International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 01,January-2015

Enhancing Quality of Search Results by Concept Based Model

Ms. W. Rubavathy
UG Scholar, Department of Information Technology
Jeppiaar Engineering College,Anna University
Chennai (Dt), Tamil Nadu, India.

Ms. P. Sharmila
UG Scholar, Department of Information Technology
Jeppiaar Engineering College,Anna University
Chennai (Dt), Tamil Nadu, India.

Mrs. R. Thilagavathy
Assistant Professor, Department of Information Technology
Jeppiaar Engineering College,Anna University
Chennai (Dt), Tamil Nadu, India.

Abstract: **Text Mining techniques are mostly based on Vector space model(term frequencies). The statistical analysis of a term frequency captures the importance of the term without a document only. But two terms can have the same frequency in the same document. But the meaning that one term contributes might be more appropriate than the meaning contributed by the other term. Hence, the terms that capture the semantics of the text should be given more importance. Here, a new concept-based mining is introduced. It analyses the terms based on the sentence, document and corpus level. 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. The sentence-based concept analysis calculates the conceptual term frequency(ctf), document-based concept analysis finds the term frequency (tf),corpus-based concept analysis determines the document frequency (df) and concept-based similarity measure. The process of calculating ctf, tf, df, measures in a corpus is attained by the proposed algorithm which is called Concept-Based mining model. By doing so we cluster the web documents in an efficient way and the quality of the clusters achieved by this model significantly surpasses the traditional single-term-base approaches.**

Index Terms—Concept-based mining model, sentence-based, document-based, corpus-based, concept analysis, conceptual term frequency, concept-based similarity.

I.INTRODUCTION

Text mining attempts to discover new, previously unidentified information by applying techniques from natural language processing and data mining. Clustering is one of the traditional data mining techniques and it is an unproven learning model where clustering methods identifies intrinsic groupings of the text documents. Set of clusters is formed in text documents clusters show high intracluster resemblance and low intercluster similarity in text documents. The Recent techniques used in text mining are concept-based which engages natural language processing and statistical analysis. Grouping of existing documents and upcoming documents can be made by the mining functionalities Such as clustering and classification[1][2]. It is vital and efficient if the classification functionality is entrenched into the existing clustering functionality. natural language processing avoids the uncertainty of different senses of a single word and multiple representation for a single sense that may varies on the author's vocabulary. Synonym based mining model inherits all the benefits of existing concept based mining model. In addition to that it flavors the fundamental nature of synonym based matching. Text mining is defined as the process of extracting Knowledge from textual databases or documents. This text mining may appears to be different from other databases due to its unstructured form and dimensions. Each word in the document is a dimension. So the primary things for text mining are, giving a structure to the data and reducing the dimensions .Giving structure to the data comes under natural language processing. Verb argument structure is one of the approaches for giving structure to a sentence. In this approach each word is given with a label (e.g. arg0, arg1.etc) There are different notations for these labels. This labeling can be done by semantic role parsing.The label tells the semantic role of the word in that particular sentence. Most of the text mining methodologies are based on vector space model [3]. Recent document clustering methods are based on the Vector Space Model (VSM) [4], [5], which is a extensively used data depiction for text categorization 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 calculated by one of numerous similarity actions that are based on such a feature vector. Normally, in text mining techniques, the term frequency of a term (word or phrase) is calculated to find out the value 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 significant to remind that extracting the relations between verbs and their arguments in the same sentence has the possible for analyzing terms within a sentence. The information concerning who is doing what to whom clarifies the donation of each term in a sentence to the meaning of the chief theme of that sentence.In this paper, a novel concept-based mining model is planned. The future model captures the semantic collection of each term inside a sentence and document somewhat than the

frequency of the term inside 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 add to the sentence semantics linked 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 absolutely reliant on the semantic structure of the sentence.When a new document is introduced to the system, the proposed mining model can identify a concept equal from this document to all the before processed documents in the data set by scanning the new document and extracting the similar concepts.

A new concept-based similarity measure which makes utilize of the concept analysis on the sentence, document, and corpus levels is proposed.This similarity measure outperforms further similarity measures that are based on term analysis models of the document only. The similarity between documents is based on a mixture of sentence-based, document-based, and corpus-based concept analysis. Similarity based on similar concepts between document pairs, is revealed to have a more important outcome on the clustering value due to the similarity's inattentiveness to loud terms that can direct to an incorrect similarity. The concepts are less responsive to sound when it comes to manipulating document similarity. This is due to the reality that these concepts are firstly extracted by the semantic role labeler and analyzed with reverence to the sentence, document, and corpus levels. Thus, the matching among these concepts is few probable to be found in nonrelated documents. The clustering results produced by the sentence-based, document-based, corpus-based, and the collective 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.

