Concept-based Mining Model for Web Document Clustering

K Munivelu Reddy
M.Tech (Computer Science),
Department of CSE,
JNTUACE, Anantapur, A.P.
E-mail:munivelureddyk@gmail.com

Dr B Eswara Reddy
Associate Professor & HOD,
Department of CSE,
JNTUACE, Anantapur, A.P.
E-mail:eswarcsejntua@gmail.com

***Abstract*—Most of the document clustering techniques are based on statistical analysis of a term, either a word or phrase.The statistical analysis of a term frequency captures the importance of the term within the document only. Thus, theunderlying mining model should indicate terms that capture the semantics of the text. In this case, Themining model can capture terms that present the concepts of the sentence, which leads to the discovery of the topic of document. A new concept-based mining model focuses on the web document clustering;the model consists of three components: concept-based statistical analyzer, COG and concept extractor.The statistical analyzer is to analyze terms on the sentence and document levels. The COG is to extract the most important terms with respect to the meaning of the text. Theconcepts that have maximum weights are selected by the concept extractor.The similarity between documents is calculated based on the Concept-based document similarity measure; It is the combination of ctf, $tf$and df.The experimental results demonstrate extensive comparison between the concept-based analysis and thestatistical analysis.**

***Keywords*—Concept-based mining model, COG, web mining, clustering, document similarity.**

I. Introduction

Natural LanguageProcessing is both a modern computational technology and a method of investigating and evaluating claims about human language itself.NLP is a term that links back into the history of Artificial Intelligence. By applying the text mining in the web domain, the process becomes what is known as Web mining [1]. It can be divided into Web structure mining, Web usage mining and Web content mining. We are mainly focusing on the last type. The Web document clustering is a part of Web content mining and it is concerned broadly because it can decrease the search space, speed up the search and improve the search precision.

Document clustering is one of the traditional data mining techniques, it is an unsupervised learning paradigm where clustering methods try to identify inherent groupings of the documents so that a set of clusters are produced in which clusters exhibit high intracluster and low intercluster similarity[2].Generally, web document clustering methods attempt to segregate the documents into groups where each group repre-sents some topic that is different to those topics represented by the other groups[3][4].

Methods used for document clustering include decision trees, partitioning, hierarchical, and density-based clustering's.Mostdocument clustering methods are based on theVSM [3] [4] which is a widely used data representation for clustering.

The VSM represents each document as a feature vector of the terms that appear in all the document set. Each feature vector contains term weightsof the terms appearing in that 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.

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.

In this paper, a new concept-based mining model is proposed. The proposed model captures the semantic structure of each term within a sentence and document rather than the frequency of the term within a document only. The system consists of five components: statistical analyzer, COG,concept extractor, concept-based similarity, and clustering Model.

The first component is concept-based statistical analyzer that analyzes each term on the sentence, document and corpus levels. After each sentence is labeled by a semantic role labeler, each term is statistically weighted based on its contribution to the meaning of the sentence.

The second component is the COG representation; captures the semantic structure of each term within a sentence and a document only. After each sentence is labeled by a semantic role labeler, all the labeled terms are placed in the COG representation according to their contribution to the meaning of the sentence. Each concept in the COG representation is weighted based on its position in the representation. The important concepts to the statistical analyzer could be non-important to the COG and vice versa. Therefore,the third component is the concept extractor; it combines two different weights computed by the concept-based statistical analyzer and the COG representation to denote the important concepts with respect to the two techniques.The extracted top concepts are used to build VSM and find out concept-based similarity measure [5] among the documentsbased on , and .

The K-NN[6] clustering model has produced different clustering results by concept-based mining model in web document clustering and has higher quality than those produced by a single term analysis only. These results are evaluated using two quality measures, the F-measure and the Entropy. Finally both of these results demonstrate extensive comparison betweenthe concept-based analysis and statistical analysis.

The rest of this paper is organized as follows. SectionII presents the related work. The concept-based mining model which includes statistical analyzer, COG,concept extractor and concept-based document similarity measure is presented in SectionIII. The experimental results are presented in sectionIV. The final section summarizes the conclusionand suggests future work.

II. Related Work

In this section,it is shown how to represent the document and present a brief overview of thematic roles background .

