IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 3, Ver. II (May-Jun. 2014), PP 90-95
www.iosrjournals.org

Self Appreciating Concept Based Model For Cross Domain Document Classification

Dipak A. Sutar[1]
[1](*Department of CSE, AMGOI Vathar, Shivaji University, MS, India)*

Abstract : *In text mining, text categorization is an important technique for classifying the documents. Most of the times statistical approaches that are based on analysis of the term in the form of frequency of the term, that is the number of occurrences of one or more words in the document are used for classification. Even statistical analysis indicates the importance of the term, but it is hard to analyze when multiple terms have the same frequency value, but one term is more important in terms of meaning than the other. Also, there are a wide variety of documents being generated that belongs to different domains which differ in formats, writing styles, etc. These domains can be news articles, e-mails, online chats, blogs, wiki articles, twitter posts, message forums, speech transcripts, etc. Often a classification method that works well in one domain does not work as well in another. The proposed system tries to implement a concept based text classification model that classifies the cross-domain text data based on the semantics or theme of the text data. Also the proposed approach makes the training system stronger and stronger at all possible positive tests of the categorizer. This system is called as a Self Appreciating Concept Based Classifier (SACBC).*

Keywords: *Document Classification, Concept Mining, Cross Domain Document classification,*

I. INTRODUCTION

Data mining is the process of analyzing and extracting knowledge by applying machine learning techniques to stored data. After internet and data warehousing, data mining is the third hottest field in the digital world. The objective of data mining is to transform extracted knowledge into a human understandable structure.

The term text mining is used when mining process is applied to the text data. It is the process of gathering high quality information from text. Text mining includes different tasks such as text categorization (text classification), text clustering, concept/entity extraction, document summarization and sentiment analysis.

The task of text categorization is to assign one or more predefined categories to a text document. It is a supervised learning problem. Generally, statistical approaches are used for a single domain text categorization. The statistical approaches are based on analysis of the term in the form of frequency of the term which is the number of occurrences of one or more words in the document. The statistical approach shows the importance of the term, but the problem with it is when multiple terms have the same frequency, but one term is more important in terms of more meaning than the other [1] [5]. That is the contribution of the one term is more than the other terms even though they has the same frequency. For better classification, a good classifier should identify such most contributable terms of the document. Thus text mining needs to consider not only statistical analysis but also semantic analysis of the text. By using concept mining we can identify most contributable term in terms of meaning from the document [1]. These terms present the concepts which capture semantics of the text. Finally, this results in the discovery of the topic of the document.

The development in the digital world and the dynamic behavior of the web produces a wide variety of text documents that is a result of documents being generated in various domains. These domains can be news articles, e-mails, online chats, blogs, wiki articles, twitter posts, message forums, speech transcripts, etc. [2] [4]. The domains represent data information in various ways, each serving a particular purpose. Often a classification scheme that works well in one domain does not work as well in another.

Here in the proposed approach, Concept Based Classifier (CBC) tries to identify concepts which represent the theme of the document by identifying terms of a document with the help of the semantic role parsing [7]. Shallow semantic parsing is the technique of labeling the terms of a sentence with semantic roles [7] [8]. By doing this the contribution of each term, so called the concept could be analyzed, which will enhance the quality of classification. The CBC consists of two stages, the training and testing. On the training part of the system, the performance of the classifier depends. The proposed approach makes the training system stronger and stronger at all possible positive tests of the categorizer. This new system of concept based classifier is named as a Self Appreciating Concept Based Classifier (SACBC) [1].

II. RELATED WORK

Concept mining is the process of extracting concepts from the text. It is the combination of artificial intelligence techniques and natural language processing. Concept based mining is very rapidly growing area.

Daniel Gildea and Daniel Jurafsky presented a system that identifies semantic roles or relationships [8] [6]. For an input sentence and a target word, the system assigns the labels such as Agent or Patient.

In an enhanced text clustering using a concept based mining [5], the classifier finds the most meaningful term in each sentence using the semantic structure of each term. The concept is analyzed at sentence, document and corpus levels. Prop Bank [6] presents the algorithm of semantic annotation. It is very helpful in the process of assignment of roles. Shallow semantic parsing [7] presents the technique of assigning the structure in the form of semantic markup to the text.

Hele-Mai Haav and Tanel-Lauri Lubi surveyed different concept based information retrieval techniques and software tools [9].

