Document Clustering and Topic Discovery based on Semantic Similarity in Scientific Literature

J. Jayabharathy[1], S. Kanmani[2] and A. Ayeshaa Parveen[1]
[1]Department of Computer Science & Engineering
[2]Department of Information Technology
Pondicherry Engineering College
Puducherry, India
E-mail: {bharathyraja, kanmani, ayeshaa}@pec.edu

***Abstract*—Unlabeled document collections are becoming increasingly common and mining such databases becomes a major challenge. It is a major issue to retrieve relevant documents from the larger document collection. By clustering the text documents, the documents sharing similar topics are grouped together. Incorporating semantic features will improve the accuracy of document clustering methods. In order to determine at a sight 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. Most of the existing topic discovery methods often assign labels to clusters based on the terms that the clustered documents contain. In this paper a modified semantic-based model is proposed where related terms are extracted as concepts for concept-based document clustering by bisecting k-means algorithm and topic detection method for discovering meaningful labels for the document clusters based on semantic similarity by Testor theory. The proposed method is compared to the Topic Detection by Clustering Keywords method using F-measure and purity as evaluation metrics. Experimental results prove that the proposed semantic-based model outperforms the existing work.**

Keywords-Document clustering; Topic discovery; Semantic similarity; Concept; Testor theory.

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.

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.

Incorporating semantic features improve the accuracy of document clustering methods. Our proposed model is a modification of the existing semantic-based model proposed by Shehata et al [3] which aims to cluster documents by meaning. As in [3] the proposed model captures the semantic structure of each term within a sentence and document rather than the frequency of the term within a document only. Each sentence in a document is labeled by a semantic role labeller considering only the arguments in the verb-argument structure unlike the work in [3]. Based on the semantic-based analysis [3], each labeled term is assigned a weight. The terms that have maximum weights are extracted as top terms. Related terms of each word (in the top terms) are added to the term vector unlike the work in [3] where synonyms/hypernyms of each word are added to the term vector. The reason for extracting related terms of a term instead of synonyms/hypernyms has been discussed in section 3.These concepts are analyzed on the sentence and document levels and used in document clustering and topic discovery. Cosine similarity is similarity measure used. The dataset used for experiments consists of scientific journal articles. In order to extract concepts, a domain specific science dictionary consisting of scientific terms is created unlike the work in [3] where WordNet [4] lexical database is used for synonyms/hypernyms extraction. Domain specific dictionary is used for concept extraction as it eliminates the need for word sense disambiguation (WSD) [5] which is not the scope of this work.

The following section discusses about the existing work and section 3 about the proposed work for document clustering and topic discovery. Section 4 gives the experimental results. Section 5 concludes the paper and discusses about the future enhancements.

978-1-61284-486-2/11/$26.00 ©2011 IEEE

II. EXISTING WORK

Most existing document clustering methods are based on the Vector Space Model (VSM) [8] which is a widely used data representation for text classification and clustering. The VSM represents each document as a feature vector of the terms in the document. Each feature vector contains term-weights of the terms in the document. The TF-IDF [9] weight (term frequency–inverse document frequency) is a weight often used which is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. The similarity between the documents is measured by one of several similarity measures [10] that are based on such a feature vector. Common ones include the cosine measure and the Jaccard measure.

There are various document clustering methods with topic discovery. A survey of document clustering algorithms with topic discovery is presented in [11]. The existing topic detection method used to compare our proposed work is topic detection by Clustering Keywords [6]. In [6] a topic is represented by a cluster of keywords. Therefore a set of keywords for the clustered documents under consideration is needed. The problem of finding a good set of keywords is similar to that of determining term weights for indexing documents. Terms in the cluster of documents with high TF-IDF values are taken as the keywords of that cluster. These keywords are clustered by bisecting k-means algorithm using cosine similarity as the similarity measure. After identifying clusters of keywords, the centers defined are taken as the representations of the topic.