Following are the explanations of the important terms used in this paper:

- *Verb argument structure*: (e.g., Ravi hits the ball)."hits" is the verb. "Ravi" and "the ball" are the arguments of the verb "hits,"
- *Label*: A label is assigned to an argument, e.g.: "Ravi"has subject (or Agent) label. "the ball" has object (or theme) label,.
- *Term*: is either an argument or a verb. Term is also either a word or a phrase (which is a sequence of words).
- *Concept*: in the new proposed mining model, concept is a labeled term.

II. CASE ROLE ANALYSIS

In general, the semantic structure of a sentence can be characterized by a form of verb argument structure. This underlying structure allows the formation of a compound meaning representation from the meanings of the individual concepts in a sentence. The verb argument structure allows a link between the arguments in the surface structures of the input text and their related semantic roles. Consider the following example: My son wants a bike.This example has the following syntactic argument frames:(Noun Phrase (NP) wants NP). In this case, some facts could be driven for the particular verb "wants":

1. There are two arguments to this verb.

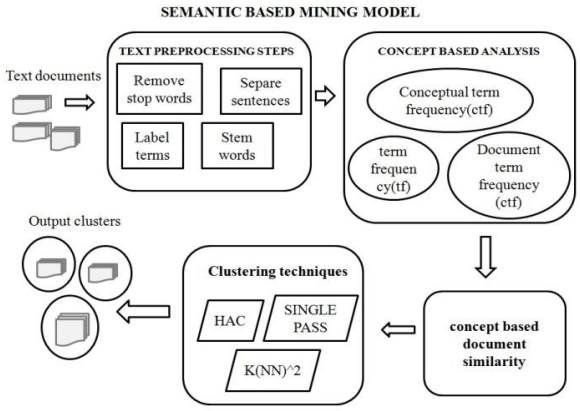
2. Both arguments are NPs.

3. The first argument "my son" is preverbal and plays the role of the subject.

4. The second argument "a bike" is a postverbal and plays the role of the direct object.

III. SEMANTIC BASED MINING MODEL

The proposed concept-based mining model contains sentence-based concept analysis, document-based concept analysis, corpus-based concept-analysis, and concept-based similarity measure, as shown in Fig. 1.



A unrefined text document is given as a input to this model each document has definite borders after successive operation of semantic role labeler, each sentences in document has one or more labeled verb argument structure this labeled verb argument structure is fully depends up on the amount of information present in each sentences of whole document. Many verbs has assigned arguments in labeled verb argument structure The verb argument structures, and the output of the labeling task, are captured and analyzed by the concept-based mining model at sentence level, document level and corpus level.In this model, the verb and the argument are measured as terms. One term can act as a argument to more than one verb in the same sentence. This indicates that this term may have

more than one semantic role in the same sentence. In those situations, this term plays vital semantic roles that contribute to the meaning of the sentence. Labeled term may be a word or phrase is represented as concepts in concept based mining model. The main responsibility here is to determine the exact investigation to concepts at three different steps such as sentence level, document level and corpus levels instead of single term evaluation in the document alone .

A. Analysis of Sentence-Based Concept:

a.*Calculation of ctf in Sentence s level:*

The ctf is the no of times concept(c) occurring in verb argument structures which is found in each sentence. The concept c, which is often appearing in different verb argument structures of the same sentence s, has the major responsibility of contributing to the meaning of s. regarding to this, the ctf is a local computation on the sentence level.

b.*Calculation of ctf in Document level:*

A concept c may have numerous ctf values in different sentences in the same document d. so, the ctf value of each concept c in whole document d is determined by the formula:

$$\text{Ctf}=\frac{\sum_{v=1}^{sv} ctfv}{sv}$$

where sv reffered as the total number of sentences which contain concept c in whole document d. calculating the average ctf value of various concepts in whole document d will measures the overall significance of concept c to the sentence meaning in document d. A concept that has ctf values in the majority of the sentences in a document, has a major role to the meaning of its sentences this will provide the chances of determining the topic of the document from the above analysis computing the average ctf values computes the overall significance of each concept to the semantics of a document through the sentences.

