RESEARCH ARTICLE OPEN ACCESS

Improving Text Clustering Quality by Concept Mining

Pradnya Randive*, Nitin Pise**
*(Department of Computer Engineering, Pune University, India)
** (Department of Computer Engineering, Pune University, India)

ABSTRACT
In text mining most techniques depends on statistical analysis of terms. Statistical analysis trances important terms within document only. However this concept based mining model analyses terms in sentence, document and corpus level. This mining model consist of sentence based concept analysis, document based and corpus based concept analysis and concept based similarity measure. Experimental result enhances text clustering quality by using sentence, document, corpus and combined approach of concept analysis.

Keywords - Text Mining, Concept Based Mining, Text Clustering.

I. INTRODUCTION

Text mining mostly used to find out unknown information from natural language processing and data mining by applying various techniques.

In this technique for discovering the importance of term in document, term frequency of term is calculated. Sometime we can notice that two terms having same frequency in document but one term leads more meaning than other, for this concept based mining model is intended. In proposed model three measures are evaluated for analyzing concept in sentence, document and corpus levels. Semantic role labeler is mostly associated with semantic terms. The term which has more semantic role in sentence, it's known as Concept. And that Concept may be either word or phrase depending on sentence of semantic structure. When we put new document in system, the proposed model discover concept match by scanning all new documents and take out matching concept. Similarity measure used for concept analysis on sentence, document and corpus level that exceeds similarity measures depending on the term analysis model of document. The results are measured by using F- measure and Entropy. This model we are going to used for web documents.

II. PREVIOUS WORK

Text mining based on statistical analysis terms. The term frequency of statistical analysis discovers the important terms within document only. This model efficiently search substantial concept within document based on new concept based similarity measure. This allows concept matching and similarity calculations in document by a robust and accurate way. Improves text clustering quality surpasses traditional single term approach. And this work extends to improve accuracy of similarity measure of document. And another way of extending this work to web document clustering [2]. Data mining contains methods for handling huge databases with resources such as memory and computation time. bEMADS and gEMADS these two algorithms are used based on Gaussian mixture model. Both resumes data into subcluster and after that generate Gaussian mixture. These two algorithms run several orders of magnitude faster than maximum with little loss of accuracy [3]. This paper shows solutions by weights on the instance points, making clustering as hardest. Here we can find out the best knowledge about formalization. Weight modifications depend on local variation. To get best result resembling algorithms and weighted version of clustering algorithms like k-means, fuzzy c-means, Expectation Maximization(EM) and k-harmonic [4].

Natural language processing play role like information extraction, question answering and summarization. Support vector machine improves performance over earlier classifier. Here task is reformulated as combined chunking and classification problem for which statistical syntactic parser may not be available. For the task of classification, head word and predicate were the most salient features, but may be difficult to estimate because of data sparsity. The portability of the system to new genres of data, we found a significant drop in performance for a text source different than the one the system was trained on [5].

III. SYSTEM ARCHITECTURE

The techniques prescribed in the previous work are used for document clustering. But it is only for documents present on system. In the proposed system we are going to use web documents and we will get the clustered output and that have shown in fig.1. This concept-based mining model consists of sentence-based concept analysis, document-based concept analysis, corpus-based concept-analysis, and concept-based similarity measure, as depicted in Figure 1. A web document is given as the input to the proposed model. Each document has well-defined sentence boundaries. Each sentence in the document is labeled automatically based on the Prop Bank notations. After running the semantic role labeler, labeled verb argument structures, the output of the role labeling task, are captured and analyzed by the concept-based mining model on sentence, document, and corpus level. In the concept-based mining model, a labeled terms either word or phrase is considered as concept.

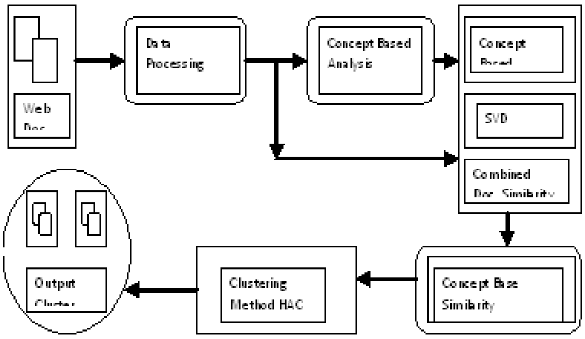

Fig 1. Concept Based Mining in Text clustering

The proposed model contains the following modules:

A. *Web Document*

Web document is given as Input to the given system. Here user can give any query to the browser. Pure HTML pages are selected by removing extra scripting. Web pages contain data such as hyperlinks, images, script. So it is necessary to remove such unwanted script if any, during the time when a page is selected for processing. The HTML code is then transferred into XML code. On that document next process is processed that is Text pre-processing or data processing.

