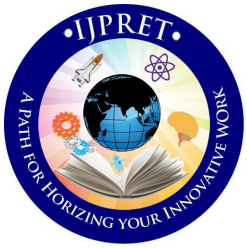
*Research Article* *ISSN: 2319-507X*
*Potnurwar AV, IJPRET, 2013; Volume 1(8): 746-753* *IJPRET*




**INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY**

A PATH FOR HORIZING YOUR INNOVATIVE WORK

# OPTIMIZATION OF DOCUMENT CLUSTERING USING PHRASES

**A.V.POTNURWAR**

**Assistant Prof., P.I.E.T**

**Accepted Date:**

**27/02/2013**

**Publish Date:**

**01/04/2013**

**Keywords**

Clustering

Phrases

**Corresponding Author**

**Mrs. A. V. Potnurwar**

## Abstract

Most of the documents clustering techniques rely on single term analysis of the document data set, such as the Vector space model. More informative features including phrases and their weights are particularly important to achieve more accurate document clustering. Document clustering is particularly useful in many applications such as automatic categorization of documents, grouping search engine results, building taxonomy of documents and others. The motivation behind the work in this paper is that we believe that document clustering should be based not only on single word analysis, but on phrases as well. Phrase based analysis means that the similarity between documents should be based on matching phrases rather than on single words Document clustering is particularly useful in many applications such as automatic categorization of documents, grouping search engine results, building taxonomy of documents and others. The motivation behind the work in this paper is that we believe that document clustering should be based not only on single word analysis, but on phrases as well. Phrase based analysis means that the similarity between documents should be based on matching phrases rather than on single words only Owing to the widely application in the fields of information retrieval, document analysis and information extraction, document cluster analysis has been concerned broadly, and gotten a great deal of research Issues It is one of essential theme in document. Most of the existing techniques for document clustering rely on a “bag of words “document representation. Each word in the document is considered as a separate feature, ignoring the word order. We investigate the use of phrases rather than words as document features for the document clustering. We present a phrase grammar extraction technique, and use the extracted phrases as the features document clustering algorithm1.0

**INTRODUCTION**

An IR system typically produces a ranked list of documents in response to a user's query. These documents are presented to the user for examination and evaluation. Although the documents are ranked, there is significant potential benefit in providing additional structure in long retrieved lists. The role of information organization becomes even more important in the interactive model of retrieval, where the focus is on the user's participation in a cycle of query formulation, presentation of search results, and query reformulation.

Traditional information retrieval systems are useful for retrieval of general documents; however these systems cannot support scenario-specific information retrieval because of:

- The lack of effective techniques to extract key features from free-text for indexing, to identify phrases with similar concepts in free-text.
- The terms used in a query are often mismatched with those from the document containing information on the same scenario.

•Relevance:

• Browsable summaries:

• Overlap:

• Snippet-tolerance:

• Speed

• Incrementality:

The objective of this paper is automatic classification of documents into specific topic areas..

Therefore, at the end of clustering analysis, semantics must be assigned to each class by an analyst. There are methods for document clustering by means of document-to-cluster similarity function (coefficient). Should the value of coefficient exceed threshold value q, the document falls into the pertinent equivalence class.

**2. MOTIVATION**

The motivation for this research is to make search engine results easy to browse. Document clustering algorithms attempt to group similar documents together.

Clustering the results of Web search engines can provide a powerful browsing tool. Document clustering has initially been investigated in Information Retrieval mainly as a means of improving the performance of search engines by pre-clustering the entire. Our modified user-cluster hypothesis is that users have a mental model of the topics and subtopics of the documents present in the result set; similar documents will tend to belong to the same mental category in the users' model. Thus the automatic detection of clusters of similar documents can help the user in browsing the result set. Clustering of search results can help users in three ways:

- to find the information
- faster that a query is poorly formulated) and to reformulate it, it can reduces the fraction of the queries on which the user gives up before reaching the desired information.

We have identified some key requirements:

- Coherent Clusters:
- Efficiently Browsable:
- Speed

**3. CLUSTER INTIALIZATION**

After surveying a range of literature experimented with the following four initialization techniques[16]:

• Random:

• Perturb:

• Marginal:

• KKZ:

Any clustering technique relies on four concepts:a data representation model,a similarity measure,a cluster model, anda clustering algorithm that builds the clusters using the data model and the similarity measure.

