iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 Issue: V Month of publication: May 2017

DOI:

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Concept-Based Document Clustering Using Bisecting K-Means Algorithm

Ananth V[1]

[1]*Department of Computer Science, Manakula Vinayagar Institute of Technology, Puducherry, India.*

Abstract: - Document Clustering has been extensively investigated as a methodology for improving document search and retrieval. Although good clustering algorithms are widely available, good solutions for labeling the clustered results to meet analysts' needs are rare. Therefore efficient topic extraction methods are essential in document clustering. Hence this paper is focused at how efficiently document clustering can be done with the help of bisecting K- means algorithm.

Keywords: - K-means Algorithm , Information retrieval (IR), Word Sense Disambiguation (WSD), Euclidean distance, F-measure ,Semantic-based Analyzer algorithm.

I. INTRODUCTION

Information extraction [1] plays a vital role in today's life. How efficiently and effectively the relevant documents are extracted from World Wide Web is a challenging issue. As today's search engine does just string matching, documents retrieved may not be so relevant according to user's query. A good document clustering approach can assist computers in organizing the document corpus automatically into a meaningful cluster hierarchy for efficient browsing and navigation, which is very valuable for overcoming the deficiencies of traditional information retrieval methods. It is a more specific technique for unsupervised document organization, automatic topic extraction and fast information retrieval or filtering [2]. It breaks down huge linear results into manageable sets. It is an automatic grouping of text documents into clusters where documents of the same cluster are more similar than the documents in different clusters.

Document clustering has been investigated for use in a number of different areas of text mining and information retrieval. Initially, document clustering was investigated for improving the precision or recall in information retrieval systems and as an efficient way of finding the nearest neighbors of a document. More recently, clustering has been proposed for use in browsing a collection of documents or in organizing the results returned by a search engine in response to a user's query. Document clustering has also been used to automatically generate hierarchical clusters of documents. For example, a web search engine often returns thousands of pages in response to a broad query, making it difficult for users to browse or to identify relevant information. Clustering methods can be used to automatically group the retrieved documents into a list of meaningful categories, as is achieved by
Enterprise Search engines such as Northern Light and Vivisimo.

By clustering the text documents, the documents sharing the same topic are grouped together. Unlike document classification, no labeled documents are provided in clustering; hence clustering is known as unsupervised learning. Topic detection deals with discovering meaningful and concise labels for the clusters which are grouped using document clustering algorithm. Searching collection of documents by choosing from the set of topics or labels assigned to the clusters becomes easy and efficient. A good descriptor for a cluster should not only indicate the main concept of the cluster, but also differentiate the cluster from other clusters. Hence a proper document clustering model is considered to consist of three phases Document pre-processing, Document clustering and Topic discovery.

Clustering is aimed at generating document groups or clusters, each one representing a different topic. However, it is not enough. Users also need to easily find out what a cluster is about for determining at a glance those of their interest. Unfortunately, common clustering techniques do not provide proper descriptions of the obtained clusters, making them difficult to interpret. The problem of determining the meaning of the clusters is simply left to the user.

II. LITERATURE SURVEY

A. Role of clustering in information retrieval

Information retrieval (IR) is the science of searching for documents, for information within documents, and for metadata about documents, as well as that of searching relational databases and the World Wide Web. Automated information retrieval systems are used to reduce what has been called "information overload". Many universities and public libraries use IR systems to provide access to books, journals and other documents. Web search engines are the most visible IR applications. There are several other applications within IR. Text clustering is one among them. A text clustering algorithm partitions a set of texts so that texts within the same group are as similar in content as possible. It is done without using any predifined catagories. Text clustering can for instance be applied to the documents retrieved by a search engine, so that they can be presented in groups according to content.

An information retrieval process begins when a user enters a query into the system. Queries are formal statements of information needs, for example search strings in web search

engines. In information retrieval a query does not uniquely identify a single object in the collection. Instead, several objects may match the query, perhaps with different degrees of relevancy. An object is an entity that is represented by information in a database. User queries are matched against the database information. Depending on the application the data objects may be, for example, text documents, images, audio, mind maps or videos. Often the documents themselves are not kept or stored directly in the IR system, but are instead represented in the system by document surrogates or metadata.

