A SEMANTIC BASED APPROACH FOR TEXT CLUSTERING USING AN ADVANCED CONCEPT-BASED MINING MODEL

[1]Reshma R, [2]Vinai George Biju
[1,2]Department of Computer Science and Engineering, Christ University Faculty of Engineering Bangalore, Karnataka, India
Email:[1]reshma.r@mtech.christuniversity.in,[2]vinai.george@christuniversity.in

***Abstract*-The most common methods for text mining are based on the statistical analysis of the terms or word, Here the frequency of the terms are considered to find out the importance of a word in the document only. But the term which contribute to the sentence semantic is significant, which leads to the discovery of the topic. In this proposed model the natural language processing technique is efficiently used to capture the semantics of the text can be applied to enhance text clustering. A concept based mining model is introduced for this. The term which contributes to the sentence semantics is analyzed on the sentence, document and corpus levels rather than the traditional analysis of document only. According to the semantics of the sentence the proposed model can identify concept match between the documents. Experiments are conducted on different datasets using this proposed concept mining model for text clustering. The experiment results are the display of extracted concepts and proposed concept based mining model can be used for the enhancement of the clustering.**

Index terms-concept based mining, Text clustering, Concept based similarity.

I. INTRODUCTION

Natural Language Processing is concerned with the interactions between computer and human languages.NLP has significant overlap with the field of computational linguistics, [1] and is often considered as a subfield of Artificial Intelligence (AI).

Clustering is one of the traditional data mining techniques. Clustering is the process of grouping the documents where the topic of the one group will be different from the other group [2],.most current document clustering methods are based on Vector Space Model(VSM)[3].VSM is widely used for text clustering and text classification.

Another common methods used for text clustering include decision trees [4], conceptual clustering [5], statistical analysis [6] etc. Usually in text mining the importance of a term in a document is identified by computing the term frequency in documents [7]. Even though 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.

Semantics deals with the meaning of the sentence. Semantic analysis is basically capturing the meaning conveyed by the sentence. It plays an important role in natural language processing [8]. In this paper a semantic based approach for text clustering using an advanced concept mining model is proposed. The proposed model captures the semantic structure of each term within a document rather than the frequency of the term within a document only. In this model, three measures for analysing concepts on the sentence, document and corpus levels are possible.

Here the concept means the semantic role of each term in the sentence. The proposed model can detect the concept in the text document by the analysis of sentence. When a new document is introduced the system can identify the matching concepts by processing introduced document and previous document. Here an extraction of matching concepts of the given documents can be done for clustering.

A new concept based similarity measure of the concept analysis on sentence and document levels are proposed. Similarity measure can be based on the sentence based, document based and corpus based concept analysis when it is applied for clustering. The concept can be a word or phrase which is totally dependent on the semantic structure of that particular sentence.

In this paper the results are demonstrated as the extracted concept list and the results are evaluated.

II. PROPOSED WORK

The proposed concept based mining model consists of sentence based and document based concept analysis. Fig .1 shows the working model of the concept based mining model.

A raw text document is input to the proposed model. Text processing is the first step applied to the raw text document. Sentences in the document are arranged with well defined boundaries, which ease the process of separation of sentences for the analysis [10], [11]. Each sentence in the document may have one or more verb argument structures. The verb arguments are labelled and the sentences which have many labelled verb argument structures have many verbs associated with their arguments. This verb argument structures are captured and analysed in document and corpus level using concept based mining model.

In this model the arguments and verbs 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. Here the term plays important semantic role that contribute to the meaning of the sentence. In the concept based mining model the labelled term that either word or phrase considered as concept [9].

