An Advanced Concept-Based Mining Model to Enrich Text Clustering

M. Yasodha[#1], Dr .P. Ponmuthuramalingam#[2]

[1]DR NGP Arts and Science College, Coimbatore , India.

[2]Government Arts Collge, Coimbatore, India.

Abstract

Abstract— 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 analyzes 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 similarity between documents is calculated based on a new concept-based similarity measure. The proposed similarity measure takes full advantage of using the concept analysis measures on the sentence, document, and corpus levels in calculating the similarity between documents. The experiments demonstrate extensive comparison between the concept-based analysis and the traditional analysis. Experimental results demonstrate the substantial enhancement of the clustering quality using the sentence-based, document-based, corpus-based, and combined approach concept analysis.

Keywords *- Concept-based mining model, sentence-based, document-based, corpus-based, concept analysis, conceptual term frequency, concept-based similarity.*

1. Introduction

Natural language processing (NLP) is a field of computer science, artificial intelligence (also called machine learning) and linguistics concerned with the interactions between computers and human (natural) languages. Specifically, the process of a computer extracting meaningful information from natural language input and/or producing natural language output.

The phrase "text mining" is generally used to denote any system that analyses large quantities of natural language text and detects lexical or linguistic usage patterns in an attempt to extract probably useful (although only probably correct) information.

Clustering can be considered the most important *unsupervised learning* problem; so, as every other problem of this kind, it deals with finding a *structure* in a collection of unlabeled data.

A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way". A *cluster* is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters.

2. Vector Space Model

Vector Space Model (or *term vector model*) is an algebraic model for representing text documents (and any objects, in general) as vectors of identifiers, such as, for example, index terms. It is used in information filtering, information retrieval, indexing and relevancy rankings.

The vector space model procedure can be divided in to three stages. The first stage is the document indexing where content bearing terms are extracted from the document text. The second stage is the weighting of the indexed terms to enhance retrieval of document relevant to the user. The last stage ranks the document with respect to the query according to a similarity measure.

The main idea of text clustering is to find which documents have many words in common, and place the documents with the most words in common into the same groups. We can illustrate how this works by using a small collection of imaginary documents. 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 other terms.

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. 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 has been s proposed in this paper. The proposed model captures the semantic structure of each term within a sentence. In the proposed model, three measures for analyzing concepts on the sentence

1. Sentence-based Concept Analysis
2. Document-based Concept Analysis
3. Corpus-based Concept Analysis

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.

2.1 Text Clustering Using Concept-based Mining Model

Following are the explanations of the important terms used in this paper:
Verb argument structure:
e.g Ramu eats banana hits: the verb
Ramu & banana: arguments of the verb "eats"
Label: assigned to an argument
Ramu: subject
banana: object
Term: either a verb or a argument, either a word or phrase
Concept: a labeled term.

2.2 Thematic Role Background

The semantic structure of a sentence can be characterized by a form of verb argument structure.

My brother wants a bicycle – NP(Noun Phrase) wants NP
In this case, some facts could be driven for the particular verb "wants":
There are two arguments to this verb.
Both arguments are NPs.
The first argument "my brother" is preverbal and plays the role of the subject.
The second argument "a bicycle" is a post verbal and plays the role of the direct object.
The study of the roles associated with verbs is referred to as a thematic role or case role analysis. Thematic roles, first proposed by Fillmore Fillmore first suggested that thematic roles are categories which help characterize the verb arguments by providing a shallow semantic language.
Recently, there have been many attempts to label thematic roles in a sentence automatically. Gildea and Jurafsky were the first to apply a statistical learning technique to the FrameNet database. In recent times, thematic roles in sentences have been tried to be labeled automatically. The first was proposed by Gildea and Jurafsky . They applied a statistical learning technique to the FrameNet Database.

