A Concept Based Mining Model For NLP Using Text Clustering

Ms Aruna Jadhav [#1]
M E Computer Engg.
MGM’s College of Engg. & Technology
Navi Mumbai
University of Mumbai, India

Dr. Subhash K. Shinde[#2]
Department of Computer Engg.
L T College of Engineering & Tech.
Koparkhairaine
University of Mumbai, India

Abstract :-
Most of 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. 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. A new concept-based mining model that analyses terms on the sentence, document, and corpus levels 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 mining model consists of sentence-based concept analysis, document-based concept analysis, corpus-based concept-analysis, and concept-based similarity measure. The term which contributes to the sentence semantics is analysed on the sentence, document, and corpus levels rather than the traditional analysis of the document only. The proposed model can efficiently find significant matching concepts between documents, according to the semantics of their sentences. The similarity between documents is calculated based on a new concept-based similarity measure.

***Keywords*—Concept-based mining model, sentence-based, document-based, corpus-based, concept analysis, Concept based similarity.**

1. Introduction

Natural Language Processing (NLP) is both a modern computational technology and a method of investigating and evaluating claims about human language itself. Text mining attempts to discover new, previously unknown information by applying techniques from natural language processing and data mining. Clustering, one of the traditional text data mining techniques, is unsupervised learning paradigm where clustering methods try to identify inherent groupings of the text documents so that a set of clusters are produced in which clusters exhibit high intra-cluster similarity and low inter-cluster similarity [1].

In this paper, a novel concept-based mining model is proposed. It captures the semantic structure of each term within a sentence and a document, rather than the frequency of the term within a document only. Each sentence is labelled by a semantic role labeller 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.

Most current document clustering methods are based on the Vector Space Model (VSM) [2] [3], 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.

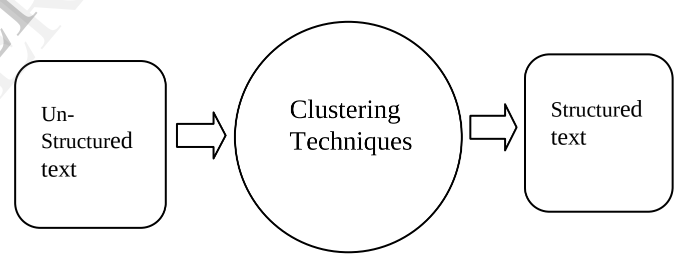


Figure 1 Common method of Text mining

Methods used for text clustering include decision trees [4], conceptual clustering [5], clustering based on data summarization [6], statistical analysis [7], neural nets [8], inductive logic programming [9], and rule-based systems [10] among others. In text clustering, it is important to note that selecting important features, which present the text data properly, has a critical effect on the output of the clustering algorithm [11]. Moreover, weighting these features accurately also affects the result of the clustering algorithm substantially [12]. The clustering results produced by the sentence-based, document-based, corpus-based, and the combined approach concept analysis have higher quality than those produced by a single-term analysis similarity only. The results are evaluated using two quality measures, the F-measure and the Entropy. Both of these quality measures showed improvement versus the use of the single-term method when the concept-based similarity measure is used to cluster sets of documents.

2. Proposed System

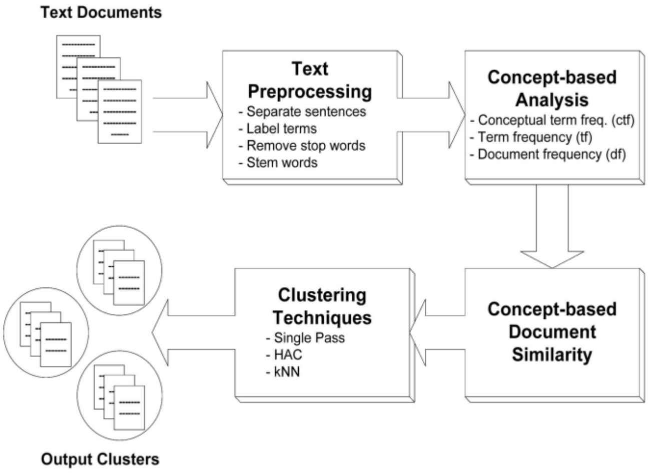


Figure 1 Concept-based mining model system.

The proposed concept-based mining model consists of concept-based term 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 labelled automatically based on the Prop Bank notations [13]. After running the semantic role labeller, each sentence in the document might have one or more labelled verb argument structures. The number of generated labelled verb argument structures is entirely dependent on the amount of information in the sentence. The sentence that has many labelled verb argument structures includes many verbs associated with their arguments. The labelled verb argument structures, the output of the role labelling task, are captured and analysed by the concept-based mining model. The objective behind the concept-based analysis task is to achieve an accurate analysis of concepts on the sentence, document, and corpus levels rather than a single-term analysis on the document only.