A. *Document Representation*

The direct mapping of the actual document into a formal representation breaks a document into a set of sentences. And sentence weights are assigned according to the mining model. A document is represented as a vector of sentences:

$d_i = \{s_{ij} : j = 1, \dots, p_i\}$,

$s_{ij} = \{t_{ijk} : k = 1, \dots, l_{ij}; w_{ij}\}$,

Where

d_i:is document *i*,

s_{ij}: is sentence *j* in document *i*,

p_i: is the number of sentences in document *i*,

t_{ijk}:is term *k* of sentence s_{ij},

l_{ij}:is the length of sentence s_{ij},and

w_{ij}:is the weight associated with sentence s_{ij}.

B.Thematic Roles Background

Generally, the semantic structure of a sentence can be characterized by a form of verb argument structure. This structure allows the creation of a composite meaning representation from the meanings of the individual concepts in a sentence.The verb argument structure permits a link between the arguments in the surface structures of the input text and their associated semantic roles. The study of the roles associated with verbs is referred to as a thematic role or case role analysis [7]. Thematic roles, first proposed by Fillmore [8], are sets of categories that provide a shallow semantic language to charac-terize the verb arguments.

Recently, there have been many attempts to label thematic roles in a sentence automatically. Gildea and Jurafsky[9] presented a discriminative model for determining the most probable role for a constituent, given the frame, predicator, and other features. These probabilities, that are to be trained depend on the verb, the head words of the constituents, the voice of the verb (active and passive), the syntactic category (S, NP, VP, PP, and so on), and the grammatical function (subject and object) of the constituent to be labeled.

A machine learning algorithm for shallow semantic parsing [10]is based on Support Vector Machines (SVMs) which results in improved performance over that of earlier classifiers by [9]. Shallow semantic parsing is formulated as a multiclass classification problem. SVMs are used to identify the argu-ments of a given verb in a sentence and classify them by the semantic roles that they play such as AGENT, THEME, and GOAL.

III. Concept-based Mining Model

The proposed concept-based mining model is an extension of the work in [11] and itaims to cluster web documents by meaning. The model consists of concept-based statistical analyzer,COG,concept extractor and concept-based document similarity measure, as depicted in Fig.1.

A. *Concept-based Statistical Analyzer*

The Objective of Concept-based statistical analyzer is to analyze terms in the sentence, document and corpus levels rather than a single-term analysis in the document set only. It consists of sentence-based concept analysis, document-based concept analysis and corpus-based concept analysis.

To analyze each concept at the sentence level, a concept-based measure, called the conceptual term frequency .The is the number of occurrences of concept *c* in verb argument structures of sentence *s*. The concept *c*, which frequently appears in the different verb argument structures of the same sentence *s*, has the principal role of contributing to the meaning of *s*. A concept *c* can have many values in different sentences in the same document d. thus, the value of concept *c* in document *d* is calculated by:

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

In equation (1), is the total number of sentences that contain concept *c* in document *d*.

To analyze each concept at the document level, a concept-based measure called term frequency .the is the number of occurrences of a concept *c* in the original document, is calculated, is a local measure on the document level.

The concept-based weighting is one of the main factors that capture the importance of a concept in a sentence and a document. is used to discriminate between non-

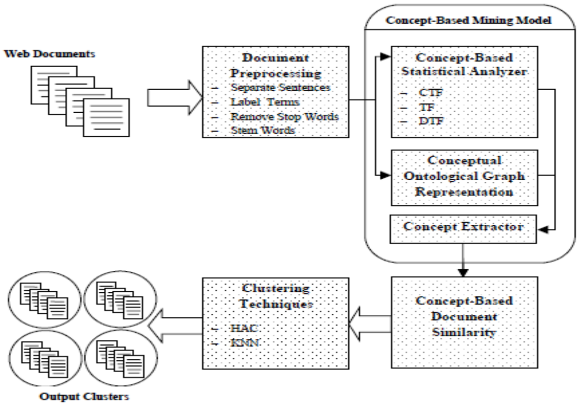

Fig.1.Concept-based Mining Model System

important terms with respect to sentence semantics and terms which hold the concepts that present the meaning of the sentence.

$$weight_{stat_i} = tfweight_i + ctfweight_i \qquad (2)$$

In equation (2), value presents the weight of concept in document *d* at the documentlevel. value presents the weight of the concept *i* in the document *d* at the sentence level based on contribution of concept to the semantics of the sentences in *d*. presents an accurate measure of the contribution of each concept to the meaning of the sentences and to the topics mentioned in a document.