Induced bisecting k-means clustering algorithm as described by [13] is used which is based on the standard bisecting k-means algorithm [12]. A simplified version of the method is as follows. Initially two elements that have the largest distance are selected as the seeds for two clusters. Next all other items are assigned to the cluster closest to one of the two seeds. After all items have been assigned to a cluster, the centers of both clusters are computed. Here a representation of items that naturally allows to define a center which typically is not an item proper but a weighted sum of items is needed. The new centers serve as new seeds for finding two clusters and the process is repeated until the two centers are converged up to some predefined precision. If the diameter of a cluster is larger than a specified threshold value, the whole procedure is applied recursively to that cluster. The algorithm therefore finds a binary tree of clusters.

III. PROPOSED WORK

The proposed model is a modification of the existing semantic-based model proposed by Shehata et al [3] which aims to cluster documents by meaning.

In this proposed work, related terms of the analyzed terms are extracted as concepts unlike the work in [3] where synonyms or hypernyms of the analyzed terms are extracted as concepts. Extracting synonyms or hypernyms as concepts will not give efficient results in the case scientific literature dataset because of the scientific terms involved. For example, the synonyms of the term *protocol* extracted from WordNet for its first sense *communication protocol* are *rule* and *prescript*. Whereas by considering related terms as concepts efficient results are produced. For example, as per the proposed model, the terms *network, protocol, transmission, communication* are extracted as a concept.

The work of this paper relies on the following:

- Extraction of related terms as concepts (discussed in subsection A).
- Usage of semantic-based term analysis [3] which analyzes terms and concepts on the sentence and document levels (discussed in subsection B).
- Applying this proposed semantic model for document clustering with sets of experiments that compare between bisecting k-means clustering algorithm [12] when the extracted features are terms only and terms and their related terms using cosine similarity (discussed in subsection C).
- Concept-based topic discovery by testor theory [7] (discussed in subsection D).

The block diagram shown in Figure 1 illustrates the sequence of steps involved in the proposed work.

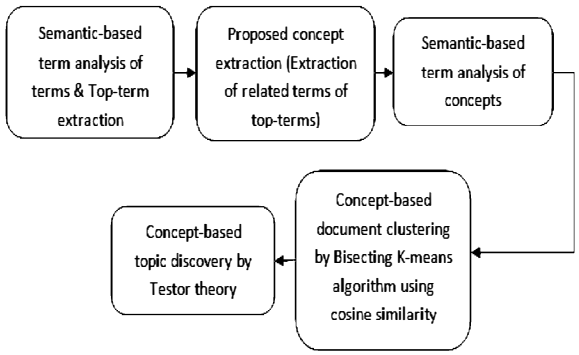

Figure 1. Block diagram of the proposed work.

A. Proposed Concept Extraction

The following proposed algorithm describes the process of related terms extraction for concepts:

1. L is an empty list of concepts
2. **for** each document d_i **do**
3. **for** each labeled term t_j in d_i **do**
4. **for** each labeled term t_k other than term t_j in d_i **do**
5. **if** term t_k is in the definition of term t_j **then**
6. **if** t_j and t_k not in concepts of L **then**
7. add t_j and t_k to new concept c_p
8. **else if** either t_j or t_k in concept c_p of L **then**
9. add t_k or t_j respectively to concept c_p
10. **else if** t_j in c_p and t_k in c_q of L **then**
11. combine c_p and c_q and add them to new concept c_r
12. remove concepts c_p and c_q from L
13. **end if**
14. **end if**

15. **end for**
16. **end for**
17. **end for**

B. Semantic-based term analysis

In the semantic-based term analysis in [3], every sentence is labeled by a semantic role labeler considering verb-argument structure i.e. term is either a verb or an argument. For example, in the statement "Bob caught the ball", *Bob* and *ball* are the arguments and *caught* is the verb. Both the verb and arguments are used for semantic-based term analysis. But in this proposed work, only arguments are taken for semantic-based term analysis. In the case of scientific literature dataset, excluding verbs have given better results. Consider the following two different statements: "magnetic field has generated laser beams" and "marble production has generated waste materials". The verb *generated* does not have any discriminative power in both the sentences. Hence it can be ignored.

