EFFECTIVE LINEAR-TIME DOCUMENT CLUSTERING IN TEXT MINING USING WEB DOCUMENT CATEGORIZATION

M. John Basha
Research Scholar, Department of CSE, Bharath University, Chennai, Tamil Nadu 600073, India

Dr. K. P. Kaliyamurthie
Professor & Head, Department of CSE, Bharath University, Chennai, Tamil Nadu 600073, India

ABSTRACT
Among data mining technique, clustering is one of the most important and traditional concept also an unsupervised learning paradigm. Similarity of a document pairs can be measured by matching of concepts. Finding or extracting the most relevant concept from the documents is a challengeable task. To address this issue, in this paper we introduce a concept of multi viewpoint based similarity measure. Our proposed methods uses multiple point of reference between document pairs to extract more relevant match concept rather than extracting only ideas based on similarity measure. Using multiple viewpoint, gathers more information about a particular topic from many different but relevant sources or concept. This strategy works well with smaller documents but is especially effective with longer documents. By gathering more relevant concepts from the documents with multiple points of reference, the document organization and retrieval can enhance the ability to make the most use of the documents held in storage and make retrieval of ideas as well as relevant task or concept much easier and faster. Experimental results shows that our proposed method efficiently extract more relevant concept.

Key words: Similarity Measure, Concept Mining, Document Clustering.

1. INTRODUCTION

Clustering in general is an important and useful technique that automatically organizes a collection with a substantial number of data objects into a much smaller number of coherent groups for further analysis. More precisely clustering is the process of organizing objects into subgroups whose members are similar in some way.

We are facing an ever increasing volume of text documents. Existence of plenty of texts flowing over the internet, repositories, some digitized personal information such as emails and blog articles are quickly piling up every day. Those things brought challenges for effective and efficient organization of text documents. To handle this issue clustering has proven to be an effective approach. With the development of World Wide Web and subsequent evolution of web2.0 this clustering technique becomes more interesting approach. Consider an example, the results returned by search engines are clustered to help the users quickly identify and focus on the appropriate set of results. Accurate clustering requires a precise definition of the closeness between a pair of objects which in turn either as a pair-wise similarity or distance. There exists a variety of similarity or distance measures have been proposed and applied widely. Some of them include cosine similarity and Jaccard correlation coefficient. Several measures such as Euclidean distance and relative entropy have been applied widely in clustering to calculate the pair wise distances. In text mining concept, the data are extracted by computing the frequency of a term to explore the importance of the term in a document.

Because of its simplicity, understandability and scalability clustering algorithm plays a vital role in many applications. It is applied in many fields like marketing, biology, libraries, insurance, city planning, earthquake studies, WWW etc.

For a relatively small collection of documents it is possible to manually perform the partitioning of documents into specific regions. And the same time while partitioning of large volumes of text it should be clustered based on similarity measure using some clustering algorithm. The main requirements that a clustering algorithm should satisfy are based on scalability, ability to deal with noise and outliers, Exhibits high dimensionality, able to deal with different types of attributes, attains insensitivity to order of input records, minimal requirements for domain knowledge to determine input parameters, discovering clusters with arbitrary shape, facilitates interpretability and usability.

The proposed method consists of sentence based concept analysis, document based concept analysis, and concept based similarity measure.
To analyze each document on concept based similarity measure the three factors needs to be calculated. The three measures include \( ct(f) \) (Conceptual term frequency), \( tf \) (term frequency), \( df \) (Document frequency). This concept based similarity measure is performed by matching concepts at the sentence, document rather than individual terms or words only. Our proposed work exploits the information extracted from the concept based analysis algorithm to better judge the similarity between the documents. The similarity measure of document and extracting more relevant concept from the raw documents can be better achieved by employing multiple reference point within the documents. We may have more accurate information regarding the extraction of relevant concept. By standing at various reference point similarity between the documents and the generation of more relevant concept would be achieved. The similarity measure to be compared includes Euclidean distance as a performance metrics.

Remaining of this paper is organized as follows: Section 2 provides the works that are related to the concept based model employed in various document or text clustering, and document categorization and extraction. Section 3 presents our proposed work for empowering the document clustering by finding more relevant concept with multi view-point based similarity measure. Section 4 demonstrates the efficiency of our work through experimental results. Then finally, section 5 concludes the paper with future work that can be carried out for improving this document clustering.