2.3 Concept-Based Analysis Algorithm

The proposed concept-based mining algorithm is composed of the sentence-based concept analysis, document-based concept analysis, the corpus-based concept analysis and the concept-based similarity measure.
The concept based mining model is depicted in the following figure

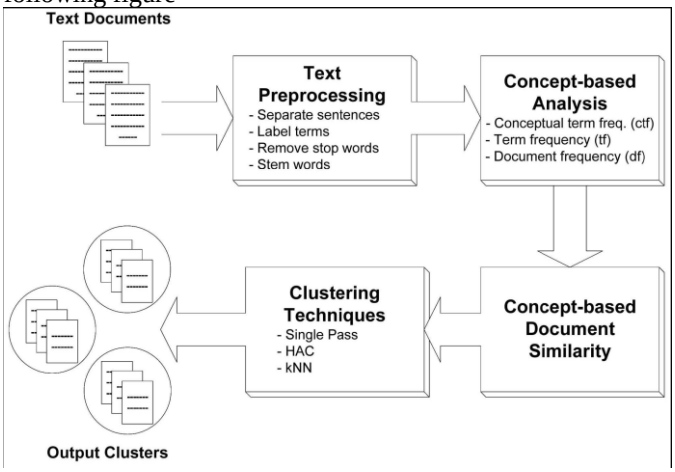

Figure1 .The concept based mining model

The sentences in the document may have one or more labeled verb argument structures. The amount of information in the sentence influences the number of verb argument structures generated by the labeler. The output of the labeler and the labeled verb argument structures are captured and analyzed by the concept based analysis model on the sentence, document and corpus levels.

The word term is used to indicate both the verb and the argument. A term can have more than one semantic role in the same sentence. This implies that a single term can be an argument to more than one verb in the same sentence. Such terms are said to play more important roles that contribute to the meaning of the sentence. In concept-based mining model, the term concept is used to describe a labeled term. A term can be either a phrase or a word.

Sentence-Based Concept Analysis

In the sentence-based concept analysis, each concept at the sentence level is analyzed and a new measure called the conceptual term frequency (ctf) is used. The ctf of a concept c in a sentence s and document d is calculated as follows:

1) Calculating ctf of concept c in Sentence s

The number of occurrences on a concept c in the verb argument structures of the sentence s is called the ctf. Here, the ctf is a local measure on the sentence level. The concept c that appears in different verb argument structures of the same sentence has s, has the principal role of contributing to the meaning of s.

2) Calculating ctf of concept c in document d

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

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

where the total number of sentences in a document d that contain the concept c is denoted by sn. The importance of a concept c to a document d is measured by taking the average of the ctf values of the concept. If a concept is found
to have ctf values in many sentences of a document, it is said to have a major contribution to the meaning of the sentences. Hence it leads to the discovery of the topic of the document.

Consider a concept c that appears twice in the document d in the sentences s1 and s2. The concept c occurs three times in the verb argument structure of the sentence s1 and five times in the verb argument structure of s2. The ctf value of the concept c is calculated as:

$$\frac{3+5}{2} = 4$$

Document-Based Concept Analysis

In order to analyze each concept at the document level, the concept based term frequency *tf* is calculated. The *tf* is calculated as the number of occurrences of a concept (word or phrase) *c* in the original document *d*. The *tf* is a local measure on the document level.

Corpus-Based Concept Analysis

The concept-based document frequency *df*, the number of documents containing concept *c*, is calculated to find the concepts that can differentiate between documents. 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.

The process of calculating *ctf, tf* , and *df* measures in a corpus is attained by the proposed algorithm which is called Concept-based Analysis Algorithm.

Concept-Based Analysis Algorithm

The concept-based analysis algorithm consists of the following steps:

1. *ddoci* is a new Document
2. L is an empty List
3. *sdoci* is a new sentence in *ddoci*
4. Build concepts list *Cdoci* from *sdoci*
5. for each concept *ci* Є *Ci* do
6. compute *ctfi* of *ci* in *ddoci*
7. compute *tfi* of *ci* in *ddoci*
8. compute *dfi* of *ci* in *ddoci*
9. *dk* is seen document, where *k*={0,1,..., doc-1}
10. *sk* is a sentence in *dk*
11. Build concepts list *Ck* from *sk*
12. for each concept *cj* Є *Ck* do
13. if (*ci* == *cj*) then
14. update *dfi* of *ci*
15. compute c*tf*weight = avg(c*tf*i, c*tf*j)
16. add new concept matches to L
17. end if
18. end for
19. end for
20. output the matched concepts list L

The proposed concept-based analysis algorithm describes the process of calculating the *ctf, tf* , and *df* of the matched concepts in the documents. The procedure begins with processing a new raw document (at line 1) which has well defined sentence boundaries.

The lengths of the matched concepts and their verb argument structures are stored for the concept-based similarity calculations.