A tremendous growth in the volume of text documents available on the Internet, digital libraries, news sources, and company-wide intranets has led to an increased interest in developing methods that can help users to effectively navigate, summarize, and organize this information with the ultimate goal of helping them to find what they are looking for. Fast and high-quality document clustering algorithms play an important role towards this goal as they have been shown to provide both an intuitive navigation/browsing mechanism by organizing large amounts of information in to a small number of meaningful clusters as well as to greatly improve the retrieval performance either via cluster-driven dimensionality reduction, term-weighting, or query expansion. The applicability of clustering is manifold, ranging from market segmentation and image processing through document categorization and Web mining.

Document Clustering has been extensively investigated as a methodology for improving document search and retrieval. The general assumption is that mutually similar documents will tend to be relevant to the same queries, and, hence, that automatic determination of groups of such documents can improve recall by effectively broadening a search request. Typically a fixed corpus of documents is clustered either into an exhaustive partition, disjoint or otherwise, or into a hierarchical tree structure. In the case of a partition, queries are matched against clusters and the contents of the best scoring clusters are returned as a result, possibly sorted by a score. In the case of a hierarchy, queries are processed downward, always taking the highest score branch, until some stopping condition is achieved. The subtree at that point is then returned as a result. Hybrid strategies are also available. These strategies are essentially variations of near neighbor search where nearness is defined in terms of the pairwise document similarity measure used to generate the clustering. Indeed, cluster search techniques are typically compared to direct near-neighbor search and are evaluated in terms of precision and recall.

B. Document Representation

The most used document data model, the Vector Space Model, was introduced by Salton [3]. Each document is represented by a vector d = tf1, tf2, tfn, where tfi is the frequency of each term (TF) in the document. In order to represent the documents in the same term space, all the terms from all the documents have to be extracted first. These results in a term space of thousands of dimensions. Because each document usually contains only at most several hundred words, these representation leads to a high degree of sparsity.

In suffix tree document model, each internal node is represented by at least two documents, meaning that the concatenation of the edge labels (the suffix) from the root to an internal node is a phrase contained in those two documents. External nodes can be represented by one or more documents. Each node contains a structure showing all the documents that ”contain” that node, as well as the number of the suffix in the document (the longest suffix has number 1, the smallest suffix has number w, where w represents the number of words in the document.

C. Clustering

Clustering algorithms group a set of documents into subsets or clusters. The algorithms' goal is to create clusters that are coherent internally, but clearly different from each other. In other words, documents within a cluster should be as similar as possible; and documents in one cluster should be as dissimilar as possible from documents in other clusters. Clustering is the most common form of unsupervised learning. No supervision means that there is no human expert who has assigned documents to classes. In clustering, it is the distribution and makeup of the data that will determine cluster membership. The concepts of similarity and distance are fundamental in data clustering, and they are a basic requirement for any algorithm. A similarity measure assesses how “close” two data points are from each other. The distance measure, or dissimilarity, does exactly the opposite. Usually, either one or the other measure is used. Anyway, each measure can be easily derived from the other.

Euclidean distance [5] is a standard metric for geometrical problems. It is the ordinary distance between two points and can be easily measured with a ruler in two- or three-dimensional space. Euclidean distance is widely used in clustering problems, including clustering text. It is also the default distance measure used with the k-means algorithm.

When documents are represented as term vectors, the similarity of two documents corresponds to the correlation between the vectors. This is quantified as the cosine of the angle between vectors, that is, the so-called cosine similarity [5]. Cosine similarity is one of the most popular similarity measure applied to text documents, such as in numerous information retrieval applications and clustering.

The Jaccard coefficient [5], which is sometimes referred to as the Tanimoto coefficient, measures similarity as the intersection divided by the union of the objects. For text document, the Jaccard coefficient compares the sum weight of shared terms to the sum weight of terms that are present in either of the two documents but are not the shared terms.

K-means is the most widely used clustering algorithm, offering a number of advantages: it is straightforward, has lower complexity than other algorithms, and it is easy to parallelize. The algorithm works as follows: first, k points are generated randomly as centres for the k clusters. Then, the

data points are assigned to the closest clusters. Afterwards, the centroids of the clusters are recomputed. Then, at each step the points are re-assigned to their closest cluster and the centroids are recomputed. The algorithm stops when there are no more movements of the data points between clusters. The main advantage of k-means clustering is that it is fast and simple to implement.

