ISSN(Online): 2319 - 8753
ISSN (Print) :2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 3, March 2015**

# Visualizing Web Search Results on the Basis of Snippet Similarity

Priyanka Daimary[1], Rachana Das[2], Sangita Kumari[3], Sakthipriya N[4]1

U.G. Student, Department of Computer Engineering, Bharath University, Chennai, Tamil Nadu, India[1, 2, 3]

Assistant Professor, Department of Computer Engineering, Bharath University, Chennai, Tamil Nadu, India[4]

**ABSTRACT***:* Search engines are programs that search documents for specified keywords and return list of documents where the keywords were found. Textual snippet based search enable us to search documents in a better way. Search tasks would be easier if users were given an overview of document clusters in an organized manner so as to show how related they are content wise. In this paper, we introduce a novel multi-viewpoint based similarity measure and two related clustering methods. Using multiple viewpoints, more informative assessment of similarity could be achieved. We also propose a visualization technique to display the results of web queries.

**KEYWORDS***:* multi-viewpoint, visualization technique, search engine

## I. INTRODUCTION

Searching for information on the internet is an important task to many users. The procedure consists in providing textual queries to a search engine, which returns a ranked list of textual snippets each containing a content summary and a link to the referred document. In our proposed, we introduce a novel multi-viewpoint based similarity measure which takes the following three cases into account: a) The feature appears in both documents, b) the feature appears in only one document, and c) the feature appears in none of the documents. For the first case, the similarity increases as the difference between the two involved feature values decreases. Furthermore, the contribution of the difference is normally scaled. For the second case, a fixed value is contributed to the similarity. For the last case, the feature has no contribution to the similarity. The proposed measure is extended to gauge the similarity between two sets of documents. Clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups.Data clustering, K-means clustering and document clustering are the three most important clustering techniques used here. Data clustering is a common technique for statistical data analysis, which is used in many fields, including machine learning, data mining, and recognition of pattern, image analysis, and bioinformatics. The computational task of classifying the data set into k clusters is often referred to as k-clustering. Document clustering aims to group, in an unsupervised way, a given document set into clusters such that documents within each cluster are more similar between each other than those in different clusters.

## II. RELATED WORK

Cole, A. J. & Wishart, D. (1970). An improved algorithm for the Jar dine-Sibson method of generating overlapping clusters. The COMPUTER JOURNAL [1] 13(2):156-163 introduces a new information theoretic divisive algorithm for feature/word clustering and is applied to text classification. Existing techniques for such “distributional clustering” of words are agglomerative in nature and result in (I) sub-optimal word clusters and (ii) high computational cost. A fast and divisive algorithm that monotonically decreases this objective function value is introduced. [1]

D’andrade, R, R. 1978, "U-Statistic Hierarchical Clustering" PSYCHOMETRIKA, 4:58-67[2] presents similarity is an important and widely used concept. Previous definitions of similarity are tied to a particular application or a form of knowledge representation. We present an information theoretic definition of similarity that is applicable as long as there is a probabilistic model. We demonstrate how our definition can be used to measure the similarity in a number of different Domains.

ISSN(Online): 2319 - 8753
ISSN (Print) :2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 3, March 2015**

Johnson, S.C. 1967, "Hierarchical Clustering Schemes" Psychometrika, 2:241-254 [3] deals with certain spherical k-means algorithm for clustering suchdocument vectors. The algorithm outputs k disjoint clusters each with a concept vector that is the centroid of the cluster normalized to have unit Euclidean norm.

Shengrui Wang and Haojun Sun. Measuring overlap-Rate for Cluster MERGING in a Hierarchical Approach to Colour Image Segmentation. International Journal of Fuzzy Systems, Vol.6, No.3, September 2004[4] presents a Concept-based document similarity to compute the similarities of documents based on the Suffix Tree Document (STD) model. By mapping each node in the suffix tree of STD model into a unique feature term in the Vector Space Document (VSD) model, the concept-based document similarity inherits the term ctf (conceptual term frequency), tf (term frequency),df (document frequency) weighting scheme in computing the document similarity with concept.
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[5] Introduces a novel algorithm called Locality Preserving Indexing (LPI) is proposed for document indexing. Each document is represented by a vector with low dimensionality. In contrast to LSI which discovers the global structure of the document space, LPI discovers the local structure and obtains a compact document representation subspace that best detects the essential semantic structure.

