Research Inventy: International Journal of Engineering And Science
Vol.8, Issue 2 (May 2018), PP -39-43
Issn (e): 2278-4721, Issn (p):2319-6483, www.researchinventy.com

Design Approach Of Data Mining Model For Text Classification

B.Vamshi Asst.Professor Dr.Kvgrao,Prof.
[12]cse Dept,Gnits
Corresponding Author: B.Vamshi

Abstract- Classification plays a vital role in many information management and retrieval tasks. This paper studies classification of text document. Text classification is a supervised technique that uses labeled training data to learn the classification system and then automatically classifies the remaining text using the learned system. In this paper, we propose a mining model consists of sentence-based concept analysis, document-based concept analysis, and corpus-based concept-analysis. 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. After extracting feature vector for each new document, feature selection is performed. It is then followed by K-Nearest Neighbour classification. The approach enhances the text classification accuracy. We define architecture that supports processing of a range of declaratively-specified KDT problems, using such general services.

Keywords: Concept Analysis, Feature Selection, Feature Vector, K-Nearest Neighbor, Supervised, Text Classification, KDT

DATE OF SUBMISSION: 31-05-2018 DATE OF ACCEPTANCE: 15-06-2018

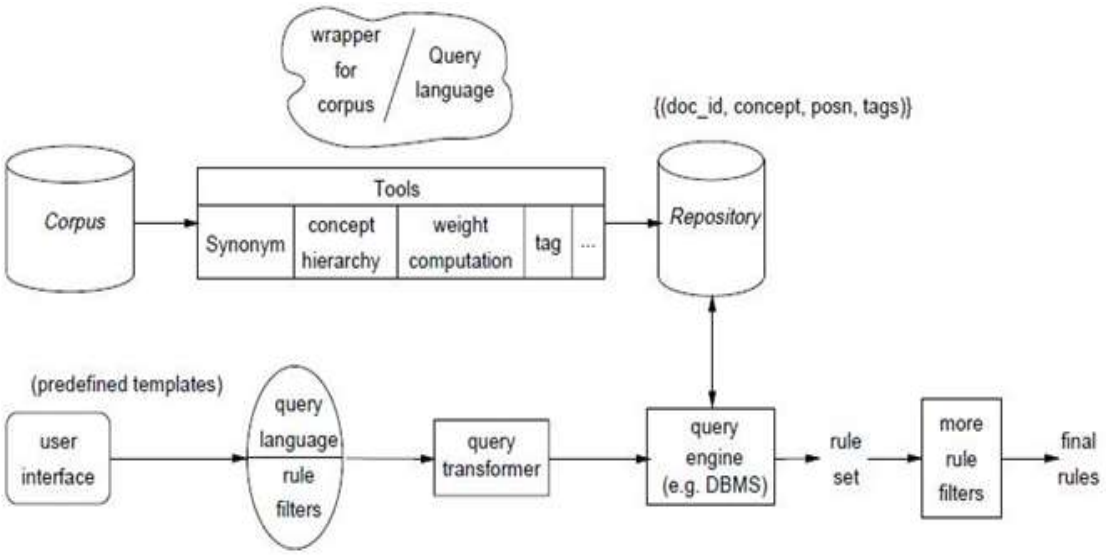
I. Introduction

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 existing techniques. Our proposed with classification algorithm (K-Nearest Neighbour) is discussed in Section 3. Performance measures are explained in the Section 4. Finally Section 5concludes the paper.

II. Data Mining On Text

Data mining technology is giving us the ability to extract meaningful patterns from large quantities of structured data. Information retrieval systems have made large quantities of textual data available. Extracting meaningful patterns from this data is difficult. Current tools for mining structured data are inappropriate for free text. We outline problems involved in Knowledge Discovery in Text, and present architecture for extracting patterns that hold across multiple documents. The capabilities that such a system could provide are illustrated. Data mining technology has created a new opportunity for exploiting the information in databases. Much of the success has been in support of marketing. Patterns in the data, such as associations among similar items purchases, enables targeting marketing to focus on what customers are likely to purchase. Current data mining solutions are highly optimized for a single pattern specification. This represents Knowledge Discovery in Text, as the variability in information in a text bases means we will need flexibility in defining the type of pattern that interests us. We want to see industry develop a base that provides general services for data mining; this will ease the development of flexible KDT (Knowledge Discovery in Text) systems. We define architecture (Figure 1) that supports processing of a range of declaratively-specified KDT problems, using such general services.

Figure 1. Text Mining System Architecture



This two-part architecture connects information retrieval systems to a generalized data mining" engine. This approach limits the need to develop KDT specific solutions. In addition, advances in both information retrieval and data mining technology will directly result in improvements in KDT capabilities.

The upper half of Figure 1 is based on information retrieval technology. Many of the tools, such as synonym matching and tagging, are already used as part of information retrieval products. As these tools improve in their support of information retrieval, they will also provide additional capabilities for KDT. The Repository may either be virtual or

may be precomputed. Some information retrieval systems already incorporate Information Extraction tools and precompute and store a repository that meets many of our requirements.

The lower half of the figure deals with the pattern detection process. Here there are significant differences from existing structured data mining tools. We view pattern generation as answering a specialized query. Existing data mining systems use a specific algorithm to find a specific type of pattern; we believe that text analysis demands greater flexibility. Instead we view the algorithmic issues as a query optimization problem. This allows us greater flexibility in the types of analysis performed. This \query optimization" approach is not limited to Knowledge Discovery in Text; such a system would support integration of data warehouses and data mining, thus we expect vendors will support a data mining approach based on query optimization.

