Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 24 (2011) 329 – 334

Procedia Engineering

www.elsevier.com/locate/procedia

2011 International Conference on Advances in Engineering

Novel top-down methods for Hierarchical Text Classification

CAO Ying[a,c];Duan run-ying[b,a*]

[a]*Modern Educational Technology and Information Center Jiangxi University of Science and Technology, Ganzhou,Jiangxi 341000 China*

[b]*Department of Computer Science and Technology GuanZhou University Sontan College, Zengcheng,GuangZhou, Guangdong,511370 China*

[c] *caoying@guigu.org*

Abstract

To classify large-scale text corpora, one common approach is using hierarchical text classification and classifying text documents in a top-down manner. Classification methods using top-down approach can scale well and cope with changes to the category trees. However, all these methods suffer from a common problem: a high level of misclassification document has unrecoverable. We define an virtual subclass for each non-leaf category to help the rejected documents go back to ancestor category ,thus improving the overall performance .Our experiments using Support Vector Machine (SVM) classifiers on the 20newsgroup collection have shown that they all could reduce blocking and improve the classification accuracy. Our experiments have also shown that the virtual category method delivered the best performance.

Keywords: hierarchical classification , virtual category, top-down approach

1. Introduction

Text classification is a crucial and well-proven method for organizing the collection of large scale documents. The predefined categories are formed by different criterions, e.g. “Economic”, “Sports”and “Computer” in news classification, “Junk Email” and “Ordinary Email” in email classification. In the literature, many algorithms [1] have been proposed, such as Support Vector Machines (SVM), k-Nearest Neighbor (kNN), Naive Bayes (NB) and so on. Empirical evaluations have shown that most of these methods are quite effective in traditional text classification applications. In past several years, hierarchical text classification has become an active research topic in database area[2][3]. Different with traditional classification, the document collections are organized as hierarchical class structure in many application fields: web taxonomies, email folders and product catalogs.

* Corresponding author. Tel.: ; fax:.
E-mail address:.

doi:10.1016/j.proeng.2011.11.2651

For hierarchical classification, there are two main methods: one is called global method (global, also known as big-bang method), global approach builds only one classifier to discriminate all categories in a hierarchy [4][5], it assigned a single document to a category directly; The local approach proceeds in a top-down fashion, category hierarchy structure were used to train the classifier for each layer. Classification of new examples is done by starting at the root node and traversing the hierarchy till correct classes are found. Recursively at each node a classifier decides which branches i.e. which edges of the graph should be traversed further down. Unlike single global classification method, Local method determines the most relevant categories of the top level and then recursively making the choice among the low-level categories[6]. The advantage of this approach is computationally more efficient, but If an upper level decision node makes a wrong decision by erroneous forwarding or not forwarding a document to a sub hierarchy, this document may be wrongly classified. Thus, these erroneous decisions may lead to invalid classifications.

Be aimed at the above-mentioned characteristic, a new hierarchical text classification method is proposed: building a virtual category on each non-leaf node, and picking out the documents which classified into the virtual category, then classify them again under its ancestor node (grandfather node) exclude the previous parent node.

The paper is organized as follows. In section 2 we describe the basic model of hierarchical classification. Then we propose our algorithm in section3. Section 4 gives experimental analysis. Section 5 concludes the paper.

2. Basic hierarchical text classification

As stated in the introduction, today's methods for text classification are usually based on flat model. Recently, some approaches for hierarchical text classification, overcoming the restrictions of flat models, were introduced. These approaches are based on classification and feature selection methods as described

The categorization algorithm used is a supervised learning procedure that uses a linear classifier based on the category levels. We are given a set of categories, organized hierarchically. We are also given a training corpus of documents already placed in one or more categories. From these, we extract vocabulary, words that appear with high frequency within a given category, characterizing each subject area. Each node's vocabulary is filtered and its words assigned weights with respect to the specific category. Then, test documents are scanned and categories ranked based on the presence of vocabulary terms. Documents are assigned to categories based on these rankings. We demonstrate that precision and recall can be significantly improved by solving the categorization problem taking hierarchy into account.

Feature selection, deciding which terms to actually include in the indexing and categorization process, is another aspect affected by size of the corpus. Some methods remove words with low frequencies both in order to reduce the number of features and because such words are often unreliable. Depending on the size of the corpus, this may still leave over 10,000 features, which renders even the simplest categorization methods too slow to be of use on very large corpora and renders the more complex ones entirely infeasible.

We describe an algorithm for hierarchical document categorization where the vocabulary and term weights are associated with categories at each level in the taxonomy and where the categorization process itself is iterated over levels in the hierarchy. Thus a given term may be a discriminator at one level in the taxonomy receiving a large weight and then become a stopword at another level in the hierarchy.