The bisecting k-means algorithm is a combination of the original k-means algorithm and hierarchical. It starts with one cluster, and splits it in two smaller ones. At the next step, one of the two clusters, either the largest one, or the least cohesive one, is split again in two clusters. The algorithm continues splitting one cluster at each step until the desired number of clusters is achieved. This algorithm can also be classified in the hierarchical divisive clustering category. Steinbach et al. [6] have proved that this algorithm is more accurate than both k-means and hierarchical agglomerative clustering.

In order to assess the performance, or quality, of different text mining algorithms, objective measures need to be established. There are three types of quality measures [6]: external, when there is a priori knowledge about the clusters, internal, which assumes no knowledge about the clusters, and relative, which evaluates the differences between different cluster solutions. The external measures are applied to both categorization and clustering, while internal and relative measures are applied only to clustering.

D. Topic Detection

In order to determine at a glance whether the content of a cluster are of user interest or not, topic discovery methods are required to tag each clusters identifying distinct and representative topic of each cluster. A topic discovery system aims to reveal the implicit knowledge present in the document clusters.

The popularity of Internet has caused an ever-increasing amount of textual documents (Web pages, news, scientific papers, etc.). This information explosion has led to a growing challenge for Information Retrieval systems to efficiently and effectively manage and retrieve this information. The standard formulation of the information access problem presumes a query, the user’s expression of an information need. However, many times it is difficult, if not impossible, to formulate such a query. Suppose, for example, a user wants to find out what events happened during a given month and/or at a given place. In this case, the information need is too vague to be described as a single topic and the user does not know the topics of her interest. Indeed, the user may wish to browse over a set of discovered topics extracted from the document collection at hand. Topic detection is an experimental method for automatically organizing search results. It could help users save time in identifying useful information from large scale electronic documents.

III. METHOD

Incorporating semantic features improve the accuracy of text clustering techniques. Hence in this proposed work, terms are considered as concepts i.e. terms along with their related terms for concept-based document clustering unlike the existing work which considers the individual terms as such for term-based document clustering. Related terms of the analyzed term are extracted using the proposed concept extraction algorithm.

Document representation by Vector Space Model and Semantic-based Analyzer algorithm [11] is used for term and concept weighting.

A. Concept-based document clustering

1) By Bisecting K-means algorithm using cosine similarity as the similarity measure.

2) By Proposed modified Bisecting K-means algorithm using semantic-based similarity [11] as the similarity measure.

Topic discovery in document clusters by Testor theory [14].

Document representation is done by Vector Space Model where a document is represented as a vector of concept weights. In this proposed work, Semantic-based analyzer algorithm is used for term and concept weighting unlike the existing work where TF-IDF is the term weighting method used. Concept-based document clustering (I/II) is done and Testor theory is applied on the clustered documents to discover the most representative concepts of the clusters which are considered as the cluster labels.

B. Proposed constant extraction algorithm

The following proposed algorithm describes the process of related terms extraction for concepts. In order to extract concepts, a domain specific science dictionary consisting of scientific terms is created as the dataset used for experiments consists of scientific journal articles. Domain specific dictionary is used for concept extraction as it eliminates the need for word sense disambiguation [16] which is not the scope of this work.

```
L is an empty list of concepts
for each document di do
for each labeled term tj in di do
for each labeled term tk other than term tj in di do
if term tk is in the definition of term tj then    if tj and tk not in concepts of L then
add tj and tk to new concept cp
else if either tj or tk in concept cp of L then
combine cp and cq and add them to new concept cr remove concepts cp and cq from L
end if
end if
end for
end for
end for
```

C. Semantic-based term analysis

The objective of the semantic-based analyzer algorithm is to analyze terms on the sentence and document levels. Then, top terms (which have maximum weights assigned by the semantic-based analysis) and their corresponding synonyms and hypernyms are used for text clustering. For instance, terms like beef and lamb are found to be similar, because they both are sub concepts of meat in WordNet. If these words are added to the term vector, the clustering technique will cluster documents that are related based on the meaning of their words rather than the words themselves. The following phases depict the process of the semantic-based term analysis:

1) Each sentence is labeled by semantic role labeller

2) For each labeled term, stop-words that have no significance are removed

3) Labeled terms are analyzed on the sentence and document levels.

4) Due to words with the same meaning appear in various morphological forms, words (in a labeled term) are normalized into a common root-form to capture their similarity,

5) Terms are sorted based on their weights (assigned by the semantic-based analysis) descendingly and top terms are extracted.

