Concept Based Labeling of Text Documents Using Support Vector Machine

K.Nithya,
Assistant Professor,
Department of CSE,
Nandha college of Technolgy,
knithya89@gmail.com.

M.Saranya,
Assistant Professor,
Department of CSE,
Nandha college of Technolgy,
Saranyamcse88@gmail.com

C.R.Dhivyaa,
Assistant Professor,
Department of CSE,
Nandha college of Technolgy,
crdhivyait@gmail.com

Abstract- Classification plays a vital role in many information management and retrieval tasks. Text classification uses labeled training data to learn the classification system and then automatically classifies the remaining text using the learned system. Classification follows various techniques such as text processing, feature extraction, feature vector construction and final classification. The proposed mining model consists of sentence-based concept analysis, document-based concept analysis, corpus-based concept-analysis, and concept-based similarity measure. 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 similarity measure. Then we analyze the term that contributes to the sentence semantics on the sentence, document, and corpus levels rather than the traditional analysis of the document only. With the extracted feature vector for each new document, Support Vector Machine (SVM) algorithm is applied for document classification. The approach enhances the text classification accuracy.

Index Terms— *Concept analysis, supervised, support vector machine, text classification.*

*****

1. Introduction

Natural language processing (**NLP**) is a field of computer science and linguistics concerned with the interactions between computers and natural languages. The need for representations of human knowledge of the world is required in order to understand human language with computers. Text Mining attempts to discover new previously unknown information by applying the techniques such as text categorization from natural language processing and data mining.

Text categorization (TC) is a supervised learning problem where the task is to assign a given text document to one or more predefined categories. It is a well-studied problem and still continues to be topical area in information retrieval (IR), because of the ever increasing amount of easily accessible digital documents on the Web, and, the necessity for organized and effective retrieval. Vector Space Model (VSM) has been used as the recent technique for text document categorization [3, 4, 8]. It represents each document as a feature vector of the terms (word or phrases) in the document. Each feature vector contains term weights of the terms in the document. In this case the high dimensionality of feature space is a major problem in TC. The number of terms present in a collection of documents, in general, is large and few are informative. Feature selection for TC helps in reducing dimensionality of feature space by identifying informative features and its primary goals are improving classification effectiveness, computational efficiency, or both. In this paper, concept based text document categorization is used with the extracted features.

The rest of the paper is structured as follows: Section 2 discusses related work that describes the various classification techniques. Our proposed work with classification algorithm (Support Vector Machine) is discussed in Section 3. Performance measures are explained in the Section 4. Finally Section 5 concludes the paper.

2. Related Works

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. In this approach quality of clustering is low.

Dhillon et al (2003) has proposed an information-theoretic feature clustering algorithm, termed as divisive clustering, and applied it to text classification. It is found that the algorithm has many good qualities. It has optimized over all clusters simultaneously. An experiment using the SVM classifiers on the 20 News groups data set has shown that the divisive clustering improves classification accuracy especially at lower number of features.

Caropreso et al (2001) has experimented with n-grams for text categorization on the Reuters dataset. It has been defined that n-gram is an alphabetically ordered sequence of n stems of consecutive words in a sentence (after stop words were removed). The authors have used both unigrams and bigrams as document features. They have extracted the top-scored features using various feature selection methods including Mutual In- formation. Their results have indicated that bigrams can better predict categories than unigrams. However, despite the fact that bigrams represent the majority of the top-scored features, the use of bigrams does not yield significant improvement of the categorization results while using the Rocchio classifier. Specifically, in 20 of the 48 reported experiments a certain increase in the accuracy is observed, while in 28 others the accuracy decreases, sometimes quite sharply.

Koster and Seutter (2003) have used Rocchio and Winnow classifiers on an EPO1A dataset. The feature induction method that they have used involves combination of single words and word pairs. It is shown that when using pairs without BOW, the results of both classifiers decrease, while when using both pairs and BOW, the results are marginally above the BOW baseline.

Scott and Matwin (1999) have applied a rule-based RIPPER classifier on Reuters and DigiTrad datasets, using document representation based on phrases. By phrases the authors has meant Noun Phrases and Key Phrases. It is identified that the rule-based classifier could benefit from the semantic power of a highly meaningful phrase. However, the results that are achieved by either scheme are roughly the same as their baseline with BOW representation. While combining the different representations with the BOW, the authors are able to improve their results, but still the maximum that they achieve is 85% of accuracy on Reuters, whereas the state-of-the-art result is close to 89%.

