ECAI 2016
G.A. Kaminka et al. (Eds.)
© 2016 The Authors and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/978-1-61499-672-9-1710

Supervised Graph-Based Term Weighting Scheme for Effective Text Classification

Niloofer Shanavas and **Hui Wang** and **Zhiwei Lin** and **Glenn Hawe**[1]

Abstract. Due to the increase in electronic documents, automatic text classification has gained a lot of importance as manual classification of documents is time-consuming. Machine learning is the main approach for automatic text classification, where texts are represented, terms are weighted on the basis of the chosen representation and a classification model is built. Vector space model is the dominant text representation largely due to its simplicity. Graphs are becoming an alternative text representation that have the ability to capture important information in text such as term order, term co-occurrence and term relationships that are not considered by the vector space model. Substantially better text classification performance has been demonstrated for term weighting schemes which use a graph representation. In this paper, we introduce a graph-based term weighting scheme, tw-srw, which is an effective supervised term weighting method that considers the co-occurrence information in text for increasing text classification accuracy. Experimental results show that it outperforms the state-of-the-art unsupervised term weighting schemes.

1 INTRODUCTION

A challenging task in text classification is the effective representation of text. The features that represent the document affect the performance of text classification. The documents for classification are usually represented in the vector space model. It assumes that the terms are independent and represents a document as an unordered set of terms and their frequencies. Although it is simple and fast, this representation does not consider structural information (order of words, relationship between words) or the semantics of text. An alternative to vector space model for representing documents is graph-based representation. A document represented as a graph instead of a vector can retain its inherent structure, thereby increasing the classification performance.

Graph-based term weighting schemes improve classification accuracy compared to traditional frequency-based term weighting methods [1, 2, 3]. The existing graph-based term weighting schemes are unsupervised, so the class-separating information is not considered for text classification where class labels are given. Supervised term weighting methods utilize the information on the membership of the training documents in the predefined classes to give higher weights to terms that are distributed differently in the classes [4]. In this paper, a supervised graph-based term weighting scheme is presented, which utilizes the rich information in text and the relationship of the terms to the predefined classes. It is a class-based function that considers the co-occurrence information in text. Experimental results show that the proposed term weighting scheme can indeed improve classification performance.

The rest of the paper is organized as follows. Section 2 explains the proposed supervised graph-based term weighting scheme. Section 3 presents the experimental results of the proposed term weighting scheme for text classification. Finally, Section 4 concludes the paper and discusses possible future work.

2 PROPOSED TERM WEIGHTING SCHEME FOR TEXT CLASSIFICATION

We have focused our study on undirected co-occurrence graphs, an effective structure-based representation of text, to capture the co-occurrence of words. The text documents are pre-processed by removing stop words and stemming before converting it to co-occurrence graphs. The nodes in the co-occurrence graph represent the unique terms in text and the edges link co-occurring terms within a sliding window of fixed length. Each document is represented by a co-occurrence graph. The terms are weighted based on the node's importance in the graph. Since degree centrality measures perform the best on co-occurrence graph [1], we used it to assign weights to terms. The degree centrality scores of the nodes are used to convert the co-occurrence graph to vector-based representation.

The research done on supervised term weighting schemes for text classification emphasizes the fact that term weighting should depend on the classification task. Text classification performance can be improved by giving more weights to terms that support in the classification of documents into the right class. This is done by utilizing the known information on the training documents in the predefined classes. In [5], a supervised term weighting scheme $tf * rf$ is proposed where tf is the term frequency and rf is the relevance frequency. In [6], the supervised term weighting scheme proposed in [5] is modified to compute a single relevance frequency value for each term for a classification task with n classes by taking the maximum of the n relevance frequency values of each term. The rf factor defined in [5] gives greater weights to terms that are more concentrated in the positive class than in the negative class. But, this weighting measure cannot handle imbalanced data. This happens when one or more classes contain more training documents than the rest of the classes. In such cases, the rf factor gives low weights to terms in the classes with few training samples (minor classes). To overcome this problem, a probability-based approach for term weighting is defined in [7] to assign better weights to terms in minor classes. However, this term weighting scheme results in overweighting of commonly occurring terms.

In order to avoid the problems that reduces the effectiveness of supervised term weighting schemes such as overweighting of commonly occurring terms, higher weights for terms in classes with large

[1] School of Computing and Mathematics, Ulster University, United Kingdom, email: shanavas-n@email.ulster.ac.uk, {h.wang, z.lin, gi.hawe}@ulster.ac.uk

number of training documents, our approach considers all the below three elements for computation of the term's relevance in the classes.

1. The term's concentration in each class as compared to its concentration in other classes.
2. The number of documents in each class that do not contain the term.
3. The average density of the term in the classes.

The new supervised term weight factor, which we name as supervised relevance weight (srw), takes into account the above three elements for its computation as discussed below.

Table 1. Notations used in the supervised term weighting scheme

Notation	Description
a	The number of documents in class C_i that contain the term t.
b	The number of documents in class C_i that do not contain the term t.
c	The number of documents not in class C_i that contain the term t.

