ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)
Vol. 2, Issue 3 (July - Sept. 2014)
International Journal of Advanced Research in Computer Science & Technology (IJARCST 2014)

Enhance the Text Clustering using an Efficient Concept-Based Mining Model

[I]M. VeeraKarthik, [II]M. Elamparithi
[I]Research Schokar, Saraswathi Thyagaraja College, Pollachi.
[II]Assistant Professor, Dept. of MCA, Saraswathi Thyagaraja College, Pollachi.

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.

Keywords
Text, Clustering, Mining, Concept, Model, Verb, Sentence.

I. Introduction

Data mining refers to extracting or mining knowledge from large amounts of data. It is becoming one of the most active and exciting research areas. Data mining is a natural result of the evolution of information technology. Our capabilities of both generating and collecting data have been increasing rapidly in the last several decades. Contributing factors include the widespread use of bar codes for most commercial products, the computerization of many businesses, scientific, and government transactions, and the advances in data collection tools ranging from scanned text and image platforms to satellite remote sensing systems.

In addition, the popular use of World Wide Web as a global information system has coded us with tremendous amount of data and information. It is impractical for human to look through all the data and discover some untapped valuable patterns. We are drowning in data, but starving for knowledge. This explosive growth in stored data has generated an urgent need for new techniques and tools that can intelligently and automatically assist us in transforming the data into useful knowledge.

Data mining is also viewed as an essential step in the process of Knowledge Discovery in Databases (KDD). It is defined as a non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns from large amount of data. Since data mining is an essential and crucial step of KDD, it is also treated as a synonym for KDD to some people.

1. Data selection where data relevant to the analysis task are retrieved from the database,
2. Preprocessing where data are cleaned and/or integrated,
3. Data transformation where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations,
4. Data mining which is an essential process where intelligent methods are applied in order to extract patterns and knowledge, and
5. Interpretation/evaluation which identifies truly interesting patterns representing knowledge based on some interestingness measures.

Data mining technologies are characterized by intensive computations on large amounts of data. The two most significant challenges in data mining are scalability and performance.

II. Text Mining

The phrase "text mining" is generally used to denote any system that analyzes 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. Text mining is a burgeoning new field that attempts to glean meaningful information from natural language text. It may be loosely characterized as the process of analyzing text to extract information that is useful for particular purposes. Compared with the kind of data stored in databases, text is unstructured, amorphous, and difficult to deal with algorithmically. Nevertheless, in modern culture, text is the most common vehicle for the formal exchange of information. The field of text mining usually deals with texts whose function is the communication of factual information or opinions, and the motivation for trying to extract information from such text automatically is compelling—even if success is only partial.

Text mining appears to embrace the whole of automatic natural language processing and, arguably, far more besides—for example, analysis of linkage structures such as citations in the academic literature and hyperlinks in the Web literature, both useful sources of information that lie outside the traditional domain of natural language processing. But, in fact, most text mining efforts consciously shun the deeper, cognitive, aspects of classic natural language processing in favor of shallower techniques more akin to those used in practical information retrieval.

Text mining is an outgrowth of this "real text" mindset. Accepting that it is probably not much, what can be done with unrestricted input? Can the ability to process huge amounts of text compensate

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

for relatively simple techniques? Natural language processing, dominated in its infancy by unrealistic ambitions and swinging in childhood to the other extreme of unrealistically artificial worlds and trivial amounts of text, has matured and now embraces both viewpoints: relatively shallow processing of unrestricted text and relatively deep processing of domain-specific material.

The Vector Space Model is widely used document clustering method and represents data for text classification and clustering. The terms in the document is represented as a feature vector. The terms can be words or phrases. Each feature vector is assigned a term weight based on the term frequency of the terms in the documents. Similarity measures that rely on the feature vector is used to find the similarity between the documents (Cosine measures and the Jaccard measure).

A. Document clustering

Text categorization is a kind of "supervised" learning where the categories are known beforehand and determined in advance for each training document. In contrast, document clustering is "unsupervised" learning in which there is no predefined category or "class," but groups of documents that belong together are sought. For example, document clustering assists in retrieval by creating links between similar documents, which in turn allows related documents to be retrieved once one of the documents has been deemed relevant to a query.