The following steps show the process of the semantic-based term analysis [3]:

- Each sentence is labeled by semantic role labeler.
- For each labeled term, stop-words that have no significance are removed.
- Labeled terms are analyzed on the sentence and document levels.
- 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.
- Terms are sorted based on their weights (assigned by the semantic-based analysis) descendingly and top terms are extracted.
- For each word (in a top term), related terms are extracted as given in Concept Extraction algorithm. Terms that have no corresponding concept are still used in text clustering. Extracting the related terms from the top terms only is a kind of pruning to reduce the dimension of the term vector.
- Term vector is extended by adding the corresponding related terms of the top terms.

Term and concept analysis:

To analyze each term at the sentence-level, a sentence-based frequency measure, called the conceptual term frequency *ctf* is utilized. The *ctf* is the number of occurrences of term *t* in arguments of sentence *s*. The term *t*, which frequently appears in different arguments 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 (*ctf*) value of its corresponding top term.

To analyze each concept at the document-level, the concept frequency *cf* is proposed, the number of occurrences of a concept *c* in the document, is calculated. At this point, each term and it's corresponding concepts have the same measures which are the *ctf* and *cf* (for concept and term) on the sentence and document levels respectively.

Semantic-based Analyzer Algorithm:

1. doc_i is a new Document where $doc_i = \{1, 2, .., N\}$ and N is a total number of documents
2. L is an empty List (L is a related terms list)
3. T is an empty List (T is a terms list)
4. **for** each sentence s in d **do**
5. **for** each labeled term t in d **do**
6. compute tf_i of t_i in d
7. compute ctf_i of t_i in s in d
8. compute the $weight_i = tf_i + ctf_i$
9. add term t with $weight_i$ to T
10. **end for**
11. **end for**
12. sort T descendingly based on *weight*
13. output the *max*(*weight*) from list T
14. **for** each term t in T that has *max*(*weight*) **do**
15. extract related terms of t by the proposed concept extraction method
16. add concept c to L
17. **end for**
18. **for** each concept c_i in L **do**
19. compute cf_i of c_i in d
20. assign ctf_i to c_i that corresponds to t_i
21. compute concept c_i $weight = cf_i + ctf_i$

The semantic-based analyzer algorithm describes the process of calculating the *cf* and the *ctf* of all the concepts in the documents. Each document is represented as a feature vector of the concepts (terms and related terms) in the document. Each feature vector contains concept-weights ($cf_i + ctf_i$) of the concepts in the document.

C. Similarity Measure

In [3] a semantic-based similarity measure is used which is calculated based on the matching concepts between two documents. Hence semantic-based similarity measure cannot be used for centroid-based clustering algorithms like K-means, Bisecting K-means, etc. In this proposed work, cosine similarity is the similarity measure used. Cosine similarity [10] is the most common similarity measure to measure the similarity between documents which is defined as:

$$sim_{cos}(d_1, d_2) = (d_1 \cdot d_2) / ||d_1|| \, ||d_2|| \quad (1)$$

where **.** indicates the vector dot product and ||d|| is the length of vector d. Given a set, S, of documents and their corresponding vector representations, we define the centroid vector c to be

$$c = 1 / |S| \, (\sum_{d \in S} d) \quad (2)$$

which is nothing more than the vector obtained by averaging the weights of the various terms present in the documents of *S*. Analogous to documents, the similarity between two centroid vectors and between a document and a centroid vector are computed using the cosine measure, i.e.,

$sim_{cos}(d, c) = (d \cdot c) / ||d|| \, ||c||$ (3)

D. Testor Theory

Testor theory [7] is used to give each cluster a tag after clusters are formed. Topic detection by testor theory [7] involves construction of a learning matrix and a comparison matrix.

Learning 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:

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:

$d(v_{ik}, v_{jk}) = 1$ *if* $v_{ik} - v_{jk} \geq \delta$

0 *otherwise* (4)

where v_{ik}, v_{jk} 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 c_k) different if the concept c_k 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.

IV. Experimental Results

The dataset used for the experimental setup contains 500 abstract articles collected from the ScienceDirect digital library. The articles are classified according to the ScienceDirect classification system into four different categories: computer networks and communications, nuclear and high energy physics, economics and econometrics, and civil and structural engineering.