6) For each word (in a top term), nouns, verbs, adjectives and adverbs are looked up in dictionary and a global list of the first two synonyms and hypernym synsets is assembled. Synonyms are extracted for nouns, verbs, adjectives and adverbs. Hypernyms are extracted for nouns and verbs. Terms that have no corresponding concept in dictionary are still used in text clustering. Extracting the synonyms and hypernyms from the top terms only is a kind of pruning to reduce the dimension of the term vector.

7) Term vector is extended by adding the corresponding synonyms or hypernyms of the top terms.

Semantic-based analyzer algorithm analyzes terms on the sentence and document levels. To analyze each term at the sentence-level, a sentence-based frequency measure, called the conceptual term frequency ct f is utilized. The ct f is the number of occurrences of term t in verb-argument structures of sentence s. The term t, which frequently appears in different verb-argument structures of the same sentence s, has the principal role of contributing to the meaning of s. To analyze each concept at the sentence-level, each concept is assigned the same (ct f) value of its corresponding top term. To analyze each concept at the document-level, the concept frequency c f is proposed, the number of occurrences of a concept c in the document, is calculated. At this point, each term and its corresponding concepts have the same measures which are the ct f and c f (for concept and term) on the sentence and document levels respectively.

D. Topic discovery by testor theory

Topic detection by testor theory involves construction of a learning matrix and a comparison matrix.For each cluster C, a learning matrix LM(C) is constructed whose columns are the most frequent concepts in the representative C, and its rows are the representatives of all clusters, described in terms of these columns. In order to calculate the typical testors, two classes are considered in the matrix LM(C). The first class is only formed by C and the second one is formed by the other cluster representatives. The goal is to distinguish cluster C from other clusters. Comparison matrix could be a matrix of similarity or a matrix of dissimilarity depending on the type of comparison criteria that are applied for each feature. In this case, the features that describe the documents are the concepts and its values are the frequency of concepts. The comparison criterion applied to all the features is:

$$d(vik, vjk) = 1 \text{ if } vik - vjk \geq \delta$$
$$= 0 \text{ otherwise}$$

Where vik, vjk are the frequencies in the cluster representative i and j in the column corresponding to the concept c respectively, and δ is a user-defined parameter. As it can be noticed, this criterion considers the two values (frequencies of the concept ck) different if the concept ck is frequent in cluster i and not frequent in cluster j. From this comparison matrix, the most representative concept c in cluster C is obtained, which is used to tag the cluster.

E. Proposed Algorithm

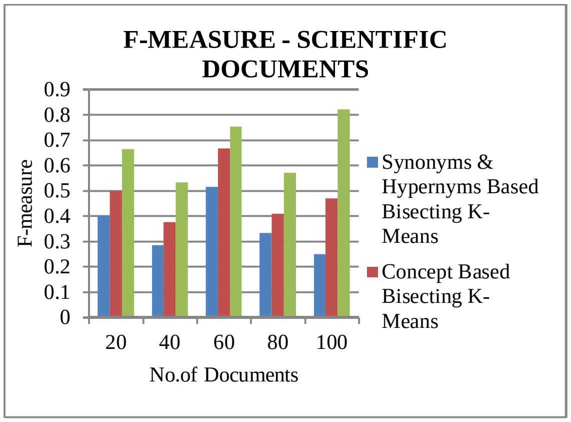
1) Input
a) Document
b) Number of clusters (K)
2) *Output:*
a) Labeled clusters of documents

F. Algorithm

```
Step 1:Document pre-processing
Decomposition of sentences
Removal of stop words
Stemming of words by Porter stemmer algorithm
Construction of weighted matrix using Semantic-based Analyzer algorithm as follows:
doci is a new Document where doci = {1, 2, ..,N} and N is a total number of documents
L is an empty List (L is a synonyms and hypernyms list)
M is an empty List (M is a matched concepts list)
 is an empty List (T is a terms list)
for each sentence s in d do
for each labeled term t in d do
compute t fi of ti in d
compute ct fi of ti in s in d
compute the weighti = t fi + ct fi
add term t with weighti to T
end for
end for
sort T descendingly based on weight
utput the max(weight) from list T
for each term t in T that has max(weight) do
xtract related terms of t using Proposed Concept Extraction algorithm
dd concept c (related terms) to L
```

```
nd for
or each concept ci in L do
compute c fi of ci in d
assign ct fi to ci that corresponds to ti
compute concept ci weight = cfi + ctfi
```

