ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.2 , No.1, Pages : 154-158 (2013)
Special Issue of ICACSE 2013 - Held on 7-8 January, 2013 in Lords Institute of Engineering and Technology, Hyderabad

Concept Based Mining Model for Text Clustering

Gousiya Begum[1], N. Musrat sultana[2]
[1]MGIT, Hyderabad, India, gousiyabegum@gmail.com
[2]MGIT, Hyderabad, India, sultanamgit@yahoo.com

Abstract : The common techniques in text mining are based on the statistical analysis of a term either word or phrase. Statistical analysis of a term frequency captures the importance of the term within a document only. Two terms can have the same frequency in their documents, but one term contributes more to the meaning of its sentences than the other term. A new concept-based mining model that analyzes terms in the sentence, document level and corpus level is introduced.

The concept based mining model can effectively discriminate between non important terms with respect to sentence semantics and terms which hold the concepts that represent the sentence meaning. The proposed model consists of sentence-based concept analysis, document-based concept analysis, corpus based concept analysis and concept-based similarity measure in calculating the similarity between documents.

Keywords: Concept-based mining model, sentence-based, document-based, corpus-based, concept analysis, concept-based similarity

1.INTRODUCTION

In text mining techniques, the term frequency of a term (word or phrase) is computed to explore the importance of the term in the document. However, two terms can have the same frequency in their documents, but one term contributes more to the meaning of its sentences than the other term. It is important to note that extracting the relations between verbs and their arguments in the same sentence has the potential for analyzing terms within a sentence [1]. The information about *who is doing what to whom* clarifies the contribution of each term in a sentence to the meaning of the main topic of that sentence. In this paper, a novel concept-based mining model is proposed which captures the semantic structure of each term within a sentence and document rather than the frequency of the term within a document only.

In the proposed model, three measures for analyzing concepts on the sentence, document, and corpus levels are computed. Each sentence is labeled by a semantic role labeler that determines the terms which contribute to the sentence semantics associated with their semantic roles in a sentence. Each term that has a semantic role in the sentence, is called a concept. Concepts can be either words or phrases and are totally dependent on the semantic structure of the sentence. When a new document is introduced to the system, the proposed mining model can detect a concept match from this document to all the previously processed documents in the data set by scanning the new document and extracting the matching concepts. This similarity measure outperforms other similarity measures that are based on term analysis models of the document only. The similarity between documents is based on a combination of sentence-based, document-based, and corpus-based concept analysis.

Generally, text document clustering methods attempt to segregate the documents into groups

where each group represents some topic that is different than those topics represented by the other groups [2], [3]. Most current document clustering methods are based on the Vector Space Model (VSM) [4], [5], which is a widely used data representation for text classification and clustering. The VSM represents each document as a feature vector of the terms (words or phrases) in the document. Each feature vector contains term weights (usually term frequencies) of the terms in the document. The similarity between the documents is measured by one of several similarity measures that are based on such a feature vector. Examples include the cosine measure and the Jaccard measure.

The similarity between the documents is measured by one of several similarity measures that are based on such a feature vector. Examples include the cosine measure and the Jaccard measure and Proximity. Proximity will check the similarity between two documents very precisely.

Hence, we propose an efficient concept based mining model to find the concepts available in the source files. The rest of the paper is described as follows. Section 2 briefs about the literature survey. The concept of thematic roles is described in Section 3 and the proposed methodology is explained with necessary diagrams in Section 4. The Results obtained in the proposed method is discussed in Section 5 and Section 6 concludes the work.

2. RELATED WORK

Pradhan et al. has cast tagging problem [6]. In his work he has researched that automatic, accurate and wide-coverage techniques that can annotate naturally occurring text with semantic argument structure can play a key role in NLP applications such as Information Extraction, Question Answering and Summarization. Shallow semantic parsing – the process of assigning a simple WHO did WHAT to WHOM, WHEN, WHERE, WHY, HOW, etc. structure to sentences in text, is the process of producing such a markup. When presented with a sentence, a parser should, for each predicate in the sentence, identify and label the predicate's semantic arguments. This process entails identifying groups of words in a sentence that represent these semantic arguments and assigning specific labels to them.

