

Integrate Text Clustering Features in Text Categorization System

ZHOU Qiang, ZHENG Yabin

Computer Science and Artificial Intelligence Division
National Laboratory for Information Science and Technology
Dept. of Computer Science and technology,
Tsinghua University, Beijing 100084
zq-lxd@mail.tsinghua.edu.cn

Abstract—The paper proposed an algorithm to integrate the text clustering features in a text categorization systems. Some primitive experimental results show the feasibility of the algorithm in the text categorization task with smaller tag set.

Index Terms—Text categorization, text clustering

I. INTRODUCTION

Automated Text Categorization (ATC) is the task of building software tools capable of classifying text documents under predefined categories or topic codes. The dominant approach is nowadays one of building text classifiers automatically by learning the characteristics of the categories from a training set of pre-classified documents. State-of-the-art machine learning methods have recently been applied to the task, leading to systems of increased sophistication and effectiveness, and stably placing ATC at the crossroads of information retrieval and machine learning.

There are two key techniques need to be solved in a typical ATC system:

- 1) How to select the suitable high discriminative features to represent the topical characteristics of a document?
- 2) How to aggregate these features to form several suitable categories for the document?

In recent years, many different machine learning models or algorithms, such as SVM and k-NN, were used to deal with the above problems. In the paper, we proposed a new method to integrate some supportive features extracted from a text clustering model to improve the classification performance of an ATC system.

II. THE INTEGRATION ALGORITHM

We selected the following two tools to implement our integration idea in Chinese texts:

- (1) FIFA text categorization tool [1]

It can output several topic tags with suitable weights for a Chinese document, based on a simple and efficient Feature Identification and Feature Aggregation (FIFA) procedure [2]. The key part of the FIFA algorithm is a large scale topical knowledge base, which consists of about 400,000 Chinese word or phrase entries manually annotated with detailed topic and domain information.

- (2) MPCA text clustering tool

It is a multinomial variation of the discrete Principal Component Analysis (PCA) [3] algorithm. The model can assign a set of independent components with weight probabilities to a document somehow representing the topical content.

The main reason to select these two tools lies in the flexible alternativeness to select different content level to integrate the text clustering features in a text categorization system. All of them can easily provide us with different document content representations in the key term or component level or the document level. In the paper, we design an algorithm to integrate them in the document level. The main idea of the algorithm is as follows:

Firstly, we assign each document with several topic tags, based on the FIFA algorithm. Secondly, we label the document clusters generated by MPCA model with suitable topic tags, based on the salient topic tags and their weights of the documents under a cluster. Finally, we relocate the topic tags of a document cluster to its member and get the relabeled topic tags for each document under a cluster.

Before giving the detailed descriptions of the algorithm, we firstly define the following useful notations in the algorithm:

- $D = \{D_1, D_2, \dots, D_m\}$ is a set of documents;
- $T = \{T_1, T_2, \dots, T_n\}$ is a set of topics;
- $C = \{C_1, C_2, \dots, C_k\}$ is a set of clusters;
- A Document-By-Topic matrix (DBT), whose element $DBT_{i,j}$ represents the weight of document D_i with topic T_j ;

- A Document-By-Cluster matrix (DBC), whose element $DBC_{i,j}$ represents the possibility that document D_i belongs to cluster C_j .
- A Cluster-By-Topic matrix (CBT), whose element $CBT_{i,j}$ represents the weight of cluster C_i with topic T_j ;
- A Feedback-Document-By-Topic matrix (FDBT), whose element $FDBT_{i,j}$ represents the feedback information given by a cluster, i.e. the feedback weight of document D_i with topic T_j .

The detailed algorithm has the following five steps:

1. Topic pre-selection:

FIFA can give us top-10 topic tags and their weights for a document. But there are some noises among them. So we should remove them to generate a good DBT for further computation.

Firstly, considering a document, if a document has a salient topic tag, other tags can be regarded as noise information. So we set a weight threshold for it. If the weight of one top-N tag is greater than this threshold, we remain it, otherwise, we remove it as a noise tag.

Secondly, considering the topics, if one topic appears in almost all the documents or in just a small part of the documents, they can be regarded as a noisy topic. So we set an up-threshold and a down-threshold, if the occurrence of topic is not in this range, we will remove it.

After the above processing, we can get a suitable DBT.

2. Document pre-selection:

In the same way, MPCA will give us a cluster and all the documents in it along with its weight. Similarly, we set a weight threshold, if the weight is less than the threshold, we will remove the document from this cluster. In our experiment, the threshold is set to the average of the weights in the cluster. After that approaching, we can get a suitable DBC.

3. Label the cluster with suitable topic tags

Its computing equation is as follow:

$$CBT = (DBC)^T * (DBT)$$

4. Topic information feedback:

Its computation equation is as follows:

$$FDBT = (DBC) * (CBT)$$

5. Information combination:

Its computation equation is as follows:

$$\text{Result} = \lambda_1 * (DBT) + \lambda_2 * (FDBT)$$

where λ_1 and λ_2 are the parameters for tuning the weight between original topic tags and clustering topic tags.

III. EXPERIMENTAL RESULTS

We made two experiments to test the performance of the integration algorithm.

