Effective Concept-Based Mining Model For Text Clustering

A.Nirmala**, K.A.Vanitha*

Abstract—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. 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. Usually in text mining techniques the basic measures like term frequency of a term (word or phrase) is computed to compute the importance of the term in the document. But with statistical analysis, the original semantics of the term may not carry the exact meaning of the term. To overcome this problem, a new framework has been introduced which relies on concept based model and synonym based approach. The proposed model can efficiently find significant matching and related concepts between documents according to concept based and synonym based approaches. The relations between verbs and their arguments in the same sentence have 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. 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. 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 nonimportant 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. Experimental results demonstrate the substantial enhancement of the clustering quality using sentence based, document based, corpus based and combined approach concept analysis. A new similarity measure has been proposed to find the similarity between a document and the existing clusters, which can be used in classification of the document with existing clusters.

Index Terms—Concept-based mining model, text clustering, conceptual ontological graph, document frequency, document level, sentence-based, document-based, corpus-based, concept analysis, conceptual term frequency and concept-based similarity.

1. INTRODUCTION

In the present work a new synonym based mining model has been proposed. It inherits all the benefits of existing concept based mining model. In addition to that it flavors the essence of synonym based matching. The present work models both clustering and classification at the same time the work shows that the same similarity measures can be used in synonym based approach also.

Usually, 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. 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.

Text mining refers to the knowledge extraction from textual databases or documents. This text mining is different from mining the other types of databases because of its unstructured form and large number of dimensions.

Each word in the document is a dimension. So the foremost things for text mining are, giving a structure to the data and reducing the dimensions. Latent semantic analysis is a most popular method used in text mining. As the text data is unstructured data and higher dimensional data, the main things that has to be done in text mining are

1. Giving a structure to the unstructured data and

2. Reduce the dimensions as much as possible.

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

**Assistant Professor, Department of Computer Applications, Dr.N.G.P Arts and Science College.

*student, Department of Computer Applications, Dr.N.G.P Arts and Science College.

labeling can be done by semantic role parsing. That is, 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. In this approach each text file is treated as a vector and the elements of vector are weight given to each word in that file. Some methods used for text clustering include decision trees, conceptual clustering [1], clustering based on data summarization [2], statistical analysis neural nets, inductive logic programming, and rule-based systems among others.

The concept-based similarity measure used in the proposed system outperforms other similarity measures. The similarity between documents depends on the concept analysis on the sentence, document and the corpus levels. The quality of clusters produced is influenced by the similarity measure used as it is insensitive to noise while calculating the similarity. This is because the concepts are analyzed in the sentence, document and corpus levels and hence the probability to find a concept match between unrelated documents is very small.

The important terms used in this paper are given below:

· Verb Argument structure: (e.g.: Adam plays the guitar). "plays" is the verb. "Adam" and "the guitar" are the arguments of the verb "plays".

· Label: An argument is assigned a label (e.g.: Adam plays the guitar). "Adam" has subject label and "the guitar" has object label.

· Term: It is either an argument or a verb. It can also be a word or a phrase.

· Concept: The concept is a labeled term.

The concepts can be identified by using natural language processing on the text document. That is by giving the structure to each sentence. This structure is called verb argument structure.

See the example for verb argument structure of a sentence.

Example:

Sentence: He *hits* a ball.

Verb: hits

Arg0: he

Arg1: a ball

These labels are according to the prop bank notations [3]. A single word may have different senses. Using this semantic role, we can get the content in which the word is being used in that sentence. Another important thing is a single sense can be represented by different words.

2. THEMATIC ROLES BACKGROUND

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

Consider the following example: My brother wants a Pen. 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 brother" is preverbal and plays the role of the subject.

4. The second argument "a pen" is a post verbal and plays the role of the direct object.

Many types of text representations have been proposed in the past. A well known one is the bag of words that uses keywords (terms) as elements in the vector of the feature space. In H. Jin, M.-L. Wong, and K.S. Leung [4], states the tf*idf weighting scheme is used for text representation in Rocchio classifiers. In addition to TFIDF, the global IDF and entropy weighting scheme is proposed in P. Kingsbury and M. Palmer [5] and improves performance by an average of 30 percent. Various weighting schemes for the bag of words representation.