Diederich et al (2003) has applied SVM on two text representations: BOW and a bag of all the functional words and bigrams of functional words in the text. The functional words mentioned are the parts of speech excluding nouns, verbs and adjectives. This document representation is supposed to preserve the style while suppressing the topic. The results have shown that the simple-minded BOW performs the sophisticated representation based on unigrams and bigrams of functional words.

3. Datamining Model

The proposed model is to achieve highly consistent result by applying a classification algorithm. Fig. 1 depicts the conceptual diagram of our model.

3.1. Preprocessing

Datasets that are chosen for this work are from Reuters 21578. In preprocessing the terms which appear too often (in every or almost every document) and thus support no information for the task are removed. Good examples for this kind of words are prepositions, articles and verbs. Stemming is a technique that can be applied for the reduction of words into their root. E.g. agreed, agreeing, disagree, agreement and disagreement can be reduced to their base form or stem agree. In this paper, Porter Stemming algorithm is used. The idea of this algorithm is the removal of all suffixes to get the root form. The main fields of application for the Porter Stemmer are languages with simple inflections, such as English. The further processing of the suffix stripping is decided by several conditions [5].

The other conditions for the Porter Stemming are:

1. *S - the stem ends with S (and similarly for the other letters).
2. *v* - the stem contains a vowel.
3. *d - the stem ends with a double consonant (e.g. -TT, -SS).
4. *o - the stem ends cvc, where the second c is not W, X or Y (e.g. -WIL, -HOP).

This rule has been applied to remove the longest matching suffix.

Table 1. Document preprocessing

Processing	Processed terms	Unique terms
Before	703818	58059
After	703818	28646

3.2. Natural Language Processing

After preprocessing, the recognition of the elements of a sentence like nouns, verbs, adjectives, prepositions, etc. is done through part of speech tagging (POS tagging). Each sentence in the document is labeled automatically based on the PropBank notations [9]. After running the semantic role labeler [9], 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 and the output of the role labeling task are captured and analyzed by the concept-based model on the sentence and document levels.

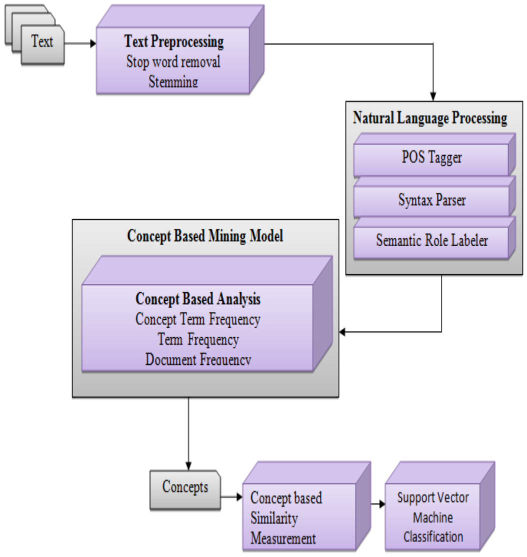

Figure 1. Conceptual diagram

3.3. Concept-Based Mining Model

The concept-based mining model consists of sentence-based concept analysis, document-based concept analysis, corpus-based concept-analysis, and concept-based similarity measure. In this model, both verb and argument are considered as terms. There are two cases. In the first case 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. In the second case 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.

3.3.1 Corpus-Based Concept Analysis

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.

3.4. Concept-Based Similarity Measure

A concept-based similarity measure, based on matching concepts at the sentence, document, corpus and combined approach rather than on individual terms (words) only, is devised. The concept-based similarity measure 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 concept is used to measure 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 discriminate among documents in calculating the similarity.

These aspects are measured by the proposed concept-based similarity measure which measures the importance of each concept at the sentence level by the ctf measure, document level by the tf measure, and corpus level by the df measure. The concept-based measure exploits the information extracted from the concept-based analysis algorithm to better judge the similarity between the documents.

