

Topic Classification Using Limited Bibliographic Metadata

Kerstin Denecke ¹, Thomas Risse ², Thomas Bähr ³

^{1,2}L3S Research Center, University of Hannover, Germany

³Technische Informationsbibliothek Hannover, Germany

¹denecke@L3S.de

Abstract. In this paper, we introduce a method for categorizing digital items according to their topic, only relying on the document's metadata, such as author name and title information. The proposed approach is based on a set of lexical resources (e.g., journal titles, conference names), on terminology extraction and traditional machine-learning technologies. Evaluation results on a real world data set show that the approach achieves promising results.

1 Introduction

Specialized public libraries, such as the British Library or the German National Library of Science and Technology (TIB), and commercial information providers offer barrier-free access to resources and an optimized, user-oriented search interface. To narrow the information space for searching and browsing, knowledge about a document's topic is required that can be provided by controlled classifications schemes and index terms. Due to the monthly increasing amount of catalogue entries, manual classification of all data items is impossible; automatic technologies are required. The development of such methods is one of the objectives of the LinSearch project (<http://www.linsearch.de>) which is partly funded by the German Federal Ministry of Economics and Technology (BMWi). In this paper we present some of the results of this project.

The considered classification problem is to assign one class out of 14 possible classes to a single catalogue entry of the TIB data collection. The objective of this classification is to support the user and also the retrieval process in restricting search results to those belonging to the domain of interest. A more fine-grained classification is not useful in this case. Our work focuses on classes that are especially relevant for the TIB whose data collection consists of documents on technology and engineering. Possible classes are 'computer science', 'mathematics', 'physics', 'architecture', 'chemistry', 'engineering', 'civil-', 'chemical-', 'electrical-', 'power-', 'production-', 'mechanical-', 'environment-', and 'process engineering'. The TIB data collection consists currently of around 15 million entries and every month between 10,000 and 30,000 new items are integrated. Each catalogue entry represents a journal paper, conference paper, a book or a research report and comprises a set of metadata. Depending on the amount of metadata available for a data item, we distinguish four different levels of data quality. Data of level I only offer document and publication information. If in addition the journal or conference information is available, the data item belongs to quality level II. Data items

of level III offer an abstract, and those of level IV provide classification information. In case of TIB this is the so-called "Basis Klassifikation (BK)" (base classification system, [1]). This hierarchical decimal classification system was originally developed in the Netherlands and is mainly used by libraries in the Netherlands and Germany.

2 Related Work

The most commonly used text representation for topic classification is the vector space representation, where each distinct word in a document collection acts as a feature [2]. Other classification features include syntactic [3], semantic and stylistic features such as character or word sequence frequencies [4] or n-gram frequencies [5]. Statistical features on words (term frequency and the like) as mainly used by existing approaches are insufficient in our context, since within document titles term frequencies may not differ significantly and the number of topic-related terms is reduced. Hulth and Megyesi use extracted keywords as classification features [6]. In our approach, also keywords are exploited but they are extracted from title information only and integrate additional bibliographic data to the feature set. So far, there is no study available that considers topic text classification for these specific conditions. Montej-Rez et al. exploit metadata in combination with extracted keywords and a multi-label classifier TECAT for text classification [7]. The keywords are extracted from the document itself, which is in our task unavailable.

3 Classification Approach

Our approach to address the previously described problem combines rule-based classification with machine learning techniques. First, the metadata of a given catalogue entry is checked for an available classification that can be mapped to one of the 14 desired classes (data of quality level IV). For this purpose, mapping rules have been established manually to map from assigned codes of the base classification system [1]. Data items of quality level II and III are processed in the second step where class-specific information on journal titles and conference names are exploited to assign a class label. For this purpose, lists with conference and journal titles have been collected from existing class-specific repositories (e.g., from DBLP) as well as from already classified data items for each class under consideration. In case the first two steps fail due to missing information as well as for data items of quality level I, core features are identified in the metadata and used by a machine-learning classifier.