The problem of the bag of words approach is how to select a limited number of features among an enormous set of words or terms in order to increase the system's efficiency and avoid over fitting states S. Pradhan, K. Hacioglu, V. Krugler, W. Ward, J. H. Martin, and D. Jurafsky [6]. In order to reduce the number of features, many dimensionality reduction approaches have been conducted by the use of feature selection techniques, such as Information Gain, Mutual Information, Chi-Square, Odds ratio, and so on. Details of these selection functions were stated in [7].The choice of a representation depended on what one regards as the meaningful units of text and the meaningful natural language rules for the combination of these units S. Pradhan, K. Hacioglu, V. Krugler, W. Ward, J. H. Martin, and D. Jurafsky [6]. With respect to the representation of the content of documents, some research works have used phrases rather than individual words.

3. TEXT MINING

A "sentence-based concept analysis", "document-based concept analysis", "corpus-based concept analysis" text clustering approach has been proposed that performs "concept-based similarity measure 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. In the proposed model, three measures for analyzing concepts on the sentence, document, and corpus levels are computed. A new concept-based similarity measure which makes use of the concept analysis on the sentence, document, and corpus levels is proposed

Sometimes alternately referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. High quality information is typically derived through the divining of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' in text mining usually refers to some combination of relevance, novelty, and interestingness. Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling (i.e., learning relations between named entities).

4. CONCEPT-BASED MINING MODEL

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 analyzed 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.

A raw text document is the input to the proposed model. Each document has well defined sentence boundaries. Each sentence in the document is labeled automatically based on the Prop Bank notations. After running the semantic role labeler [5], each sentence in the document might have one or more labeled verb argument structures. The number of generated labeled verb argument structures is entirely dependent on the amount of information in the sentence. The sentence that has many labeled verb argument structures includes many verbs associated with their arguments. The labeled verb argument structures, the output of the role labeling task, are captured and analyzed by the concept-based model on the sentence and document levels.

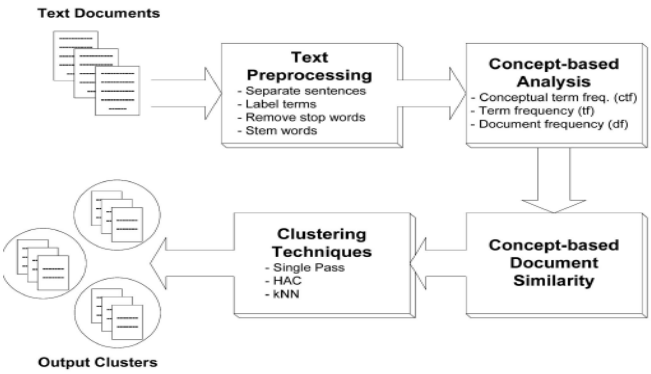

Fig.4.1 Concept-based Mining model

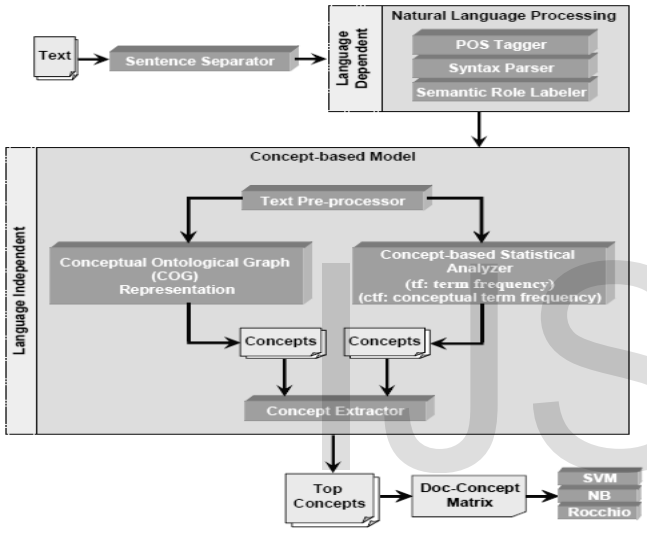

Fig.4.2 Concept-based model system

4.1 Text Clustering

Clustering is an unsupervised classification process; differently from supervised classification no a priori information about classes is required. Document clustering is an optimization process which attempts to determine a partition of the document collection so that documents within the same cluster are as similar as possible (cluster compactness) and the discovered clusters as separate as possible (cluster distinctness). Document clustering algorithms are used in a variety of tasks and applications for facilitating automatic organization, browsing, summarization, and retrieval of structured and unstructured documents.

4.2 Clustering techniques

- Agglomerative vs. divisive. The former begin by treating each text as a cluster and successively merge them until a stopping criterion is met (the bottom-up style); the latter begin by placing all texts in a single group and perform splitting until a stopping criterion is met (the top-down style).
- Hierarchical vs. partitional. This aspect relates to the structure of the clusters that are produced. The former algorithms form a hierarchy of clusters: clusters at lower levels are nested to upper level clusters. The latter produce a single at partition.
- Hard vs. fuzzy. This aspect concerns cluster membership. The former methods allocate each text to a single cluster while the latter predict its degree of membership for multiple clusters. A fuzzy method can be converted to a hard one by assigning texts to the cluster that has the highest degree of membership.