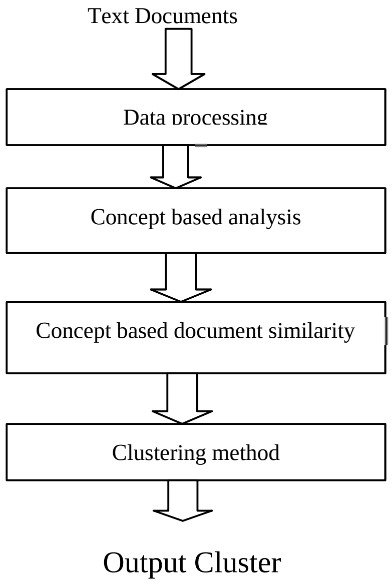


Fig .1.Concept based mining model system

A. *Sentence based concept analysis*

Each sentence in the document is analysed. A conceptual term frequency (ctf) is proposed in this model by analysing the concept of the sentences, which gives the conceptual term frequency measure.

Here in sentence based concept analysis ctf is local measure on the sentence level.

B. *Calculating ctf of sentence s*

Ctf of the sentence is calculated by counting the number of occurrence of concept c in the verb argument structures of sentence s. If the concept c appears frequently in the verb argument structure of same sentence, and then it has very much importance in the meaning of that particular sentence.

C. *Calculating ctf of document d*

The concept c might have number of *ctf* values in different sentences in the same document *d*. The *ctf* value of concept *c* in the document can be calculated using the equation below:

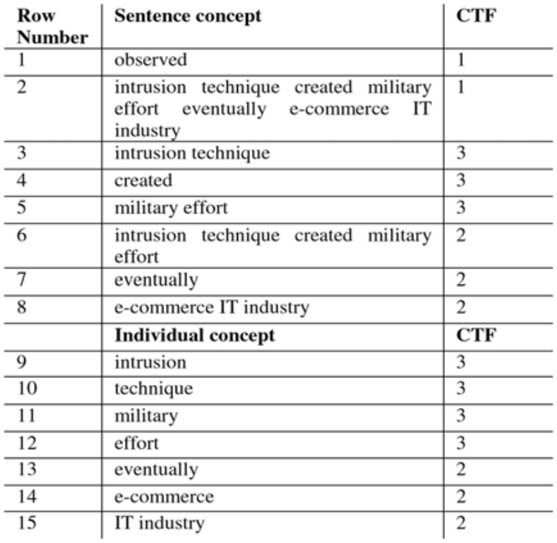
$$ctf = \frac{\sum_{n=1}^{sn} ctf_n}{sn},$$

Here *sn* is the total number of sentences that contain concept c in document d [9].

By calculation of the ctf values, this represents the overall importance of each concept to the semantics of a document through the sentences.

Table 1

Example of calculating ctf measure



Row Number	Sentence concept	CTF
1	observed	1
2	intrusion technique created military effort eventually e-commerce IT industry	1
3	intrusion technique	3
4	created	3
5	military effort	3
6	intrusion technique created military effort	2
7	eventually	2
8	e-commerce IT industry	2
	Individual concept	**CTF**
9	intrusion	3
10	technique	3
11	military	3
12	effort	3
13	eventually	2
14	e-commerce	2
15	IT industry	2

D. *Document based concept analysis*

Document based concept analysis is accomplished by calculating the number of occurrence of concept c in the original document.

E. *Corpus based concept analysis*

To identify the concepts of different documents, concept based document frequency is calculated. This concept based document frequency is global measure.

F. ***Illustration of computing the proposed conceptual term frequency (ctf).***

Sentence based concept analysis is explained using the example below:

"*We have* ***observed*** *how some intrusion techniques,* ***created*** *for the military effort, have eventually been* ***applied*** *for the ecommerce and IT industry*"

The words which marked bold are the verbs of the sentence are important for the semantic structure of the sentence. These words are **observed**, **created** and **applied**. Document cleaning is the first step performed. The stop words are removed and stemming is performed using stemming algorithm [12], [13]. These all operations are done in the sentence level.