Clustering schemes have seen relatively little application in text mining applications. While attractive in that they do not require training data to be pre-classified, the algorithms themselves are generally far more computation-intensive than supervised schemes ([Willett, 1988] surveys classical document clustering methods).

Clustering is an indispensable data mining technique, particularly for handling large scale

data. Applied to documents, it automatically groups ones with similar themes together while separating those with different topics. Creating a concise representation of a document is a fundamental problem for clustering and for many other applications that involve text documents, such as information retrieval, categorization and information extraction. Redundancy in feature space adds noise and often hurts subsequent tasks.

B. Term clustering

Term clustering uses clustering techniques to group terms based on their distributions in a given text collection. Each of the resulting term clusters consists of terms that co-occur frequently with each other. The underlying assumption is that such terms are either similar or closely related, and there is no need to distinguish between them for the clustering task. In contrast, term clusters provide compact and efficient representation of texts.

Both term clustering and dimensionality reduction depend heavily on the input data, and this cause several restrictions. First, it is difficult to generalize the latent topics term clusters for the former and term combinations for the latter to new data, especially for previously unseen terms. An extreme case is when the new texts have completely different topics from the data used for constructing the clusters: for example, when building the clusters with texts on sports when most of the new texts discuss politics.

Texts are then represented by the term clusters, and each word's weight is aggregated to the term cluster it belongs to. This connects texts with related topics yet distinct vocabularies, which is particularly helpful for short texts, such as queries and search result snippets. Indeed, term clustering has been extensively investigated in information retrieval, dating back to the work of Sparck Jones in the 1970-80s. For longer documents, however, studies show that using term clusters yields limited success in the text categorization task. In contrast, in the context of text clustering, term clustering has been shown to be quite effective. Slonim and Tishby extend their information theoretic clustering method called the information bottleneck method to term clustering. Their method works in two steps first grouping terms based on their occurrence in the texts, and then representing and clustering the texts by the resulting term clusters. Their experiments on the 20Newsgroup collection show a 17% improvement over the complete-link hierarchical agglomerative clustering with the bag-of-words model.

Term clusters reduce the dimensionality of the feature space from 2000 (words) to 10-50 (term clusters), which contributes to the method's success. Frequent itemset clustering shows a similar effect (Beil et al., 2002; Fung et al., 2003). It first applies association rule learning to identify the frequent itemsets of words in a text collection (Agrawal and Srikant, 1994). Each frequent itemset can be regarded as a term cluster consisting of words that occur together in a minimum fraction of the texts. The intuition is that texts belonging to the same category share many frequent itemsets and those from different categories (i.e., topics) share few.

III. Existing Methodology

Most of the common techniques in text mining are based on the statistical analysis of a term, either word or phrase. Statistical analysis of a term frequency captures the importance of the term within a document only. However, two terms can have the same frequency in their documents, but one term contributes more to the meaning of its sentences than the other term. Thus, the underlying text mining model should indicate terms that capture the semantics of text. In this case, the mining model can capture terms that present the concepts of the sentence, which leads to discovery of the topic of the document.

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 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. 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.

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 concepts 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.

Clustering is a well-known problem in statistics and engineering,

namely, how to arrange a set of vectors (measurements) into a number of groups (clusters). Clustering is an important area of application for a variety of fields including data mining, statistical data analysis and vector quantization.

"Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources." Text mining techniques are the fundamental and enabling tools for efficient organization, navigation, retrieval and summarization. With more and more text information are spreading around on Internet, text mining is increasing in importance. Text clustering and text classification are two fundamental tasks in text mining.

The k-means algorithm is based on the simple observation that the optimal placement of a center is at the centroid of the associated cluster.

The attractiveness of the k-means lies in its simplicity and flexibility. In spite of other algorithms being available, k-means continues to be an attractive method because of its convergence properties. However, it suffers from major shortcomings that have been a cause for it not being implemented on large datasets. The most important among these are

(i) K-means is slow and scales poorly with respect to the time it takes for large number of points;
(ii) The algorithm might converge to a solution that is a local minimum of the objective function.

Much of the related work does not attempt to confront both the before mentioned issues directly. Because of these shortcomings, K-means is at times used as a hill climbing method rather than a complete clustering algorithm, where the centroids are initialized to the centers obtained from some other methods.

Before proceeding, we explicitly define the problem that we are looking at. In a traditional k-means algorithm (for which we develop an improved solution) each of the points is associated with only one partition also called the cluster. The number of partitions is pre-specified. Each of the partitions is recognized by a cluster center, which is the mean of all the points in the partition.