4.3. Comparison

4.3.1 Agglomerative Algorithms

A new class of agglomerative algorithms by constraining the agglomeration process using clusters obtained by partitional algorithms. Our experimental results showed that partitional methods produced better hierarchical solutions than agglomerative methods, and that the constrained agglomerative methods improved the clustering solutions obtained by agglomerative or partitional methods alone. These results suggest that the poor performance of agglomerative methods may be attributed to the merging errors they make during early stages, which can be eliminated to some extent by introducing partitional constrains.

4.3.2 K-means Clustering

For K-means we used a standard K-means and a variant of K-means, bisecting K-means. Our results indicate that the bisecting K-means technique is better than the standard K-means approach and as good as or better than the hierarchical approaches that we tested. More specifically, the bisecting K-means approach produces significantly better clustering solutions quite consistently according to the entropy and overall similarity measures of cluster quality. Furthermore, bisecting K-means seems consistently to do slightly better at producing document hierarchies In addition, the run time of bisecting K-means is very attractive when compared to that of agglomerative

hierarchical clustering techniques - O (n) versus O (n2).The reason that our relative ranking of K-means and hierarchical algorithms differs from those of other researchers could be due to many factors. First we used many runs of the regular K-means algorithm. If agglomerative hierarchical clustering techniques such as UPGMA are compared to a single run of K-means, then the comparison would be much more favorable for the hierarchical techniques. Secondly, we used incremental updating of centroids, which also improves K-means. Of course, we also used the bisecting K-means algorithm, which, to our knowledge, has not been previously used for document clustering. While there are many agglomerative hierarchical techniques that we did not try, we did try several other techniques which we did not report here. The results were similar– bisecting K-means performed as well or better then the hierarchical techniques that we tested. Finally, note that hierarchical clustering with a K-means refinement is essentially a hybrid hierarchical-K-means scheme similar to other such schemes that have been used before. In addition, this scheme was better than any of the hierarchical techniques that we tried, which gives us additional confidence in the relatively good performance of bisecting K-means vis-à-vis hierarchical approaches.

We caution that the main point our paper is not a statement that bisecting K-means is "superior" to any possible variations of agglomerative hierarchical clustering or possible hybrid combinations with K-means. However, given the linear run-time performance of bisecting K-means and the consistently good quality of the clustering that it produces, bisecting K-means is an excellent algorithm for clustering a large number of documents. We argued that agglomerative hierarchical clustering does not do well because of the nature of documents, i.e., nearest neighbors of documents often belongs to different classes. This causes agglomerative hierarchical clustering techniques to make mistakes that cannot be fixed by the hierarchical scheme. Both the K-means and the bisecting K-means algorithms rely on a more global approach, which effectively amounts to looking at the similarity of points in a cluster with respect to all other points in the cluster. This view also explains why a K-means refinement improves the entropy of a hierarchical clustering solution.

4.4 Concept-based Statistical Analyzer

To analyze each concept at the sentence-level, a concept-based frequency measure, called the conceptual term frequency (*ctf*) is utilized. 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 analyze 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 concept-based weighting is one of the main factors that capture the importance of a concept in a sentence and a document. Thus, the concepts which have highest weights are captured and extracted.

$$\text{weight}_{\text{stati}} = \text{tfweight}_i + \text{ctfweight}_i \qquad (4.1)$$

In calculating the value of $\text{weight}_{\text{stati}}$ in equation (4.1), the tfweight_i value presents the weight of concept *i* in document *d* at the document-level and the ctfweight_i value presents the weight of the concept *i* in the document *d* at the sentence-level based on the contribution of concept *i* to the semantics of the sentences in *d*. The sum between the two values of tfweight_i and ctfweight_i presents an accurate measure of the contribution of each concept to the meaning of the sentences and to the topics mentioned in a document.