IV. RESULT ANALYSIS



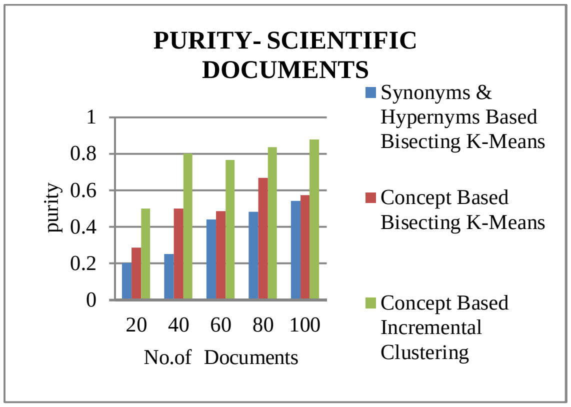


Fig 4. shows the comparison of F-measure for term-based document clustering and concept-based document clustering (I and II), varying the total number of clusters. Fig 5. shows the comparison of Purity for term-based document clustering and concept-based document clustering (I and II), varying the total number of clusters. For concept-based document clustering I, the percentage of improvement ranges from +22.00% to +43.39% increase in the F-measure quality, and +25.00% to +50.94% increase in Purity. For concept-based document clustering II, the percentage of improvement ranges from +30.76% to +86.36% increase in the F-measure quality, and +56.52% to +75.47% increase in Purity.

Fig 6. shows the comparison of F-measure for topic discovery by clustering keywords method and concept-based topic discovery by testor theory (I and II), varying the total number of documents. Fig.7. shows the comparison of Purity for topic discovery by clustering keywords method and concept-based topic discovery by testor theory (I and II), varying the total number of documents. For the concept-based topic discovery by testor theory I, the percentage of improvement ranges from +16.39% to +106.66% increase in the F-measure quality, and +11.13% to +160.00% increase in Purity. For the concept-based topic discovery by testor theory II, the percentage of improvement ranges from +21.31% to +156.66% increase in the F-measure quality, and +19.23% to +230.00% increase in Purity.

rom the experimental results, it is clearly seen that the proposed work I and II show good performance when compared to the Existing work and Proposed work II performs better than Proposed work I.

V. CONCLUSION

The key contributions of this project are the extraction of related terms of the analyzer terms as concepts for concept-based document clustering by bisecting k-means algorithm using cosine similarity measure and concept-based document clustering by proposed modified bisecting k-means algorithm using semantic-based similarity measure and concept-based topic discovery by testor theory. Concept-based document clustering is compared to the term-based document clustering and concept-based topic discovery by testor theory is compared to topic discovery by Clustering keywords method using F-measure and Purity as evaluation metrics. Experimental results prove that concept-based document clustering is more efficient than term-based document clustering and concept-based topic discovery by testor theory is more efficient than topic discovery by clustering keywords method.

One future enhancement is the inclusion of word sense disambiguation (WSD) strategy to avoid the use of domain specific dictionary. By incorporating WSD, right sense of a term can be extracted from lexical databases like WordNet. Another future work is to perform experiments on different datasets of various categories of documents to analyze the quality of the proposed related terms concept extraction and concept-based topic discovery by testor theory.

REFERENCES

[1] Jiawei Han and Micheline Kamber, “Data Mining Concepts and techniques”, Second Edition.

[2] G. Salton, A. Wong, and C. S. Yang. “A Vector Space Model for Automatic Indexing”, ACM, vol. 18 no. 11, pp. 613–620, 1975.

[3] Salton, Gerard and Buckley C, “Term-weighting approaches in automatic text retrieval”, Information Processing & Management, vol. 24 no. 5, pp. 513-523, 1988.

[4] Anna Huang, “Similarity Measures for Text Document Clustering”, NZCRSC’08, April 2008.

[5] Michael Steinbach, George Karypis and Vipin Kumar, “A Comparison of Document Clustering Techniques”, Proc. KDD-2000 Workshop Text Mining, 2000.

[6] Christian Wartena and Rogier Brussee, “Topic Detection by Clustering Keywords”, IEEE 19th *International Conference and Expert System Application*, 2008.