III. Data Mining Model

The proposed model is to achieve highly consistent result by applying a classification algorithm. Figure 3.1 depicts the conceptual diagram of our model. A. Preprocessing Datasets that are chosen for this work are from Reuters 21578.In preprocessing the terms which appear too often 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.

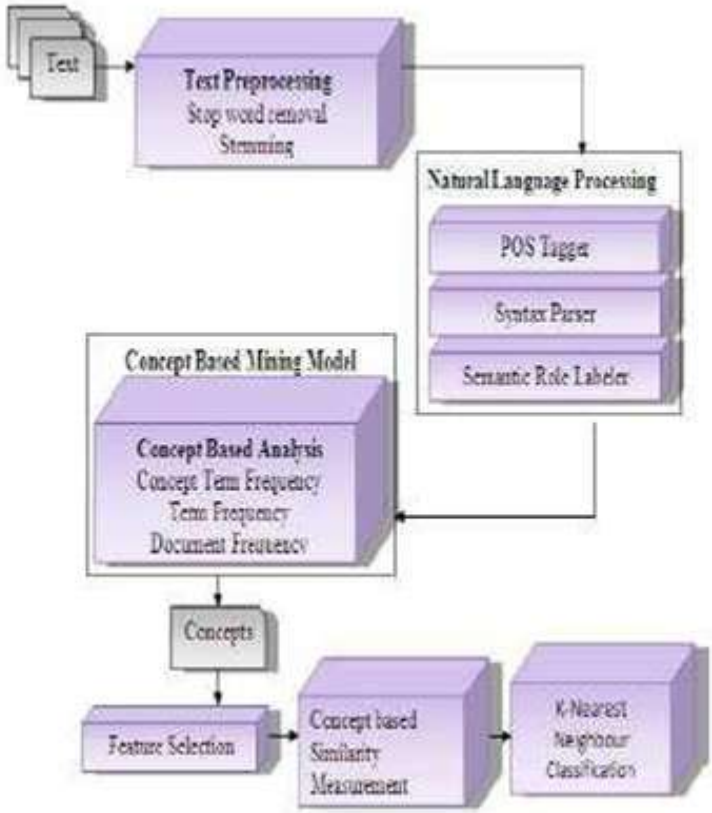
TABLE 1. DOCUMENT PREPROCESSING

Reuters 21578	Processed terms	Unique terms
Before Preprocessing	703818	58059
After Preprocessing	703818	28646

B. 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).

Figure 2 Conceptual Diagram



Each sentence in the document is labeled automatically based on the PropBank notationsAfter 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.

C. 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. 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.

D. Feature Selection

Feature selection is performed on extracted concepts to improve efficiency and accuracy of text categorization algorithms. It is the process of selecting a subset of the terms occurring in the training set and using only this subset as features in text classification. In the selection process, each feature (term or single word) is assigned with a score according to a score-computing function. The mathematical definitions of these score computing functions are often defined by some probabilities which are estimated by some statistic information in the documents across different categories. For the convenience of description, we give some notations of these probabilities below.

(i).Information Gain (IG)

Here both the class membership and the presence/absence of a particular term are seen as random variables, and one computes how much information about the class membership is gained by knowing the presence/absence statistics. Indeed, if the class membership is interpreted as a random variable C with two values, positive and negative, and a word is likewise seen as a random variable T with two values, present and absent, then using the information-theoretic definition of mutual information, Information Gain is defined as:

$IG(t)H(C) \square H(C \square T) \square \sum \square , cP(C \square c, T \square \square$
$In[P(C \square c, t \square \square) / P(C \square c)P(T \square \square)]$
here, ranges over {present, absent} and c ranges over {c+, c–}.

E. 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.

F. K-Nearest Neighbour

K-nearest neighbor algorithm is one of the machine learning algorithm. It simply assigns the property value for the object to be the average of the values of its *k*- nearest neighbors. It can be useful to weight the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. (A common weighting scheme is to give each neighbor a weight of 1/*d*, where *d* is the distance to the neighbor. This scheme is a generalization of linear interpolation.) The best choice of *k* depends upon the data; generally, larger values of *k* reduce the effect of noise on the classification, but make boundaries between classes less distinct. A good *k* can be selected by various heuristic techniques, for example, cross validation. The special case where the class is predicted to be the class of the closest training sample (i.e. when $k = 1$) is called the nearest neighbor algorithm. The accuracy of the *k*-NN algorithm can be improved by various feature selection techniques.

IV. Performance Measures

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

A) 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 = Precision(i,j) = Mij / Mj (1)

R = Recall(i,j) = Mij / Mi	(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.

$F(i) = 2PR / (P + R)$

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

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

V. Conclusion

The proposed work on concept based mining model for text classification 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 K-Nearest Neighbour classification.

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.

[5]. Joachim T. (2002) 'Learning to Classify Text Using Support Vector Machines', Methods Theory and Algorithms, Kluwer/Springer.

[6]. 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

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

[8]. P. Kingsbury and M. Palmer. 'Propbank: the next level of treebank', In Proceedings of Treebanks andLexical Theories, 2003.

[9]. 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.

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

[11]. 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.

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

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

B.Vamshi "Design Approach of Data Mining Model for Text Classification" Research Inventy: International Journal of Engineering And Science, vol. 08, no. 02, 2018, pp. 39–43.