B. Analysis of Document-Based Concept:

To examine the each concept at the document level, the calculation of concept based term frequency tf and the no of occurrences of concept c in whole document have to be done. The tf is a local computation on the document level.

C. Analysis of Corpus-Based Concept:

To take out concepts that can separate between documents, the calculation of concept-based document frequency df, the no of documents that contains concept c have to be evaluate.The df is a global measure on the corpus level. This measure is used to recompense the concepts that only emerge in a tiny number of documents as these concepts can separate their documents among others. Concept-based Analysis Algorithm is used to calculate the ctf, tf and df measures in a corpus level.

D. Concept-Based Analysis Algorithm

1. ddoci is a new Document
2.L is an empty List (L is a matched concept list)
3. sdoci is a new sentence in ddoci
4. Build concepts list Cdoci from sdoci
5. for each concept ci ∈ Ci do
6. compute ctfi,tfi,dfi of ci in ddoci
7. sk is a sentence in seen document dk where k={1. . . . doci _ 1}
8. construct concepts list Ck from sk
9. for each concept cj ∈ Ck do
10. if (ci == cj) then
11. update dfi of ci
12. compute ctfweight =avg (ctfi, ctfj)
13. add new concept matches to L
14. end if
15. end for
16. end for
17. output the matched concepts list L

The 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 document that has definite sentence borders. Each sentence is semantically labeled according to [6]. matched concept length of each sentence and their verb argument structures are stored for the concept-based similarity calculations .The concept in the verb argument structures represents the semantic structure of each sentence is processed sequentially. Each concept in the current document is coordinated with the other concepts in the previously processed documents. To match the concepts in previous documents is gifted by keeping a concept list L,that holds the entry for previous documents which shares a concept with the existing document. After the document is processed, L contains all the matching concepts between the current document and previous document. So previous processed documents shares at least one concept with the new document. Finally, L is output as the list of documents with the matching concepts and the required information about them. The concept-based analysis algorithm is capable of matching each concept in a new document with all the previously processed documents in certain time.

E. Tables, Examples and Equations and Example of Calculating the Proposed Conceptual Term Frequency (ctf) Measure

Consider the following sentence:

Ravi **watched** movie yesterday ,his friend lakshman **played** hockey ,tomorrow they may **changed** their idea of entertainment.

In this sentence, the semantic role labeler identifies three 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 Ravi] have [TARGET watched] [ARG1 movie yesterday ,his friend lakshman played hockey , tomorrow they may changed their idea of entertainment.].

Ravi **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** hockey , 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. Thus, this sentence consists of the following three verb argument structures:
1. First verb argument structure for the verb watched:
[ARG0 Ravi]
[TARGET watched]
[ARG1 movie yesterday ,his friend lakshman played hockey, tomorrow they may changed their idea of entertainment.].
2. Second verb argument structure for the verb played:
[ARG1 lakshman]
[TARGET played]
[ARG2 hockey , 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].
. 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:
. Ravi
. watched.
.movie lakshman played hockey , tomorrow they changed their idea of entertainment.].
2. Concepts in the second verb argument structure of the verb played:
. lakshman
. played
.hockey, tomorrow they changed idea of entertainment.
3. Concepts in the third verb argument structure of the verb changed.
. they
. changed
. 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. The terms generated after this step are called concepts

TABLE 1
Example of Calculating the Proposed ctf Measure

Row number	Sentence Concepts	ctf
(1)	Ravi	1
(2)	watched	1
(3)	movie lakshman played hockey , tomorrow they changed their idea of entertainment	1
(4)	lakshman	2
(5)	played	2
(6)	hockey, tomorrow they changed idea of entertainment	3
(7)	They	3
(8)	Changed	
(9)	Idea of entertainment	3
	Individual concepts	
(10)	Ravi	1
(11)	Movie	1
(12)	Lakshman	1
(13)	hockey	1
(14)	they	1
(15)	their	1
(16)	idea	3
(17)	entertainment	3

F. A Concept-Based Similarity Measure

Concepts express local context information, which is vital in determining an exact similarity between documents. A concept-based similarity measure, based on matching concepts at the sentence level, document level, corpus level and combined approach rather than on individual terms (words) only, is devised..The concept-based measure exploits the information extracted from the concept-based analysis algorithm to better judge the similarity between the documents.This similarity measure is a function of the following factors:
1. the number of matching concepts in the verb argument structures in each document
2. the total number of sentences that contain matching concept in each document
3. the total number of the labeled verb argument structures, v, in each sentence s,
4. the ctf of each concept in each document
5. the tf of each concept in each document
6. the df of each concept
7. the length of each concept in the verb argument structure in each document
8. the length of each verb argument structure which contains a matched concept
9. the total number of documents in the corpus.