E.M. Voorhees Implementing agglomerative hierarchical clustering algorithms for use in document retrieval. Information Processing and MANAGEMENT, 22(6):465–476, 1986[6] Introduces an algorithm which deals with k-centre problems. This algorithm tends to extend to metrics in discrete time problems. A simplex approximation algorithm for L-capacitated k-centre is provided.

Sun Da-fei, Chen Guo-li, Liu Wen-ju. The discussion of maximum like hood parameter estimation based on EM algorithm. Journal of Henan University. 2002, 32(4):35~41 [7] suggests two techniques for feature or term selection along with a number of clustering strategies. The selection techniques significantly reduce the dimension of the vector space model. Examples that illustrate the effectiveness of the proposed algorithms are provided.

Khaled M. Hammouda, Mohamed S. KAMEL, efficient phrase-based document indexing for web document clustering, IEEE transactions on knowledge and data engineering, October 2004[8] presents two improvements of Lazy Learning. Both methods include input selection and are applied to long-term prediction of time series. First method is based on an iterative pruning of the inputs and the second one isperforming a brute force search in the possible set of inputs using a k-NN.
Haojun sun, Zhuhai liu, lingjun Kong, A Document Clustering Method Based on Hierarchical Algorithm with Model Clustering, 22nd international conference on advanced information networking and applications[9] describes several machines learning machine algorithm for this task and promising initial result with a prototype system that has created a knowledge base describing university people, course, and research project.

Shi zhong, joy deepGhosh, Generative Model-Based Document Clustering: A Comparative Study, The University of Texas[10] investigates an online version of the spherical k-means algorithm based on a well-known Winner-Take-All competitive learning

## III. SYSTEM ANALYSIS

### A. Previous work

In the existing System, it greedily picks the next frequent item set which represent the next cluster to minimize the overlapping between the documents that contain both the item set and some remaining item sets. In other words, the clustering result depends on the order of picking up the item sets, which in turns depends on the greedy heuristic. This method does not follow a sequential order of selecting clusters.

### B. A novel Hierarchical algorithm

In this project, the main work is to develop a novel hierarchical algorithm for document clustering which provides maximum efficiency and performance. A novel way to evaluate similarity between documents is made and

ISSN(Online): 2319 - 8753
ISSN (Print) :2347 - 6710

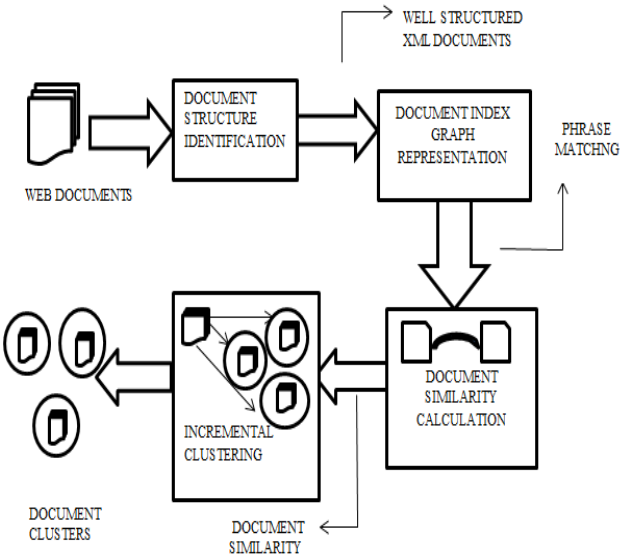
International Journal of Innovative Research in Science, Engineering and Technology

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 3, March 2015**

consequently new criterion is formulated for functions for document clustering. The purpose of this test is to check how much a similarity measure coincides with the true class labels. It is particularly focused in studying and making use of cluster overlapping phenomenon to design cluster merging criteria. Experiments in both public data and document clustering data show that this approach can improve the efficiency of clustering and save computing time. Hierarchical clustering algorithm creates decomposition of the given set of data objects and similarities between clusters are more researched.[9]

**C.SYSTEM ARCHITECTURE**




**The followings are the modules involved in this project. They are**

**A. HTML Parser**

Parsing is the first step done when the document enters the process state.

Parsing is defined as the separation or identification of Meta tags in a HTML document. Here, the raw HTML file is read and it is parsed through all the nodes in the tree structure.

