# Dynamic Threshold Selection Method for Multi-label Newspaper Topic Identification

Lucie Skorkovská

University of West Bohemia, Faculty of Applied Sciences, Dept. of Cybernetics
Univerzitní 8, 306 14 Plzeň, Czech Republic
lskorkov@kky.zcu.cz
www.kky.zcu.cz

**Abstract.** Nowadays, the multi-label classification is increasingly required in modern categorization systems. It is especially essential in the task of newspaper article topics identification. This paper presents a method based on general topic model normalisation for finding a threshold defining the boundary between the "correct" and the "incorrect" topics of a newspaper article. The proposed method is used to improve the topic identification algorithm which is a part of a complex system for acquisition and storing large volumes of text data. The topic identification module uses the Naive Bayes classifier for the multiclass and multi-label classification problem and assigns to each article the topics from a defined quite extensive topic hierarchy - it contains about 450 topics and topic categories. The results of the experiments with the improved topic identification algorithm are presented in this paper.

**Keywords:** topic identification, multi-label text classification, language modeling, Naive Bayes classification.

## 1 Introduction

The goal of the text classification (or topic identification) is to categorize a set of documents into predefined set of topic classes or categories. Usually in the field of text classification we are considering only the multiclass classification, where unlike in the binary classification there is more than two possible classes. The simplest task of the text classification is to assign one topic to each document, but real world applications including e-mail routing, web content topical organization or news topic identification require the multi-label classification - each document can belong to more than one topic.

Our topic identification algorithm is a part of a complex system for acquisition and storing large volumes of text data [1]. The system was implemented to gather the training data for the estimation of the parameters of statistical language models for natural language processing (automatic speech recognition, machine translation, etc.). Since it has been shown that not only the size of the training data is important, but also the right scope of the language models training texts is needed [2], the topic identification algorithm is used for large scale language modeling data filtering [3].

Two main approaches to the text classification can be identified - the discriminative techniques like support vector machines(SVMs) [4][5], decision trees [6] and neural

I. Habernal and V. Matousek (Eds.): TSD 2013, LNAI 8082, pp. 209–216, 2013.

networks; and generative techniques like Naive Bayes classifier (NBC) [7][8] and Expectation Maximization based methods.

This paper describes a method based on general topic model normalisation for finding a threshold defining the boundary between the "correct" and the "incorrect" topics of a newspaper article in the generative classification techniques. The generative classifier outputs a distribution of probabilities (or likelihood scores) and a method for processing this distribution into the sets of the "correct" and the "incorrect" topics is needed. The proposed method is used to improve the results of the NBC in the topic identification module.

## 2 Multi-label Text Classification

The existing methods for multi-label classification can be divided into two main categories - *data transformation (DT) methods* and *algorithm adaptation methods*. The methods of the first group transform the problem into the single-label classification problem and the methods in the second group extend the existing algorithms to handle the multi-label data directly. According to [9] we can divide the existing *data transformation methods*:

First two methods, marked as *DT*1 and *DT2*, simply transforms the multi-label data set into single-label [10]. Method DT1 selects only one label from the multiple labels for each data instance and method DT2 discards every multi-label data instance from the set. These methods cannot be really used in a multi-label classification since they remove all the multi-label information from the data set.

The third data transformation method *DT3* considers each set of labels as one label together. The single label classifier then could be used, choosing for each data item one of the predefined sets of labels. The disadvantages of this methods are clear - first, we can end with large number of label sets with only few examples of training data for each set; and second, we cannot assign different combination of labels to the classified data than those previously seen in the training data. This method was used in the works [10][8].

The most common data transformation method *DT4* trains a binary one-vs.-rest classifier for each class. The labels for which the binary classifier yields a positive result are then assigned to the tested data item. The disadvantage of this method is that you have to transform the data set into $|L|$ data sets, where $L$ is the set of possible labels, containing only the positive and negative examples. The second disadvantage is that you have to find the threshold for each binary classifier. This method was used in [4][11][12][5] and also is often used as a baseline for other methods testing [10][8][5].

