Available online at www.sciencedirect.com

ScienceDirect

Procedia Technology 24 (2016) 1220 – 1225

Procedia Technology

International Conference on Emerging Trends in Engineering, Science and Technology (ICETEST - 2015)

An Efficient Text Classification Scheme Using Clustering

Anisha Mariam Thomas[a]* , Resmipriya M G[b]

[a] *PG Scholar, Amal Jyothi College of Engineering, Kerala, India*
[b] *Assistant Professor, Amal Jyothi College of Engineering, Kerala, India*

Abstract

Text classification method that uses efficient similarity measures to achieve better performance is being proposed in this paper. Semi-supervised clustering is used as a complementary step to text classification and is used to identify the components in text collection. Clustering makes use of labeled texts to capture silhouettes of text clusters and unlabeled texts to adapt its centroids. The category of each text cluster is labeled by the label of texts in it. Thus here the text clustering is used to generate the classification model for the next text classification step. When a new unlabeled text is incoming, measure its similarity with the centroids of the text clusters and give its label with that of the nearest text cluster. The similarity is calculated using different similarity measures. Results and evaluations are summarized and it is found that the system provides better accuracy when a Similarity Measure for Text Processing (SMTP) used for the distance calculation.

Peer-review under responsibility of the organizing committee of ICETEST – 2015

Keywords: Data Mining; Classification; Semi-supervised Clustering; Similarity Measures.

1. Introduction

The amount of data is increasing day by day so there is a need for efficient data processing applications. Data can be in different forms it can be either in text, image, spatial form etc. Among this the most common form of data that

* Corresponding author. Tel.: +91-828-181-7916.
E-mail address: anishamariam91@gmail.com

Peer-review under responsibility of the organizing committee of ICETEST – 2015
doi:10.1016/j.protcy.2016.05.095

we are handling enormously is the text data. The news stories we are reading, postings and messages on social media all are mainly in text form. So now a day there is great significance for the text mining process.

Text mining is the process of analyzing the written text, a lot of research works are engaged in this area. It has got a wide variety of applications in government, research, business etc. So here in this paper we are dealing with text data. The major text mining tasks include text clustering, text classification, text categorization, sentiment analysis, document summarization, production of granular taxonomies etc.

Classification of text documents is an important operation in the text processing field and it is the process of assigning class labels to the unseen documents based on the model generated in the training phase. Many applications require operations such as ``find class labels''. For example a bank officer wants to analyze the loan data to know which customer loan applications are risky and which are safe. In this example the class labels are safe and risky.

With ever increasing applications of text processing there is a need for an efficient classification scheme. In this work we summarize a text classification method using clustering. Thus the aim of this work involves developing a method for text classification in which the classification model is created using semi-supervised clustering. The two main steps of the data classification process are: building the classifier (Classification Model) and using the classifier for classification. Out of these two steps the classification model creation is an important step.

The organization of the paper is as follows. The first section gives you a brief introduction about the system and also about text mining. The second section discuss about the related works. And the third section gives a detailed overview of the implemented system. Detailed experiments and results are discussed in the fourth section. Finally the last section tells you about the concluding remarks and also gives the future scope of the implemented system.

2. Related Works

Text clustering and text classification are the two important text mining task as we already discussed. The clustering can be used to aid the text classification either as an alternative approach to term selection for dimensionality reduction or as a technique to enhance the training set. In the second approach clustering is used to discover the kind of structure in training examples. The model for classification is constructed using the extracted clusters.

TESC (Text classification using Semi-supervised Clustering) is an approach for text classification using clustering proposed by Zhang et al. [1]. In this work the task of constructing classification model is done using a semi-supervised clustering and this model is then further used to assign the correct class labels to the new document of the domain.

A tremendous amount of side information is available and can be used for the text mining to improve the performance. Side information means the extra information or the meta-data provided in the documents. This includes the document provenance information, locations, web logs, hyperlinks etc. Charu et al. [3] used these side information to improve the performance of classification and clustering. But there are problems with the proper handling of side information and the noisy side information may affect the performance of the mining.