In the first experiment, we used 3600 documents extracted from the 863 test set to test the categorization precision of the algorithm. 37 domain tags from Chinese library classification system are used to annotate the correct tags for the documents.

At the preprocessing stage, we used the FIFA algorithm to assign 10 tags with weights for each document, and clustered all the documents into N clusters, where we set the parameter N of MPCA algorithm as 18. Figure 1 and figure 2 show the performance improvement of the integration algorithm after the topic pre-selection stage and the combination stage. We can find that after removing some noise information in the FIFA's outputs, the precision of the top-1 categorization tag can be improved from 46.4% to 54.9%. Then, after the computation through feedback and combination, the precision of the top-1 tag can be further improved to 57.4%.

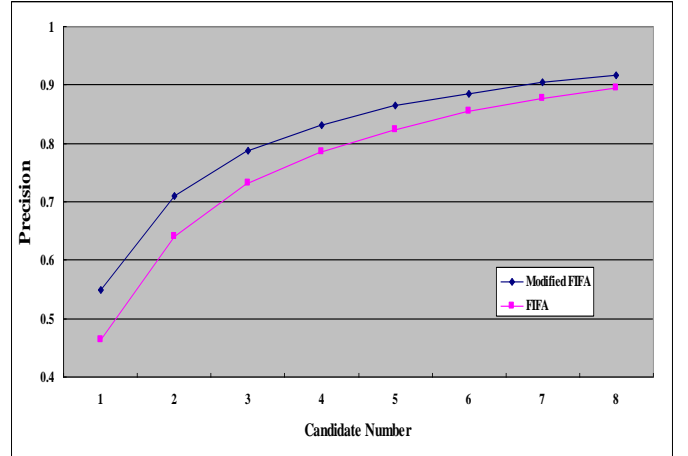


Figure 1. Performance improvement after the topic pre-selection stage

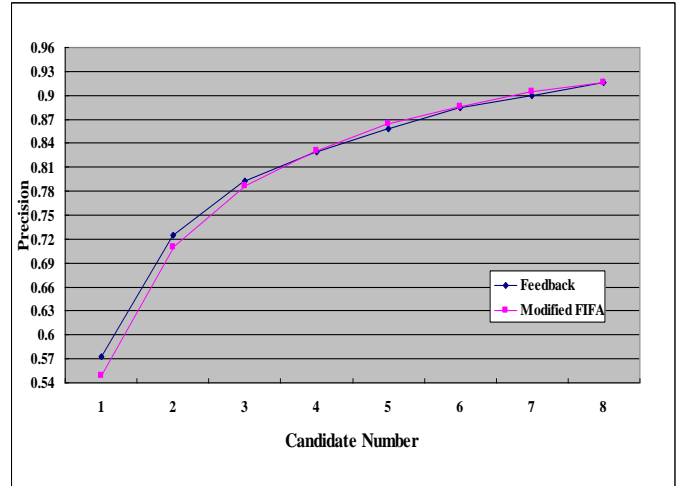


Figure 2. Performance improvement after feedback and information stages

The result is not very good and far below our expectation. A possible reason is due to the sensibility of the MPCA to its clustered parameter N. In most cases, the MPCA algorithm can not output more than 20 clusters properly. Therefore, it may not be suitable to deal with our current text categorization task with

37 classes. So we designed another experiment to see whether our current integration algorithm can get better performance in some smaller tag set.

In the second experiment, we selected 2615 documents annotated with the correct domain tag from a tag set with the following 9 tags: computer, traffic, education, economy, military, sports, medicine, arts and politics. We separated them into a training set with 450 documents and a test set with 2165 documents.

After that, we used the MPCA algorithm to cluster all the 2615 documents into 9 clusters, labeled them with suitable tags based on correct tag information of the training documents among them and feedback the cluster tag to all the test documents in the cluster.

The commonly-used micro precision, recall and F-measure are used to evaluate the text categorization performance on the test set. The processed results are also compared with the results of the kNN and SVM algorithms upon the same training and test sets. Table 1 shows the results. We can see that our algorithm get the better performance with 87% Micro-F1, approaching to the best result provided by SVM algorithm. This result proves the feasibility of our integration algorithm in the text categorization task with smaller tag set.

Table 1. Performance comparison with other algorithms

	MicroP	MicroR	Micro-F1
KNN	81.8%	81.8%	81.8%
Our Method	87.4%	87.4%	87.4%
SVM	91.2%	91.2%	91.2%

IV. ACKNOWLEDGEMENTS

The research was supported by National Natural Science Foundation of China (NSFC) (Grant No. 60573185, 60520130299).

REFERENCES

- [1] JB Zhu and TS Yao (2002) "FIFA-based text classification" *Journal of Chinese Information Processing*, 16(3), 20-26.
- [2] JB Zhu, TS Yao (2002). "FIFA: a simple and effective approach to text topic automatic identification", In : *Proceedings of International Conference On Multilingual Information Processing*, Shenyang, China , Feb. 2002 ,207 – 215
- [3] W. Buntine, A. Jakulin, "Discrete Components Analysis", in *Subspace, Latent Structure and Feature Selection Techniques*, Springer-Verlag, edited by C. Saunders and M. Grobelnik and S. Gunn and J. Shawe-Taylor, 2006