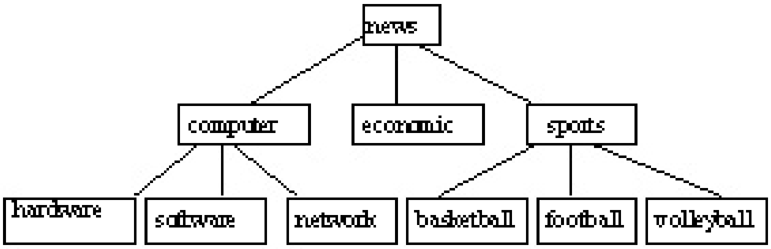

Fig.1 the topology of hierarchical topics

For example, in Figure 1, on the first level, according to the category of "computer”, we can more easily distinguish whether the document belong to “sports" or "computer"; but when we want to further determine which sub-category this document belong to, e.g. "hardware" or "software" , the feature states for "computer" feature does not have a strong differentiation, because the documents to be classified all belongs to this category of computer. In this case, we should weaken its proper weight, and even to remove these features. Based on these characteristics, we will use appropriate level of features to represent documents on different level.

In hierarchical classification learning process, classifier is trained for each non-leaf node respectively at any category-level. emerging the training documents belong to the same category into one document, and extracting various models using the training documents only in the same category; in classifying stage, when a new document come, assigned it to the best appropriate category according to the root node classifier, and then use the best category classifier decide which path will continue, so go on, until the document assigned to a leaf class. .

3. Hierarchical text classification with virtual category

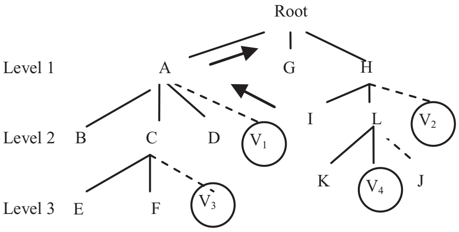

Fig.2 the topology of hierarchical contains virtual category

As mentioned before, error propagation is a problem concerning hierarchical text classification. If a wrong classification is made in top inner nodes, a document may be assigned to a class in the according sub hierarchy. Thus, one has to deal with wrongly propagated documents along the decision path. Within this paper this problem is overcome by adding a virtual child category for each parent category, for example when a document is assigned to a parent category *A*, but is rejected by its subclass *B*, *C* and *D*, this causes the document does not belong to any of the leaf class. We assume that these documents are belonging to the virtual category V_1.The virtual category only created under the non-leaf node except root, for figure 2, V1,V2,V3,V4 all are virtual categories, they are connected with their parent node using a dotted line ,indicated that they are not belong to his parent category, but were misclassified into the parent category, the documents in virtual category are the “blocked ”documents mentioned before. We pick out these documents back to the upper level, and re-assigned it to upper level category except its own parent node.

The virtual category is called VC for short, and our proposed algorithm referred to as VCHC. Following is the detailed algorithm explanation.

Algorithm

1. Construct the classifier of each parent category (PC) using flat classification method, then classify document, and get the document set D , $\{\exists d \mid d \in D, d \notin child(PC)\}$;
2. Construct the classifier of each virtual category (VC),it is trained together with the node under the same parent category ,the training set is $\{\forall d \mid d \in child(PC)\} \cup \{\forall d \mid d \in VC\}$;
3. At test stage, when a test document is classified into PC, but the document may belong to pc, may not;
4. If the document $\{\forall d' | d' \rightarrow PC\}$ also assigned into the real child category, then goto step 1 until reach leaf category ,then output the leaf category;
5. If the document $\{\forall d' | d' \rightarrow PC\}$, and $d' \in VC$, then go back to PC level category, and re-assigned it to the category except PC, then goto step 3;

Among them ,PC is short for parent category of VC, $d \in C$ indicates document d is belong to category C, $d \rightarrow C$ indicates d is classified into category C , $child(PC)$ indicates the set of true child class of node PC.

4. Experiment

Our experiments have been carried out using a dataset provided by 20NewGroups , This database contains about 20,000 news distributed in 20 different newsgroups, （download address ：http://people.csail.mit.edu/jrennie/20Newsgroups/） , each newsgroup represents a category. This dataset version we used is that the documents is sorted by date, the training set and test set are divided proportionally, the number of documents in which the training set, 60%, a total of 11,314 articles document, test set 40% the number of documents, a total of 7532 documents . Table 1 list the hierarchy of data sets, there are seven categories on the first layer, 16 categories on second layer and the 20 categories on third layer.

Table 1 the hierarchical topics on 20NewsGroup

First level	Second level	Third level	First level	Second level	Third level
alt	alt.ethesim	alt.ethesim	misc	misc.forsale	misc.forsale
comp	comp.graphics	comp.graphics	sci	sci.electronics	sci.electronics
	comp.os	comp.os		sci.med	sci.med
	comp.sys	comp.sys.ibm		sci.space	sci.space
		comp.sys.mac		sci.crypt	sci.crypt
	comp.windows	comp.windows	soc	soc.religion	soc.religion
rec	rec.autos	rec.autos	talk	talk. politics	talk.politics.misc
	rec.motocycles	rec.motocycles			talk.politics.guns
	rec.sport	rec.sport.baseball			talk.politics.mideart
		rec.sport.hockey		talk.religion	talk.religion.misc

In our experiments for single category classification rate ,F1 value is used to test the effectiveness of the classifier, which took into account the precision and recall rate factors; for the overall performance of the classifier we used micro-average F1 (Micro-averaging F1) and macro average F1 (Macro -averaging

F1) both evaluation.