All the points in a partition are closer to its center than to any other cluster center. The accuracy of a clustering approach is defined by the distortion criterion, which is the mean squared distance of a point from its cluster center. Thus the objective is to get the clustering with minimum distortion and the specified number of clusters.

IV. Proposed Scheme

The proposed mining model is an extension of the work in. The proposed concept-based mining model consists of sentence-based concept analysis, document-based concept analysis, corpus-based concept-analysis, and concept-based similarity measure along with synonym based approach. 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 parts of speech. 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 mining model on sentence, document, and corpus levels. In this model, both the verb and the argument are considered as terms. One term can be an argument to more than one verb in the same sentence. This means that this term can have more than one semantic role in the same sentence. In such cases, this term plays important semantic roles that contribute to the meaning of the sentence. In the concept-based mining model, a labeled terms either word or phrase is considered as concept. 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 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 proposed mining model can be depicted as shown in figure 4.1.

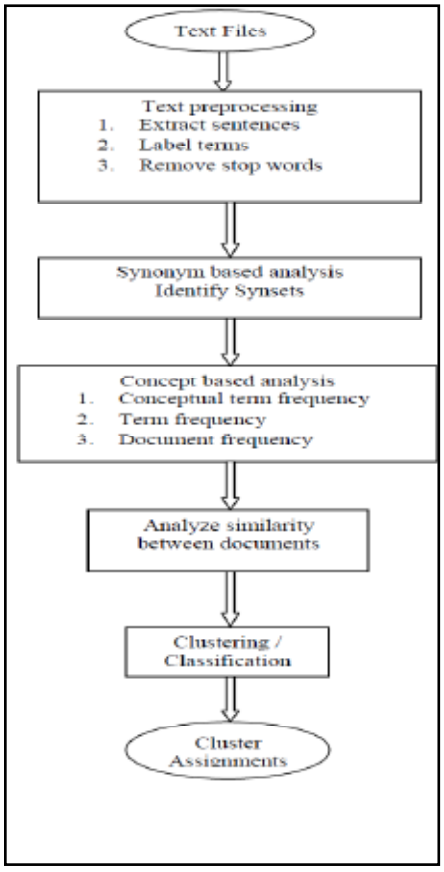

Fig. 4.1: Proposed Mining Model

Step wise process:

- Preprocessing of text.
- Identify the concepts.
- Calculating conceptual term frequency at.
  - Sentence level.
  - Document level.
  - Corpus level.
- Find Synsets for each concept.
- Identify significant concepts based on frequency.
- Cluster the documents.

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

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

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. thematic roles in a sentence automatically.

Gildea and Jurafsky [17] were the first to apply a statistical learning technique to the Frame Net database. The similarity between two features, *ft*and *fs*, is measured as the similarity between the class variable *C* distribution they induce: *P(C|ft)* and *P(C|fs)*. In the case of text classification, the similarity of two features is the similarity between their joint distributions with the category variable. For clustering this means that features with similar distributions over the classes (should) belong to the same cluster. Intuitively, if two different features have similar distributions over the classes, they will play a similar role in the classification process, and thus might as well be clustered together.

Two classifiers (NB and SVM) are trained on the training examples of each cluster and the testing examples are classified and assigned the label of the class of the cluster (all training examples in each cluster are supposed to have the same class label). The comparison of the method with distributional clustering (Baker & McCallum, 1998) and feature clustering on Reuters-21578 and 20NG shows prom. Text has always been the default way of storing information for hundreds of years, and mainly time, personal and cost constraints prohibit us from bringing texts into well structured formats (like data frames or tables).

V. Proposed Mining Model & Its Functions

The proposed concept-based mining model consists of sentence-based concept analysis, document-based concept analysis, corpus-based concept-analysis, and concept-based similarity measure along with synonym based approach. 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 parts of speech. The number of generated labeled verb argument structures is entirely dependent on the amount of information in the sentence.

Step wise process:

- Preprocessing of text.
- Identify the concepts.
- Calculating conceptual term frequency at
  - ➢ Sentence level.
  - ➢ Document level.
  - ➢ Corpus level.
- Find Synsets for each concept.
- Identify significant concepts based on frequency.
- Cluster the documents.