B. Conceptual Ontological Graph

The VSM does not represent any relation among the terms. Therefore, the sentences are broken down into individual components without any representation of either the sentence structure or the sentence semantic structure.The COG representation is a graph G=(C,R) where the concepts are represented as vertices(C).The relations among the concepts such as agents, objects and actions are represented as (R).C is a set of nodes , where each node *c* represents a concepts in the sentence or a nested conceptual graph G; R is a set of edges ,such that each edge *r* is the relation between an ordered pair of nodes .

First, the COG representation captures the concepts and the relations between concepts. Then, it presents the concepts with their relations into one sentence-based conceptual graph representation; it provides different nested levels of concepts in a hierarchical manner based on the importance of the concepts in a sentence.The concepts are placed in the COG according to the amount of overlapping between these terms with respect to the words. To present the levels of the COG hierarchy, five types of verb argument structures are utilized and assigned to their corresponding conceptual graphs:

- One: one generated verb argument structure.
- Main: more than one generated verb argument structure and it has the maximum number of terms that refer to the other verb argument structures.
- Container: more than one generated verb argument structure and it refers to the other arguments and at the same time, the structure does not have the maximum number of referent terms.
- Referenced: the terms referenced by terms in either the main or the container structure.
- Unreferenced: the terms are not referred by any other terms.

This scheme creates a conceptual graph for each verb argumentstructure. Each type of verb argument structure is assigned to its corresponding conceptual graph. TheL_{COG} is proposed to rank concepts with respect to the sentence semantics in the COG. TheL_{COG} measure is assigned to One, Unreferenced, Main, Container and Referenced levels in the COG representation with values 1,2,3,4 and 5 respectively.

The is assigned to each concept presented in theCOG and itis calculated by:

$$weight_{COG_i} = tfweight_i * L_{COG_i} \qquad (3)$$

In equation (3), value presents the importance of the concept *i* in the document *d* at the sentence level based on the contribution of concept *i* to the semantics represented by the levels of the COG. ranks the concepts in document *d* with respect to the contribution of each conceptto the meaning of the sentences and to the topics mentioned in a document.

C. Concept Extractor

The process of selecting the top concepts from the concepts extracted by the concept-based statistical analyzer and the COG is called Concept Extractor. and into one new combined weight is called . It is calculated by:

$$weight_{comb_i} = weight_{stat_i} * weight_{COG_i} \qquad (4)$$

In equation (4), is computed by the concept-based Statistical analyzer and is computed by the COG representation.

D. Concept-based Document Similarity

Concepts convey local context information, which is essential in determining an accurate similarity between documents. A concept-based similarity measure, based on matching concepts at the sentence, document and corpus levels and it relies on three critical aspects. First, the analyzed labeled terms are the concepts that capture the semantic structure of each sentence. Second, the frequency of a concepts is used to measures 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 calculating similarity.

The importance of each concept at the sentence level by the measure, document level by the measure, and corpus level by the measure. The concept-based measure exploits the information extracted from the concept-based statistical analysis and COG algorithms to judge better the similarity between the documents. The concept-based similarity between two documents d1 and d2 is calculated by:

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

$$\times weight_{i_2}, \qquad (5)$$

$$weight_i = (tfweight_i + ctfweight_i) * \log(\qquad (6)$$

In equation (5),*l* is length of concept, is length of verb argument structure, *m* is number of matching concepts, *N* is number of documents.In equation (6), value rewards the weight of the concept *i* on the corpus level, is document frequency, and are computed based on maximum weighted concepts of values, and it is computed using equation (4). value presents the weight of concept *i* at sentencelevel, value presents the weight of concept *i* at document level.

$$ctfweight_i = \frac{ctf_{ij}}{\sqrt{\sum_{j=1}^{cn}(ctf_{ij})^2}}, \qquad (7)$$

$$tfweight_i = \frac{tf_{ij}}{\sqrt{\sum_{j=1}^{cn}(tf_{ij})^2}}, \qquad (8)$$

In equation (7) and (8),cn is total number of concepts, ctf is conceptual term frequency and tf is term frequency.