**3.1 A SIMILARITY MEASURE**

As mentioned earlier, phrases convey local context information,which is essential in determining an accurate similarity between documents..The phrase similarity between two documents is calculated based on the list of matching phrases between the two documents. From an information theoretic point of view, the similarity between two objects is regarded as how much they share in common. The cosine and the Jaccard measures are indeed of such nature, but they are essentially used as single-term based similarity measures. Lin gave a

formal definition for any information theoretic similarity measure in the form of:

$$sim(x,y) = \frac{x \cup y}{x \cap y}$$

The basic assumption here is that the similarity between two documents is based on the ratio of how much they overlap to their union, all in terms of phrases. This definition still coincides with the major assumption of the cosine and the Jaccard measures, and to Lin's definition as well.

**3.2 CLUSTER MODEL**

Any clustering algorithm assumes a certain cluster structure. Sometimes the cluster structure is not assumed explicitly, but is inherent due to the nature of the clustering algorithm itself. For example, the k-means clustering algorithm assumes spherical shaped (or generally convex shaped) clusters. This is due to the way k-means finds cluster centers and updates object memberships. Also if care is not taken, we could end up with elongated clusters, where the resulting partition contains a few large clusters and some very small clusters. Wong and Fu [13] proposed a strategy to keep the cluster sizes in a certain range, but it could be argued that forcing a limit on cluster size is not always desirable. A dynamic model for finding clusters irrelevant of their structure is CHAMELEON, which was proposed by Karypis et al [10].. A good example of overlapping document cluster generation is the tree-based STC system proposed by Zamir and Etzioni [2]. Another way for generating overlapping clusters is through fuzzy clustering where objects can belong to different clusters to different degrees of membership [5].

**3.2. DOCUMENT ACQUISTION**

The documents are retrieved from several popular search engines, including Google, Teoma, MSN Search, AltaVista. A wrapper, communicating with each search engine, translates the query into an acceptable request to the search engine, submits the request, retrieves snippets, and parses the results to extract noun phrases from each snippet. After obtaining the snippets from individual sources, the system removes duplicates to form an aggregated document set. A relevancy indicator, which is the average of the rank orders of a document in

all sources, used to sort the documents. The Fig 1 below shows

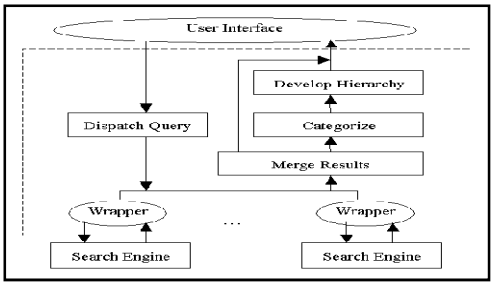



**Fig1 Flow Chart For Document acquisition**

### 3.3 PHRASE EXTRACTION[7]

The phrase extraction is given as follows:

- Make a list of all the bigrams in the input sequence.
- Find a bigram *(wi, w j )* with the maximum association weight *A(wi, w j )* in the list of rules.
- In the sequence, replace all the instances of the bigram ***(wi*** , wj) with a new symbol ***wk*** (taken from the rule). This is now the new sequence.
- for each replaced instance output the expanded phrase.
- Update the list of all the bigrams in the sequence

### 3.4 PHRASE RANKING

Phrase features are used to calculate a score for each phrase. Another method for scoring phrases was used, which is based on individual word weights.

