

Identifying Information with Environmental Relevance in Web Sites of the Public Administration Using Automatic Classification

Franz Schenk¹, Daniel Meyerholt²

¹Coordination Center PortalU at the Lower Saxony Ministry of Environment and Climate Protection

franz.schenk@portalu.de

²Department of Business Information Systems/Very Large Business Applications at the University of Oldenburg

daniel.meyerholt@uni-oldenburg.de

Abstract

The German environmental information portal *PortalU* is a publicly financed information infrastructure. It offers a single point of entry for searches on all kinds of information with environmental relevance held by public authorities. From the point of view of the query facility in PortalU, all information sources can be treated uniformly based on the common structure of the search results. One of the key problems in PortalU is how to ensure the relevance of data that is available in a distributed index structure. Whereas some data sources, e.g. metadata catalogues, typically contain only highly relevant material with respect to the content model of PortalU, other data sources like Web indexes tend to contain more heterogeneous material. In this paper we describe how automatic classification of Web pages can be used in order to provide information with a high degree of environmental relevance.

1 Introduction

The index of Web pages in *PortalU*¹ comprises about 4 million Web pages. This collection is compiled in a collaborative process. Users from a large number of public authorities specify entry points to the Web content of their institutions. Based on this information, a Web crawler (built with *Apache Nutch*² technology) retrieves all the Web pages that can be reached from these entry points. Additional URL-patterns for the specification of inclusion or exclusion patterns can be given. In this way the search-space is manually specified by expert users. Based on this specification the web crawler of *PortalU* fetches all Web pages from the search-space and an index structure is generated from this set of documents using *Apache Lucene*³. Moreover, experts from the public administration assign smaller sets of Web pages most specific to one each of the available 21 environmental topics in *PortalU*. For example, the topics water, waste or energy are available. These sets of Web pages allow for user queries based on environmental themes.

Technically, the Web crawler gathers the Web pages for the index using pattern-based URL extraction that is based on given path schemas (i.e. *umwelt.niedersachsen.de*). This works only in situations where these patterns are used consistently. However, many Web sites do not show pattern-based path-schemas in

¹ <http://www.portalu.de/>

² <http://lucene.apache.org/nutch/>

³ <http://lucene.apache.org/>

their structure. For example, it is often the case that content management systems do not use structured path names within the URLs (i.e. *hessen.de*) but assign numeric identifiers to the pages⁴. In such cases, where no inherent semantics are available in the structure of the Web site, it is not possible to derive what area of the public administration a page belongs to. While it is obvious that pages extending the pattern *umwelt.sachsen.de/xxx* all belong to an environmental department, the same information cannot be derived from an URL like *saarland.de/79685.htm* (which in fact also belongs to an environmental department). This means that it is not always possible to define the search-space by giving path-patterns. In such cases, the search-space has to be explicitly defined. This is not possible for Web sites with thousands of pages and frequent changes to both content and structure.

A solution to this problem can be automatic classification of Web pages. The idea behind automatic classification is to teach machines how to distinguish one class of information from another class. A trained machine would be able to discriminate useful web pages with environmental relevance from un-relevant material. Instead of giving distinct path names it is then sufficient to specify entry points like *niedersachsen.de*. The effort of specifying the relevant parts of a Web site is left to the machine-based classification.

In the next section we explain the technical background of automatic classification and will show in the following section how we used this technique in our experiments. We conclude this paper with a discussion and an outlook on future work.

2 Technical Background

To approach the problem of automatic classification in the context of a web crawler like *Apache Nutch*, it was necessary to provide a comprehensive software tool that reuses some of the architectural concepts of the crawler and offers an expert user with an appropriate experimentation environment.

The used web crawler is built on top of the *Apache Hadoop*⁵ infrastructure that provides an open source implementation of Google's *MapReduce* approach (Dean/Ghemawat, 2004). In general it allows distributed storage and processing of very large data sets on a cluster of computers. Amongst the core storage and processing functionalities, several software components were developed under the umbrella of *Hadoop* to provide for example a scalable database system (*HBase*⁶) or data warehouse concepts (*Hive*⁷). Nowadays *Hadoop* is leveraged by several technological solutions that are found in the field of cloud computing and used by companies like Yahoo!, Facebook, LinkedIn, a9.com, Twitter or Last.FM which have to deal with huge amounts of data.

Nutch can be deployed on a *Hadoop* cluster and all of its operations like database processing, fetching documents from the internet, parsing the documents' contents or creating an index are executed and processed using a *MapReduce* implementation⁸. All data storages used by *Nutch* (databases, indexes and web

⁴ i.e., http://www.hessen.de/irj/RPDA_Internet?cid=b81a4975a5b727c59032e1c3ab8f42cd is a Web page about protected sites. The path name gives no indication concerning the nature of the content.