2.1 Sentence-Based Concept Analysis

To analyse each concept at the sentence level, a new concept-based frequency measure, called the conceptual term frequency ctf is proposed. The ctf calculations of concept c in sentence s and document d are as follows: The ctf is the number of occurrences of concept c in verb argument structures of sentence s. The concept c, which frequently appears in different verb argument structures of the same sentence s, has the principal role of contributing to the meaning of s. In this case, the ctf is a local measure on the sentence level.

2.2 Document-Based Concept Analysis

To analyse each concept at the document level, the concept based term frequency tf, the number of occurrences of a concept (word or phrase) c in the original document, is calculated. The tf is a local measure on the document level.

2.3 Corpus-Based Concept Analysis

To extract concepts that can discriminate between documents, the concept-based document frequency df, the number of documents containing concept c, is calculated. The df is a global measure on the corpus level. This measure is used to reward the concepts that only appear in a small number of documents as these concepts can discriminate their documents among others.

2.4 Concept-Based Term Analysis

The objective of this task is to achieve a concept-based term analysis (word or phrase) on the sentence and document levels rather than a single-term analysis in the document set only. To analyse each concept at the sentence-level, a concept based frequency measure, called the conceptual term frequency (ctf) is proposed. The ctf is the number of occurrences of concept c in verb argument structures of sentence*s*. The concept c, which frequently appears in different verb argument structures of the same sentence s, has the principal role of contributing to the meaning of *s*. To analyse each concept at the document-level, the term frequency tf , the number of occurrences of a concept (word or phrase) *c* in the original document, is calculated. The process of calculating tf and ctf measures in a set of documents is attained by the proposed algorithm which is called (Concept-based Term Analyser).

2.5 A Concept-Based Similarity Measure

Concepts convey local context information, which is essential in determining an accurate similarity between documents. A new concept-based similarity measure, based on matching concepts at the sentence and document levels rather than on individual terms (words) only, is devised. The concept-based similarity measure relies on two critical aspects. First, the analysed labelled terms are the concepts that capture the semantic structure of each sentence. Secondly, 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. These aspects are measured by the proposed concept-based similarity measure which measures the importance of each concept at the document-level by the tf measure and at the sentence-level by the ctf measure. The concept-based measure exploits the information extracted from the concept-based term analyser algorithm to better judge the similarity between the documents. This similarity measure is a function of the following factors: the number of matching concepts (m) in the verb arguments structures in each document (d), the total number of sentences (s) in each document d, the total number of the labelled verb argument structures (v) in each sentences, the tfi of each concept ci in each document *d* where (i = 1, 2, ...,m), the ctfi of each concept *ci* in *s* for each document d where (i = 1, 2, ...,m), the length (l) of each concept in the verb argument structure in each document d, and the length (s) of each verb argument structure which contains a matched concept. The conceptual term frequency (ctf) is an important factor in calculating the concept-based similarity measure between documents. The more frequent the concept appears in the verb argument structures of a sentence in a document, the more conceptually similar the documents are. This similarity measure is a function of the following factors :-

1. the number of matching concepts, m, in the verb argument structures in each document d,
2. the total number of sentences, sn, that contain matching concept ci in each document d,
3. the total number of the labelled verb argument structures, v, in each sentence s,
4. the ctfi of each concept ci in s for each document d, where i = 1; 2; . . .;m,
5. the tfi of each concept ci in each document d, where i = 1; 2; . . .;m.
6. the dfi of each concept ci, where i = 1; 2; . . .;m,
7. the length, l, of each concept in the verb argument structure in each document d,
8. the length, Lv, of each verb argument structure which contains a matched concept, and
9. the total number of documents, N, in the corpus.

3.1 Example of Calculating the Proposed Conceptual Term Frequency (ctf) Measure.

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 sentence, the semantic role labeller identifies three target words (verbs), marked by bold, which are the verbs that represent the semantic structure of the meaning of the sentence. These verbs are created, made, and lead. Each one of these verbs has its own arguments as follows:

. [ARG0 Texas and Australia researchers] have [TARGET created] [ARG1 industry-ready sheets of materials made from nanotubes that could lead to the development of artificial muscles].

. Texas and Australia researchers have created industry-ready sheets of [ARG1 materials] [TARGET made] [ARG2 from nanotubes that could lead to the development of artificial muscles].

. Texas and Australia researchers have created industry-ready sheets of materials made from [ARG1 nanotubes] [R-ARG1 that] [ARGM-MOD could] [TARGET lead] [ARG2 to the development of artificial muscles].