The notations a, b and c used in the supervised term weighting scheme are explained in Table 1. In Equation 1, $(a/max(1,c))$ determines the concentration of the term t in class C_i as compared to its concentration in other classes and $(a/max(1,b))$ helps to reduce the higher weights assigned to terms in classes having more training documents by considering the number of documents in class C_i that do not contain the term t.

$$class\_rel\_prob(t, C_i) = \log_2\left(2+\frac{a}{max(1,c)}\right) * \log_2\left(2+\frac{a}{max(1,b)}\right) \quad (1)$$

$class\_rel\_prob(t, C_i)$ defined in Equation 1 is computed for each class C_i. The maximum of $class\_rel\_prob(t, C_i)$ values of the term t, $max(class\_rel\_prob(t, C_i))$, is calculated to obtain a single value for the term t. In order to reduce the overweighting of commonly occurring terms, the average density of the term t in the classes is determined by dividing the sum of densities of the term t in the classes by the number of classes. The computation of average density of the term t is shown in Equation 2 where N_i is the total number of documents in class C_i and C is the total number of classes.

$$avg\_density(t) = \frac{\sum_{i=1}^{C}\left(\frac{a}{N_i}\right)}{C} \quad (2)$$

The supervised relevance weight (srw) of each term t determines the relevance of the term in the classes. It gives higher weights to terms that help in distinguishing the documents in different classes. It is calculated as shown below in Equation 3.

$$srw = max(class\_rel\_prob(t, C_i)) * log_{10}\left(\frac{1}{avg\_density(t)}\right) \quad (3)$$

The new improved term weighting measure (tw-srw) for a term t represented by a node in the co-occurrence graph is defined as the product of tw and srw i.e. $tw * srw$ where tw is the term weight determined by the centrality score for the term t in the graph.

3 EXPERIMENTS AND RESULTS

We have experimented with four pre-processed datasets[1], WebKB, R8, R52, 20 Newsgroups, to evaluate the proposed supervised term weighting scheme for text classification. The documents in the training and testing set are converted to co-occurrence graphs that capture terms that co-occur within a sliding window of size 2. The performance of the supervised graph-based term weighting measure (tw-srw), traditional term weighting measure i.e. term frequency - inverse document frequency (tf-idf) and graph-based term weighting measures such as tw and tw-idf are evaluated for text classification with SVM. tw is determined by the degree centrality score and tw-idf is computed as the product of tw and inverse document frequency (idf). We used the linear support vector machine implementation in scikit-learn called the LinearSVC as the classifier and set the penalty parameter C of the error term to its default value of 1 and the loss function to hinge. The proposed weighting scheme consistently outperforms the unsupervised term weighting schemes for the four datasets tested as shown in Table 2, Table 3 and Table 4.

Table 2. Precision scores for different term weighting schemes

Dataset	tf-idf	tw	tw-idf	tw-srw
WebKB	0.8459	0.8961	0.8757	**0.9111**
R8	0.9622	0.9689	0.9758	**0.9802**
R52	0.9200	0.9031	0.9389	**0.9549**
20 Newsgroups	0.7813	0.7845	0.8353	**0.8462**

Table 3. Recall scores for different term weighting schemes

Dataset	tf-idf	tw	tw-idf	tw-srw
WebKB	0.8467	0.8933	0.8746	**0.9112**
R8	0.9621	0.9685	0.9758	**0.9799**
R52	0.9190	0.9186	0.9435	**0.9552**
20 Newsgroups	0.7763	0.7814	0.8335	**0.8441**

Table 4. F1 scores for different term weighting schemes

Dataset	tf-idf	tw	tw-idf	tw-srw
WebKB	0.8462	0.8902	0.8693	**0.9108**
R8	0.9618	0.9681	0.9756	**0.9799**
R52	0.9146	0.9037	0.9370	**0.9520**
20 Newsgroups	0.7760	0.7740	0.8301	**0.8430**

4 CONCLUSION

Effective representation that considers the structure information in text and an appropriate term weighting measure that takes into account the relationship between terms and the term's relevance to the classification task increases the performance of text classification. An interesting future work is to explore graph-based term weighting measures that consider more complex dependencies in text for classification task.

REFERENCES

[1] K. Valle and P. Ozturk, 'Graph-based Representations for Text Classification', India-Norway Workshop on Web Concepts and Technologies, Trondheim, Norway, 2011.

[2] F. D. Malliaros and K. Skianis, 'Graph-Based Term Weighting for Text Categorization', Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2015, pp. 1473-1479, 2015.

[3] S. Hassan, R. Mihalcea, and C. Banea, 'Random-Walk Term Weighting for Improved Text Classification', Proceedings of the IEEE International Conference on Semantic Computing, ICSC 2007, pp. 242-249, 2007.

[4] F. Debole and F. Sebastiani, 'Supervised Term Weighting for Automated Text Categorization', Text Mining and its Applications: Results of the NEMIS Launch Conference, Springer Berlin Heidelberg, pp. 81-97, 2004.

[5] M. Lan, C. L. Tan, J. Su, and Y. Lu, 'Supervised and Traditional Term Weighting Methods for Automatic Text Categorization', IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 721-735, 2009.

[6] N. P. Xuan and H. L. Quang, 'A New Improved Term Weighting Scheme for Text Categorization', Knowledge and Systems Engineering: Proceedings of the Fifth International Conference KSE 2013, Springer International Publishing, pp. 261-270, 2014.

[7] Y. Liu, H. T. Loh, and A. Sun, 'Imbalanced text classification: A term weighting approach', Expert Systems with Applications, pp. 690-701, 2009.

[1] http://ana.cachopo.org/datasets-for-single-label-text-categorization