International Journal of Advanced Trends in Computer Science and Engineering, Vol.2 , No.1, Pages : 154-158 (2013)
Special Issue of ICACSE 2013 - Held on 7-8 January, 2013 in Lords Institute of Engineering and Technology, Hyderabad

Gruber and Fillmore proposed Thematic roles. Generally, the semantic structure of a sentence can be characterized by a form of verb argument structure. The study of the roles associated with verbs is referred to a thematic role or case role analysis [1]. Thematic roles, are sets of categories that provide a shallow semantic language to characterize the verb arguments.

Marcus et al., 1994 proposed that predicate argument relations are marked for part of the verbs. He has done his work on results using PropBank1 (Kingsbury et al., 2002), a 300k-word corpus in which predicate argument relations are marked for part of the verbs in the Wall Street Journal (WSJ) part of the Penn Tree- Bank (Marcus et al., 1994) [6]. The arguments of a verb are labeled ARG0 to ARG5, where ARG0 is the PROTOAGENT (usually the subject of a transitive verb) ARG1 is the PROTO-PATIENT (usually its direct object), etc. PropBank attempts to treat semantically related verbs consistently. In addition to these CORE ARGUMENTS, additional ADJUNCTIVE ARGUMENTS, referred to as ARGMs are also marked. Some examples are ARGMLOC, for locatives, and ARGM-TMP, for temporal.

Fillmore described a shallow semantic interpreter [7] based on semantic roles that are less domain specific than to airport or joint venture company. These roles are defined at the level of semantic frames (1976), which describe abstract actions or relationships, along with their participants.

Gildea and Jurafsky were the first to apply a statistical learning technique to the FrameNet database [6]. They presented a discriminative model for determining the most probable role for a constituent, given the frame, predicator, and other features.

S. Y. Lu we proposed a syntactic clustering procedure, in which each formed cluster is

characterized by a pattern grammar [8]. Therefore, the procedure yields not only the clustering results grammar for each cluster. In order to do so, a grammar must be inferred when a new cluster is initiated, and later it is updated whenever an input pattern is added to the same cluster. Error-correcting parsers are employed to measure the distance between an input pattern and the languages generated from the inferred grammars. The input pattern is then classified according to the nearest neighbor syntactic recognition rule. The emphasis of the syntactic clustering

procedure is the use of grammar in which the hierarchy of the structure of patterns is described.

S. Kaski et al., in his work discussed that One of the traditional methods of searching for texts that match a query is to index all the words (hereafter called *terms*) that have appeared in the document collection [9]. The query itself, typically a list of appropriate keywords, is compared with the term list of each document to find documents that match the query. Terms can be combined by Boolean logic in order to control the breadth of matching. The following three fundamental problems in applying Boolean logic to text retrieval (see, e.g., [39]) make it an unsatisfactory solution. (1) Recall and precision2 of retrieval are sensitive to small changes in the formulation of a query. For Boolean queries there is no simple way of controlling the size of the output, and the output is not ranked in the order of relevance. (2) The results of a query offer no indication on how many valuable documents were *not* retrieved, especially if the document collection is unfamiliar. (3) If the domain of the query is not known well it is difficult to formulate the query, i.e., to select the appropriate keywords.

3. THEMATIC ROLES

All human languages at their core semantic structure seem to have a form of predicate argument layout. This underlying structure allows the creation of a composite meaning representation from the meaning of the individual parts of a linguistic input. Grammar has an important role for supporting the predicate-argument structure. Consider the following examples:

- My daughter wants a doll.
- My son wants to play outside.

These two examples can be classified as having one of the following syntactic argument frames: (Noun Phrase(NP) wants NP) or (NP wants Infinitive Phrase (Inf-NP)). In this case, some facts could be driven for the particular predicate ”wants”

- There are two arguments to this predicate
- Both arguments must be NPs
- The first argument is pre-verbal and plays the role of the subject
- The second argument is a post-verbal and plays the role of the direct object.