⁵ <http://hadoop.apache.org/>

⁶ <http://hbase.apache.org/>

⁷ <http://hive.apache.org/>

⁸ <http://hadoop.apache.org/mapreduce/>

page contents) are stored in the *Hadoop Distributed Filesystem (HDFS)*⁹ which allows redundant distribution of data among several nodes in a *Hadoop* cluster.

The software tool used during the experimentations of this paper uses the web page storage (segments) and the Lucene indexes created by Nutch so duplicated web page fetching, parsing and indexing could be avoided. Furthermore Lucene's Analyzer and Tokenizer components were leveraged by the software tool as these are proven for the given tasks. During the experiments it was discovered that the handling of multiple Nutch instances greatly eases the management of different web page sets so the software tool is able to access different Nutch segments and Lucene Indexes for example to divide the examined web pages according to their classes (e.g. the 21 topics of PortalU).

Apart from the integration of Nutch, a web based user interface was created to provide a comprehensive and easy to use experimentation environment. By that user interface it is possible to control each aspect of the text classification process:

- Definition of web page sets to be examined
- Association of web page sets with classes or labels
- Creation of training/test sets for further processing
- Creation of dictionaries using different feature selection methods (simple frequency based approaches, X^2 -test based and using mutual information of the classes)
- Configuring and executing classification algorithms
- Testing of the algorithm
- Application of the trained classifier models on a given set of web pages or another Nutch instance.

The creation of dictionaries is a crucial step in the application of information retrieval methods. Dictionaries are composed of features, which are mostly represented using single words or tokens. A huge dictionary is not desired and even lowers the quality of classifications due to overfitted classifiers. Additionally a dictionary that is composed of several thousand features significantly lowers the performance of algorithms applied. The selection of a smaller dictionary is done by applying a feature selection algorithm to all features extracted from the training document set. These selection algorithms assign a score to each feature and then select only the top ranking features for the creation of the final dictionary. For our experimentations we decided to create dictionaries consisting of a maximum of 200 features per class as it was shown, that a dictionary of 100-200 features performs best (Manning/Raghavan/Schütze, 2008, 254). The features are selected using the mutual information of each feature with respect to the different classes of the training documents and was developed in the software tool following (Manning/Raghavan/Schütze, 2008, 252). Feature selection is very important for information retrieval techniques and many different approaches can be used and differ in terms of applicability and quality (Rogati/Yang, 2002).

After creating the dictionary it is possible to examine the dictionary with the provided user interface and to delete features as needed or put them in a blacklist to ignore them in subsequent dictionary generations. The final dictionary is used to train the desired classification algorithm. Currently supported algorithms for text classification include a Multinomial Naïve Bayes implementation, an algorithm based on Rocchio relevance feedback, k Nearest Neighbour and Support Vector Machines. For the experiments done in this paper we decided to use Naïve Bayes and Support Vector Machines as these are two major algorithms used in the field of text classification.

The Multinomial Naïve Bayes implementation follows the proposed algorithm found in (Manning/Raghavan/Schütze, 2008, 241) and thus applies several optimizations like Laplace Smoothing and avoiding floating point underruns by using sums of logarithmic values instead of multiplying all probability factors of features.

⁹ <http://hadoop.apache.org/hdfs/>

The second classification method used for the experimentations of this paper was using Support Vector Machines (SVM). Compared to traditional Naïve Bayes methods these algorithms are more complex and sophisticated so it was decided to include the commonly used LIBSVM SVM library (Chang/Lin, 2011). LIBSVM provides a wide range of usable SVM kernels, several command line tools as well as a Java based implementation that has been used in the created software tool.

The different classification algorithms are used in a generic manner by the software tool, so it is easy to add different algorithms or implementations of a specific one. Training and testing of a selected algorithm is done using the same functionalities of the tool and the output of the quality of the test run of a configured algorithm is also presented in a common form that is used for example for the values given later in this paper.

Apart from the different algorithms to be used, it is possible to classify using a one class of all or a multi class approach: A web page that is being classified could be assigned only one label (e.g. environmentally related or not) or can have multiple labels (e.g. waste, soil, chemicals). The current implementation does not allow the multiple label approach when using Support Vector Machine classification.

Another unit of the created software system that is currently under development is a Plug-In for Nutch that allows the application of trained classifiers to web pages that are indexed by a Nutch instance. By using this Plug-In automatic document classification is implemented.

3 Results

As automatic classification is a very large field of research in artificial intelligence, we have limited our experiments to two of the most popular algorithms: Multinomial Naive Bayes and Support Vector Machines. We tested these algorithms in several different scenarios. In this section we describe how we used these algorithms and what results could be found.