The *DT5* method decomposes each training data with $n$ labels into $n$ data items each with only one label. One classifier with the distribution of probabilities or likelihoods for all labels is then learned from the transformed data set. The distribution is then processed to find the correct labels of the data item. This approach is used in the work [13] and also in our experiments. The problem of finding the border between correct and incorrect topics is further addressed in Section 2.1.

The last method *DT6* decomposes each training data item into $|L|$ data items each with only one label $l$ and a value $Y(l)$, where $Y(l) = 1$ for the labels which belong to the data item and $Y(l) = -1$ otherwise.

The *algorithm adaptation methods* are methods handling the multi-label data directly or methods that somehow combine one of the DT methods with an existing classification method. For example work [14] uses the adapted C4.5 algorithm; the two extensions of AdaBoost - AdaBoost.MH and AdaBoost.MR were implemented with the combination of DT6 method in the work [6]; in the work [8] the DT3 method is used in combination with the Naive Bayes classifier, the distribution for the sets of label is estimated with the expectation maximization algorithm; the work [5] improves the DT4 method in the combination with SVMs; the adaptation of kNN classifier (ML-kNN) with combination of DT4 method was used in [11].

### 2.1 Threshold Definition for DT5 Method

As the topic identification module in our system uses a Naive Bayes classification algorithm (the motivation for choosing the NBC is described in Section 3.1) we tried to find out some related work on the problem how to select the set of correct topics from the output distribution of the NBC. A straightforward approach is to select the labels for which the likelihood is greater than a specific threshold (e.g. 0.5) or select a predefined number of topics. In the work [7] the training data with only one label was selected (methods DT1 or DT2) and only the one best label is assigned to each news article, therefore it could not be considered a multi-label classification. In our later work, we selected 3 topics for each article [3]. To our knowledge, the only work concerning the finding of a threshold for choosing the correct topics in the output of a distribution classifier is described in [13]. The classifier used in this work outputs a likelihood distribution of topics for the tested article and the dynamic threshold is set as the mean plus one standard deviation of the topic likelihoods. The assumption is that topics that have a likelihood greater than this threshold are the best choices for the article. The method for finding a threshold proposed in this paper is described in Section 3.1.

## 3 System for Acquisition and Storing Data

The topic identification module is a part of a system designed for collecting a large text corpus from Internet news servers described in [1]. The system consists of a SQL database and a set of text processing algorithms which use the database as a data storage for the whole system. One of the important features of the system is its modularity - new algorithms can be easily added as modules.

For the topic identification experiments the most important parts of the system are the text preprocessing modules. Each new article is obtained as a HTML page, then the *cleaning* algorithm is applied - it extracts the text and the metadata of the article. Then the *tokenization* and *text normalization* algorithms are applied - text is divided into a sequence of tokens and the non-orthographical symbols (mainly numbers) are substituted with a corresponding full-length form. The tokens of a normalized text are processed with a *vocabulary-based substitution* algorithm. Large vocabularies prepared

by experts are used to fix the common typos, replace sequences of tokens with a multi-word or to unify the written form of common terms. *Decapitalization* is also performed - substitutes the capitalized words at the beginning of sentences with the corresponding lower-case variants. The output of each of the preprocessing algorithm is stored as a text record in the database.

Lemmatization has been shown to improve the results when dealing with sparse data in the area of information retrieval [15] and spoken term detection [16] in highly inflected languages, on that account the experiments on the effects of lemmatization in the field of topic identification was performed [17]. As a result of these experiments the automatic *text lemmatization* is also applied in our work. The lemmatization module uses a lemmatizer described in the work [18]. The lemmatizer is automatically created from the data containing the pairs full word form - base word form. A lemmatizer created in this way has been shown to be fully sufficient in the task of information retrieval [18].

### 3.1 Topic Identification Module

The purpose of the topic identification module in our system is to filter the huge amount of data according to their topics for the future use as the language modeling training data. So far, the topic identification module (which is further described in [3]) used a Naive Bayes based classification algorithm and assigned 3 topics chosen from a hierarchical system - a "topic tree" to each article.

