (similarities) between the new and old clusters in step three.

Repeat steps 2 and 3 up to the point where every object has been gathered into a single N-sized cluster. Single-linkage clustering differs from complete-linkage clustering and average-linkage clustering in that Step 3 can be finished in a variety of ways. In single-linkage clustering, take into account that the distance between any two members of one cluster is equal to the smallest distance between any two members of the other cluster (also known as the connectedness or minimal technique).

If the data consists of similarities, take into account that the similarity between two clusters is equal to the highest similarity between any two members of either cluster. In complete-linkage clustering, take into account that the distance between any two members of one cluster is equal to the largest distance between any two members of the other cluster (also known as the diameter or maximum technique).

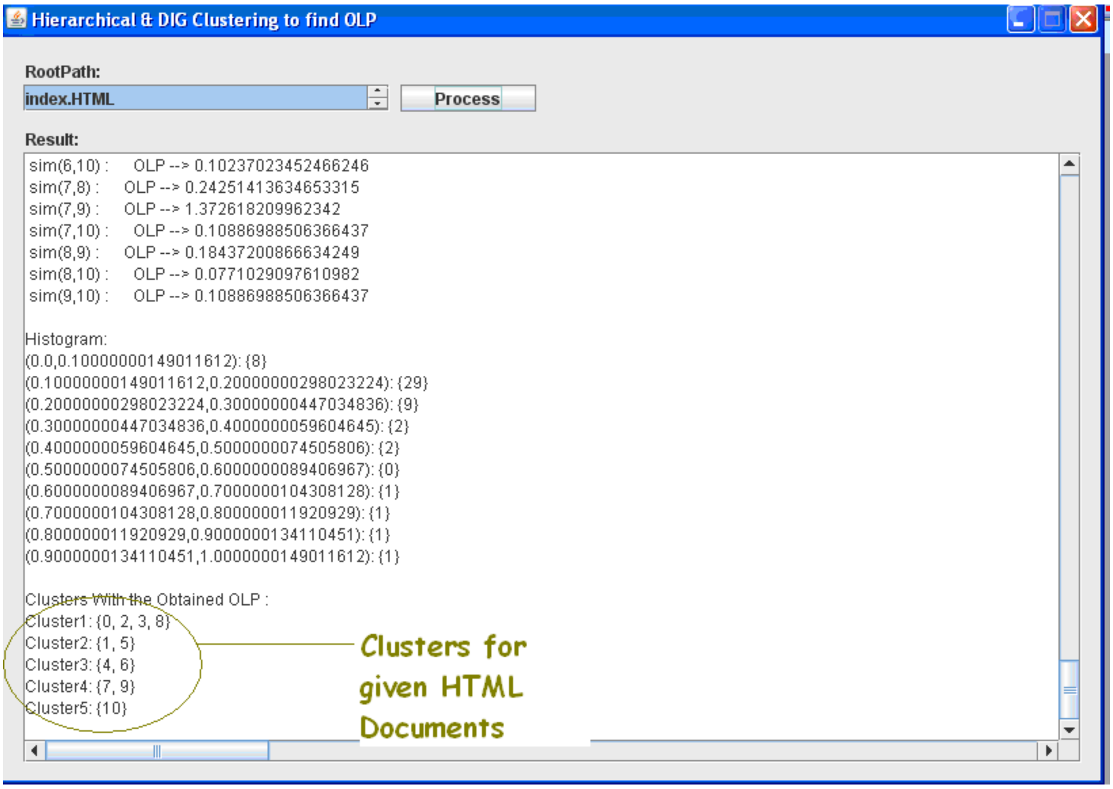
When utilising average-linkage clustering, the distance between two clusters is taken to be equal to the average distance. This divisive Fuzzy clustering method is referred to as agglomerative since it combines groups repeatedly.