In the concept-based mining model, a labeled terms either word or phrase is considered as concept. 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.

Preprocessing of Text

In this module each document is read from the corpus. In each document, the sentences are separated. As the raw text data is unstructured data, we have to give a proper structure to each sentence. So each sentence is given a verb argument structure.

To get the verb argument structure, each word in the sentence is tagged with the parts of speech of that word in that sentence. Using these parts of speech for each term the verbs are identified. And then for each verb arguments are identified. These arguments are labeled as ARG0, ARG1, ARG2 etc. basing on the number of verbs for which the term is argument. For example if a term is argument for a single verb then the argument is ARG0. If it is argument for two verbs then the label is ARG1.

Another important technique in text mining is reducing the dimensionality of the text. That is we have to remove some unnecessary words. This can be done using standard stop lists. Each word is checked against the standard stop word list. If it is a stop word, then it is treated as insignificant word and it is removed from the process.

Identify the concepts

After completion of first step, we are remained with the labeled terms which are significant for matching that is to find the similarity. So each labeled term is treated as a concept.

Calculating conceptual term frequency

To analyze each concept at the sentence level, a new concept-based frequency measure, called the conceptual term frequency *ctf* is used. The ctf calculations of concept c in sentence s and document d are as follows:

➔ At sentence level:

The *ctf* is the number of occurrences of concept c in verb argument structures of sentence s. The concept c, which frequently appears in different verb argument structures of the same sentence s, has the principal role of contributing to the meaning of s. In this case, the *ctf* is a local measure on the sentence level. A concept c can have many ctf values in different sentences in the same document d. Thus, the ctf value of concept c in document d is calculated by

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

Where sn is the total number of sentences that contain concept c in document d. Taking the average of the ctf values of concept c in its sentences of document d measures the overall importance of concept c to the meaning of its sentences in document d. A concept, which has ctf values in most of the sentences in a document, has a major contribution to the meaning of its sentences that leads to discover the topic of the document. Thus, calculating the average of the ctf values measures the overall importance of each concept to the semantics of a document through the sentences.

➔ At document level: -

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

➔ At corpus level:

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

can discriminate their documents among others.

Find Synsets for each concept

Depending on the author's vocabulary to have the same semantics different words may be used. For example "intelligent" and "brilliant". Both are of same meaning but words are different. To identify theses type of words we have to find the synonyms for each concept. We have an efficient data base called WordNet which gives synonyms for words. The set of synonyms for a word is called synset. So by using word net database we can get synset for each concept. While finding the matching between documents words are being compared. When there is no exact word matching then the corresponding synsets will be checked for matching. So the original semantic are preserved.

Identify significant concepts based on frequency:

Based on the frequency at three levels, weightage will be given to each concept. The more significant concept will have more weight. The weights can be calculated as follows.

$$weight_i = (tf\ weight_i + ctf\ weight_i) * \log\left(\frac{N}{df_i}\right) \quad --(3)$$

$$tf\ weight_i = \frac{tf_{ij}}{\sqrt{\sum_{j=1}^{cn}(tf_{ij})^2}} \quad ----(4)$$

$$ctf\ weight_i = \frac{ctf_{ij}}{\sqrt{\sum_{j=1}^{cn}(ctf_{ij})^2}} \quad -----(5)$$

1. The total number of documents, N, in the corpus.
2. The *ctfij* of each concept ci in s for each document dj.
3. The *tfij* of each concept ci in each document dj.
4. The *dfi* of each concept ci.
5. *cn* is the total number of the concepts which has a term frequency value in document d.

Cluster Documents

To cluster the documents we need a clustering algorithm. K Nearest Neighbors algorithm is a most popular algorithm that can be used for any type of clustering process. In text clustering each text file is considered as a data point and similarity between two documents is treated as distance between two data points. So for each pair of documents in the corpus, similarity is calculated. A similarity matrix is built to represent the similarity between each pair of documents in the corpus. The similarity between two documents d1 and d2 is calculated using the formula.

$$sim_v(d_1, d_2) = \sum_{i=1}^{m}\left(\frac{l_{i_1}}{Lv_{i_1}}, \frac{l_{i_2}}{Lv_{i_2}}\right) \times weight_{i_1} \times weight_{i_2} \quad (2)$$

Where d1 and d2 two documents,
The number of matching concepts, m, in the verb argument structures in each document d,
The total number of sentences, sn, that contain matching concept ci in each document d,
The total number of the labeled verb argument structures, v, in each sentence s,
The length, *l*, of each concept in the verb argument structure in each document d,
The length, *Lv*, of each verb argument structure which contains a matched concept, and weighti can be calculated using equation (2).
The algorithm for finding the similarity is given below. This algorithm takes two documents as input. And it calculates the similarity between these documents using the formula given in equation (5).