F-measure and purity are the performance measures used to evaluate the quality of document clustering and topic discovery. F-measure [12] combines the precision and recall ideas from information retrieval. Each cluster is treated as if it were the result of a query and each class as if it were the desired set of documents for a query. The recall and precision of that cluster for each given class are calculated. More specifically, F-measure for cluster j and class i is calculated as follows: $Recall(i, j) = n_{ij} / n_i$, $Precision(i, j) = n_{ij} / n_j$ *and* $F(i, j) = (2 * Recall(i, j) * Precision(i, j)) / ((Precision(i, j) + Recall(i, j))$ where n_{ij} is the number of members of class i in cluster j, n_j is the number of members of cluster j and n_i is the number of members of class i. For each class, only the cluster with the highest F-measure is selected. Finally, the overall F-measure of a clustering solution is weighted by the size of each cluster:

$F(S) = \sum_j n_j/n \, max(F(i, j))$ (5)

The purity measure [10] evaluates the coherence of a cluster, that is, the degree to which a cluster contains documents from a single class. Given a particular cluster C_i of size n_i, the purity of C_i is formally defined as:

$P(C_i) = 1 / n_i \, max(n_{ih})$ (6)

where $max(n_{ih})$ is the number of documents that are from the dominant class in cluster C_i and n_{ih} represents the number of documents from cluster C_i assigned to class h. The overall purity of a clustering solution is:

$Purity(S) = 1/n \sum_i max(n_{ih})$ (7)

Figure 2 shows the comparison of F-measure for document clustering by single-term only and document clustering by term and related terms, varying the total number of clusters. Figure 3 shows the comparison of Purity for document clustering by single-term only and document clustering by term and related terms, varying the total number of clusters. For the term and related terms weighting, the percentage of improvement ranges from +22.81% to +47.15% increase in the F-measure quality, and +25.00% to +48.83% increase in Purity.

Figure 4 shows the comparison of F-measure for topic discovery by clustering keywords method and concept-based topic discovery by testor theory, varying the total number of documents. Figure 5 shows the comparison of Purity for topic discovery by clustering keywords method and concept-based topic discovery by testor theory, varying the total number of documents. For the concept-based topic discovery by testor theory, the percentage of improvement ranges from +23.06% to +77.89% increase in the F-measure quality, and +48.66% to +75.62% increase in Purity.

V. Conclusion and Future Enhancements

The key contributions of this paper are the extraction of related terms of the analyzer terms as concepts for concept-based document clustering and concept-based topic discovery by testor theory. Document clustering by term and related terms is compared to the document clustering by single-term 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 document clustering by term and related terms is more efficient than document clustering by single-term only 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 strategy to avoid the use of domain specific dictionary. Another future work is to perform experiments on different datasets of various categories of documents.

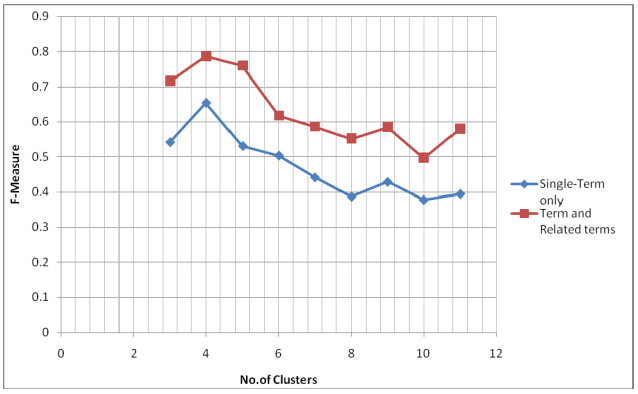

Figure 2. Document Clustering comparison of F-measure: Single-Term only Vs Term and Related terms.

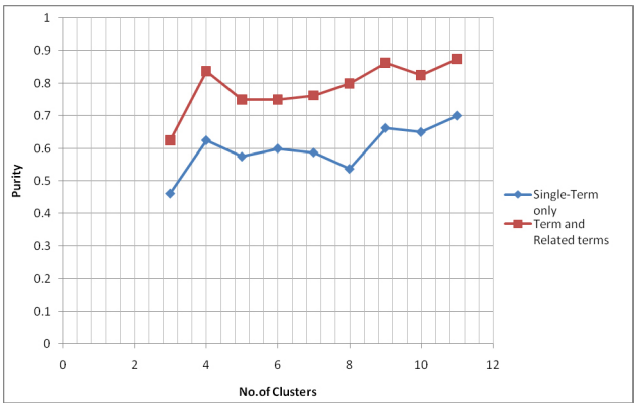

Figure 3. Document Clustering comparison of Purity: Single-Term only Vs Term and Related terms.