Instead of starting with a single cluster and breaking it down into smaller fragments, divvying up Fuzzy clustering works in the other manner. Divisive techniques are rare and have only been applied a few times. Naturally, it is useless to have all N items grouped together into a single cluster, but after the entire Fuzzy tree has been collected and k clusters are needed, the k-1 longest linkages are eliminated.

4. RESULT AND DISCUSSION

CLUSTER FORMATION

The development of clusters is the ultimate step. This is depicted in the diagram below. As a result, Agglomerative Fuzzy clustering was used to cluster the documents, and the causes were reported.



Hierarchical & DIG Clustering to find OLP
Process
Clusters for given HTML Documents

RootPath: index.HTML

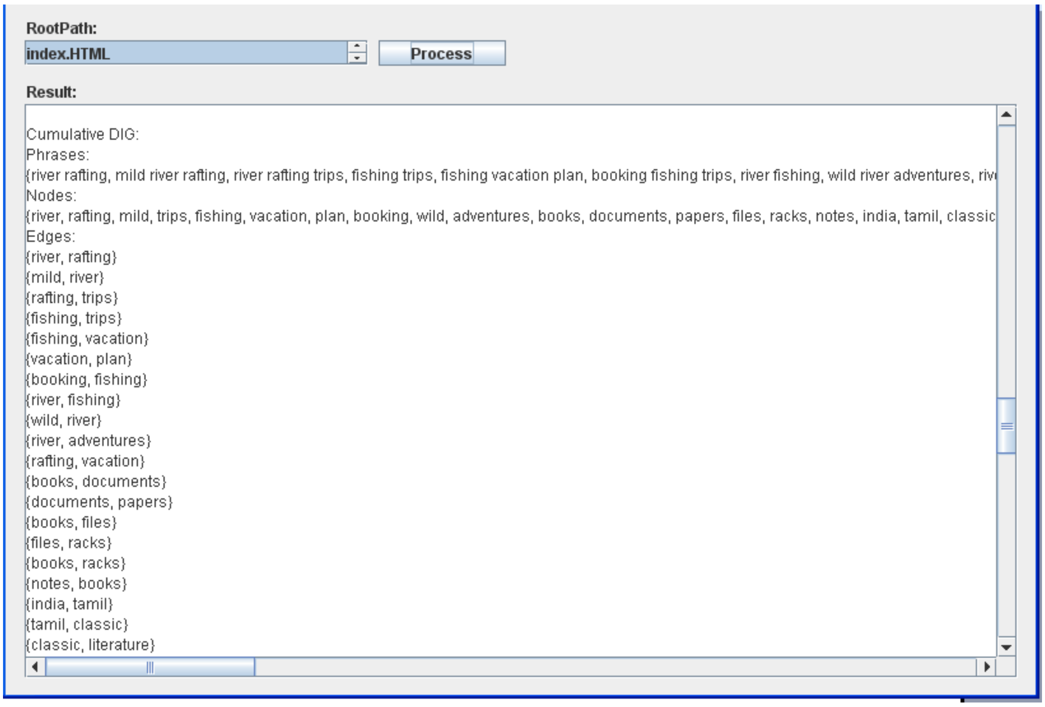
Result:

```
 sim(6,10) :    OLP --> 0.10237023452466246
 sim(7,8) :    OLP --> 0.24251413634653315
 sim(7,9) :    OLP --> 1.372618209962342
 sim(7,10) :    OLP --> 0.10886988506366437
 sim(8,9) :    OLP --> 0.18437200866634249
 sim(8,10) :    OLP --> 0.0771029097610982
 sim(9,10) :    OLP --> 0.10886988506366437

Histogram:
(0.0,0.1000000014901161 2): {8}
(0.1000000014901161 2,0.20000000298023224): {29}
(0.20000000298023224,0.30000000447034836): {9}
(0.30000000447034836,0.4000000059604645): {2}
(0.4000000059604645,0.5000000074505806): {2}
(0.5000000074505806,0.6000000089406967): {0}
(0.6000000089406967,0.7000000104308128): {1}
(0.7000000104308128,0.800000011920929): {1}
(0.800000011920929,0.9000000134110451): {1}
(0.9000000134110451,1.0000000149011612): {1}

Clusters With the Obtained OLP :
Cluster1: {0, 2, 3, 8}
Cluster2: {1, 5}
Cluster3: {4, 6}
Cluster4: {7, 9}
Cluster5: {10}
```