In each of these cases, the pre-verbal argument always acts as the entity doing the wanting, while the post-verbal plays the role of the entity that is wanted. More generally, the predicate argument structure permits the link between the arguments in surface structures of the input text and their associated semantic roles. The study of roles associated with verbs is usually referred to thematic role or case role analysis [10].

Thematic role is a synonym of semantic role. A Semantic role is the role played by a participant in a situation. Thematic roles, first proposed by Gruber and Fillmore are set of categories that provide a shallow semantic language to characterize the verb arguments.

4. CONCEPT BASED MINING MODEL PROCESS

The proposed concept-based mining model consists of sentence-based concept analysis, document-based concept analysis, corpus-based concept-analysis, and concept-based similarity measure. A raw text document is the input to the proposed model. Each document has well-defined sentence boundaries. Each sentence in the document is labeled automatically based on

parser. After running the semantic role labeler, each sentence in the document might have one

or more labeled verb argument structures. In this model, both the verb and the argument are considered as terms. One term can be an argument to more than one verb in the same sentence. This means that this term can have more than one

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.2 , No.1, Pages : 154-158 (2013)
Special Issue of ICACSE 2013 - Held on 7-8 January, 2013 in Lords Institute of Engineering and Technology, Hyderabad

semantic role in the same sentence. In such cases, this term plays important semantic roles that contribute to the meaning of the sentence. In the concept-based mining model, a labeled term either word or phrase is considered as concept.

The System architecture consists of the following main modules:

- Text preprocessing
- Concept Analysis and
- Concept based similarity measure

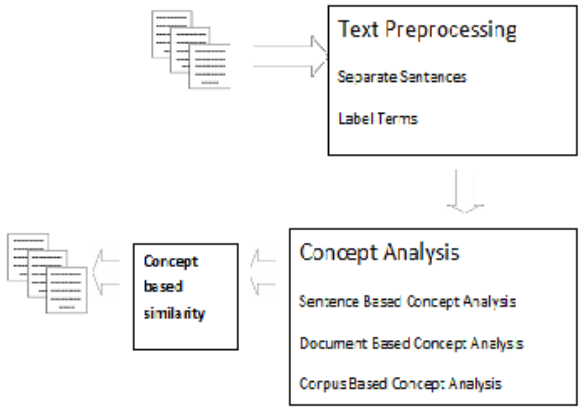

Figure 4.1 Architecture of Concept Based Model

Figure 4.1 is an Architecture of Concept Based model and it consists of sentence-based concept analysis, document-based concept analysis and concept-based similarity measure.

4.1. Text Preprocessing

- **Label Terms**

A raw text document is the input to the proposed model. Each document has well defined sentence boundaries [1]. Each sentence in the document is labeled automatically based on the parser. After running the semantic role labeler, each sentence in the document might have one or more labeled verb argument structures. The labeled verb argument structures, the output of the role labeling task, are captured and analyzed by the concept-based mining model on sentence, document levels. In this model, both the verb and the argument are considered as terms. One term can be an argument to more than one verb in the same sentence. This means that this term can have more than one semantic role in the same sentence. In such cases, this term plays important semantic roles that contribute to the meaning of the sentence. In the concept-based mining model, a labeled term either word or phrase is considered as concept.

- **Removing stop words**

In computing **stop words** are words which are filtered out prior to, or after, processing of natural language data (text). It is controlled by human input and not automated. There is not one definite list of stop words which all tools use, if even used. Some tools specifically avoid using them to support phrase search.

- **Stem words**

In linguistic morphology, **stemming** is the process for reducing inflected (or sometimes derived) words to their stem, base or root form – generally a written word form. The stem need not be identical to the morphological root of the word; it is usually sufficient that related words map to the same stem, even if this stem is not in itself a valid root. Algorithms for stemming have been studied in computer science since 1968. Many search engines treat words with the same stem as synonyms as a kind of query broadening, a process called **conflation**. Stemming programs are commonly referred to as **stemming algorithms** or **stemmers**.

4.2 Concept Analysis

To analyze each concept at the sentence level is called as sentence based concept analysis.