Liu et al. [4] generalize a boosting framework for improving the supervised learning algorithm with unlabeled data known as semi-boost. It improves the classification accuracy iteratively. Similarity measures play a significant role in classification and clustering [5]. So proper selection of the suitable measure is an important step in text mining. Similarity measure is a real valued function that measures the similarity between two objects. SMTP [2] is an efficient similarity for text classification and clustering and it satisfies all the desirable properties for a good similarity measure. Joris Dhondt et al. [6] adopted a pairwise adaptive dissimilarity measure for large high dimensional document datasets.

There are, as can be seen, many different similarity measures that are used to perform text clustering and classification. The development in the various approaches has been an evolving process involving newer and better techniques that provide higher accuracy. In this paper we also present a study of different similarity measures that are used for text processing.

3. System Description

A classification approach in which the model for classification is generated using semi-supervised clustering is proposed in this paper. A large amount of unlabeled data is widely available and also human labeling is a time consuming and costly process. The use of semi-supervised approaches has got its own advantages. The proposed

system mainly contains four modules and are: preprocessing, semi-supervised clustering, similarity calculation and classification.

Data preprocessing is one of the important steps in text mining activities. Text preprocessing system consist of activities like extraction, stemming, stop word removal and vector formation. Detailed block diagram for the preprocessing is shown in figure 1. Extraction is done through tokenization of the file. Tokenization is the process of breaking a stream of text into words, phrases, symbols or other meaningful elements called tokens.

Stop word removal is the process of removing stop words. Stop words are those words that are used frequently in English and these words are useless for text mining. Stop words are language specific words which carry no information. The most commonly used stop words in English are eg: is, a, the etc.

Stemming is another important preprocessing step used to find the root/stem of a word. For example, the word walks, walking, walked all can be stemmed into the root word ``WALK''. The purpose of this preprocessing step is to remove various suffixes, to reduce the no. of words, to save memory space and time. This can be done using various algorithms like Porter stemmer, Lovins algorithm etc. In our system we used the famous Porter stemming algorithm.

The next important step is the vector formation. In order to perform any text mining activity we should first covert the text contents into numeric formats for further processing. For that we either use the term frequency (tf), inverse document frequency (idf), term frequency-inverse document frequency (tf-idf). In our experiment tf-idf is used to generate the document vectors and finally a term document matrix is formed.

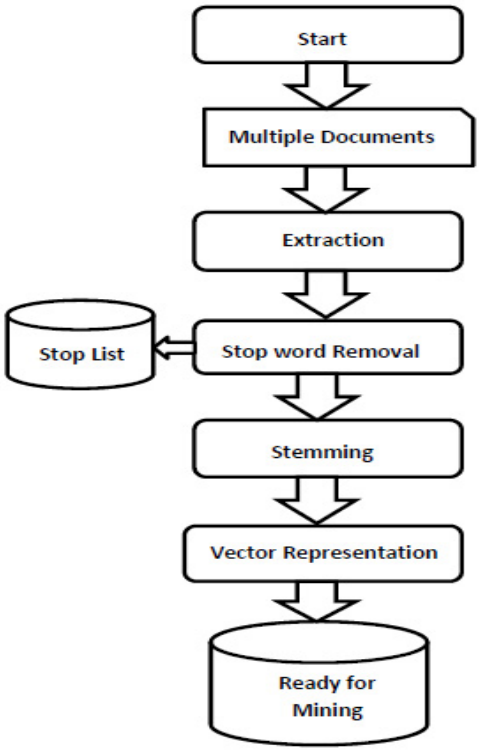

Fig 1. The Preprocessing Steps.

Now the text data is ready for mining process and this data is clustered using a semi-supervised clustering algorithm. Here semi-supervised clustering is used to create the classification model. Semi-supervised clustering means we are using both labeled and unlabeled data for doing clustering. The labeled texts are used to form the silhouettes of text clusters and the unlabeled texts are used to capture the centroids of the text components.

The detailed process of clustering includes three steps: initialization, clustering and output formation. The clustering process is semi-supervised that we are using both labeled and unlabeled text as input. In the initialization step each text that are associated with a label are given that label, if a text has got no label then it is marked as ``unlabeled''. So in the initialization step we are forming candidates for the next clustering step.