In [2], the different cross domain text categorization techniques and their comparison are given.

III. CONCEPT ANALYSIS

The goal of concept analysis is to identify the concepts in the document which contribute to the semantics of the document. The concept is a term or phrase which represents the theme or idea that is being hidden behind the text data. The concepts are analyzed with the help of semantic role Labeler. The semantic role Labeler assigns the semantic tags to the terms in the sentence. Each sentence might have one or more verb argument structures. The verb argument structure of a sentence includes verb associated with its arguments [8] [5]. For example: "*Sachin hits the ball"*. Here "*hits*" is the verb. "*Sachin*" and "*the ball*" are the arguments associated with the verb "hits". The sentence that has multiple verb argument structures includes multiple verbs associated with their arguments.

IV. SYSTEM ARCHITECTURE

4.1 Concept-Based Classifier (CBC)

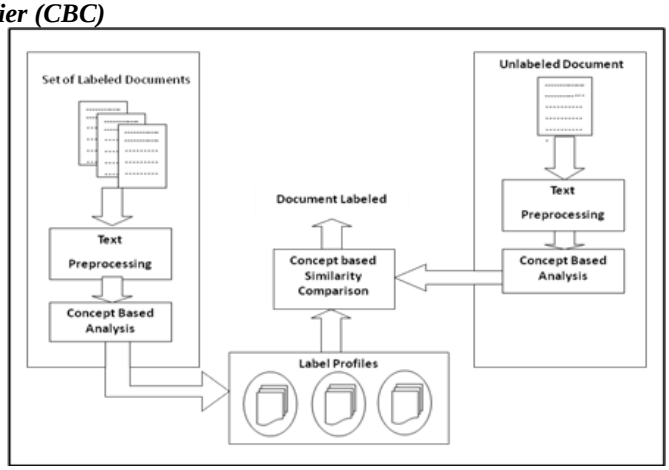

Fig. 1. Concept Based Classifier (CBC)

The system consists of different blocks, first is an input block which will supply input to the system. The input is tokenized into sentences and it is given further to semantic role Labeler (SRL). The SRL assigns semantic roles to the terms according to PropBank notations [6]. After labeling, preprocessing is done on the data. Preprocessing is done because there are terms which occurs too often, but carries no information for the classification task. Such terms need to be removed. Here, data preprocessing block consists of two processes 1) Stop words removal 2) Stemming.

Then the data are analyzed at different levels viz. Sentence, document and corpus levels. The category profile is created for each category using a concept based analysis algorithm.

The similar process is applied to test the document. Then the similarity between category profile and test document profile is calculated using a concept based similarity algorithm. After that, the appropriate category label is assigned to the test document using the similarity value.

4.1.1 Sentence-based Concept Analysis

To analyze concepts at sentence level the conceptual term frequency (*ctf*) measure is used. The ctf value in a sentence is number of occurrences of the concept 'c' in verb argument structures of sentence 's'.

The *ctf* value of concept 'c' in document 'd' is calculated by:

$$ctf = \frac{\sum_{n=1}^{sn} ctf_n}{sn} \tag{1}$$

Where, S_n is the total number of sentences that contain concept c in document d. This average of *ctf* values of concept 'c' measures the overall contribution of the concept to the semantics of the document 'd' through the sentences.

4.1.2 Document-based Concept Analysis

To analyze each concept at the document level, the concept based term frequency *tf* is used. It is the number of occurrences of a concept c in the original document.

4.1.3 Corpus-based Concept Analysis

The concept based document frequency *df* is used for corpus based analysis, which measures the number of documents containing concept c. The *df* measure is used to extract concepts that appear only in a small number of documents, hence can discriminate between documents from one another.

4.1.4 Corpus-based Analysis Algorithm

The *tf*, *ctf* and *df* values are calculated using a concept based analysis algorithm.

1. d_{doci} is a new Document
2. L is a matched concept list (Initially empty)
3. s_{doci} is a new sentence in d_{doci}
4. Build concepts list C_{doci} from s_{doci}
5. for each concept c_i € C_i do
6. compute ctf_i of c_i in d_{doci}
7. compute tf_i of c_i in d_{doci}
8. compute df_i of c_i in d_{doci}
9. d_k is seen document, where $k = \{0; 1; \ldots; doci - 1\}$
10. s_k is a sentence in d_k
11. Build concepts list C_k from s_k
12. for each concept c_j € C_k do
13. if ($c_i == c_j$) then
14. update df_i of c_i
15. compute $ctfweight = \text{avg}(ctf_i , ctf_j)$
16. add new concept matches to list L
17. end if
18. end for
19. end for
20. output the matched concepts list L
21. End.

