Greedy Term Selection for Document Classification with Given Minimal Precision

Kristóf Csorba, István Vajk

Department of Automation and Applied Informatics
Budapest University of Technology and Economics
Goldmann György tér 3, H-1111 Budapest, Hungary
kristof@aut.bme.hu, vajk@aut.bme.hu

Abstract: In this paper we present a new term selection technique for document clustering. Classifiers are trained to recognize documents of one single topic. The technique is designed to maintain a predefined minimal precision, while maximizing the recall of the classification. The most useful terms are collected in a greedy way. The minimal precision is ensured due to a minimal document score for selection, which is calculated based on a separate training set after the term set optimization.

Keywords: document classification, term selection, minimal precision

1 Introduction

Document clustering is the classification of natural language documents into groups, based on a given classification criteria, like topic similarity. In a supervised learning scenario, the system extracts features from labeled examples and learns to identify documents of the same topics. A large family of methods is based on vector spaces, where documents are represented by vectors in a space of features, like occurrences of various terms. Every used term (not every term is useful for the classification) is assigned to a feature, and the coordinate of the document along this dimension is a function of the occurrence of the term in the documents. A frequently used weighting scheme family is the TFIDF (term-frequency, inverse document frequency) scheme [1].

Most document classification systems aim to achieve the best classification accuracy in terms of the well known F-measure. Unlike those, our system was designed to achieve a fixed minimal precison and to maximize the recall under these conditions. That means that it does not have to maximize both recall and precision: precision has to be kept only above a predefined limit. This approach is motivated by information retrieval systems, where there is no strict need to classify all the documents, but discarding an ambiguous case is much more

acceptable, than a misclassification. For instance if we are searching for documents in a given topic on the web, we do not want to see all the documents in the target topic. But seeing some other topics among the selected documents is very disturbing.

The system is designed to identify documents of one given target topic at a time. If someone wants to classify documents into multiple categories with this Single Class Identification approach, a separate classifier will have to be employed for every topic.

The identification of documents in a given target topic (document class) is performed based on a scoring system. Every document is assigned a score, which is the number of observed terms appearing in the document. Observed terms are those terms, which are used by the system. As there are many words, which do not provide significant information about the target topic, the most important part of the training of a classifier is the identification of the terms, which are worth observing in the documents to classify. (The term selection method searches usually more useful term, than needed, as the optimal number of terms is decided in a later step of the training.) Those documents, which have a score higher than a minimal score limit, get selected. The score limit is calculated during the training phase and is used to ensure the predefined minimal precision.

The remaining part of the paper is organized as follows: Section 2. describes the document representation used by the classification system, Section 3. explains the selection of the observed terms and Section 4. describes the method for the minimal score calculation and the optimization of the number of observed terms. Measurements are presented in Section 5. and conclusions are made in Section 6.

2 Document Selection Technique

Documents are represented in the form of a term document matrix. Each document is represented by a column and each term (appearing at least in one document) is represented by a row. Binary values are used in the matrix, as we use only the information, that a given term is present in a document or not. The exact appearance number is not needed. If the *i*-th term is present in the *j*-th document, than $x_{i,j}=1$.

Binary representation was selected, as we belive, that if a word is mentioned in a document, than the topic is in connection with this word and the number of occurrences takes only noise into the term document matrix. Further more the score calculation uses only the presence as well.

3 Term Selection Technique

The term selection phase of the classifier training aims to select the term set, which can be used to achieve the highest recall, while maintaining the given minimal precision limit. As theoretically every possible term set is a candidate, a brute force searching method cannot be applied. Instead of a brute force search, the optimal set is searched in two steps: at first as many as possible useless terms are removed to reduce the search space. Useless terms are identified on the basis of two conditions:

- If the relative frequency of a term among the documents in the target topic is not much higher, than in the other topics, than it will not be able to help in the separation of the target document class.
- If the frequency of a term in the target topic is below a given limit (like 1%), it is not frequent enough to capture documents.

These two condiditions are easy to check in the term document matrix and are capable to remove many useless terms.