2. RELATED WORK
Searching of related information about the performance improvement on document clustering is an increasing and also an interesting research area. Many of the research scholars are trying until a date for a method of retrieving more relevant concept on text clustering. Here, we present some of the relevant works carried out earlier by various scholars.

Document clustering has been investigated for use in a number of different areas of text mining and information retrieval.

Document clustering is an efficient way of finding the nearest neighbors of a document. Scholars of an article in [22] designed such efficient document clustering by the combination of standard k-means algorithm, “bisecting” K-means and a variant of K-means. Bisecting K-means technique is better than standard K-means approach. K-means clustering quality can be achieved to replace the batch updates of cluster centroids by an online learning strategy. Online update leads to improve the clustering results. This approach was proposed by scholars in [14].

Previous work shows that k means clustering is far better than hierarchical clustering of documents. But the authors of [19] designed a framework that utilizes hierarchical topic structure to decompose the classification task into a set of simple problems. Each of these smaller problems can be solved accurately by focusing only on a very small set of features. The set of relevant features varies widely throughout the hierarchy; classifier is employed to examine a small subset. This is a standard computational task but possesses difficulties in robustness.

To measure the quality of cluster hierarchy, quality metrics like F-measure and the results compared with various clustering algorithm. This approach was dealt by authors in [9] which are a fast and effective text mining using linear time document clustering. The evaluation considers some feature selection parameters like tfidf, feature vector length. But this technique has limitations of scalability. A robust hierarchical clustering algorithm designed in [17] that employed link and not distances to measure the similarity between a pair of data points when merging the clusters. This method is useful in situations where a domain expert/similarity table is the only source of knowledge. Efficient document clustering can help automatically organize the document corpus into a meaningful cluster hierarchy for efficient browsing and navigation. Such efficiency can be achieved by the scholars of through non-negative matrix factorization.
which is combined form of K-means clustering. This method uses an online NMF algorithm to efficiently handle very large scale and streaming datasets. Conventionally this algorithm requires the data matrix to reside in the memory during the solution process leads to problematic approach when the datasets are very large.

Similarity of a document can be identified by various clustering algorithm as in [18] and [20]. Similarity between two objects is defined to be the amount of information contained in the commonality between objects divided by the amount of information in the descriptions of the objects. This word similarity was proposed by the scholars in [2]. The evaluation of automatically generated lexical resources is a difficult problem.

In [11], an algorithm for automatic semantic analysis was designed. It assigns a semantic role to constituents in a sentence. This approach is used to treat the problem of semantic role labeling like the similar problems of parsing. Identifying the semantic roles filled by constituents of a sentence can provide a level of shallow semantics analysis useful in solving a number of natural language processing tasks. The similarity between the sentences is expressed in terms of the distance between their corresponding sentences. This work was proposed by authors in [15]. According to this approach membership of a sentence is determined with respect to the nearest neighbor or K-nearest neighbor rule. Works on clustering presented in [5] proposed off-line algorithm for clustering static information and uses on-line version of algorithm for clustering dynamic information. Authors of [16] analyze the behavior of clustering algorithms on weighted data. Under such algorithms classify the clustering methods into three broad categories such as weight responsive, weight robust, weight considering. With weighted data, the consideration is easily captured by having the weights of the point represent signal frequency. This weighted setting enables convenient clustering.

The term that captures the semantic structure of the text would be given more importance. This leads to extract concept and enhance the document clustering as more efficiently. One such model was designed in [12]. This model takes three factors for concept analysis. The quality of the clusters achieved by this model significantly improves the performance. An efficient phrase based matching of web document clustering is employed to match the similarity between the documents. This algorithm works better by carefully watching the pair-wise document similarity distributed within the clusters. This robust method was proposed in [1] provides accurate document similarity that leads to improved results in web document clustering. Many of the algorithm breakdown the clustering process owing to the size of the document space. So whenever the dimensionality of the feature space becomes high relative to the document space, the problem can be solved by the generalization of graph partitioning which are automatically capable of discovering documents similarity which was proposed in [6].

Similarity of a xml document can also be detected by the technique developed in [4]. XML documents can be compared to their structural similarity and group them into the clusters so that different storage organization, information retrieval can be exploited more effectively. The designed approach completely detect the structural similarity between xml documents which significantly differs from standard methods also allow a significant reduction of the computational costs. Based on the paper proposed by author in [7] the goal of cluster analysis is to partition a data set of N objects into subgroups such that those in each particular group are more similar to each other than to those of other groups. This algorithm enables to uncover group of objects that have preferentially close values on different more possibly overlapping, and an attribute of subsets.