Throughout the experiments, SVMLight classifier is used for classifier model, and chi-square statistic (X2) is used for feature selection. We use LTC to compute the feature's weight.

In this paper, the feature selection method is novel, which the features for each training classifier are different, we need secondary feature selection for each category level. For example, we believe that all training documents belong to one category of comp, rec… talk. First , select features under all dataset ,and represent documents with these features, then classify the test documents into the appropriate sub-category, such as comp ; at second category level ,we have to select another features for comp category, which significant less than the features on first level , then represent documents in comp with these secondary selected features, classify them to sub-category , we also follow a similar operation in turn.

We compare VCHC with another two algorithms flat (flat classification) and BH(basic hierarchal),the result are shown in Table 2 and figure 3 .Table 2 is the micro-average and macro average using three algorithms on different dimensions. Flat indicate flat classification, BH indicate the basic hierarchical mentioned before in section 2 , VCHC indicate this improved hierarchical classification algorithm.

Table 2 the results on 20NewsGroup

	flat		BH		VCHC	
	micF1	macF1	micF1	macF1	micF1	macF1
800	0.659	0.733	0.756	0.743	0.761	0.745
1500	0.722	0.748	0.777	0.765	0.812	0.801
2000	0.733	0.750	0.783	0.772	0.826	0.816
3000	0.767	0.755	0.791	0.779	0.831	0.825
4000	0.772	0.760	0.794	0.784	0.839	0.829
5000	0.777	0.779	0.801	0.791	0.847	0.831
6000	0.785	0.785	0.803	0.792	0.842	0.839
7000	0.789	0.788	0.803	0.793	0.845	0.836

From the table2 we can clearly find that VCHC is performance much better than the other two methods, and show excellent characteristics in each dimension, BH methods followed, flat method is the worst. With the increase in dimension, each classification algorithm has improved, on 6000 dimension tended to be the best value, but there is almost no growth when the dimension grows up to 7000.

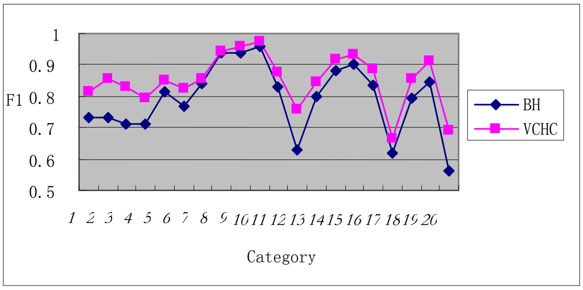

Fig3.The f1 between BH and VCHC on twenty topics

Figure3 is the F1 values comparison chart between BH and VCHC on 6000 dimension in all 20 categories, the category on the horizontal axis category correspond to in Table 1, from this chart , the good performance of VCHC can be seen even more intuitively, and behave stable in all categories. The F1 value is smaller in BH, the more the percentage increase, when BH has already a good performance in categories such as category10 which was rec. sport. hockey, while almost no growth. This phenomenon is expected, the classification accuracy rate is already very high , which means very little misclassification of documents in inner parent category, do not need to go back to the ancestor class, the text block number is almost zero, so there is no growth at all .

5. Conclusion

In this paper we propose an algorithm to reduce the blocking in top-down hierarchy classification, which define an virtual subclass for each non-leaf category to help the rejected documents go back to its ancestor category. Experiment shows that the NH approach is useful in this classification problem, since it always provide much better results than both the flat classifier model and basic hierarchy classifier model.

References

[1] F. Sebastiani. 2002. Machine learning in automated text categorization. ACM computing surveys, 34(1):1–47.

[2] Juho Rousu, Craig Saunders, Sandor Szedmak, and John Shawe-Taylor. 2006. Kernel-based learning of hierarchical multilabel classification models. In Journal of Machine Learning Research.

[3] L. Cai and T. Hofmann. 2007. Exploiting known taxonomies in learning overlapping concepts. In Proceedings of International Joint Conferences on Artificial Intelligence.

[4] Youdong Miao and Xipeng Qiu. 2009. Hierarchical centroid-based classifier for large scale text classification. In Large Scale Hierarchical Text classification (LSHTC) Pascal Challenge.

[5]Xipeng Qiu, Jinlong Zhou, and Xuanjing Huang. 2011.An effective feature selection method for text categorization.In Proceedings of the 15th Pacific-Asia Conference on Knowledge Discovery and Data Mining.

[6] T.Y. Liu, Y. Yang, H. Wan, H.J. Zeng, Z. Chen, and W.Y.Ma. 2005. Support vector machines classificationwith a very large-scale taxonomy. ACM SIGKDD Explorations Newsletter, 7(1):43.