Consider the following sentence:

"Texas and Australia researchers have created industry-ready sheets of materials made from nanotubes that could lead to the development of artificial muscles"

In this example, stop words are removed and concepts are shown without stemming for better readability as follows:

1. Concepts in the first verb-argument structure:
• Texas
• created
• industry-ready sheets of material nanotubes lead development of artificial muscles
2. Concepts in the second verb-argument structure:
• materials
• nanotubes lead development artificial muscles
3. Concepts in the third verb-argument structure:
• nanotubes
• lead
• development artificial muscles

It is imperative to note that these concepts are extracted from the same sentence. Thus, the concepts mentioned in this example sentence are:

• Texas
• Australia
• researchers
• created
• industry
• ready
• sheets
• materials
• nanotubes
• lead
• development
• artificial
• muscles

After finding the concepts at sentence level, concepts are also found at document level.

4.3 Concept Based Similarity Measure

A concept-based similarity measure, based on matching concepts at the sentence, document is devised. The concept-based similarity measure relies on three critical aspects.

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.2 , No.1, Pages : 154-158 (2013)
Special Issue of ICACSE 2013 - Held on 7-8 January, 2013 in Lords Institute of Engineering and Technology, Hyderabad

First, the analyzed labeled terms are the concepts that capture the semantic structure of each sentence. Second, the frequency of a concept is used to measure the contribution of the concept to the meaning of the sentence, as well as to the main topics of the document. Last, the number of documents that contains the analyzed concepts is used to discriminate among documents in calculating the similarity.

In order to calculate the similarity Jaccard Distance and Proximity measures are used.

Jaccard Distance measure shows the dissimilarity between two objects whereas Proximity measure shows the similarity between two objects.

5. RESULTS AND DISCUSSION

The proposed concept based mining model has been implemented in the workingplatform of JAVA.

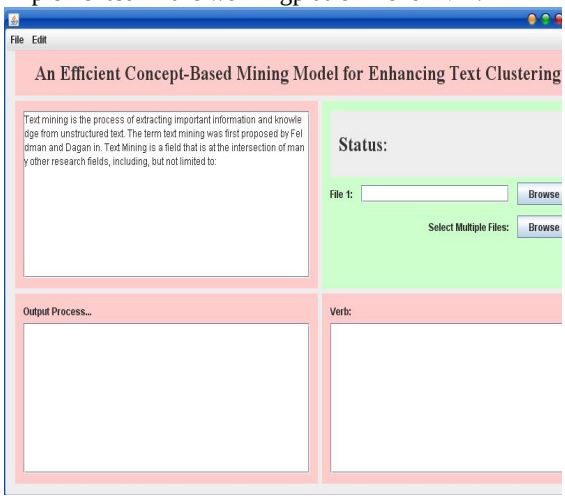

Figure 5.1 Reading the text

Figure 5.1 is used by the user enter the data through the keyboard which is read and after reading the data from the user it has to be saved to a file from which the further processing of finding the concepts will carry on. This is the basic screen.

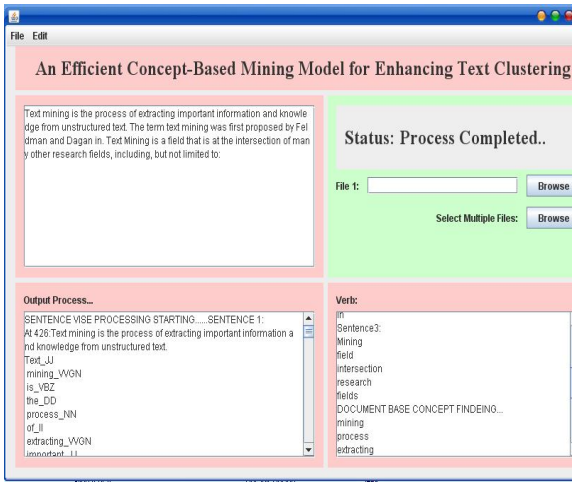

Figure 5.3 Sentence and document based concepts

Figure 5.3 will show how the processing is done when the user select the check option from the Edit menu and the processing is shown in the output process frame where as the final output i.e., the concepts are displayed in the verb frame. The evaluation of the implemented Concept Based Model is performed on a estimating a Hidden Markov model based on a standard corpus, parsing, tokenization, chunking and sentence model.