The final feature set for the classification consists of the following attributes: (1) the first five author names, (2) the publisher information, (3) the name of the corporate creator, and (4) a class-specific score for each category. Attributes 1 to 3 are directly derived from the metadata information. To calculate the scores (attribute 4) the document title and - if available - its abstract is exploited. A class-specific score corresponds to the number of matches of keyphrases extracted from a document and a class-specific term list. For each category under consideration such a class-specific term list of domain-relevant terms and phrases has been established semi-automatically from existing resources and from already classified material. Keywords and -phrases, i.e. word groups

with a maximum of 5 words that are neither stop words nor start or end with a stop word are extracted from title and abstract information. Each extracted keyphrase is looked up in the class-specific term lists; matches are counted per class resulting in a class-specific score per category.

The resulting feature set is exploited for document classification by a machine-learning classifier. Different algorithms that are implemented in the WEKA library [8] have been tested. The LogitBoost classifier based on logistic regression performed best and is therefore used in our experiments.

4 Evaluation

In this section, we describe the results when applying the introduced algorithm to the TIB dataset. In a first experiment, manually classified documents derived from different publishers and of quality level I or II (i.e., abstracts and BK-codes are unavailable) have been exploited. In a 10-fold-cross validation with 1500 documents per category the method achieved an accuracy of 86.7%. We also tested the algorithm with different feature sets and conclude that class-specific scores are well suited as features while exploiting additional metadata such as author names or publisher is not helpful in the given context. A possible explanation for this is that in the given data collection, documents of the same author are very rare.

In an additional evaluation, a second data set with unclassified data entries was manually evaluated. The classifier was trained on 1500 documents per category. Five employees of the TIB manually evaluated documents of their special field. Specialists for the categories 'mathematics', 'mechanical engineering', 'chemistry', 'physics' and 'engineering' were involved. The other categories will be considered in future evaluations. From the 1180 data items, 82.7% were correctly classified by our approach. Documents of the domains 'chemistry' and 'physics' were almost completely correct classified (99% accuracy). Accuracy results of 70% and 87% were achieved for documents of the domains 'engineering' and 'mechanical engineering'. The worst results were achieved for the categories 'mathematics' (58% accuracy). For a more comprehensive description of the evaluation we refer to the full paper on our approach. Evaluation results of Level III and IV data will be represented in a different paper.

5 Conclusion

In this work, a text classification approach is introduced that relies only on bibliographic metadata. We show that despite this reduced semantic information good classification results are achieved. The best results are obtained when relying only on the title and abstract information. The lexical resources used by the approach can be easily extended and allow in this way an easy modification to similar classification problems. Furthermore, these term lists can be used to support the indexing process. In future work, we will test whether the assigned categories can be exploited to improve user satisfaction in document retrieval. For this purpose, the method will be integrated into the processes of the TIB to improve search facilities and support the restriction of search results.

References

1. Common Library Network GBV: Basisklassifikation (2008) <http://www.gbv.de/vgm/info/mitglieder/02Verbund/01Erschliessung/02Richtlinien/05Basisklassifikation/index>.
2. Sebastiani, F.: Machine learning in automated text categorization. In: ACM Computing Surveys, 34(1), Kluwer Academic Publishers (2002) 1–47
3. Moschitti, A., Basili, R.: Complex linguistic features for text classification: A comprehensive study. In: Proceedings of the 26th European Conference on Information Retrieval Research. (2004) 181–96
4. N. Cancedda, E.G., Goutte, C., Renders, J.: Word sequence kernels. In: Journal of Machine Learning Research. (2003) 1059–82
5. Peng, D., Schuurmans, F., Wang, S.: Language and task independent text categorization with simple language models. In: Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. (2003)
6. Hulth, A., Megyesi, B.B.: A study on automatically extracted keywords in text categorization. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL. (2006) 537–44
7. Montejo-Raez, A., Urena-Lopez, L., Steinberger, R.: Text categorization using bibliographic records: Beyond document content. In: Procesamiento del Lenguaje Natural, 35. (2005) 119–262
8. Witten, I., Frank, E.: Data Mining. Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco (2000)