Document classification is an important process for helping the information retrieval systems to organize the vast amount of data. To make the document classification process automatic clustering techniques have been employed by the author in [13]. A tree structure called DC-tree was introduced to cluster the documents without any training set. This algorithm
is less sensitive to the document insertion order and is more tolerant towards noise documents. The co-clustering concept proposed by scholars in [8] aims to treat non-negative contingency table as a joint probability distribution between two discrete random variables that take values over the rows and columns. Co-clustering is pair of maps from rows to row-clusters and from columns to column clusters. Theoretical formulation is used to co-cluster the documents.

According to the paper [10] to process large document collections quickly and to search collections that total in the order of billions to trillions of words then the information retrieval should needs to be efficient. Since the collection of online data has grown at least as quickly as the speed of computers, managing and retrieval of data is very essential. [21] Focus on document similarity based on concept tree distance. Each document is taken as a concept tree and employs a tree similarity measure based on tree edit distance to compute similarities between concept trees. This technique constructs the concept trees representing the documents and applies the tree edit distance algorithm to calculate the document similarity.

3. PROPOSED METHODOLOGY

In this section, the proposed method for finding the more relevant concept with multi viewpoint based similarity measure is detailed. The overall process flow diagram is also depicted in this section. The algorithm that shows the detailed computation of finding relevant concept using our proposed method is represented in this section.

Similarity measure

The similarity between the documents is based on the combination of sentence-based, corpus-based, and document-based analysis. Document similarity can be measured by matching of concepts between document pairs. Significant effects on the clustering quality due to similarity’s insensitivity to noisy terms lead to incorrect similarity. Extracting concepts are less sensitive to noise when it comes to calculating document similarity. The concepts are originally extracted by the semantic role labeler and these concepts are analyzed with respect document, sentence and corpus levels. The proposed concept based similarity measure extracts more relevant concept in document pairs.

Semantic role labeler

Semantic structure of a sentence can be characterized by a form of verb argument structure. Consider an e.g. subject-verb-object. Now, before u object and saying the words have some meaning but not structure. Assume two sentences,

“Dog bites man”

“Man bites dog”

In the above sentences, the words have meaning, also the same time two sentences containing the same words in the same structure but can have different meaning depending on where in the structure, the words appear. Move to 1st case, dog that bites man, then 2nd case the man who bites the dog. By changing the position of “words” otherwise said to be “terms” within the structure changes the meaning of structure. Each term has a semantic role in the sentence is called “concept”. Labels are assigned to each sentence using semantic role labeler. Concepts can be either words or phrase and are totally dependent on the semantic structure of the sentence. Our proposed method output the matched concepts by this semantic role analysis in a sentence.
Concept mining
The objective behind our proposed model is to achieve an accurate analysis of concept on the sentence, corpus level and also the document and finding more relevant and matched concept using multi viewpoint based similarity measure. The following factors needs to be computed to find similarity based on concept analysis.

Sentence based concept analysis
This approach analyzes each concept at the sentence level using concept based frequency measure which is denoted by conceptual term frequency (ctf). The ctf calculation of concept c in sentence s and document d is expressed as,

\[ \text{ctf} = \frac{\sum_{n=1}^{sn} ctf_n}{sn} \]  

(a) Calculating ctf factor of concept c in sentence s:
The ctf is the number of occurrence of concept c in verb argument of sentence s. If the concept c appears frequently in different verb argument of same structure of same sentence s then it plays a principal role of contributing to meaning of s.

(b) Calculating ctf factor of concept c in document d:
The ctf value of concept c in document is computed if the concept c takes many ctf values in different sentences in the same document d. It is denoted in equation (1) as,

Document based concept analysis
The concept based term frequency tf is computed to analyze each concept at the document level. It denotes the number of occurrence of a concept c in original document.

Corpus based concept analysis
The number of documents containing concept c is finding out by calculating document frequency df which extract concept that can discriminate between documents.

The overall flow of our proposed model is depicted in fig 2.

Similarity between the documents is checked by matching of relevant concept. The output may be exact match or partial match but the efficiency can be achieved by extracting exact concepts within the documents. The proposed method attains overall accuracy and efficiency by using multiple reference point between the document pairs. The similarity between the documents \((d_i, d_j)\) is obtained with respect to the angle between the documents.
By standing at various reference point $d_h$ to view document similarity based on concept between two documents $d_i$ and $d_j$, more exact matching of concepts can be extracted within the document. By this way our proposed method achieves efficiency.