The major work is getting the tags based on different methods like firstBest, nBest, confidence of tokenization. These methods are used to get the efficient tags based on the probability of the tag compared to neighboring words. After getting tags based on verb and its corresponding arguments getting the concepts

Distance and Proximity is calculated between the text documents in order to check the similarity between the documents.

Proximity is calculated so that the concepts are retrieved based on Precision.

The concepts which are displayed gives the semanticness of the text in the document and later on these concepts can be used to cluster the documents.

6. CONCLUSIONS

This work bridges the gap between natural language processing and text mining disciplines. A new concept based mining model composed of four components i.e sentence based concept analysis, documents based concept analysis, corpus based concept analysis and concept based similarity measure is proposed to improve the text clustering quality. By exploiting the semantic structure of the sentences in documents, a better text clustering result is achieved. By combining the factors affecting the weights of concepts on the sentence, document, and corpus levels, a concept-based similarity measure that is capable of the accurate calculation of pairwise documents is devised. This allows performing concept matching and concept-based

similarity calculations among documents in a very robust and accurate way. The quality of text clustering achieved by this model significantly surpasses the traditional single term based approaches.

There are a number of possibilities for extending this work. One direction is to link this work to Web document clustering. Another direction is to apply the same model to text classification. The intention is to investigate the usage of such model on other corpora and its effect on classification, compared to that of traditional methods.

REFERENCES

[1] **[1]** Shady Shehata, Fakhri Karray and Mohamed S. Kamel, "An Efficient Concept-Based Mining Model for Enhancing Text Clustering", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 22, No.10, pp. 1360 – 1371, October 2010.

[2] B. Frakes and R. Baeza-Yates, Information Retrieval: Data Structures and Algorithms. Prentice Hall, 1992.

[3] K. Aas and L. Eikvil, "Text Categorisation: A Survey," Technical Report 941, Norwegian Computing Center, June 1999.

[4] G. Salton, A. Wong, and C.S. Yang, "A Vector Space Model for Automatic Indexing," Comm. ACM, vol. 18, no. 11, pp. 112-117, 1975.

International Journal of Advanced Trends in Computer Science and Engineering, Vol.2 , No.1, Pages : 154-158 (2013)
Special Issue of ICACSE 2013 - Held on 7-8 January, 2013 in Lords Institute of Engineering and Technology, Hyderabad

[5] G. Salton and M.J. McGill, Introduction to Modern Information Retrieval. McGraw-Hill, 1983.
[6] S. 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] C. Fillmore, "The Case for Case," Universals in Linguistic Theory, Holt, Rinehart and Winston, 1968.
[8] S.Y. Lu and K.S. Fu, "A Sentence-to-Sentence Clustering Procedure for Pattern Analysis," IEEE Trans. Systems, Man, and Cybernetics, vol. 8, no. 5, pp. 381-389, May 1978.
[9] T. Honkela, S. Kaski, K. Lagus, and T. Kohonen, "WEBSOM—Self-Organizing Maps of Document Collections," Proc. Workshop Self-Organizing Maps (WSOM '97), 1997.
[10] D. Jurafsky and J.H. Martin, Speech and Language Processing. Prentice Hall, 2000.
[11] U.Y. Nahm and R.J. Mooney, "A Mutually Beneficial Integration of Data Mining and Information Extraction," Proc. 17th Nat'l Conf. Artificial Intelligence (AAAI '00), pp. 627-632, 2000.
[12] L. Talavera and J. Bejar, "Generality-Based Conceptual Clustering with Probabilistic Concepts," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 23, no. 2, pp. 196-206, Feb. 2001.
[13] H. Jin, 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.
[14] T. Hofmann, "The Cluster-Abstraction Model: Unsupervised Learning of Topic Hierarchies from Text Data," Proc. 16th Int'l Joint Conf. Artificial Intelligence (IJCAI '99), pp. 682-687, 1999.
[15] M. Steinbach, G. Karypis, and V. Kumar, "A Comparison of Document Clustering Techniques," Proc. Knowledge Discovery and Data Mining (KDD) Workshop Text Mining, Aug. 2000.