The conceptual term frequency (ctf) is an important factor in calculating the concept-based similarity measure between documents. .The concept-based matching consists of either an accurate match or partial match between two concepts. accurate match means that both concepts have the same words.Partial match means that one concept includes all the words thatappear in the other concept.

IV.TRIAL RESULTS

To analyse the signifiacance of concept matching in computing an exact measure of the similarity between documents extensive sets of trials using the concept based term analysis and similarity measures are conducted.The trial setup consists of three datasets .The first dataset consists of 20,000 documents from Thrombin datasets.The second dataset consists of 30,000 documents from AWS public dataset.The third dataset consisits of 25,000 documents from Letor4.0 dataset.In these datasets the text is examined directly instead of using a metadata (ie) associated with the text documents.The similarity measures which are computed by using the document ,sentence,corpus and combined approach concept analysis are used to compute four similarity matrices in documents. Document clustering techniques include HAC,Single pass and (K-NN)^2[7][8].The concept based weighting is one of the essential factors that absorbs the significance of a concept in a document.

The weight for tf is computed by

$$weight_i = \text{tf}\ weight_i$$

The weight for ctf is determined by

$$weight_i = \text{ctf}\ weight_i$$

The weight for both tf and ctf is calculated by

$$weight_i = \text{tf}\ weight_i + \text{ctf}\ weight_i$$

To analyse the quality of the clustering,two quality measures are mainly used in literature of the text mining for the use of document clustering.F-measure is the first measures which joins the precision and recall measures.the precision P and recall R of a cluster l with respect to a class k is given as

$$P=\text{Precision}(k,l)=\frac{M_{kl}}{M_l}$$

$$R=\text{Recall}(k,l)=\frac{M_{kl}}{M_k}$$

Where M_{kl} is the number of members of the class in cluster l.$M_l$is the number of members in the cluster l. M_k is the number of members of clas s k.

The F-measure of class k is given as

$$F(k)=\frac{2PR}{P+R}$$

The overall F-measure for clustering result C is given as

$$F_C = \frac{\sum_k (|k| XF(k))}{\sum_k |k|}$$

Where |k| is the number of objects in class k.

DATA SETS	SINGLE -TERM		CONCEPT BASED (CTF)	
	F-measure	*Entropy*	*F-measure* *Entropy*	
AWS public Data sets	0.72±0.31	0.25±0.08	0.89±0.13	0.6±0.4
Thrombin Data sets	0.69±0.21	0.31±0.10	0.88±0.14	0.5±0.3
Letor 4.0 Data sets	0.58±0.20	0.38±0.11	0.87±0.12	0.8±0.2

Entropy simply measures the quality of unnested clusters and it also measures how homogeneous the cluster is and it is calculated by

Entropy simply measures the quality of unnested clusters and it also measures how homogeneous the cluster is and it is calculated by

$$E_C = \sum_{l=1}^{n} \left(\frac{M_i}{M} \times E_l\right)$$

Where $M_l is$ the size of cluster l and M is the data objects.

V. CONCLUSION

This work bridges opening between natural language processing and text mining disciplines. four components of new concept based mining model improves the text clustering quality. By exploiting the semantic structure of the sentences in documents, a better text clustering result is achieved. The first component (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 (corpus based analysis) 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 significance 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 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 strong and accurate way. The excellence of text clustering achieved by this model considerably surpasses the traditional single term- based approaches.There are a number of possibilities for extending this paper. One chance is to apply this concept to web document clustering. Another chance 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. 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
2. 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.
3. K.J. Cios, W. Pedrycz, and R.W. Swiniarski, Data Mining Methods for Knowledge Discovery. Kluwer Academic Publishers, 1998.
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. M.F. Porter, "An Algorithm for Suffix Stripping," Program, vol. 14, no. 3, pp. 130-137, July 1980.
7. A.K. Jain and R.C. Dubes, Algorithms for Clustering Data. Prentice Hall, 1988.
8. L.V Bijuraj ,clustering and its applications in Published in Proceedings of National Conference on New Horizons in IT - NCNHIT 2013.