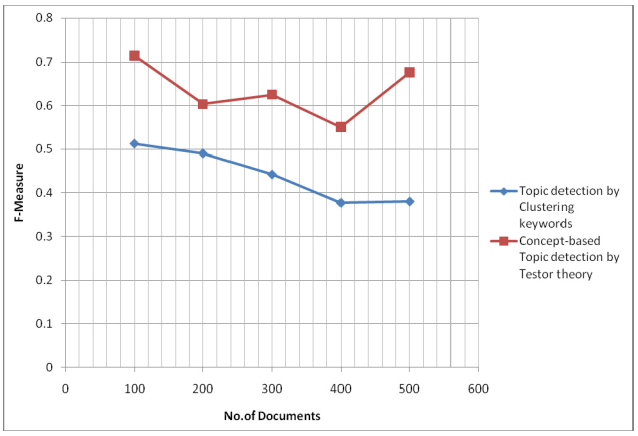

Figure 4. Topic Discovery comparison of F-measure: Topic detection by Clustering keywords Vs Concept-based Topic detection by Testor theory.

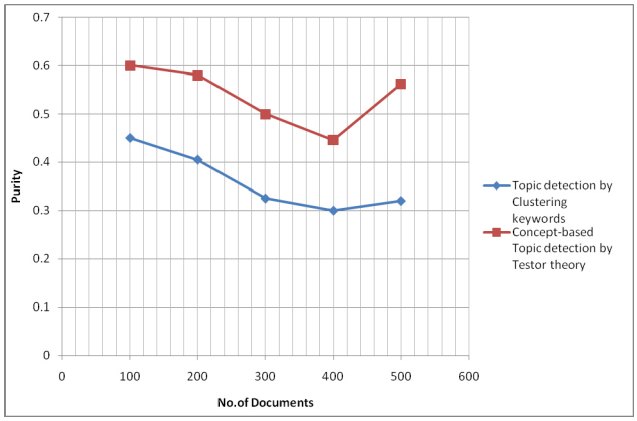

Figure 5. Topic Discovery comparison of Purity: Topic detection by Clustering keywords Vs Concept-based Topic detection by Testor theory.

REFERENCES

[1] http://en.wikipedia.org/wiki/Information_extraction.

[2] Jiawei Han and Micheline Kamber, "Data Mining Concepts and techniques", Second Edition.

[3] Shady Shehata, "AWordNet-based Semantic Model for Enhancing Text Clustering", *IEEE International Conference on Data Mining Workshops*, 2009.

[4] G. A. Miller, "Wordnet: a lexical database for english," *Commun. ACM*, vol. 38, no. 11, pp. 39-41, 1995.

[5] http://en.wikipedia.org/wiki/Word_sense_disambiguation

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

[7] Fang li, Qunxiong zhu and Xiaoyong lin, "Topic Discovery in Research literature Based on Non-negative Matrix Factorization and Testor theory", *IEEE Asia-Pacific Conference on Information Processing*, 2009.

[8] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval. McGraw-Hill*, 1983.

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

[10] Anna Huang, "Similarity Measures for Text Document Clustering", *NZCRSC '08*, April 2008.

[11] J. Jayabharathy, S. Kanmani and A. Ayeshaa Parveen, "A Survey of Document Clustering Algorithms with Topic Discovery", *Journal of Computing*, Vol. 3, Issue 2, February 2011.

[12] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," in Knowledge Discovery and Data Mining (KDD) Workshop on TextMining, August 2000.

[13] F. Archetti, P. Campanelli, E. Fersini, and E. Messina, "A hierarchical document clustering environment based on the induced bisecting k-means", In H. L. Larsen, G. Pasi, D. O. Arroyo, T. Andreasen, and H. Christiansen, editors, FQAS, volume 4027 of *Lecture Notes in Computer Science, Springer*, pp 257–269, 2006.