Arguments labels1 are numbered ARG0, ARG1, ARG2, and so on depending on the valency of the verb in sentence. The meaning of each argument label is defined relative to each verb in a lexicon of Frames Files [13]. Despite this generality, ARG0 is very consistently assigned an Agent-type meaning, while ARG1 has a Patient or Theme meaning almost as consistently [13]. Thus, this sentence consists of the following three verb argument structures :

1. First verb argument structure for the verb created:
. [ARG0 Texas and Australia researchers]
. [TARGET created]
. [ARG1 industry-ready sheets of materials made from
nanotubes that could lead to the development of artificial muscles].

2. Second verb argument structure for the verb made:
. [ARG1 materials]
. [TARGET made]
. [ARG2 from nanotubes could lead to the development of artificial muscles].

3. Third verb argument structure for the verb lead:
. [ARG1 nanotubes]
. [R-ARG1 that]
. [ARGM-MOD could]
. [TARGET lead]
. [ARG2 to the development of artificial muscles].

A cleaning step is performed to remove stop words that have no significance, and to stem the words using the popular Porter Stemmer algorithm [14]. The terms generated after this step are called concepts. 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 of the verb created:
. Texas Australia researchers
. created
. industry-ready sheets materials nanotubes lead development artificial muscles

2. Concepts in the second verb argument structure of the verb made:
. materials
. nanotubes lead development artificial muscles

3. Concepts in the third verb argument structure of the verb lead:
. nanotubes
. lead
. development artificial muscles.

Row Number	Sentence Concepts	CTF
(1)	texas australia researchers	1
(2)	created	1
(3)	industry ready sheets materials nanotubes lead development artificial muscles	1
(4)	materials	2
(5)	nanotubes lead development artificial muscles	2
(6)	nanotubes	3
(7)	lead	3
(8)	development artificial muscles	3
	Individual Concepts	CTF
(9)	texas	1
(10)	australia	1
(11)	researchers	1
(12)	industry	1
(13)	ready	1
(14)	sheets	1
(15)	development	3
(16)	artificial	3
(17)	muscles	3

TABLE1 Ex. of Calculating the Proposed ctf Measure

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,
. materials,
. nanotubes lead development artificial muscles,
. nanotubes,
. lead, and
. development artificial muscles.

The traditional analysis methods assign the same weight for the words that appear in the same sentence. However, the concept-based mining model discriminates among terms that represent the sentence concepts using the proposed ctf measure. This analysis is entirely based on the semantic analysis of the sentence. In this example, some concepts have higher conceptual term frequency ctf than others, as shown in Table 1. In such cases, these concepts (with high ctf) contribute to the meaning of the sentence more than other concepts (with low ctf).

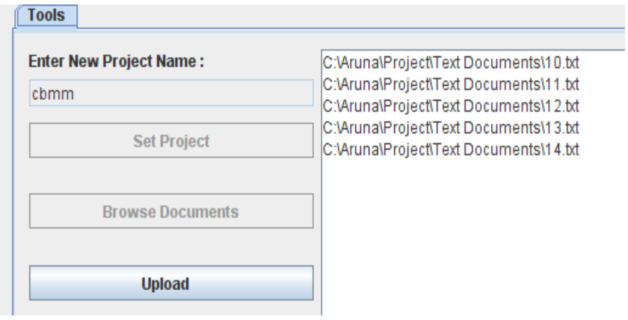
As shown in Table 1, the concept-based analysis computes the ctf measure for:-

1. The concepts which are extracted from the verb argument structures of the sentence, which are in Table 1 from row (1) to row (8).
2. The concepts which are overlapped with other concepts in the sentence. These concepts are in Table 1 from row (4) to row (8).
3. The individual concepts in the sentence, which are in Table 1 from row (9) to row (17).

In this example, the topic of the sentence is about materials made from nanotubes which could lead to the development of artificial muscles. The nanotubes, lead, and development artificial muscles concepts, which present this meaning, have the highest ctf value with 3. In addition, the concept Texas Australia researchers, which has the lowest ctf, has no major significant effect on the main topic of the sentence. Thus, the concepts with high ctf such as nanotubes, lead, and development artificial muscles present indeed the topic of the sentence.

4 Experimental Results

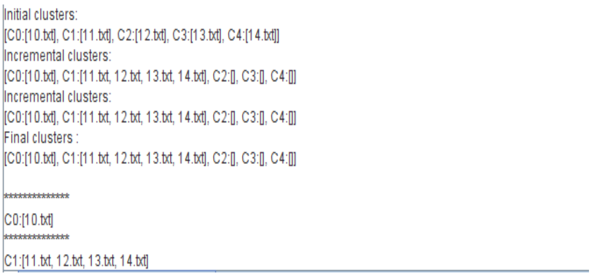
To test the effectiveness of concept matching in determining an accurate measure of the similarity between documents, extensive sets of experiments using the concept-based term analysis and similarity measure are conducted. A group of unstructured text documents are selected. Concepts are retrieved from the documents . The cosine similarity for the text documents is calculated and depending on the value of the similarity clusters are formed for the given input.