Each concept (in the for loop, at line 5) in the verb argument structures, representing the semantic structures of the sentence, is processed in a sequential manner. Each concept in the current document is matched with the other

concepts in the previously processed documents. A concept list L is maintained to find the match to previous documents. The concept list L holds the entry for each of the previous documents that shares a concept with the current document.
After processing, L contains all the matching concepts between the current document and any previous document that shares at least one concept with the new document. Finally, L is output as the list of documents with the matching concepts and the necessary information about them. The concept-based analysis algorithm is capable of matching each concept in a new document *d* with all the previously processed documents in *O(m)* time, where *m* is the number of concepts in *d*.

3. Tables, Examples and Equations

Example of Calculating the Proposed Conceptual Term Frequency (ctf) Measure
Consider the following sentence:

Ramu **watched** movie yesterday ,his friend lakshman **played** cricket , tomorrow they may **changed** their idea of entertainment.
In this sentence, the semantic role labeler 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 watched , played and changed. Each one of
these verbs has its own arguments as follows:

[ARG0 Ramu] have [TARGET watched] [ARG1 movie yesterday ,his friend lakshman played cricket , tomorrow they may changed their idea of entertainment.].
Ramu **watched** movie yesterday , his friend [ARG1 lakshman] [TARGET played] [ARG2 cricket , tomorrow they may **changed** their idea of entertainment].
Ramu **watched** movie yesterday ,his friend lakshman **played** cricket , tomorrow [ARG1 they] [ARGM-MOD may] [TARGET changed] [ARG2 their idea of entertainment].
Arguments labels1 are numbered ARG0, ARG1, ARG2,and so on depending on the valiancy of the verb in sentence. The meaning of each argument label is defined relative to each verb in a lexicon of Frames Files.
Despite this generality,ARG0 is very consistently assigned an Agent-type meaning, while ARG1 has a Patient or Theme meaning almost as consistently [23]. Thus, this sentence consists of the following three verb argument structures:
1. First verb argument structure for the verb watched:
[ARG0 Ramu]
[TARGET watched]
[ARG1 movie yesterday ,his friend lakshman played cricket , tomorrow they may changed their idea of entertainment.].

2. Second verb argument structure for the verb played:
[ARG1 lakshman]
[TARGET played]
[ARG2 cricket , tomorrow they may **changed** their idea of entertainment]
3. Third verb argument structure for the verb changed:
[ARG1 they]
[ARGM-MOD may]
[TARGET changed]
[ARG2 their idea of entertainment].
A cleaning step is performed to remove stop words that have no significance, and to stem the words using the popular Porter Stemmer algorithm [24]. 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 watched:
. Ramu
. watched.
.movie lakshman played cricket , tomorrow they changed their idea of entertainment.].
2. Concepts in the second verb argument structure of the verb played:
. lakshman
. played
.cricket, tomorrow they changed idea of entertainment.
3. Concepts in the third verb argument structure of the verb changed.
. they
. changed
. idea of entertainment.

TABLE 1
Example of Calculating the Proposed ctf Measure

Row number	Sentence Concepts	ctf
1	Ramu	1
2	watched	1
3	movie lakshman played cricket , tomorrow they changed their idea of entertainment.	1
4	lakshman	2
5	played	2
6	cricket, tomorrow they changed idea of entertainment.	2
7	They	3
8	Changed	3

9	Idea of entertainment	3
Individual Concept		
10	Ramu	1
11	Movie	1
12	Lakshman	1
13	Cricket	1
14	They	1
15	Their	1
16	Idea	1
17	Entertainment	1

It is imperative to note that these concepts are extracted from the same sentence. Thus, the concepts mentioned in this example sentence are:
. Ramu
. watched.
.movie lakshman played cricket , tomorrow they changed their idea of entertainment.].
. lakshman
. played
. cricket, tomorrow they changed idea of entertainment.
. they
. changed
. idea of entertainment.

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 (9).
2. The concepts which are overlapped with other concepts in the sentence. These concepts are in Table 1 from row (4) to row (9).
3. The individual concepts in the sentence, which are in Table 1 from row (10) to row (17).