$Sim(d_i, d_j) = \prod_{d_h} \sum_{d_i, d_j} \frac{sim(d_i - d_h, d_j - d_h)}{n - n_r}$

(2)

Figure 2 Overall Flow of our proposed model
Algorithm 1. Retrieval of relevant matched concepts.

An algorithm that depicts our computation of match concepts is given above:

4. EXPERIMENTAL EVALUATION

The goal of our proposed method that uses multiple reference point between the document pairs is to maximize the retrieval of more relevant and matched concepts. The experiment was conducted to test the effectiveness of relevant concept matching in determining an accurate measure of the similarity between documents. The effectiveness of our proposed approach are evaluated based on precision, recall, the prediction accuracy and the time taken for processing and retrieving the relevant match concepts from the documents. The following figures show comparison between our proposed and existing method. The existing method used concept based mining model to retrieve the match concepts whereas our proposed method uses multiple reference point within the concept based mining model to retrieve the more relevant exact match concepts and cluster the documents based on exactly matched concepts.

Precision is a measure that expresses the fraction of returned or retrieved documents that are more relevant and matched concepts. This measure shows the consistent results which are purely based on the measure and understanding of relevance. Precision can be calculated using the equation (3),

\[
\text{Precision} = \frac{\text{No. of relevant document retrieves}}{\text{Total no. of document retrieved}}
\]  

Fig 3. illustrates that our proposed method has higher precision value than the existing one for matched concept retrieval Figure 4 reveals the recall values for proposed and existing techniques.

Similarly the existing and proposed approaches are compared using Recall which is another measuring factor. The estimation of recall value is calculated through equation (4).

\[
\text{Recall} = \frac{\text{No. of relevant document retrieved}}{\text{Total no. of existing relevant document}}
\]
An efficiency document extracting technique’s accuracy depends on how accurately it retrieves the documents based on more relevant matched concepts. In order to predict the proposed methods efficiency we calculate its overall prediction accuracy. Three factors are considered for finding match concepts in documents. Proposed and existing methods are experimented with same factors and the results are expressed in figure 5. Prediction accuracy is carried out using the equation (5).

\[ \text{Prediction Acc} = \frac{N_{TP} + N_{TN}}{N_{TP} + N_{FP} + N_{FN} + N_{TN}} \]  

(5)

The values $N_{TP}$, $N_{FP}$, $N_{FN}$, $N_{TN}$ are the number of true positive, false positive, false negative and true negative.

Fig 5. Comparison of prediction accuracy existing vs. proposed.

Fig 5. Illustrates that the prediction accuracy of our approach which is evident that the proposed is superior to the existing. Experimental results shows that our approach retrieves 79% of the relevant matched concepts from the document accurately, whereas the existing approach extract 65.3% of related documents.

Running time of the approach plays a vital role in the evaluation of the effectiveness of the system. We measured the running time for both the proposed and existing methods to output the list of documents with matching relevant concept.
Figure 6 portrays the time required for both proposed and existing techniques for retrieving more relevant matched concept. It also reveals that the proposed method consumes less time than the existing method. This explicitly denotes that proposed method retrieves more relevant matched concept in documents faster than the existing method. The computation time is calculated in seconds.

The experimental results shows that our proposed method retrieves more relevant matched concepts in document pairs with an accuracy rate of 79% which is 13.7% greater than the existing method. It is also shown that our proposed method extracts relevant matched concepts in much fast seconds when compare to the existing technique.

5. CONCLUSION
A new technique to empower the document clustering by extracting the more relevant concept can be achieved by employing multiple reference point on the documents. The proposed methodology gains more efficiency by gathering more relevant match concepts from the documents with multiple points of reference at the document pair, both the document organization and retrieval can enhance the ability to make the most use of the documents held in storage and make retrieval of ideas as well as relevant task or concept much easier and faster.

The experimental results proves that our proposed method extract document with an accuracy of 79% which are more relevant and matched concepts. It also consumes less time than the existing method. Our proposed method outperforms the existing method. To enhance the proposed approach, carry out the same work to web document clustering.

REFERENCES

M. John Basha and Dr. K. P. Kaliyamurthie

Conference on Information and Knowledge Management, Bethesda, Maryland, USA, Pages 208-217, November3-7, 1998