After the term prefiltering, a greedy term set creation is started: in every iteration we check every candidate term for the best achievable recall if it is included in the term set. Temporarily every term is added to the set and the minimal score is calculated based on the training set to maintain the minimal precision. After this, recall is measured in the training set and registered for the term. The term with the highest achieved recall is added permanently to the term set and a new iteration is started. The procedure ends, if there are no more suitable candidate terms or a predefined maximal term set size is achieved.

Pseudo code for an iteration of the greedy term set creation:

```
T = 0                       // Set of terms
A = AllAvailableTerms       // Collect all availablt terms
while |A| > 0               // While there are candidate terms
         foreach tau in A   // For every candidate term
                  R(tau)=recall(T+{tau})
         end foreach
         best = arg max{R(tau)}
         T = T + best
         A = A - best
end while
return T
```

In the pseudo-code the function *recall* returns the best recall value, which is achievable by setting the minimal score limit to ensure the predefined minimal precision.

4 Score Limit and Term Number Optimization

The term set is created based on the training set. As the increment in the minimal score limit leads usually to a decrement in the recall (less documents achieve the new score limit, although a new term has been added), selected terms tend to have almost 100% precision in the training set. This is trivial, as the best terms are selected. (Especially in the beginning of the greedy iteration.) This enables the usage of a low minimal score and allows a high recall. Unfortunately these selected terms have high precision in the testing set as well, but often lower, than 100%. That means, that the minimal score limit will be underestimated, because the calculation relies too strongly on the documents in the training set. This effect leads to lower precisions, which means the system cannot achieve the predefined minimal precision.

To avoid this effect, after the terms are collected, the minimal score is recalculated, based on a new training set. (This is usually done be dividing the training set into two parts: one for the term set creating and one for the calculation of the minimal score.) This leads to a classifier, which is much less sensitive to the dissimilarities between the training and the testing document sets. The minimal score calculated during the greedy term set creation and the final recalculation is shown in Fig. 1.

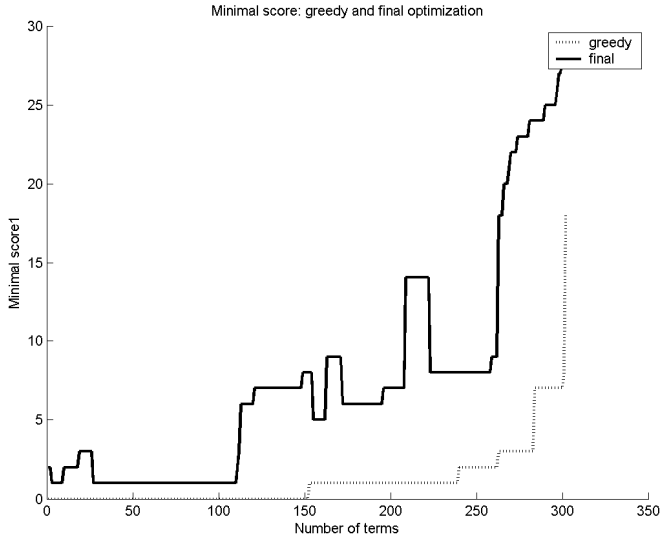

Figure 1
Estimated, and finally calculated minimal score

After creating the term set and the minimal score limit (for every term number), the only value remaining to be optimized is the number of terms. As it can be seen in the example in Fig. 2, every time the minimal score has to be increased, the recall drops. If the limit remains the same while new terms are added, recall increases again, until the limit has to be increased again to preserve the precision above the predefined limit. The question is, how can we determine the optimal term number based only on the minimal score values.