In the clustering step the cluster candidates that are having the same labels are grouped together into one cluster and it is now identified by its centroid and the label associated with it. The unlabeled documents are added to a list. Then we take one unlabeled document from the list and calculate its distance from all the other unlabeled documents in the list and also from the labeled cluster centroids. If that unlabeled document has got minimum distance to labeled cluster then it is added to that cluster and it is given the label of that cluster. The centroid of that cluster is updated and the unlabeled document is removed from the list. If the unlabeled document has minimum distance with another unlabeled document then that two unlabeled documents are merged together added to the list of unlabeled documents as one. And it is marked as unlabeled itself. The process continues until all the unlabeled documents are given a label. Finally after the clusters formed, if any clusters has less than 3 documents as member then that cluster is marked as irrelevant cluster and is deleted from the model. That means that cluster doesn't contribute much to our classification process.

After semi-supervised clustering process a collection of labeled text clusters are produced as output. And these labeled text clusters will act as a classification model for the text classification step. The basic method used here is to label an unlabeled text with the label of the text cluster that is nearest to the unlabeled text. The nearness of an unlabeled document is found out using efficient similarity measures. The measures that we use here in this paper are Euclidean distance, Cosine similarity measure, SMTP (Similarity Measure for Text Processing) [2] and Dice coefficient.

A good similarity measure should satisfy certain properties such as the presence or absence of a feature is more essential. The similarity between the objects should increase when the difference between two non-zero values of a specific feature decreases. The similarity values should decrease when the number of presence-absence features is high. Two documents are least similar to each other if none of the features have non-zero values in both documents, the measure should be symmetric and the standard deviation of the feature is also taken into account. These properties are satisfied by the SMTP and the similarity for two documents $d_1 = <d_{11}, d_{12}, \ldots, d_{1m}>$ and $d_2 = <d_{21}, d_{22}, \ldots, d_{2m}>$ define a function F as:

$$F(d_1, d_2) = \frac{\sum_{j=1}^{m} N_*(d_{1j}, d_{2j})}{\sum_{j=1}^{m} N_\cup(d_{1j}, d_{2j})} \tag{1}$$

Where

$$N_*(d_{1j}, d_{2j}) = \begin{cases} 0.5\left(1+\exp\left\{-\left(\frac{d_{1j}-d_{2j}}{\sigma_j}\right)^2\right\}\right), \\ \quad if\ \ d_{1j}d_{2j} > 0 \\ 0, \ \ if\ d_{1j} = 0\ and\ d_{2j} = 0 \\ -\lambda, otherwise, \end{cases} \tag{2}$$

$$N_\cup(d_{1j}, d_{2j}) = \begin{cases} 0, \ \ if\ d_{1j} = 0\ and\ d_{2j} = 0 \\ 1, otherwise. \end{cases} \tag{3}$$

Then SMTP for two documents is given by:

$$S_{SMTP}(d_1, d_2) = \frac{F(d_1, d_2) + \lambda}{1 + \lambda} \tag{4}$$

Mainly three cases are being considered here. For the first case, set a lower bound 0.5 and decrease the similarity as the difference between the feature values of the two documents increases, scaled by a Gaussian function as shown in Eq.(2) where σ_j is the standard deviation of all non-zero values for features in the training data set. For the second case, set a negative constant - λ disregarding the magnitude of the non-zero elements value. For the last case, there is no contribution to the similarity. The similarity between two documents is calculated using the Eq.(4).

4. Experiments And Results

The main aim of this work is to improve the performance of classification not the clustering. The data set that we are using for the experiments is Reuters-21578, a collection of news stories and it contains both labelled and unlabelled documents. We took 800 documents for our experiments and that consist of both labelled and unlabelled documents. Labelled documents are those documents that got class information on the other hand unlabelled documents are those documents in which there is no class information is provided. Out of the 800 documents 560 documents are unlabeled and rest of them are labeled. Unlabeled data are inexpensive and are easily available. The classes that we are using in our experiment are earn, acq, grain, coffee, ship, crude and the distribution are 60, 80, 40, 32, 28 documents respectively.