The concept based analysis algorithm takes a document in which the sentences are labeled semantically by semantic role labeler (SRL) tool. Then the sentences of the current documents are processed sequentially. Each concept in the sentence is matched with the concepts in the already processed documents. The *ctf ,tf I df* values are updated if the concepts are matched. Finally the concept list L contains all the matched concepts and its corresponding measures.

4.1.5 Training the Classifier

The algorithm is used for training the data and uses concept based analysis algorithm for category profile creation.

Input: A set of labeled documents.
Output: Concept based profile Pi for each category i.
1. For each document call Concept based analysis algorithm.
2. Build common concepts list, i.e., profile Pi for each category with corresponding *ctf* measure based on the Li of each document.
3. End.

4.1.6 Testing the Classifier

The algorithm is used for testing the data and uses concept based similarity algorithm for assigning category to test documents.

Input: An unlabeled document.
Outcome: Assignment of class label to the unlabeled document.
1. For the test document call Concept based analysis algorithm and prepare Li for the current test document.
2. Do similarity measure between Li and each of the profiles Pi prepared (during training) by calling *Procedure_Similarity_Measure*.
3. Fix-up the label based on the similarity measure values.
4. End

4.1.7 Concept-Based Similarity Measure
Procedure_Similarity_Measure:
Input: Test document D and a label profile P
Output: Similarity measure value.
1) Do similarity Measure between the D & P by:

$$\cos\_sim(D,P) = \frac{dot(D,P)}{|D||P|} \tag{2}$$

Where,

$$dot(D,P) = \sum_{i=1}^{v} d_i . p_i \tag{3}$$

$$|D| = \sqrt{\sum_{i=1}^{v} d_i^{\,2}} \tag{4}$$

$$|P| = \sqrt{\sum_{i=1}^{v} p_i^{\,2}} \tag{5}$$

2) End.

4.2 Self-Appreciating Concept-Based Classifier (SACBC)

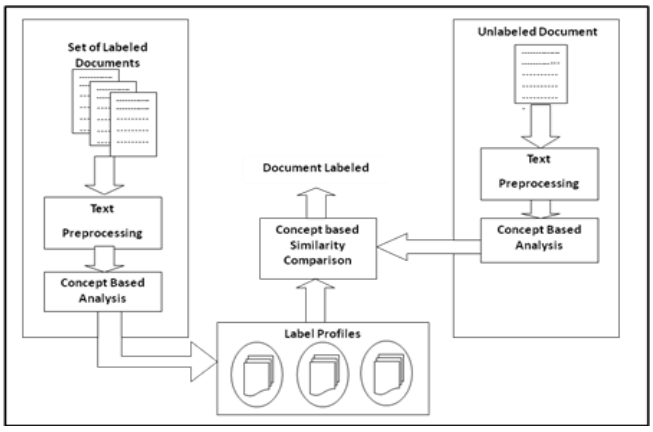

Fig. 2. Self-Appreciating Concept Based Classifier (SACBC)

Mostly the training is mostly done with the labeled documents which are expensive to collect as an expert (human like) is needed that takes much effort and time for labeling. Here an alternate approach to reduce the resource utilization is used. Whenever the test result is true positive, the profile of this document can be added up with the existing profile of the particular category. Thus the prepared profile is strengthened throughout the lifetime of the system.

The algorithm of SACBC is given below,
Input: Concept list of test document with positive result.

Output: Updating the concepts of the Corresponding label profile (Let Each Ci is a concept in the test document and Each Cj is a concept in the label profile Li)
1) Repeat the following steps for each concept Ci of
the test document
a) If (*Ci* == *Cj*) then
i) Update *ctfi* of Ci
ii) ctfweight = avg(*ctfi*,*ctfj*)
iii) Update existing concepts or add a new concept matches to Li.
b) End If
2) End

V. EXPERIMENTS AND RESULTS

5.1 Dataset

The dataset used for experimentation contains 1000 documents of different categories and domains. The dataset has four categories,

a) Artificial Intelligence
b) Operating Systems
c) Databases
d) Computer Networks.