**B.CUMULATIVE DOCUMENT**

The cumulative document is the sum of all the documents, which contain meta-tags from all the documents. We find the references (to other pages) in the input base document and read other documents and then find references in them and so on. Thus in all the documents their meta-tags are identified, from the base document.

**C. DOCUMENT SIMILARITY**

The similarity between two documents is found by the cosine-similarity measure technique. The weights in the cosine-similarity are found from the TF-IDF measure between the phrases (meta-tags) of the two documents.TF-stands for Term Frequency; IDF-stands for Inverse Document Frequency.

• This is done by computing the term weights involved.
• TF = C / T
• IDF = D / DF.

D -> 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**

ISSN(Online): 2319 - 8753
ISSN (Print) :2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 3, March 2015**

**D.CLUSTERING**

Clustering is dividing the data into similar group of objects. If we represent the data by fewer clusters it loses certain fine details, but achieves simplification. The similar documents are grouped together in a cluster, if their cosine similarity measure is less than a specified threshold [9].

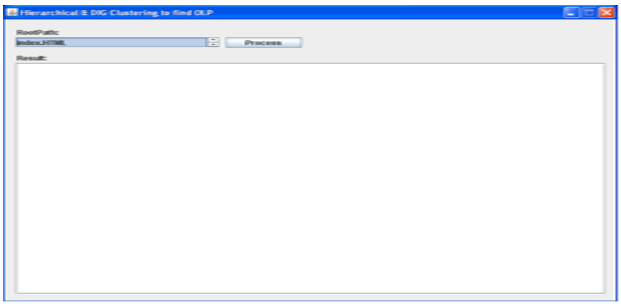
## IV. IMPLEMENTATION AND RESULT

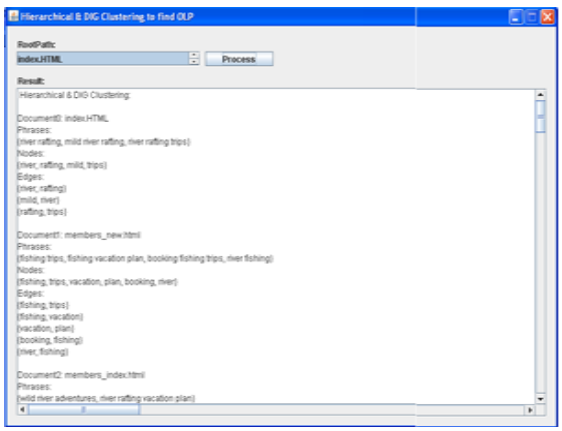
The implementation part of this document Hierarchical clustering involves the following steps.

- Input of HTML documents
- HTML parsing
- finding the Cumulative document with the help of Meta tags
- Finding similarity & OLP
- Clustering based on OLP rate.

**1. HTML DOCUMENTS**

The sample HTML documents are used to find the similarity and clustering efficiency. In this example many such documents are provided



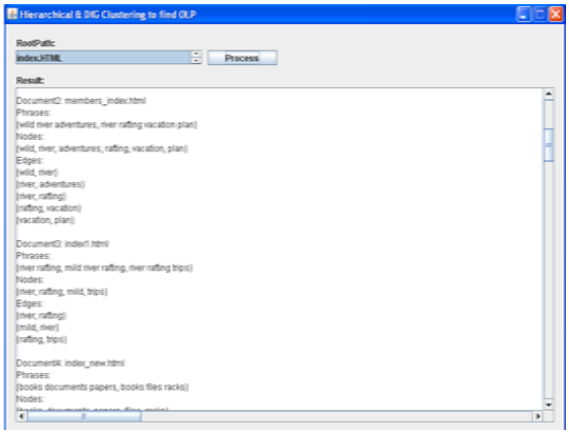



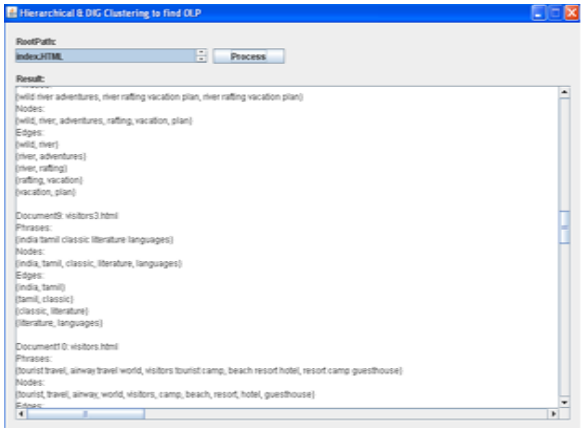
ISSN(Online): 2319 - 8753
ISSN (Print) :2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