B. *Data Processing*

First step is separate sentences from the documents. After this label the terms with the help of Prop Bank Notation. With the help of Porter algorithm remove the stem word and stop words from the terms.

C. *Concept Based Analysis*

This is important module of the proposed system. Here we have to calculate the frequencies of the terms. Conceptual term frequency (ctf), Term frequency (tf) and Document frequency (df) are calculated

1) *Sentence Based Concept Analysis*

For analyzing every concept at sentence level, concept based frequency measure; called conceptual term frequency is used.

a) *Calculating ctf in sentence s*

Ctf is the number of occurrences of concept c in verb structure of sentence s. If concept c frequently appears in structure of sentence s then it has principal role of s.

b) *Calculating ctf in document d*

A concept c can have many ctf values in different sentences in the same document d. Thus, the ctf value of concept c in document d is calculated by:

$$ctf = \frac{\sum_{n=1}^{sn} ctf_n}{sn},$$

Where sn: total number of sentences containing concepts in document d

2) *Document Based Concept Analysis*

For analyzing concepts at document level term frequency tf in original document is calculated. The tf is a local measure on the document level.

3) *Corpus Based Concept Analysis*

To calculate concepts from documents, document frequency df is used. Document Frequency df is the global measure. With the help of Concept based Analysis Algorithm we can calculate ctf, tf, df.

D. *Similarity Approach*

This module mainly contains three parts. Concept based similarity, Singular Value Decomposition and combined based similarity it contains. Here we get that how many percentage of concept math with the given web document.

E. *Concept Based Similarity*

A concept-based similarity measure depends on matching concept at sentence, document, and corpus instead of individual terms. This similarity measure based on three main aspects. First is analyzed label terms that capture semantic structure of each sentence. Second is concept frequency that is used to measure participation of concept in sentence as well as document. Last is the concepts measured from number of documents.

Concept based similarity between two document is calculated by :

$$sim_c(d_1, d_2) = \sum_{i=1}^{m} max\left(\frac{l_{i_1}}{Lv_{i_1}}, \frac{l_{i_2}}{Lv_{i_2}}\right) \times weight_{i_1} \times weight_{i_2}$$

Term frequency is calculated by following formula:

$$tf\ weight_i = \frac{tf_{ij}}{\sqrt{\sum_{j=1}^{cn} (tf_{ij})^2}},$$

F. *Clustering Techniques*

This module used three main basic techniques like Single pass, Hierarchical Agglomerative Clustering, and K-Nearest Neighbor. With the help of these techniques we can get that which cluster is having highest priority.

G. *Output Cluster*

Last module is the output Cluster. After applying the clustering techniques we get clustered document. That will help to find out main concepts from the web document

IV. SYSTEM IMPLEMENTATION

Proposed system model illustrates flow of implementation as shown in Fig.1. First, web document given input to the system where, HTML pages are collected and their XML conversion is carried out. In the second module that is in Text Processing carried out separate sentences, label terms, and removing stop words and stem words. Third module Concept based analysis measures conceptual term frequency (ctf), term frequency (tf), and document frequency (df). Next module concept based document similarity find out how many percentage of concept is similar to the given concept. After this applying the clustering techniques like single pass, HAC algorithm and KNN algorithm on calculated frequencies. So finally we get clustered output with matching concepts.

V. RESULT ANALYSIS

To test the effectiveness of concept matching in determining an accurate measure of the similarity between documents, extensive sets of experiments using the concept-based term analysis and similarity measure are conducted. In the data sets, the text directly is analyzed, rather than, using metadata associated with the text documents. Following are some graphs for the implementation of this system:

Precision (i,j)= Mij / Mj

Recall (i,j)= Mij / Mi

F-Score = 2PR / P+R

Where Mij is the numbers of members of class i in cluster j, Mj is the number of members of cluster j, and Mi is the number of members of class i.