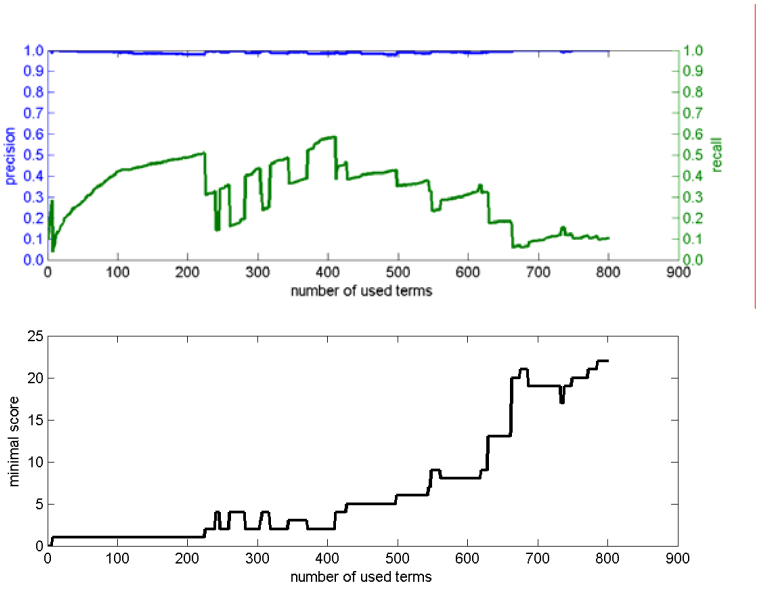

Figure 2
Recall and the minimal score, which maintains the high precision

The first strategy is the following: we assume, that a term number under 10 cannot be optimal, as it is unlikely, that a classifier observing less, than 10 words can identify the documents of a given topic. So we look for the forst term number over 10, which after the minimal score limit is increased. We assume, that this term number will be the one of the first local maximum of the recall. Based on our observations, recall values higher than this value usually need much more terms, which would cause the system to have worse speed performance.

The second strategy is an extended form of the first one: after finding the first increment in the score limit, we scan the whole search space for the highest term number, which has a lower or equal limit, than the one of the first increment. (In the example on Fig. 3, this value is around 290.)

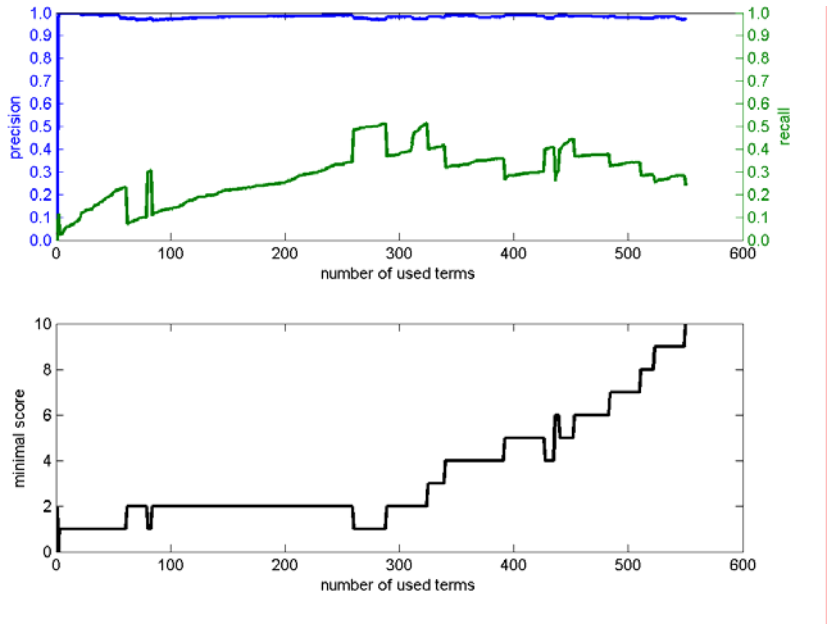

Figure 3
Optimizing the term number: low minimal score may return for higher term numbers

The classifier training is complete, if the term set has been created and the optimal term number and minimal score has been calculated. If there are multiple document topics to separate, each of them has to be a classifier trained for.

5 Experiments

The system was tested on the 20Newsgroups data set [2] by employing the topic grouping suggested in [4]. The grouping of the topics means the merging of some topics being very close to each other.

Figures 4 and 5 show the results achieved with a minimal precision of 98% and 90% to maintain. The high precision limit allows only lower recall. Some categories seem to be easy to identify (recall is over 60%), but some of them are hard cases. The imprevement of these hard cases is subject to further research. The term number optimization is performed based on the second strategy described earlier.

The number of term used by the system varies strongly, depending on the number of important terms shared between the target topic and the other topics.