Identify the verb argument structures in the sentence are the important step in the first stage. There by the semantic analysis [14], [15] of the each sentence will be possible. Then the term or word generated after all these steps are called concepts. In this example, the concepts generated after all steps are shown below without stemming for better understanding:

1. Concepts in the first verb argument structure of the verb **observed**:
   - "observed", "intrusion technique created military effort eventually e-commerce IT industry"
2. Concepts in second verb argument structure of the verb **created:**
   - **"**Intrusion technique", "created" and "military effort".
3. Concepts in the third verb argument of the word **applied:**
   - "Intrusion technique created military effort","eventually", and "e-commerce IT industry".

These concepts obtained from the same sentence. The concepts extracted are "observed", "intrusion technique created military effort eventually e-commerce IT industry", "intrusion technique", "created", "military effort", "intrusion technique created military effort", "eventually", "e-commerce IT industry".

In traditional analysis same weight is given for the words that present in the same sentence. In concept based analysis which analyse the semantics of the sentence.

Table1 shows the calculation of conceptual term frequency measure (ctf) of this particular sentence. Higher value of the conceptual term frequency measure is important which contributes more to the meaning of that sentence.

The table shows the ctf measures of the sentence. The rows 1 to 8 show the extracted sentence concepts. The rows 3 to 8 show the concepts which are overlapped with the other concepts. The individual concepts are in row 9 to 15.

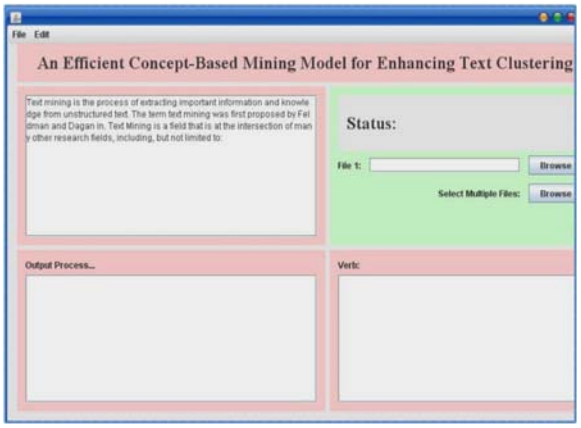
Here topic of the sentence will be the concept which has the highest ctf value. In this example "observed" has the very lowest ctf value, so it has very less importance in the meaning of the sentence." Intrusion technique”,”created”,”military” and effort have the importance in the meaning of the sentence.

G. Concept based similarity measure

To enhance clustering concept based similarity measure of the input documents can be calculated. The efficient algorithm developed from the proposed concept based mining model can efficiently be used for the clustering purpose. The concept based mining model identifies matching concepts in sentence and document level. Finally the clustering of documents is possible by extracting the concepts of different documents and the documents have the same concept can put together.

III. EXPERIMENT RESULTS

To demonstrate the improvement in clustering the different documents given as the input and which processed with proposed concept mining algorithm. The figure 2.1 shows the window where user can give the inputs. Using the browse button user can input the text document.



An Efficient Concept-Based Mining Model for Enhancing Text Clustering

Text mining is the process of extracting important information and knowledge from unstructured text. The term text mining was first proposed by Feldman and Dagan in. Text Mining is a field that is at the intersection of many other research fields, including, but not limited to:

Status:

File 1:

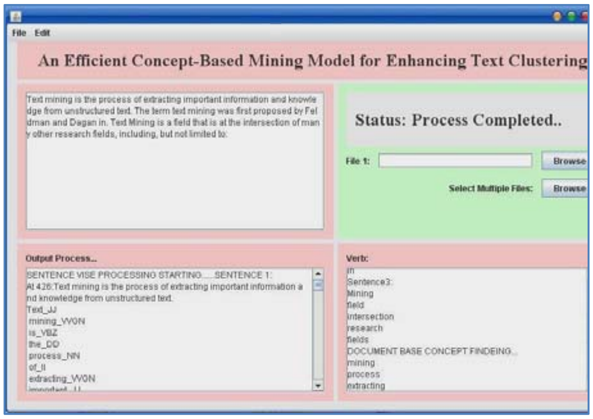
Select Multiple Files:

Output Process...