3.5. Support Vector Machine

Support vector machine (SVM) is a binary classifier that performs classification task by constructing hyper planes in multidimensional space that separates cases of different class labels. To do a multi-class classification pair wise classification can be used. It classifies the text documents under the correct category based on the class labels. Multi-class classification of text documents are shown in the Fig. 2

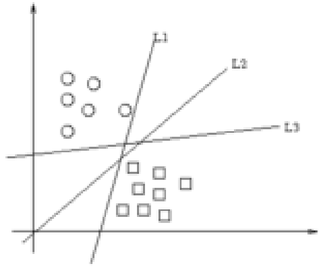

Figure.2 Multi-class classification using SVM

4. Performance Measures

Performance of the classifier is measured with the help of F-Measure.

4.1. F-Measure

F-measure is the combination of Precision and Recall. Precision is the percentage of retrieved documents that are in fact relevant to the query (i.e., “correct” responses).Recall is the percentage of documents that are relevant to the query and were, in fact, retrieved. The precision P and recall R of a cluster j with respect to class i are defined as,

$$P = \text{Precision}(i,j) = Mij / Mj \qquad (1)$$

$$R = \text{Recall}(i,j) = Mij / Mi \qquad (2)$$

Where Mij is the number of class i in cluster j, Mj is the number of cluster j, and Mi is the number of members of class i.

$$F(i) = \frac{2PR}{P + R} \qquad (3)$$

The overall F-measure for the clustering result C is the weighted average of the F-measure for each class i,

$$F_C = \frac{\sum_i (|\, i \,| \times F(\, i\,))}{\sum_i |\, i \,|} \qquad (4)$$

5. Conclusion

The proposed work on perception based Labeling of Text Documents proves to be an effective as well as an efficient method. By exploiting the semantic structure of the sentences in documents, a better text categorization result is achieved. Further accuracy is improved by employing Support Vector Machine classification. The proposed unsupervised learning method is compared with supervised learning and ensured that the F-measure is maximum to achieve high-quality clustering.

6. References

[1] Al-Mubaid H. and Umair S.A. (2006) “A New Text Categorization Technique Using Distributional Clustering and Learning Logic”, IEEE Transactions on Knowledge and Data Engineering , vol. 18, no. 9

[2] Ajoudanian S. and Jazi D.M. (2009) “Deep Web Content Mining”, World Academy of Science, Engineering and Technology 49.

[3] G. Salton, A. Wong, and C.S. Yang, “A Vector Space Model for Automatic Indexing”, Comm. ACM, vol. 18, no. 11, pp. 112-117, 1975.

[4] G. Salton and M.J. McGill, “Introduction to Modern Information Retrieval”, McGraw-Hill, 1983.

[5] Harb B., Drineas P., Dasgupta A., Mahoney W.M., and Josifovski V. (2007) “Feature Selection Methods for Text Classification”, SanJose, California, USA.

[6] Joachim T. (2002) “Learning to Classify Text Using Support Vector Machines”, Methods Theory and Algorithms, Kluwer/Springer.

[7] Kim H., Howland P., and Park H. (2005) “Dimension Reduction in Text Classification with Support Vector Machines”, Journal of Machine Learning Research 6 pp.37-53

[8] K. Aas and L. Eikvil. “Text categorisation: A survey technical report”, Technical report, Norwegian Computing Center, June 1999.

[9] P. Kingsbury and M. Palmer. “Propbank: the next level of Treebank”, In Proceedings of Treebanks andLexical Theories, 2003.

[10] Nahm Y.U and Mooney J.R. (2001) “A Mutually Beneficial Integration of Data Mining and Information Extraction”, University of Texas, Austin, TX 78712-1188.

[11] Sebastiani F. (2002) “Machine Learning in Automated Text Categorization”, ACM Computing Surveys, vol. 34, no. 1, pp. 1-47.

[12] Sheata S., Karray F., and Kamel M. (2010) “An Efficient Concept Based Mining Model for Enhancing Text Clustering”, Proceedings of Sixth IEEE International Conference Data Mining, vol. 22, no.10.

[13] Sheata S., Karray F., and Kamel M. (2006) “Enhancing Text Clustering Using Concept Based Mining Model”, Proceedings of Sixth International Conference. Data Mining, 0-7695-2701-9.

[14] Steinbach M., Karypis G., and Kumar V. (2000) “A Comparison of Document Clustering Techniques”, Proceedings of Knowledge Discovery and Data Mining Workshop Text Mining.