VI. Results & Discussions

The experimental setup consisted of four data sets. The first data set contains 23,115 ACM abstract articles collected from the ACM digital library. The ACM articles are classified according to the ACM computing classification system into five main categories: general literature, hardware, computer systems organization, software, and data. The second data set has 12,902 documents from the Reuters 21,578 data set. There are 9,603 documents in the training set, 3,299 documents in the test set, and 8,676 documents are unused. Out of the five category sets, the topic category set contains 135 categories, but only 90 categories have at least one document in the training set. These 90 categories were used in the experiment. The third data set consisted of 361 samples from the Brown corpus. Each sample has 2;000þ words. The Brown corpus main categories used in the experiment were press: reportage; press: reviews, religion, skills and hobbies, popular lore, belles-letters, and learned; fiction: science; fiction: romance and humor. The fourth data set consists of 20,000 messages collected from 20 Usenet newsgroups.

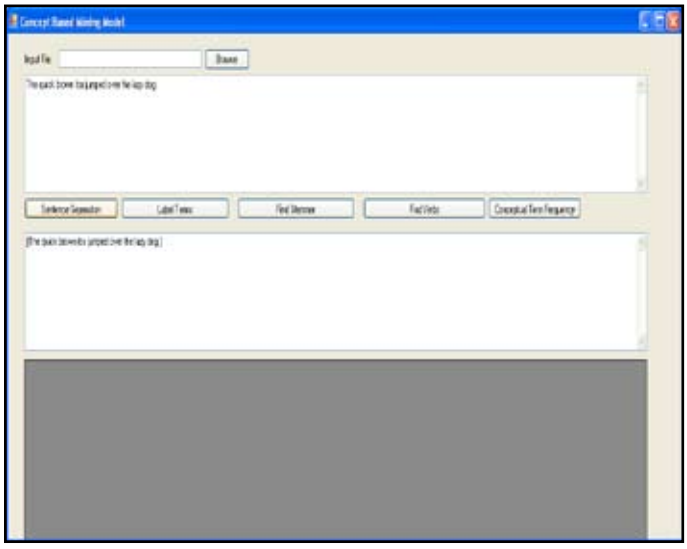

Fig 6.1: Senetence Seperation

Tokenization is the process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens. The list of tokens becomes input for further processing such as parsing or text mining. Tokenization is useful both in linguistics (where it is a form of text segmentation), and in computer science, where it forms part of lexical analysis.

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

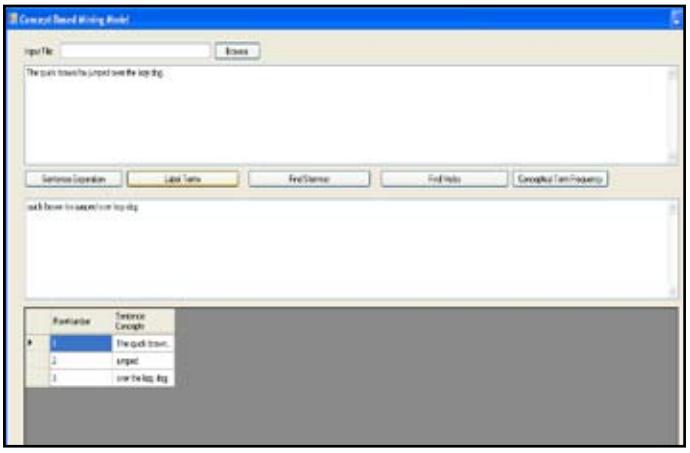

Fig. 6.2: Label Terms

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 parser. After running the semantic role labeler, each sentence in the document might have one or more labeled verb argument structures. In this model, both the verb and the argument are considered as terms. One term can be an argument to more than one verb in the same sentence.