The model construction phase is one of the most important steps in classification. In order to obtain better classification accuracy good model construction methods should be adopted. The use of semi-supervised clustering for the model creation is a better approach. Most of the classification algorithms need labeled documents in the training phase for the classification model generation. And in this model both labeled and unlabeled documents are used for model creation and unlabeled documents are given for classification. The cluster centroids acts as the model for classification and these cluster centroids are used for classifying the documents. The clusters formed and the corresponding number of documents in each cluster is given in the table 1.

Table 1. Clusters formed in the classification model creation phase.

Clusters	Number of Documents in each Clusters
Earn	120
Acq	121
Ship	186
Coffee	122
Crude	136
Grain	115

In the classification phase of the system unlabeled documents are given as input to classifier and we give 100 unlabeled documents as input to the system. Labels are given to this unlabeled document by calculating the distance of the unlabeled document to the cluster centroids. And the unlabeled document is given the label of the cluster to which it has got minimum distance. To calculate the accuracy value the number of correctly classified documents are summed up and divided by the total number of documents considered. In this experiment, the performance is evaluated by the classification accuracy, AC, which compares the predicted label of each document with that provided by the document corpus:

$$AC = \frac{\sum_{i=1}^{n} E(ci,ci\prime)}{n} \quad (5)$$

where *n* is the number of testing documents, c_i and c_i' are the target label and the predicted label, respectively, of the i^{th} document. $E(c_i; c_i') = 1$ if ci = c_i' and $E(c_i; c_{i'}) = 0$ otherwise.

When Euclidean distance is used as the similarity measure for classification out of the total 100 documents given as input for the classification 88 documents are correctly classified and that gives an accuracy of 0.88. And when Cosine similarity 89 documents are correctly classified and that gives an accuracy of 0.89. On the other hand when Dice coefficient is used it correctly classifies 81 documents giving an accuracy value of 0.81. When SMTP is used as the similarity measure it gives an accuracy value of 0.93 that means out of the 100 documents 93 documents are correctly classified. From the experiments we have done it is quite clear that SMTP gives better performance. The results obtained are shown in the table 2.

Table 2. Accuracy values obtained when different similarity measures are used for classification.

Similarity Measures	Accuracy
Euclidean	120
Cosine	121
SMTP	186
Dice Coefficient	122

5. Conclusion and Future Work

An approach to text classification using for semi-supervised clustering has been proposed and the accuracy values obtained by applying the similarity measures in classification algorithms are been analysed. The basic assumption is that each category of documents comes from multiple components, which can be identified by clustering. In order to elaborate the clustering process, unlabeled documents are utilized to adapt the centroids of cluster candidates iteratively and labeled documents are utilized to capture the silhouettes of the clusters. This is a search-based approach to semi-supervised clustering. It provides better classification results. Different similarity measures are applied in classification process and the results are taken out of that SMTP gets better accuracy values. As a future enhancement dimensionality reduction techniques can be used. And the dimension of the term document matrix can be reduced. So that better execution time can be achieved and also large number of documents can be handled easily.

References

[1] Wen Zhang, Xijin Tang, Taketoshi Yoshida, `TESC: An approach to Text classification using Semi-supervised Clustering', Knowledge-Based Systems, November 2014.

[2] Yung-Shen Lin, Jung-Yi Jiang, and Shie-Jue Lee, `A Similarity Measure for Text Classification and Clustering, IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 7, July 2014.

[3] Charu C. Aggarwal, Yuchen Zhao, and Philip S. Yu, `On the Use of Side Information for Mining Text Data', IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 6, June 2014.

[4] Pavan Kumar Mallapragada, Rong Jin, Member, Anil K. Jain, Fellow, and Yi Liu,`SemiBoost: Boosting for Semi-supervised Learning', IEEE Transctions On Pattern Analysis And Machine Intelligence, Vol. 31, No. 11, November 2009.

[5] Hung Chim and Xiaotie Deng, `Efficient Phrase-Based Document Similarity for Clustering, IEEE Transactions On Knowledge And Data Engineering, Vol. 20, No. 9, September 2008.

[6] Joris Dhondt, Joris Vertommen, Paul-Armand Verhaegen, Dirk Cattrysse, Joost R. Duon, `Pairwise-Adaptive Dissimilarity Measure for Document Clustering, Inf. Sci., vol. 180, no. 12, pp. 23412358, 2010.