The topic hierarchy built in a form of a topic tree is based on our expert findings in topic distribution in the articles on the Czech favorite news servers like *ČeskéNoviny.cz* or *iDnes.cz*. The topic tree has 32 generic topic categories like `politics` or `sports`, each of this main category has its subcategories, the deepest path in the tree has a length of four nodes. Totally it contains about 450 topics and topic categories, which correspond to the keywords assigned to the articles on the mentioned news servers. The articles with these "originally" assigned topics are used as training texts for the identification algorithm.

**Identification Algorithm.** Current version of the topic identification module uses a multinomial Naive Bayes classifier (NBC), chosen due to the results of experiments published in [3]. NBC is known to be the fastest learning classifier [5], although having worse accuracy than support vector machines (SVMs), for our task is the best possible choice. As mentioned before, our topic identification runs in a real application. The articles are stored in a database, so the "training" of the identification is done simply by counting the statistics containing the number of occurrences of each word in the whole collection, number of occurrences of each word in each document and the number of occurrences of each word in the documents belonging to a topic.

New articles are downloaded every day and they are instantly processed - the articles which we use as training data since they have the "originally" assigned keywords are used to update the word occurrence statistics tables - as a result, our topic training data update every day. To the rest of the downloaded articles the topic identification module assigns the topics from our topic hierarchy. Every day more than 600 new articles are

downloaded to our database and they contain more than 130 new topic training articles, so we had to choose the topic identification algorithm which will be fast and can use the easily updatable statistics stored in the database tables as the trained classifier data. This is why we have chosen to use the NBC over the SVMs.

In the Naive Bayes classifier the probability $P(T|A)$ of an article $A$ belonging to a topic $T$ is computed as

$$P(T|A) \propto P(T) \prod_{t \in A} P(t|T) \tag{1}$$

where $P(T)$ is the prior probability of a topic $T$ and $P(t|T)$ is a conditional probability of a term $t$ given the topic $T$. The probability is estimated by the maximum likelihood estimate as the relative frequency of the term $t$ in the training articles belonging to the topic $T$:

$$\hat{P}(t|T) = \frac{tf_{t,T}}{N_T} \tag{2}$$

where $tf_{t,T}$ is the frequency of the term $t$ in $T$ and $N_T$ is the total number of tokens in articles of the topic $T$. The uniform prior smoothing was used in the estimation of $P(t|T)$.

The goal is to find the most likely or the maximum a posteriori topic (or topics) $T$ of an article $A$ - for each article the topics with the highest probability $P(T|A)$ are chosen:

$$T_{map} = \arg\max_T \hat{P}(T|A) = \arg\max_T \hat{P}(T) \prod_{t \in A} \hat{P}(t|T) \ . \tag{3}$$

The prior probability of the topic $\hat{P}(T)$ was implemented as the relative frequency of the articles belonging to the topic in the training set, but we found out that it has only small to no effect on the identification results.

**General Topic Model Normalisation Method for Finding the Dynamic Threshold.** In our topic identification module we use the combination of the data transformation method DT5 (the article is used as training data for each topic label it has) and the threshold for the selection of the topics to assign to an article. So far we have been selecting the best 3 topics for each article. This is not the best way, because some short articles can concern only one topic, on the other hand some long articles, especially from the politics category often incorporate many other topics. The right way to select the "correct" topics for an article would be setting a dynamic threshold, which should be somehow dependent on the article topic likelihood distribution.

The *General topic model normalisation method (GTMN)* for finding the threshold we propose is inspired by the World model normalisation technique (WMN) used in the speaker recognition task [19][20]. The multinomial NBC is formally equal to the language modeling approach in the information retrieval [21], each topic is described by an unigram language model. In addition to the different topic models, a general topic model is also created as a language model of the whole collection.