(a)



RootPath: index.HTML

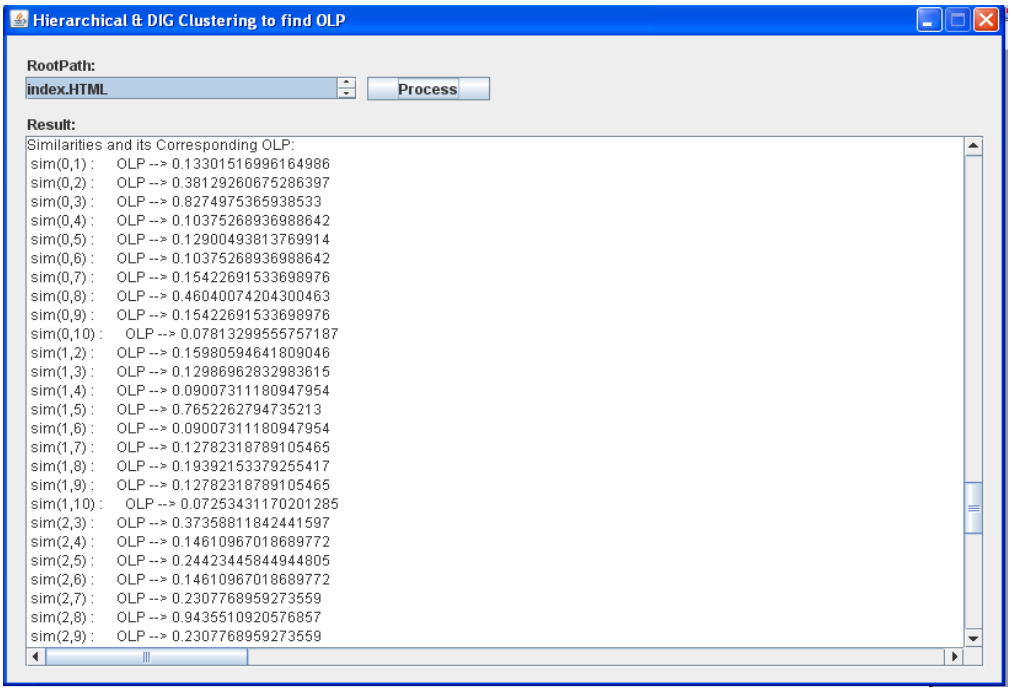
Result:

```
Cumulative DIG:
Phrases:
{river rafting, mild river rafting, river rafting trips, fishing trips, fishing vacation plan, booking fishing trips, river fishing, wild river adventures, riv
Nodes:
{river, rafting, mild, trips, fishing, vacation, plan, booking, wild, adventures, books, documents, papers, files, racks, notes, india, tamil, classic
Edges:
{river, rafting}
{mild, river}
{rafting, trips}
{fishing, trips}
{fishing, vacation}
{vacation, plan}
{booking, fishing}
{river, fishing}
{wild, river}
{river, adventures}
{rafting, vacation}
{books, documents}
{documents, papers}
{books, files}
{files, racks}
{books, racks}
{notes, books}
{india, tamil}
{tamil, classic}
{classic, literature}
```

(b)