3.1 The mathematical framework

The mathematical framework of the concept based mining model is explained below:
· A concept c is a string of words, *c=" wi1,wi2,. . .,win "* where *n* is the total number of words in concept *c*.
· A sentence s is a string of concepts, *s="ci1,ci2,. . .,cim "* where *m* is the total number of concepts generated from the verb argument structures in sentence *s*.
· A document *d* is a string of words, d=" *wi1,wi2,. . .,wit*" where *t* is the total number of words in document *d*.
· The function *freq(strsub, strtotal)* is the number of times that substring strsub appears in string strtotal.
· The concept-based term frequency of document *d* is *tf=freq(ci,d).*
· The conceptual term frequency of sentence *S* is *ctfs=freq(ci,s)*
· The conceptual term frequency *ctf* of document d is calculated by (1).
The concept-based weighting of a concept is as in (3).
· The concept-based similarity between documents *d1* and *d2* using concepts is

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

4. Conclusions

The system consists of a new concept-based mining model which is composed of four components, is used to improve the text clustering quality. A better quality in clustering is achieved by exploiting the semantic structure of the sentences in documents. The first component analyzes the semantic structure of each sentence to capture the sentence concepts using the proposed conceptual term frequency *ctf* measure (sentence-based concept analysis). Then, the second component, analyzes each concept at the document level using the concept based term frequency *tf* (document-based concept analysis). The third component analyzes concepts on the corpus level using the document frequency *df* global measure (corpus-based concept analysis). 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. A concept-based similarity measure that is capable of the accurate calculation of pair wise documents is devised by combining the factors affecting the weights of concepts on the sentence, document, and corpus levels. 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 is considerably better than the traditional single term-based approaches. Further research can be done in this paper to use the same model for text classification.

References

[1] A. Strehl, J. Ghosh, and R. Mooney, "Impact of Similarity Measures on Web-Page Clustering," Proc. 17th Nat'l Conf. Artificial Intelligence: Workshop Artificial Intelligence for Web Search (AAAI), pp. 58-64, 2000.
[2] C. Fillmore, "The Case for Case," Universals in Linguistic Theory, Holt, Rinehart and Winston, 1968.
[3] D. Gildea and D. Jurafsky, "Automatic Labeling of Semantic Roles," Computational Linguistics, vol. 28, no. 3, pp. 245-288, 2002.
[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.
[5] G. Salton and M.J. McGill, Introduction to Modern Information Retrieval. McGraw-Hill, 1983.
[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] Ian H. Written, "Text Mining," Computer Science, University of Waikato, Hamilton, New Zealand.
[8] K.J. Cios, W. Pedrycz, and R.W. Swiniarski, Data Mining Methods for Knowledge Discovery. Kluwer Academic Publishers, 1998.
[9] 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.
[10] P. Kingsbury and M. Palmer, "Propbank: The Next Level of Treebank," Proc. Workshop Treebanks and Lexical Theories, 2003.
[11] R. Feldman and I. Dagan, "Knowledge Discovery in Textual Databases (KDT)," Proc. First Int'l Conf. Knowledge Discovery and Data Mining, pp. 112-117, 1995.
[12] S. Soderland, "Learning Information Extraction Rules for Semi- Structured and Free Text," Machine Learning, vol. 34, nos. 1-3, pp. 233-272, Feb. 1999.
[13] 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.
[14] T. Honkela, S. Kaski, K. Lagus, and T. Kohonen, "WEBSOM—Self- Organizing Maps of Document Collections," Proc. Workshop Self-Organizing Maps (WSOM '97), 1997.
[15] 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.

AUTHOR BIOGRAPHIES

M. Yasodha is working as an Assistant Professor in the Department of Computer Science, Dr. N.G.P. Arts and Science College, Coimbatore and doing Ph.D., in Bharathiar University, Coimbatore. She has done her M.Phil., in the area of Data Mining in Bharathiar University, Coimbatore. She has done her post graduate degree MCA in Bharathiar University, Coimbatore. She has presented and published a number of papers in reputed journals. She has four years of teaching and research experience and her research interests include Data Mining, Web mining, Semantic Web mining and Text mining.

DR. P. **Ponmuthuramalingam** received his Masters Degree in Computer Science from Alagappa University, Karaikudi in 1988 and the Ph.D. in Computer Science from Bharathiar University, Coimbatore. He is working as Associate Professor and Head in Department of Computer Science, Government Arts College(Autonomous), Coimbatore. His research interest includes Text mining, Semantic Web, Network Security and Parallel Algorithms.