First, the NBC classifier is used to output a likelihood topic distribution. Then, the topic likelihood scores $\hat{P}(T|A)$ are normalised with the score of the general model $\hat{P}(G|A)$:

$$\hat{P}(T|A)_{GTMN} = \frac{\hat{P}(T|A)}{\hat{P}(G|A)} \tag{4}$$

Now we have a list of the likelihoods normalised by the general topic model, specifically we have the list of how better the topics describe the article in comparison with the general topic model. We select only the topics which are better scoring than the general topic model and we make the assumption that the topics which have at least 80 percent of the normalised score of the best scoring topic are the "correct" topics to be assigned.

## 4 Evaluation

In this section the proposed General topic model normalisation method for finding the threshold is compared to the previously used selection of 3 topics for each article and also to selection only one topic as used in [7] and setting the threshold as the mean plus one standard deviation (MpSD) of the topic likelihoods used in the work [13]. For the experiments the smaller collection containing the articles from the news server *ČeskéNoviny.cz* separated from the whole corpus was used [17]. The collection contains 31 419 articles, divided into 27 000 training and 4 419 testing articles.

The evaluation of the result of the multi-label classification requires different metrics than those used in evaluation of single-label classification. We have chosen the metrics somewhat similar to the evaluation used in the field of information retrieval (IR), where each newly downloaded article is considered to be a query in IR and precision and recall is computed for the answer topic set. Similar measures was used in [5] and [10]. For the article set $D$ and the classifier $H$ precision ($P(H, D)$) and recall ($R(H, D)$) is computed:

$$P(H, D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{T_C}{T_A}, \qquad R(H, D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{T_C}{T_R} \tag{5}$$

where $T_A$ is the number of topics assigned to the article, $T_C$ is the number of correctly assigned topics and $T_R$ is the number of relevant reference topics. The $F_1(H, D)$-measure is then computed from the $P(H, D)$ and $R(H, D)$ measures:

$$F_1(H, D) = 2\frac{P(H, D) \cdot R(H, D)}{P(H, D) + R(H, D)}. \tag{6}$$

The results of our experiments are shown in Table 1, from which we can draw following conclusions:

– When choosing only one topic, the precision is quite high, because the first topic is usually correct.

**Table 1.** Comparison of different threshold finding methods

| metric / method(H) | 1 topic | 3 topics | MpSD | GTMN |
|---|---|---|---|---|
| $P(H, D)$ | 0.8123 | *0.5859* | 0.0554 | **0.5916** |
| $R(H, D)$ | 0.3191 | *0.6155* | 0.9611 | **0.6992** |
| $F_1(H, D)$ | 0.4582 | *0.6003* | 0.1048 | **0.6409** |

- The MpSD method achieves high recall, because it selects about 50 topics for each article, on the other hand precision is really low. We believe it is because the method was proposed for the document collection with only 10 topics, unfortunately in our case (450 topics) the method fails.
- The proposed GTMN method achieved the best results and we believe it is more universal than the MpSD method, since, thanks to the general topic model normalisation, the topic set can be of any size.

## 5 Conclusions and Future Work

The performed experiments with the topic likelihood threshold finding for distribution classifiers suggest that the new proposed General topic model normalisation method for finding the threshold performs better than other previously published tested methods. We have done the same evaluation on a different collection of documents separated from our database and the results were the same. In the future work, we will test the proposed method on other collections with different number of topic categories to confirm the universality of this method.

The advantage of the hierarchical organization of the topics is currently used only for the selection of documents to be used as the training data for the estimation of the parameters of statistical language models for natural language processing. For the future work, we would like to take the advantage of hierarchical topic tree and the relations between the topics also in the topic identification algorithm as described in [5].

**Acknowledgments.** The work has been supported by the Ministry of Education, Youth and Sports of the Czech Republic project No. LM2010013 and by the University of West Bohemia, project No. SGS-2013-032.