Figure 2: Cluster formation

DOCUMENT SIMILARITY AND OLP



RootPath: index.HTML

Result:

```
Similarities and its Corresponding OLP:
 sim(0,1) :    OLP --> 0.13301516996164986
 sim(0,2) :    OLP --> 0.38129260675286397
 sim(0,3) :    OLP --> 0.8274975365938533
 sim(0,4) :    OLP --> 0.10375268936988642
 sim(0,5) :    OLP --> 0.12900493813769914
 sim(0,6) :    OLP --> 0.10375268936988642
 sim(0,7) :    OLP --> 0.15422691533698976
 sim(0,8) :    OLP --> 0.46040074204300463
 sim(0,9) :    OLP --> 0.15422691533698976
 sim(0,10) :    OLP --> 0.07813299555757187
 sim(1,2) :    OLP --> 0.15980594641809046
 sim(1,3) :    OLP --> 0.12986962832983615
 sim(1,4) :    OLP --> 0.09007311180947954
 sim(1,5) :    OLP --> 0.7652262794735213
 sim(1,6) :    OLP --> 0.09007311180947954
 sim(1,7) :    OLP --> 0.12782318789105465
 sim(1,8) :    OLP --> 0.19392153379255417
 sim(1,9) :    OLP --> 0.12782318789105465
 sim(1,10) :    OLP --> 0.07253431170201285
 sim(2,3) :    OLP --> 0.37358811842441597
 sim(2,4) :    OLP --> 0.14610967018689772
 sim(2,5) :    OLP --> 0.24423445844944805
 sim(2,6) :    OLP --> 0.14610967018689772
 sim(2,7) :    OLP --> 0.2307768959273559
 sim(2,8) :    OLP --> 0.9435510920576857
 sim(2,9) :    OLP --> 0.2307768959273559
```

Figure 3: Document Similarity

This comprises analysing document similarity and calculating the Overlapping Rate as a result (OLP Rate).

CONCLUSION

This study suggests a novel divisive Fuzzy clustering technique that uses the overlap rate for cluster merging. According to experiments with general data sets and a document set, the unique technique may reduce the time cost, reduce the space complexity, and boost the accuracy of clustering. Results from the recently suggested technique, in particular, show significant improvements in document clustering. This work might be expanded upon and improved upon in the future in a number of different ways. Using several similarity calculation techniques to improve the precision of document similarity computations is one direction this study may go. Even if the new approach outperforms more established ones, there is undoubtedly room for improvement.

REFERENCES

1) Cole, A. J. & Wishart, D. (1970). An improved algorithm for the Jardine-Sibson method of generating overlapping clusters. The Computer Journal 13(2):156-163.
2) D'andrade,R. 1978, "U-Statistic Hierarchical Clustering" Psychometrika, 4:58-67.
3) Johnson,S.C. 1967, "Hierarchical Clustering Schemes" Psychometrika, 2:241-254.
4) Shengrui Wang and Haojun Sun. Measuring Overlap-Rate for Cluster Merging in a Hierarchical Approach to Color Image Segmentation. International Journal of Fuzzy Systems, Vol.6, No.3, September 2004.
5) Jeff A. Bilmes. A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. ICSI TR-97-021, U.C. Berkeley, 1998.
6) E.M. Voorhees. Implementing agglomerative hierarchical clustering algorithms for use in document retrieval. Information Processing and Management, 22(6):465–476, 1986.
7) Sun Da-fei,Chen Guo-li,Liu Wen-ju. The discussion of maximum likehood parameter estimation based on EM algorithm. Journal of HeNan University. 2002,32(4):35~41
8) Khaled M. Hammouda, Mohamed S. Kamel, efficient phrase-based document indexing for web document clustering, IEEE transactions on knowledge and data engineering, October 2004
9) Haojun sun, zhihui liu, lingjun kong, A Document Clustering Method Based on Hierarchical Algorithm With Model Clustering, 22nd international conference on advanced information networking and applications,
10) Shi zhong, joydeep ghosh, Generative Model-Based Document Clustering: A Comparative Study, The University Of Texas.