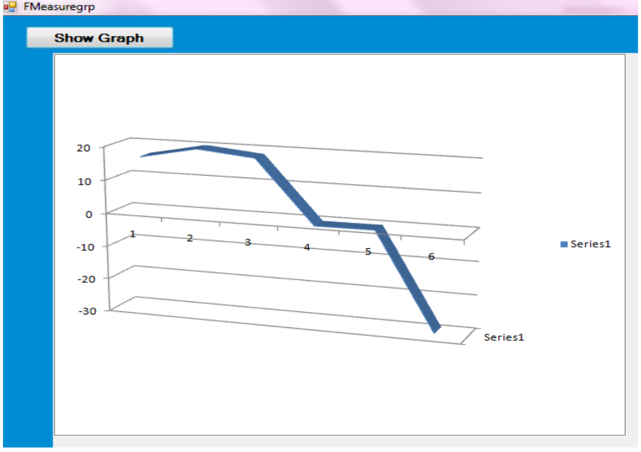

Fig 2: Graph for F-Measure

Entropy is calculated by following formula:

$$E_j = -\sum_i p_{ij} log(p_{ij})$$

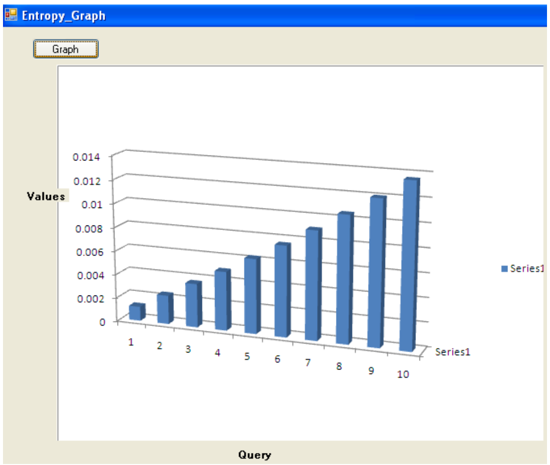

Fig 3: Graph for Entropy

VI. CONCLUSION

Concept based mining model is composed of four components and that improves text clustering quality. The first component is the sentence-based concept analysis which analyzes the semantic structure of each sentence to capture the sentence concepts using the proposed conceptual term frequency ctf measure. Then, the second component, document-based concept analysis, analyzes each concept at the document level using the concept-based term frequency tf. The third component analyzes concepts on the corpus level using the document frequency df global measure. The fourth component is the concept-based similarity measure which allows measuring the importance of each concept with respect to the semantics of the sentence, the topic of the document, and the discrimination among documents in a corpus. We can apply the same model for the Text Classification.

REFERENCES

[1] Shehata, Fakhri Karray, and Mohamed S. Kamel, “An Efficient Concept-Based Mining Model for Enhancing Text Clustering”, IEEE Transaction on Knowledge and Data Engg, VOL. 22, NO. 10, OCTOBER 2010

[2] S. Shehata, F. Karray, and M. Kamel, “Enhancing Text Clustering Using Concept-Based Mining Model,” Proc. Sixth IEEE Int’l Conf. Data Mining (ICDM), 2006

[3] M.-L. Wong, and K.S. Leung, “Scalable Model-Based Clustering for Large Databases Based on Data Summarization,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 27, no. 11, pp. 1710-1719, Nov. 2005.

[4] R. Nock and F. Nielsen, “On Weighting Clustering,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 28, no. 8, pp. 1223-1235, Aug. 2006

[5] S. Pradhan, K. Hacioglu, V. Krugler, W. Ward, J.H. Martin, and D. Jurafsky, “Support Vector Learning for Semantic Argument

Classification," Machine Learning, vol. 60, nos. 1-3,pp. 11-39, 2005.

[6] Pradhan, W. Ward, K. Hacioglu, J. Martin, and D. Jurafsky, "Shallow Semantic Parsing Using Support Vector Machines," Proc. Human Language Technology/North Am. Assoc. for Computational Linguistics (HLT/NAACL), 2004

[7] S. Pradhan, K. Hacioglu, W. Ward, J.H. Martin, and D. Jurafsky, "Semantic Role Parsing: Adding Semantic Structure to Unstructured Text," Proc. Third IEEE Int'l Conf. Data Mining (ICDM), pp. 629-632, 2003.

[8] P. Mitra, C. Murthy, and S.K. Pal, "Unsupervised Feature Selection Using Feature Similarity," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, no. 3, pp. 301-312, Mar. 2002.