3.1 Dictionaries

The initial step in automatic classification is the training of the algorithm. The training needs two sets of Web pages that can be used for the creation of two separate dictionaries. Each dictionary contains a vocabulary that is typically for one set of documents and makes it different from the other set. The score of a term in a dictionary depends on its frequency in both document sets. If the frequency is high in one set but low in the other set, the score will be high. In order to compute as many of these high-scored terms it is very important to compile the document sets carefully. We needed one set comprising relevant information about environmental topics (set A) and another set comprising non-relevant information (set B).

We decided to compile set A from all Web pages of the Web site *umweltbundesamt.de*. This site covers a wide range of different information, all with environmental relevance. Moreover, the information is provided by a public administration. This is important because the trained algorithm is also intended for the classification of Web pages from the public administration. A source with comparable vocabulary and structure is an ideal candidate. The site comprises more than eight thousand documents; roughly five thousand of them are Web pages. We limited our experiments to Web pages, mainly for performance reasons.

The compilation of set B was harder to achieve. We started with a large set of pages derived from the german pages of the Open Directory Project¹⁰. This collection of Web pages covers a wide range of topics

¹⁰ The pages have been extracted from the constantly updated RDF representation of the directory which can be found at <http://www.dmoz.org/rdf.html>

and seemed suitable because of its different structure and broader scope compared to the site *umwelt-bundesamt.de*. The resulting dictionary was disappointing, though. It contained only a small number of useful words and the preliminary testing results of the trained algorithms were poor.

We derived that the set of irrelevant documents must contain as little information about environmental topics as possible while still covering a broad range of topics and comprising a large number of documents. The main problem here is that “environment” is a cross-sectoral topic and is often subsumed under other topics. Our analysis of several alternatives, looking for Web sites covering a broad range of topics and providing relevant content brought us to several German news sites (e.g. *zeit.de*, *spiegel.de*). But none of these sites did allow for a distinction between environmental and non-environmental information. This, however, is the most important prerequisite for the training, as set B has to be as disjunctive from set A as possible.

Eventually, we found a suitable candidate in the site *sachsen.de* of the federal state Saxonia. This site consists of approximately eleven thousand Web pages and shows a highly structured content in so far as administrative areas can be distinguished by subdomain names. Hereby, subdomains with environmental relevance could be excluded from the set (see Appendix B for a list of the excluded subdomain names). The resulting dictionary showed more relevant terms compared to the previous test.

One goal in our experiments was to test the influence of set B on both the quality of the dictionary and the classification. Therefore, we extended our initial set B by adding the Web site *rlp.de* of the federal state Rhineland-Palatinate. Now we had two sets to compare: set B₁ comprising the Web pages of *sachsen.de* and set B₂ comprising set B₁ and the additional Web pages of *rlp.de*.

	A	B ₁	A	B ₂
Number of Web pages	5131	11527	5131	21368
Size of dictionary	176	85	153	69

Table 1: Sets of Training Documents

For the combinations A-B₁ and A-B₂ four different dictionaries are created while classifying the training documents. Note that there are two different dictionaries for set A although the document set remains the same. The differences between these two dictionaries of the Set A as well as words that B₁ and B₂ do not share are listed in Appendix A. Future experiments will inspect differences between dictionaries in more detail but are out of scope of this paper.

Although the initial dictionaries are created automatically by the software they can be edited by hand afterwards. We call this phase manual training of the dictionary. During this phase we removed ordinary entries like numbers and dates (e.g. *2011*, *Januar*), but also some phrases or names (e.g. *nicht*¹¹, *sachsen*). There are always words that are very specific to one of the training sets but cannot be found in the other set. These would be ideal candidates for a dictionary unless they are very specific for the training set only. A good example is the name *Kurt Beck*¹². Both parts of this name have been included in the B₂ dictionary because they are very common in the documents of set B₂ but they do not occur at all in set A. Nevertheless, it is not likely that these words will be useful during future classifications on different sets of Web pages (e.g. *niedersachsen.de*). Only words that are both common and typical for non-environmental documents should be included in the dictionaries. Manual training has to add this kind of knowledge that is

¹¹ not (engl. transl.)

¹² the Minister President of Rhineland-Palatinate

not available to a machine having no further knowledge about the world. The resulting sizes of the dictionaries are shown in Table 1; the words from the dictionaries are given in Appendix A.

We used the dictionaries that were obtained from combinations of set A with sets B₁ and B₂ in the training of the algorithms, which is described in the next section.

3.2 Training and Testing of the Algorithms

With the given dictionaries the algorithms can be trained. Training is an automatic process and needs no interaction. In the case of *Support Vector Machines*, however, there are numerous parameters for controlling the training. The influences of these parameters on the result of classification are very complex and would justify for an experiment of its own. In our tests, we followed the default approach as given in LIBSVM's accompanying literature (Hsu/Chang/Lin, 2010).