## References

1. Švec, J., Hoidekr, J., Soutner, D., Vavruška, J.: Web text data mining for building large scale language modelling corpus. In: Habernal, I., Matoušek, V. (eds.) TSD 2011. LNCS, vol. 6836, pp. 356–363. Springer, Heidelberg (2011)
2. Psutka, J., Ircing, P., Psutka, J.V., Radová, V., Byrne, W., Hajič, J., Mírovský, J., Gustman, S.: Large vocabulary ASR for spontaneous Czech in the MALACH project. In: Proceedings of Eurospeech 2003, Geneva, pp. 1821–1824 (2003)
3. Skorkovská, L., Ircing, P., Pražák, A., Lehečka, J.: Automatic topic identification for large scale language modeling data filtering. In: Habernal, I., Matoušek, V. (eds.) TSD 2011. LNCS, vol. 6836, pp. 64–71. Springer, Heidelberg (2011)

4. Joachims, T.: Text categorization with support vector machines: Learning with many relevant features. In: Nédellec, C., Rouveirol, C. (eds.) ECML 1998. LNCS, vol. 1398, pp. 137–142. Springer, Heidelberg (1998)
5. Godbole, S., Sarawagi, S.: Discriminative methods for multi-labeled classification. In: Dai, H., Srikant, R., Zhang, C. (eds.) PAKDD 2004. LNCS (LNAI), vol. 3056, pp. 22–30. Springer, Heidelberg (2004)
6. Schapire, R.E., Singer, Y.: Boostexter: A boosting-based system for text categorization. In: Machine Learning, pp. 135–168 (2000)
7. Asy'arie, A.D., Pribadi, A.W.: Automatic news articles classification in indonesian language by using naive bayes classifier method. In: Proceedings of the 11th International Conference on Information Integration and Web-based Applications & Services, iiWAS 2009, pp. 658–662. ACM, New York (2009)
8. McCallum, A.K.: Multi-label text classification with a mixture model trained by em. In: AAAI 1999 Workshop on Text Learning (1999)
9. Tsoumakas, G., Katakis, I.: Multi-label classification: An overview. Int. J. Data Warehousing and Mining, 1–13 (2007)
10. Boutell, M.R., Luo, J., Shen, X., Brown, C.M.: Learning multi-label scene classification (2004)
11. Zhang, M.L., Zhou, Z.H.: A k-nearest neighbor based algorithm for multi-label classification. In: 2005 IEEE International Conference on Granular Computing, vol. 2, pp. 718–721 (2005)
12. Yang, Y.: An evaluation of statistical approaches to text categorization. Journal of Information Retrieval 1, 67–88 (1999)
13. Bracewell, D.B., Yan, J., Ren, F., Kuroiwa, S.: Category classification and topic discovery of japanese and english news articles. Electron. Notes Theor. Comput. Sci. 225, 51–65 (2009)
14. Clare, A., King, R.D.: Knowledge discovery in multi-label phenotype data. In: Siebes, A., De Raedt, L. (eds.) PKDD 2001. LNCS (LNAI), vol. 2168, pp. 42–53. Springer, Heidelberg (2001)
15. Ircing, P., Müller, L.: Benefit of Proper Language Processing for Czech Speech Retrieval in the CL-SR Task at CLEF 2006. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 759–765. Springer, Heidelberg (2007)
16. Psutka, J., Švec, J., Psutka, J.V., Vaněk, J., Pražák, A., Šmídl, L., Ircing, P.: System for fast lexical and phonetic spoken term detection in a czech cultural heritage archive. EURASIP J. Audio, Speech and Music Processing (2011)
17. Skorkovská, L.: Application of lemmatization and summarization methods in topic identification module for large scale language modeling data filtering. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2012. LNCS, vol. 7499, pp. 191–198. Springer, Heidelberg (2012)
18. Kanis, J., Skorkovská, L.: Comparison of different lemmatization approaches through the means of information retrieval performance. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2010. LNCS, vol. 6231, pp. 93–100. Springer, Heidelberg (2010)
19. Sivakumaran, P., Fortuna, J., Ariyaeeinia, M.A.: Score normalisation applied to open-set, text-independent speaker identification. In: Proceedings of Eurospeech 2003, Geneva, pp. 2669–2672 (2003)
20. Zajíc, Z., Machlica, L., Padrta, A., Vaněk, J., Radová, V.: An expert system in speaker verification task. In: Proceedings of Interspeech, vol. 9, pp. 355–358. International Speech Communication Association, Brisbane (2008)
21. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, New York (2008)