E. *Concept-based Analysis Algorithm*

```
1.  $d_{doci}$is a new Document
2.  Lis an empty List(L is a concept list)
3.  Tis an empty List( T is a top concept list)
4.  Mis an empty List( M is a matched concepts list)
5.  for each labeled sentence $s_{doci}$ in $d_{doci}$ do
6.   create $COG_i$ for each sentence $s_{doci}$
7.  foreach concept $c_i$in sentence $s_{doci}$do
8.        compute $ctf_i$of $c_i$ in $d_{doci}$
9.        compute$tf_i$of $c_i$in $d_{doci}$
10. compute $weight_{stat_i}$for each concept $c_i$
11. compute $weight_{COG_i}$for each concept $c_i$
12. compute $weight_{comb_i} = weight_{stat_i} * weight_{COG_i}$
13.      add concept $c_i$ to L
14. end for
15. end for
16. sort  L descendingly  based on $weight_{comb_i}$
17. output the $max(weight_{comb_i})$from List L to List T
18. for each concept $c_i$in T do
19.  if ($c_i$= = $c_j$) where $j = \{1,2,\ldots,N\}$ and
      N is total number of documents then
20.      compute $ctf_{weight} = avg(ctf_i, ctf_j)$
21.      compute $tf_{weight} = avg(tf_i, tf_j)$
22.      update $df_i$of $c_i$
23.       add new concept matches to M
24.  end if
25.  end for
26. output the matched concepts  list M
```

The procedure begins with processing a new document (at line 1) which has well defined sentence boundaries. Each sentence is semantically labeled according to [12]. Each concept in the current document is matched with the other concepts in the previously processed documents. After the document processed, M contains all the matching concepts between the current document and all previous documents. Finally, M is output as the list of documents with the matching conceptsand necessary information about them. The concept-based analysis algorithm is capable of matchingeach concept in a new document (d) with all the previously processed documents in $O(m)$ time, where m is the number of concepts.

IV. Experimental Results

In order to test the effectiveness of the Web document clustering, we conducted a set of experiments using our proposed mining model, similarity measureand K-NN clustering method.

The availability of web document datasets suitable for clustering is limited. However, we used four datasets. The first dataset is collected from 4-Universities dataset and it is classified into department, course, staff, student andproject categories. The second dataset is collected from Bank dataset and it has 11 categories. The third dataset contains abstracts collected from ACM digital library. The ACM abstracts areclassified into three main categories: data mining, computer organization and software engineering. The fourth dataset is collected from Reuters-21578 dataset.

The similarities which are calculated by the concept based model are used to compute a similarity matrix among documents. One standard document clustering technique [13] is chosen for testing the effect of the concept-based similarity on clustering: K-NN. To evaluate the quality of the clustering, two quality measures widely used in the text mining literature for the purpose of document clustering.

The first is the *F-measure*, which combines the Precision and Recall measures from the Information Retrieval literature. The precision and recall of a cluster with respect to a class are defined a i and

where isthe number of members of classes in cluster , is the number of members of cluster j, and is the number of members of class . The F-measure of a class is defined as

The second measure is the *Entropy*, which measures how homogeneous a cluster is. The higher the homogeneity of a cluster, the lower the entropy is, and vice versa. For every cluster in the clustering result , the probability that a member of cluster belongs to class is computed.

Basically, the aim is to maximize the F-measure, and minimize the Entropy of clusters to achieve high qualityclustering. The results are listed in Table 1 show the improve-ment on the clustering quality using the concept-based mining model. Fig.2 and Fig.3 shows the Table 1 results more clearly, showing the achieved improvement in comparison with other method. The percentage of improvement range from +42.13% to +91.72% increase in the F-measure quality, and -47.78% to -63.47% drop in Entropy.

Table 1: Clustering Improvement

Dataset	Single-Term Similarity		Concept-Based Similarity		Improvement
	F-Measure	Entropy	F-Measure	Entropy	
Bank	0.617	0.252	0.877	0.123	+42.13%,-51.19%
Reuters-21578	0.508	0.282	0.904	0.103	+77.95%,-63.47%
4 Universities	0.435	0.293	0.834	0.153	+91.72%,-47.78%
ACM	0.468	0.288	0.866	0.131	+85.04%,-54.51%