*(An ISO 3297: 2007 Certified Organization)*

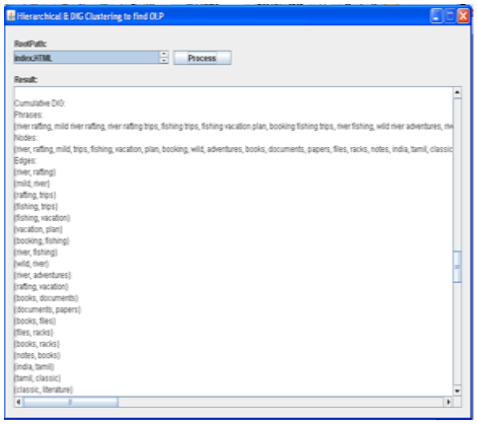
**Vol. 4, Issue 3, March 2015**




## 2. CUMULATIVE DOCUMENTS

The next step is finding the cumulative document. With the help of the documents provided, the edges, nodes and phrases of each document are analysed and listed as shown below.




## 3. DOCUMENT SIMILARITIES AND OLP

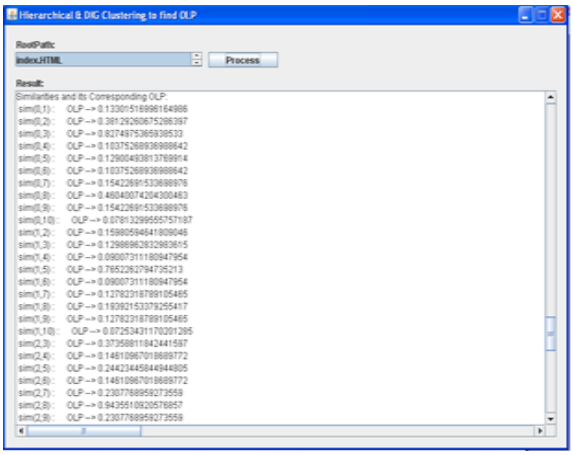
This involves the document similarity analysis and thereby finding the Overlapping Rate (OLP Rate).

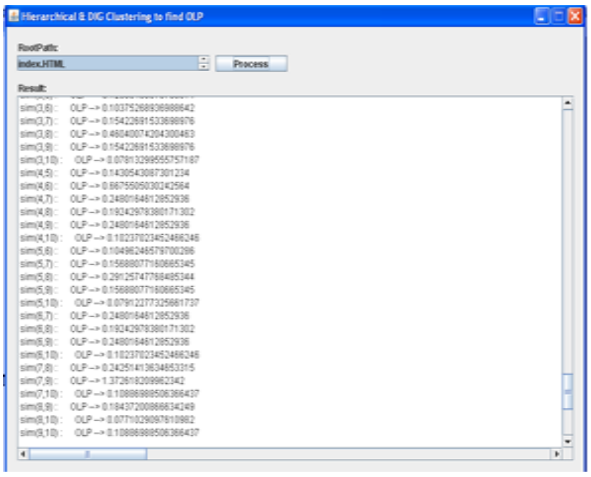
ISSN(Online): 2319 - 8753
ISSN (Print) :2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

*(An ISO 3297: 2007 Certified Organization)*

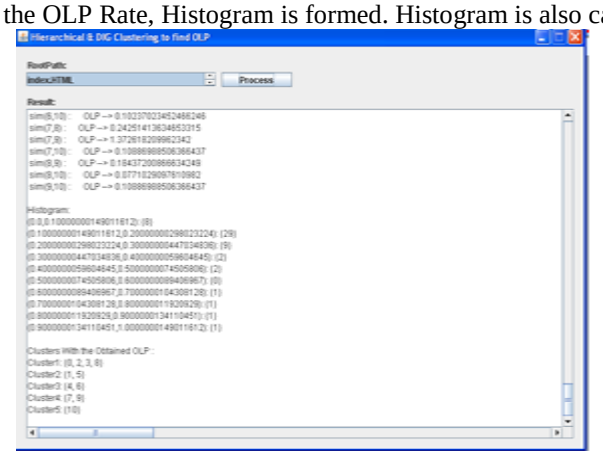
**Vol. 4, Issue 3, March 2015**




## 4. HISTOGRAM FORMATIONS

After finding the similarity and the OLP Rate, Histogram is formed. Histogram is also called as Dendogram.