4.5 Conceptual Ontological Graph (COG)

The COG representation is a conceptual graph $G = (C, R)$ where the concepts of the sentence, 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 ($c_1, c_2, \ldots, c_n$), where each node c represents a concept in the sentence or a nested conceptual graph G; and R is a set of edges ($r_1; r_{2,\ldots,} r_m$), such that each edge r is the relation between an ordered pair of nodes ($c_i, .., c_j$). The output of the role labeling task, which are verbs and their arguments are presented as concepts with relations in the COG representation. This allows the use of more informative concept matching at the sentence-level and the document-level rather than individual word matching.

$$\text{tfweight}_i = (tf_{ij}) / (\sqrt{\textstyle\sum_{j=1}^{cn}} (tf_{ij})^2) \qquad (4.2)$$

This scheme creates a conceptual graph for each verb argument structure. Each type of verb argument structure is assigned to its corresponding conceptual graph. The COG presents the conceptual graphs as levels, which are determined according to their types. A new measure L_{COG} is proposed to rank concepts with respect to the sentence semantics in the COG representation. The proposed L_{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. Instead of selecting concept from only one level in the COG representation, concepts in the entire levels of the COG representation are considered and weighted. The proposed $weight_{COG}$ is assigned to each concept presented in the COG representation and is calculated by:

$$weight_{COGi} = tfweight_i * L_{COGi} \quad (4.3)$$

In equation (4.3), the *tfweighti* value presents the weight of concept *i* in document *d* at the document-level as shown in equation (4.2). The L_{COGi} 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 of the sentences represented by the levels of the COG representation. The multiplication between the two values of $tfweight_i$ and L_{COGi} ranks the concepts in document *d* with respect to the contribution of each concept to the meaning of the sentences and to the topics mentioned in a document. For implementation and performance purposes, it is imperative to note that the COG representation maintains the identification number of each concept and each relation node, rather than, the values of the nodes.

5. EXPERIMENTAL RESULTS

5.1 Porter Stemmer Algorithms

. Porter's algorithm was developed for the stemming of English-language texts but the increasing importance of information retrieval in the 1990s led to a proliferation of interest in the development of conflation techniques that would enhance the searching of texts written in other languages. By this time, the Porter algorithm had become the standard for stemming English, and it hence provided a natural model for the processing of other languages. In some of these new algorithms the only relationship to the original is the use of a very restricted suffix dictionary (Porter, 2005), but Porter himself has developed a whole series of stemmers that draw on his original algorithm and that cover Romance (French, Italian, Portuguese and Spanish), Germanic (Dutch and German) and Scandinavian languages (Danish, Norwegian and Swedish), as well as Finnish and Russian (Porter, 2006).

These stemmers are described in a high-level computer programming language, called Snowball (Porter, 2006) that has been developed to provide a concise but unambiguous description of the rules for a stemmer. Some non-English stemmers can operate effectively using simple sets of rules, with Latin being perhaps the best example of a language that is defined in what is essentially algorithmic form (Schinke *et al*., 1996). However, this level of regularity and simplicity is by no means common; in such cases, Snowball provides a concise but powerful description that can then be processed by a compiler to give a C or Java implementation of the algorithm for the chosen language (Porter, 2001). In passing, it is worth noting that this paper by Porter contains an extremely illuminating discussion of stemming and the structures of words that are very well worth reading, even if one does not wish to obtain any of the downloadable programs. These developments of the Porter algorithm can only serve further to increase the level of knowledge and understanding of the original, English-language version; this level is already considerable as is evidenced by the following simple citation analysis. While the precise relationship between citation and significance is a matter of some dispute, it does seem reasonable to regard the 1980 Program paper as being a significant contribution to the literature since a search of the *ISI Web of* Knowledge database on 21st March 2006 yielded 442 citations.

Porter's algorithm is important for two reasons. First, it provides a simple approach to conflation that seems to work well in practice and that is applicable to a range of languages. Second, it has spurred interest in stemming as a topic for research in its own right, rather than merely as a low-level component of an information retrieval system. The algorithm was first published in 1980; however, it and its descendants continue to be employed in a range of applications that stretch far beyond its original intended use.

5.2 Implementation of Porter Stemming algorithm

A consonant will be denoted by c, a vowel by v. A list ccc... of length greater than 0 will be denoted by C, and a list vvv... of length greater than 0 will be denoted by V. Any word, or part of a word, therefore has one of the four forms:

CVCV ... C
CVCV ... V
VCVC ... C
VCVC ... V

These may all be represented by the single form [C] VCVC ... [V] where the square brackets denote arbitrary presence of their contents. Using (VC) {m} to denote VC repeated m times, this may again be written as

[C](VC){m}[V].

'm' will be called the measure of any word or word part when represented in this form.

The rules for removing a suffix will be given in the form

(Condition) S1 -> S2

This means that if a word ends with the suffix S1, and the stem before S1 satisfies the given condition, S1 is replaced by S2. The condition is usually given in terms of m.

The `condition' part may also contain the following:

*S - the stem ends with S (and similarly for the other letters).

v - the stem contains a vowel.