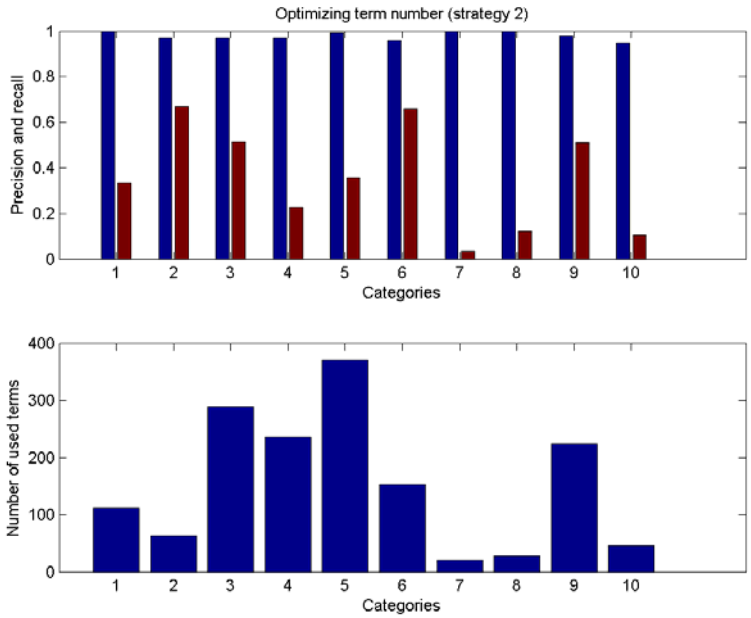

Figure 4
Results for the 10 topics: precision (kept over 98%), recall, and the number of used terms

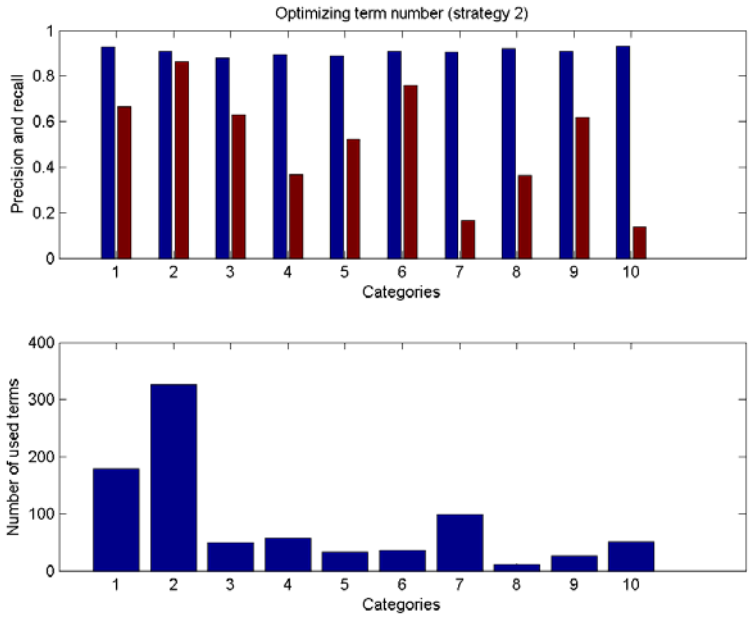

Figure 5
Results for the 10 topics with lower precision limit (90%)

Conclusions

The document classification system presented in this paper aims to maintain a fixed minimal level of precision, while searching for the optimal term set for the recall maximization. Based on the results on the 20Newsgroups data set, the achieved recall for at least 98% precision varies between 20% and 60%. Lower minimal precision allows higher recall values, but the most important capability of the proposed method is the ability to keep precision above a predefined limit.

Acknowledgments

This work has been fund of the Hungarian Academy of Sciences for control research and the Hungarian National Research Fund (grant number T042741).

References

[1] A. Singhal: Modern Information Retrieval: A Brief Overview, in IEEE Data Engineering Bulletin, 2001, Vol. 24, No. 4, pp. 35-43

[2] K. Lang: Newsweeder: Learning to Filter Netnews, in ICML, 1995, pp. 331-339

[3] L. Li, W. Chou: Improving Latent Semantic Indexing-based Classifier with Information Gain, Tech. Rep., May 16, 2002

[4] R. Bekkerman, R. El-Yaniv, N. Tishby, Y. Winter: On Feature Distributional Clustering for Text Categorization, ACM SIGIR'01, 2001