Verb:

Figure.2.1 inputting the text

The figure 2.2 shows the output results where the verb frame displays the concept of the sentences. Here the concept which has the highest value will be the considered as the concept of the document.



An Efficient Concept-Based Mining Model for Enhancing Text Clustering

Text mining is the process of extracting important information and knowledge from unstructured text. The term text mining was first proposed by Feldman and Dagan in. Text Mining is a field that is at the intersection of many other research fields, including, but not limited to:

Status: Process Completed..

File 1:

Select Multiple Files:

Output Process...

```
SENTENCE VISE PROCESSING STARTING.....SENTENCE 1:
At 426:Text mining is the process of extracting important information a
nd knowledge from unstructured text.
Text_JJ
 mining_VVGN
 is_VBZ
 the_DD
 process_NN
 of_II
 extracting_VVGN
```

Verb:

```
in
Sentence3:
Mining
field
intersection
research
fields
DOCUMENT BASE CONCEPT FINDEING...
mining
process
extracting
```

Figure 2.2 The Concept List

IV. CONCLUSIONS

The proposed concept mining model is an efficient clustering technique which makes use of the natural language processing. The semantic based approach of this model improved the clustering quality. The proposed model analyzed the documents in the sentence level, document level. The concept term frequency of the documents are calculated, by that the concept of document are identified.

The documents allowed to process with the algorithm and the concept lists are showed in the frame.

The proposed model can be efficiently used in different areas where clustering is important. Applying the proposed model to the web documents is one of the extensions for this work. This can be efficiently used for different web applications; search engines are one of the examples where we can use this clustering method. Another extension for this concept based mining model is which can be efficiently used for text classification.

REFERENCE

[1] A. k. S. R.Potet, Natural Language processing and text mining, springer, 2007.

[2] L. D. M. M. K. M. F. F. D. S. Nadia Nedjah, Intelligent Text Catogorization and Clustering, 2009.

[3] J. Kogan, Introduction to Clustering Large and High Dimrnsional Data, Cambridge University, 2006.

[4] K. Han Jiawei, Data Mining Concepts and Techniques, USA Morgan Kaufmann, 2001.

[5] M. W. Berry, Survey of text Mining Clustering,classification and retrieval, Springer, 2004.

[6] J. A. a. Balazsfeif, Cluster Analysis for Data Mining and System Identification, Birkha , 2007.

[7] G. R. Reddy, "A Frequent based text clustering approach using novel similarity measure," *Advance computing conference(IACC),* 2014.

[8] U. T. Tanveer Siqqiqui, Natural Laguage

Processing and Information Retrieval, Oxford University Press, 2008,Aprill.

[9] F. K. S. Shady Shehata, "An Efficient Concept Based Mining Model for Enhancing Text Clustering," *IEEE Transactions on knowledge and data engineering,* vol. 22, 2010.

[10] J. K. Michaelw Berry, Text Mining application and theory, Kli-Blackwell, 2010.

[11] J. L. H. Salvador Garcia, Data Preprocessing in Data Mining, Springer, 2014.

[12] M. A. G. Jivani, "A Comapritive Study of Stemming Algorithms," *International Journel of computer Technology and Applications(IJCTA),* 2011.

[13] G. G. David A.Hall, "A Detailed Analysis of Stemming Algorithms," 1996.

[14] R. R. Pillai, "Enhanced Semanitic Preserved concept based mining model for enhancing document clustering," *International Journel of intentive Engineering and Services,* 2014.

[15] K. H. w. H. J. Sameer Pradhan, "Semantic Role Parsing:Adding Semantic structure to unstructured text," *IEEE International Conference in Data Mining,* 2003.

[16] j.-Y. J.-J. L. Yung-Shen Lin, "Similarity

Measure of Text classification and clustering," *IEEE transactions on Knowledge aand Data Engineering,* 2014.