### 3.5COLLECTING PHRASE STASTICS

For each phrase appearing in the cluster, collect the following statistics: document frequency (DF), and term frequency (TF) with respect to the cluster S, the parent cluster P and the general English corpus E. Document frequency of a phrase p with respect to a cluster C, denoted by DFC, is the number of documents in the cluster that contain p. Term frequency of a phrase p in a cluster C, denoted by TFC, is total number of occurrences of p in the cluster We have identified three key requirements for a post-retrieval document clustering system **Relevance**.,**Browsable Summaries, Speed.**

Therefore the clustering system ought to be able to cluster up to one thousand documents in a few seconds

### ALGORITHM

Algorithm implemented in this paper is having following steps:

1. Accept thequery of the user
2. Create an UI to accept the directory which contains all the document
3. After accepting the document, they will be converted into numerical document by using vector array ..
4. The query is express as numerical expression
5. The occurrence of individual words is done in the document.
6. After getting the words the stops words are eliminated
7. The frequency of individual words is calculated
8. According to the occurrence of the words, the ranking of the documents is done.
9. This is called as second base cluster
10. After getting the cluster the occurrences of words with respect to the grammatical position is calculated
11. The words in noun position the given most priority and then the rest of the grammatical positions

The third cluster are prepare with the maximum accuracy

**5. APPLICATIONS**

Document clustering has been used in a number of applications .In information retrieval systems, it has been used to improve the precision and recall performance [14], and as an efficient way to find similar documents [15] It has also been proposed in browsing document collections and organizing the results of a search engine query] Document clustering has also been used to automatically generate hierarchical grouping of documents[18]

**6. CONCLUSION**

We presented a novel technique for calculating the probability and expected document frequency of any given noncontiguous lexical cohesive relation.. We further described a method that compares observed and expected document frequencies through a statistical test as a way to give a direct numerical

evaluation of the intrinsic quality of a multi-word unit (or of a set of multi-word units). This technique does not require work of a human expert, and it is fully language and application independent. It is generally accepted that, in English, two words at a distance five or more are not connected. We can attempt to deal with this by using short documents, for example sentences, or even comma-separated units We presented a novel technique for calculating the frequency of word occurrence such a statistical test as a way to give a direct numerical evaluation of the intrinsic quality of a multi-word unit (or of a set of multi-word units)

**7. REFERENCES**

1. Evan ER ijsbergen.“Information Retrieval”.

2. Zamir, Etzioni, Madani,and R. M.Karp.“Fast and intuitive clustering of web documents” In KDD’97, pp. 287 - 290, 1997.

3. M. Steinbach, G. Karypis, V. Kumar, A comparison of document clustering techniques. Proc. KDD Workshop on Text Mining, 1-20, 2000.

4. B. Luke. K-means clustering tutorial.http://ncisgi.ncifcrf.gov/lukeb/kmeans.html

5. Taeho Jo, “ Single Pass Algorithm for Text Clustering by Encoding Documents into Tables ”, IT Convergence, KAIST Institute

6. Oren Zamir and Oren Etzioni, “ Web Document Clustering”

7. Mrs. K. P. Supreethi ,“Web Document Clustering Technique Using CaseGrammar Structure”

8. MING HEI TSUI, BRESLEY LIM, DAMING ,“ Web Search Result Refinement by Document clustering”

9. Juhyun Han, Taehwan Kim, Joongmin Choi, “Web Document Clustering By Using Automatic Keyphrase Extraction” Dept. of Computer Science and Engineering Hanyang University1271 Sa-3-Dong, Ansan, Gyeonggi-Do, Koreafjhhan, kimth,jmchoig@cse.hanyang.ac.

10. Ying Zhao and George Karypis “Criterion Functions for Document Clustering Experiments and Analysis” University of Minnesota, Department of Computer Science / Army HPC Research Center Minneapolis, MN 55455Technical Report #01-40

11. Xiaohui Cui, Thomas E. Poto "Document Clustering Analysis Based on Hybrid PSO+K-means Algorithm" kApplied Software Engineering Research Group, Computational Sciences and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6085, USA cuix, potokte@ornl.gov

12. Dell Zhang ,"Semantic, Hierarchical, Online Clustering of Web Search Results" , Department of Computer Science School of ComputingS15-05-24, 3 Science Drive National University of Singapore Singapore 1175432 Singapore-MIAllianceE4-04-10, 4 Engineering Drive 3Singapore

13. Kam-Fai Wong, "Improving Document Clustering by Utilizing Meta-Data" by Department of Systems Engineering and Engineering Management, Samuel Sambasivam and Nick theodosopoulos

14. C. J. var Rijsbergen,Information Retrieval. Buttersworth,London, UK, 1989.

15. M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques", KDD Workshop on Tat Mining, 2000."