The dataset contains the documents of the following domains,

a) IEEE Papers (PDF Files)
b) Wiki Articles
c) Web Blogs
d) Mails

5.2 Results

The results are obtained by taking different training and testing sets. The result is measured in terms of precision, recall and F1 score.

TABLE 1. Testing of 300 documents and training set of 500 documents

Class	Precision	Recall	F1
Artificial Intelligence	**0.66**	**0.86**	**0.746**
Operating Systems	**0.70**	**0.79**	**0.742**
Computer Networks	**0.71**	**0.6**	**0.65**
Databases	**0.75**	**0.58**	**0.654**

5.3 Comparison of CBC and SACBC

The comparison of concept based classifier (CBC) and self appreciating concept based classifier (SACBC) is done in terms of accuracy.

TABLE 2. Accuracy for test documents

Sr. No	No. of test Documents	No. of training documents	Accuracy(CBC)	Accuracy(SACBC)
1	**300**	**500**	**70.56**	**73.41**
2	**200**	**800**	**78.7**	**80.2**

The Accuracy of the system is calculated for test data containing only small size files and it is compared with the accuracy of test data containing an average file size.

TABLE 3. Accuracy of the system after varying size of test documents

Sr. No.	Size of the documents	Accuracy(CBC)	Accuracy(SACBC)
1	**Average size (pdf up to 400 KB, text up to 30 KB)**	**78.7**	**80.2**
2	**Small size (1,2,3 KB)**	**65.4**	**65.4**

The Accuracy of the system is calculated for test data containing a less number of documents that is a small testing dataset. The accuracy of CBC and SACBC is compared.

TABLE 4. Accuracy of the system after varying Number of test documents

Sr. No.	Number of documents	Accuracy(CBC)	Accuracy(SACBC)
1	**50**	**73.9**	**73.9**
2	**75**	**76.2**	**76.2**

VI. CONCLUSION

From the results it is concluded that, we can achieve accuracy up to 80% of the system which is quite appreciable and very good for cross-domain document classification. This improvement in accuracy is result of concept based analysis of the documents. The self appreciating concept based classifier (SACBC) works good with less number of training documents as compared to simple concept based classifier (CBC). SACBC takes much more time than CBC, but accuracy is not that much improved compared with CBC. Also, this model works better with average size cross domain documents. If the size of the test documents is very small, that is if the file contains only 5-6 sentences then accuracy is decreased.

As a future scope, we can use the system to work a lot better with documents having a very small size, so the overall accuracy of the system can be improved.

REFERENCES

[1] Arul Deepa K & Deisy C,"A Self Appreciating Approach of Text Classifier Based on Concept Mining", 2012 International Conference on Computer Communication and Informatics (*ICCCI -2012), Jan. 10 – 12, 2012, Coimbatore, INDIA*

[2] M. Ramakrishna Murty, J.V.R Murthy, Prasad Reddy PVGD, S. C. Satapathy "A Survey of Cross-Domain Text Categorization Techniques", *I*st *Int'I Conf*'. *on Recent Advances in Information Technology , RAIT -2012.*

[3] K. Nithya, P C D. Kalaivaani, R. Thangarajan, "*An Enhanced Data Mining Model for Text Classification*"

[4] Anjum Gupta,"New Framework for Cross-domain Document Classification", *Ph.D. thesis, NPS, Montery, California, March 2011.*

[5] Shehata, F. Karray, and M. Kamel, "An Efficient Concept Based Mining Model for Enhancing Text Clustering*" IEEE Transactions on Knowledge and Data Engineering, Vol. 22, No. 10, October 2010.*

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

[7] Sameer Pradhan, Kadri Hacioglu, Wayne Ward, James H. Martin, Daniel Jurafsky, **"**Semantic Role Parsing: Adding Semantic Structure to Unstructured Text"

[8] Daniel Gildea, Daniel Jurafsky, University of California, Berkeley, and International Computer Science Institute, University of Colorado, Boulder "Automatic Labeling of Semantic Roles"

[9] Hele-Mai Haav, Tanel-Lauri Lubi, Institute of Cybernetics at Tallinn Technical University Akadeemia. "A Survey of Concept-based Information Retrieval Tools on the Web"