Enter New Project Name : cbmm

C:\Aruna\Project\Text Documents\10.txt
C:\Aruna\Project\Text Documents\11.txt
C:\Aruna\Project\Text Documents\12.txt
C:\Aruna\Project\Text Documents\13.txt
C:\Aruna\Project\Text Documents\14.txt

Fig 4.1 Browse Unstructured Text Documents



```
Initial clusters:
[C0:[10.txt], C1:[11.txt], C2:[12.txt], C3:[13.txt], C4:[14.txt]]
Incremental clusters:
[C0:[10.txt], C1:[11.txt, 12.txt, 13.txt, 14.txt], C2:[], C3:[], C4:[]]
Incremental clusters:
[C0:[10.txt], C1:[11.txt, 12.txt, 13.txt, 14.txt], C2:[], C3:[], C4:[]]
Final clusters :
[C0:[10.txt], C1:[11.txt, 12.txt, 13.txt, 14.txt], C2:[], C3:[], C4:[]]

**************
C0:[10.txt]
**************
C1:[11.txt, 12.txt, 13.txt, 14.txt]
```

Fig 4.2 Clusters of input documents

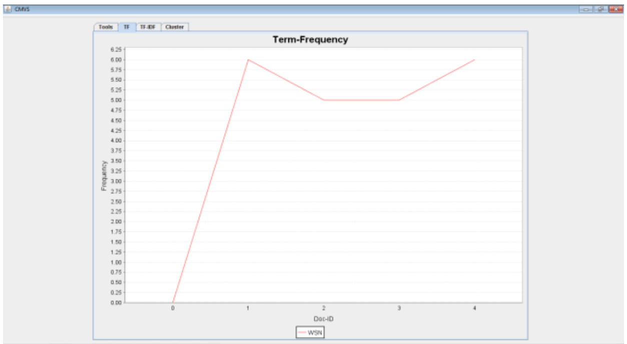


Fig 4.3 Term Frequency

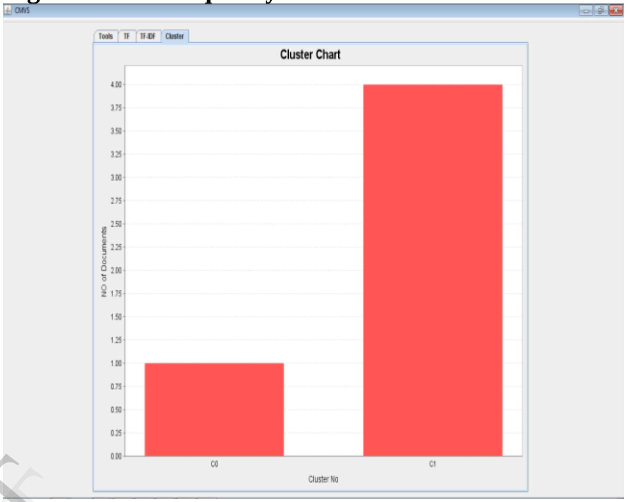


Fig 4.4 Bar Chart - Clusters

The figure 4.3 shows the graph of occurrence of concepts in the entire set of documents. The results also gives the information that the term contributing to the meaning of the document is WSN. The document describes the information of WSN. The figure 4.2 and 4.4 shows the final output clusters with each cluster number grouping the total number of documents.

5 Conclusion

This work bridges the gap between natural language processing and text mining disciplines. A new concept based mining model composed of four components, 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. The first component is the sentence-based concept analysis which analyses 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, analyses each concept at the document level using the concept-based term frequency tf. The third component analyses concepts on the corpus level using the document frequency df global measure. The fourth component is the concept-based s

imilarity 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. 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.

References

[1] K.J. Cios, W. Pedrycz, and R.W. Swiniarski, Data Mining Methods for Knowledge Discovery. Kluwer Academic Publishers, 1998.

[2] 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.

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] T. Honkela, S. Kaski, K. Lagus, and T. Kohonen, “WEBSOM—Self-Organizing Maps of Document Collections,” Proc. Workshop Self-Organizing Maps (WSOM ’97), 1997.

[9] M. Junker, M. Sintek, and M. Rinck, “Learning for Text Categorization and Information Extraction with ILP,” Proc. First Workshop Learning Language in Logic, 1999.

[10] S. Soderland, “Learning Information Extraction Rules for Semi-Structured and Free Text,” Machine Learning, vol. 34, nos. 1-3, pp. 233-272, Feb. 1999.

[11] 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.

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

[13] P. Kingsbury and M. Palmer, “Propbank: The Next Level of Treebank,” Proc. Workshop Treebanks and Lexical Theories, 2003.

[14] M.F. Porter, “An Algorithm for Suffix Stripping,” Program, vol. 14, no. 3, pp. 130-137, July 1980.