*d - the stem ends with a double consonant (e.g. -TT, -SS).

*o - the stem ends cvc, where the second c is not W, X or Y (e.g. -WIL, -HOP).

Step 1a

SSES -> SS	caresses -> caress
IES -> I	ponies -> poni ties -> ti
SS -> SS	caress -> caress
S ->	cats -> cat

Step 2

(m>0) ATIONAL -> ATE	relational - > relate
(m>0) TIONAL -> TION	conditional -> condition rational -> rational
(m>0) ENCI -> ENCE	valenci -> valence
(m>0) ANCI -> ANCE	hesitanci -> hesitance
(m>0) IZER -> IZE	digitizer -> digitize
(m>0) ABLI -> ABLE	conformabli -> conformable
(m>0) ALLI -> AL	radicalli -> radical
(m>0) ENTLI -> ENT	differentli -> different
(m>0) ELI -> E	vileli -> vile
(m>0) OUSLI -> OUS	analogousli -> analogous
(m>0) IZATION -> IZE	vietnamization -> vietnamize
(m>0) ATION -> ATE	predication -> predicate
(m>0) ATOR -> ATE	operator -> operate
(m>0) ALISM -> AL	feudalism -> feudal
(m>0) IVENESS -> IVE	decisiveness -> decisive
(m>0) FULNESS -> FUL	hopefulness -> hopeful
(m>0) OUSNESS -> OUS	callousness -> callous
(m>0) ALITI -> AL	formaliti -> formal
(m>0) IVITI -> IVE	sensitiviti -> sensitive
(m>0) BILITI -> BLE	sensibiliti -> sensible

The test for the string S1 can be made fast by doing a program switch on the penultimate letter of the word being tested. This gives a fairly even breakdown of the possible values of the string S1. It will be seen in fact that the S1-strings in step 2 are presented here in the alphabetical order of their penultimate letter. Similar techniques may be applied in the other steps.

ACKNOWLEDGEMENT

With all humility and submissiveness I surrender myself at the diving feet of god and submit my foremost gratitude and indebtedness of having gracefully blessed me with knowledge, skill and enthusiasm.

I own my heart felt gratitude to Dr.Nalla G Palaniswamy, M.D, A.B (USA), and Chairman, KMCRET (Kovai Medical Centre and Research & Educational Trust) Chairman, Dr N.G.P arts and Science College for providing me him wishes for doing this dissertation.

I take a great pleasure in thanking Dr.Thavamani D Palaniswami, M.D., A.B (USA), Secretary, Dr N.G.P Arts and Science College for her blessing provided for the completion of this dissertation.

I express my sincere thanks to Dr.P.R.Muthuswamy, M.A., M.B.A., FDPM (IIM.A), Ph.D., Principal, Dr N.G.P Arts and Science College for giving me a chance for doing this dissertation as a part of my degree.

My deep gratitude and thanks to Mrs.R.KOUSALYA, M.C.A., M.Phil, Ph.D., Head, Department of COMPUTER APPLICATIONS, Dr.N.G.P Arts and Science College, Coimbatore for her irrational, through provoking aspects towards our Research.

It is a great privilege and immence pleasure to express my profound sense of gratitude to Mrs.A.NIRMALA M.C.A., M.Phil, Ph.D., Research Guide in Computer Applications, Dr.N.G.P Arts and Science College, Coimbatore for her eminent guidance, innovative ideas, and immense help rendered at various stages of this study.

6. CONCLUSION

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 ensures on the sentence, document, and corpus levels in calculating the similarity between documents. Large sets of experiments using the proposed concept-based mining model on different data sets in text clustering are conducted. 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.

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 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 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 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 pair wise 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] R. Feldman and I. Dagan. Knowledge discovery in textual databases (kdt). In Proceedings of First International Conference on Knowledge Discovery and Data Mining, pages 112{117, 1995.

[2] C. Fillmore. *The case for case. Chapter in: Universals in Linguistic Theory*. Holt, Rinehart and Winston, Inc., New York, 1968.

[3] W. Francis and H. Kucera. *Manual of information to accompany a standard corpus of present-day edited american english, for use with digital computers*, 1964.

[4] 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.

[5] P. Kingsbury and M. Palmer. Propbank: the next level of treebank. In *Proceedings of Treebanks and Lexical Theories*, 2003.

[6] S. Pradhan, K. Hacioglu, V. Krugler, W. Ward, J. H. Martin, and D. Jurafsky. Support vector learning for semantic argument classification. *Machine Learning*, 60(1-3):11{39, 2005.

[7] 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.