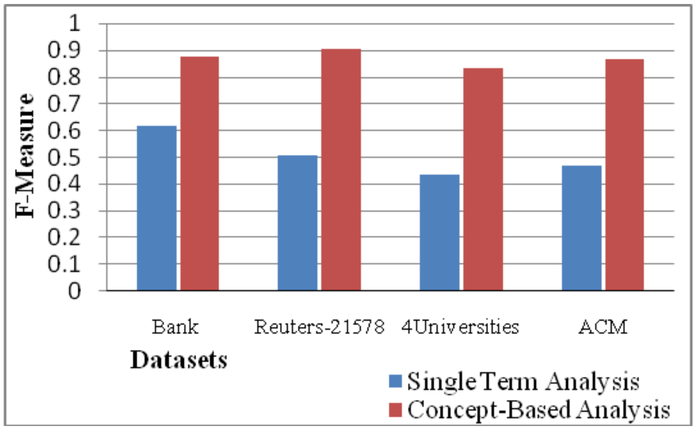

Fig.2.Clustering Quality-F-Measure

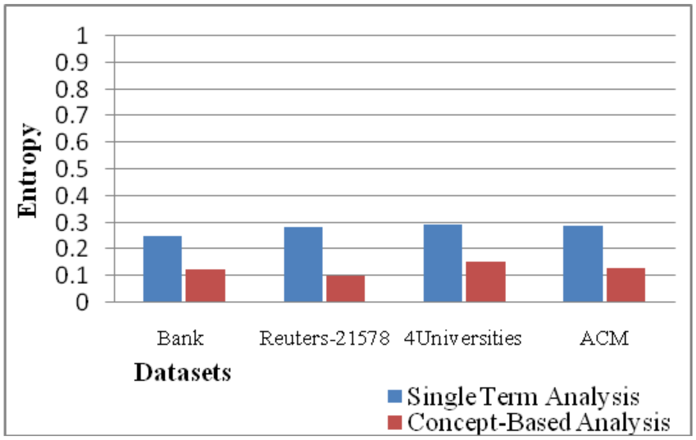

Fig.3.Clustering Quality-Entropy

V. Conclusion

This paper bridges the gap between natural language pro-cesssing and web mining disciplines. A new concept-based mining model composed of four components, is proposed to improve the document clustering quality.

The first component is the Concept-based Statistical Analyzer which analyzes the semantic structure of each sentence to capture the sentence concepts using the measure.

The second component is the COG. This representation captures the structure of the sentence semantics represented in the COG hierarchical levels. Thethirdcomponent is the concept extractor which combines the weights of concepts extracted by the Concept-based Statistical Analyzer and the COG into one top conceptlist.The extracted top concepts are used to build standard normalized feature vectors using standard Vector Space Model for the purpose of document clustering.The fourth component is the concept-based simil-arity measure which allows measuring the importance of each concept with respect to the semantics of the sentence, and the topic of thedocument, and the discrimination among documents in a corpus.

The combining of the factors affects the weights of concepts on concept-based statistical analyzer, COG and concept-based similarity measure that is capable of the accurate calculation of pair-wise documents is devised. The quality of document clustering achieved by this model significantly surpasses the traditional single-term based approaches.

Acknowledgment

The authors would like to express their cordial thanks to their esteemed patrons Prof. A.AnandaRao, Principal, JNTUACE,Anantapur, Prof.V.Sankar, Registrar and Prof.

B.Krishna Gandhi, Vice-Chancellor, JNTU Anantapur, Andhra Pradesh.

REFERENCES

[1] R. Kosala and H. Blockeel, "Web Mining Research: A Survey",ACM SIGKDD Explorations Newsletter, vol. 2, no. 1, pp. 1-15, 2000.

[2] J.Han, M.Kamber,"Data Mining–Concepts and Techniques", Morgan Kaufmann Publishers, CA,USA, 2003.

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

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

[5] S. Shehata,F. Karray,and M. Kamel,"Enhancing Text Clustering Using Concept-based mining model,"Proc. Sixth IEEE Int'l Conf. Data Mining (ICDM), 2006.

[6] A.K.Jain, M.N. Murty and P.J. Flynn,"Data Clustering : A Review", ACM Computing Surveys, Vol. 31, No. 3, September 1999

[7] D. Jurafsky and J.H. Martin,"Speech and Language Processing",Prentice Hall, 2000.

[8] C. Fillmore, "The Case for Case," Universals in Linguistic Theory,Holt, Rinehart and Winston, 1968.

[9] D. Gildea and D. Jurafsky,"Automatic Labeling of SemanticRoles,"Computational Linguistics, vol. 28, no. 3, pp. 245-288, 2002.

[10] S. Pradhan, W. Ward, K. Hacioglu, J. Martin, and D. Jurafsky,"Shallow Semantic Parsing Using SVM,"HLT/NAACL, 2004.

[11] 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, October 2010.

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

[13] M.Steinbach, G.Karypis, and V.Kumar, "A Comparison of Document Clustering Techniques," Proc. Knowledge Discovery andData Mining (KDD) Workshop Text Mining, Aug. 2000.