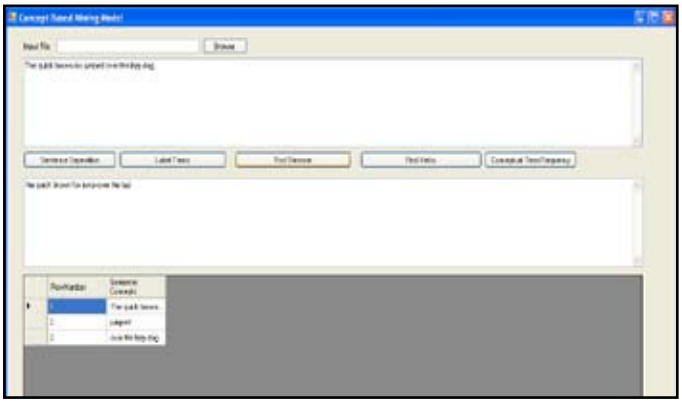

Fig. 6.3: Find Stemmer

In linguistic morphology, stemming is the process for reducing inflected (or sometimes derived) words to their stem, base or root form – generally a written word form. The stem need not be identical to the morphological root of the word; it is usually sufficient that related words map to the same stem, even if this stem is not in itself a valid root.

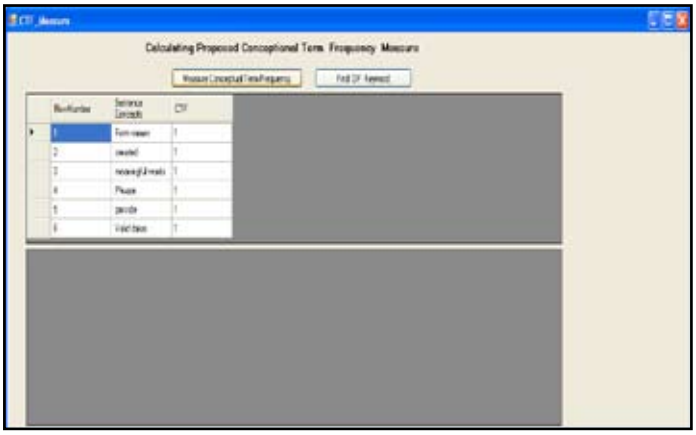

Fig. 6.4: Calculating proposed conceptional term frequency measure

The conceptual term frequency (ctf) is an important factor in calculating the concept-based similarity measure between documents. The more frequent the concept appears in the verb argument structures of a sentence in a document, the more conceptually similar the documents.

VII. Conclusions

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 omponent, 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] *K. Aas and L. Eikvil. Text categorisation: A survey technical report 941. Technical report, Norwegian Computing Center, June 1999.*

[2] *M. Collins. Head-Driven Statistical Model for Natural Language Parsing. PhD thesis, University of Pennsylvania, 1999.*

[3] *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.*

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

[5] *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.*

[6] *D. Gildea and D. Jurafsky. Automatic labeling of semantic roles. Computational Linguistics, 28(3):245{288, 2002.*

[7] *T. Joachims. Text categorization with support vector machines: learning with many relevant features. In C. N¶edellec and C. Rouveirol, editors, Proceedings of ECML-98, 10th European Conference on Machine Learning, number 1398, pages 137{142, Chemnitz, DE, 1998. Springer Verlag, Heidelberg, DE.*

[8] *D. Jurafsky and J. H. Martin. Speech and Language Processing. Prentice Hall Inc., 2000.*

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

[10] *M. F. Porter. An algorithm for su±x stripping. Program, 14(3):130{137, July 1980.*
[11] *S. Pradhan, K. Hacioglu, V. Krugler, W. Ward, J. H. Martin, and D. Jurafsky. Support vector learning for semantic argument classi¯cation. Machine Learning, 60(1-3):11{39, 2005.*
[12] *S. Pradhan, K. Hacioglu, W. Ward, J. H. Martin, and D. Jurafsky. Semantic role parsing: Adding semantic structure to unstructured text. In Proceedings of the 3th IEEE International Conference on Data Mining (ICDM), pages 629{632, 2003.*
[18] *Anna Nick Wreden, Communications Week Interactive, February 17, 1997.*
[19] *Arsitk Karen Watterson, Datamining poised to go mainstream October 1999.*
[20] *Barbaros A.,Information and Privacy Commissioner/ Ontario, Data Mining: Staking a Claim on Your Privacy, January 1998.*