## 5. CLUSTER FORMATIONS

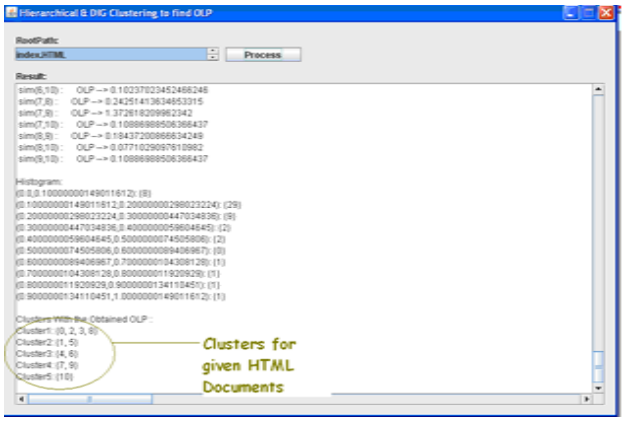
Then the final step is the formation of clusters. This is shown in the below figure. Thus the Document clustering using Hierarchical Clustering is done and the causes are documented.

ISSN(Online): 2319 - 8753
ISSN (Print) :2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 3, March 2015**




## III.CONCLUSION

Given a data set, the ideal scenario would be to have a given set of criteria to choose a proper clustering algorithm to apply. Choosing a clustering algorithm, however, can be a difficult task. Even ending just the most relevant approaches for a given data set is hard. Most of the algorithms generally assume some implicit structure in the data set. One of the most important elements is the nature of the data and the nature of the desired cluster. Another issue to keep in mind is the kind of input and tools that the algorithm requires. This report has a proposal of a new hierarchical clustering algorithm based on the overlap rate for cluster merging. The experience in general data sets and a document set indicates that the new method can decrease the time cost, reduce the space complexity and improve the accuracy of clustering. Specially, in the document clustering, the newly proposed algorithm measuring result show great advantages. The hierarchical document clustering algorithm provides a natural way of distinguishing clusters and implementing the basic requirement of clustering as high within-cluster similarity and between-cluster dissimilarity.

## IV.FUTURE WORKS

In the proposed model, selecting different dimensional space and frequency levels leads to different accuracy rate in the clustering results. How to extract the features reasonably will be investigated in the future work.
There are a number of future research directions to extend and improve this work. One direction that this work might continue on is to improve on the accuracy of similarity calculation between documents by employing different similarity calculation strategies. Although the current scheme proved more accurate than traditional methods, there are still rooms for improvement.

## REFERENCES

1. Cole, A. J. & Wishart, D. An improved algorithm for the Jar dine-Sibson method of generating overlapping clusters. The Computer Journal 13(2):156-163.(1970)
2. D'andrade, R., "U-Statistic Hierarchical Clustering" Psychometrika, 4:58-67. (1978)
3. Johnson, S.C. "Hierarchical Clustering Schemes" Psychometrika, 2:241-254. (1967)
4.Shengrui Wang and Haojun Sun. Measuring overlap-Rate for Cluster Merging in a Hierarchical Approach to colour 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 like hood parameter estimation based on EM algorithm. Journal of Henan University. 2002, 32(4):35~41
8. Khaled M. Hammouda, Mohamed S. Kamal, efficient phrase-based document indexing for web document clustering, IEEE transactions on knowledge and data engineering, October 2004
9. Haojun sun, Zhuhai 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, joy deepGhosh, Generative Model-Based Document Clustering: A Comparative Study, The University of Texas.
11. Erick Gomez-Nieto, Frizzy San Roman, Paulo Pagliosa, Wallace Casaca, Elias S. Helou, Maria Cristina F. de Oliveira, Member, IEEE, and Luis Gustavo Nonato, Member, IEEE: Similarity preserving snippet-based visualization of web search results.