The trained algorithm can be tested on subsets of the training sets A and B₁ or B₂. The typical procedure is to use only two thirds of the training sets for training so that the other third can be used for testing the quality of the trained algorithm. Hereby, a comparable but still independent set of Web pages is available. The results for the *Naïve Bayes (NB)* algorithm are shown in Table 2:

	A + B ₁		A + B ₂	
	A	B ₁	A	B ₂
Precision	0.969	0.871	0.918	0.927
Recall	0.656	0.991	0.726	0.982
F1	0.783	0.927	0.811	0.954
Micro averaged F1	0.891		0.925	
Macro averaged F1	0.855		0.882	

Table 2: Training Results for Naïve Bayes

Precision and *recall* are typical indicators in the evaluation of information retrieval processes. They are inversely proportional to the numbers of *false-positives* and *false-negatives*, respectively. *F1* is the weighted harmonic mean (or F measure) and trades off precision versus recall. Macro averaged *F1* values are based mainly on the arithmetic mean of the *F1* values of all classified classes. The micro averaged *F1* value is calculated from the summarized *fp*, *fn*, and *tp* values from all classified classes. The micro averaged *F1* value better suited if there is a great difference in size between the classes of documents.

The test results were quite good for all sets of documents when compared to typical values found in the literature (Manning/Raghavan/Schütze, 2008, 261). It is planned to improve these results further by applying additional techniques to the algorithm. Some promising approaches are given in literature (Rennie/Shih/Teewan/Karger, 2003, 616-623).

Looking at the sets of documents separately we see that values for sets B₁ and B₂ are better compared to the results for set A. This supports our intuition that environmental information is a strongly cross-sectional topic. Often, environmental information can also be regarded in a more general sense, i.e. as a social or economic topic. Therefore, it is easier to determine the general nature of a document than to proof its environmental relevance.

Results for *F1 values* are better for sets B₁ and B₂ than for set A. We discuss this observation in Section 4.1.

More important is the comparison of the test results of the combinations $A + B_1$ with $A + B_2$. All *F1 values* are better after the training with the larger set B_2 compared to the training with its subset B_1 . The micro-averaged means are higher than the macro-averaged means because of the great difference in size between set A and both sets B_1 and B_2 .

We used the same procedure for the training of the *Support Vector Machines (SVM)* algorithm. There are a number of parameters available that can be used to influence the training process. Moreover, there are some optimization strategies that automatically adjust parameters in the experiment. Namely we created document vectors using the *tf-idf* (term frequency, inverse document frequency) calculation scheme for the elements of the vectors which correspond to the used dictionary. The values are then scaled for the whole training set to values between -1 and +1 and afterwards fed to the *SVM* training process which uses a radial basis function (*RBF*) kernel. The needed parameters for the training process have been derived using the tool “svm-grid” of the LIBSVM distribution. The characteristics of the trained *SVM* algorithm on the sets of test documents are shown in Table 3.

	$A + B_1$		$A + B_2$	
	A	B_1	A	B_2
Precision	0.798	0.761	0.855	0.934
Recall	0.321	0.964	0.713	0.978
F1	0.458	0.851	0.789	0.955
Micro averaged F1	0.766		0.926	
Macro averaged F1	0.654		0.872	

Table 3: Training Results for Support Vector Machine

Surprisingly, *SVM* did not perform better than *NB*. Although all *F1*-values are similar in the experiment $A+B_2$ when comparing *SVM* and *NB*, there is a big difference in the experiment $A+B_1$. Although the trend in the differences between the experiments is the same between the algorithms, the absolute values of the differences are much higher for *SVM* (see Table 4:). This shows that *SVM* was much more sensible to alterations of the dictionaries in our experiments. The main contribution to the bad performance in the experiment with document set B_1 and *SVM* came from a high number of *false-negatives*, therefore the low recall-value (see Table 3).

	NB			SVM		
	$A + B_1$	$A + B_2$	$(A+B_2)-(A+B_1)$	$A + B_1$	$A + B_2$	$(A+B_2)-(A+B_1)$
Micro F1	0.891	0.925	0.034	0.766	0.926	0.160
Macro F1	0.855	0.882	0.028	0.654	0.872	0.218

Table 4: Comparison of F1-values

After training the algorithms we evaluated the trained algorithms using the combination of sets A with B_2 and A with B_1 . The results of this evaluation are shown in the following section.

3.3 Application of the Trained Algorithms

We tested our trained algorithms on real world data in order to verify the quality of the automatic classification process. Therefore, we needed a further set of web pages in order to examine the quality of the

trained algorithms. This set C consists of the Web site *niedersachsen.de*, which is the Web site of the federal state Lower Saxonia and comprises more than 140.000 Web pages. It shows, like the sites *sachsen.de* and *rlp.de*, an evident structure that is reflected in useful URL patterns. Altogether, we identified 19 subdomains (e.g. *umwelt*, *naturschutzgebiete*, and *numis*, see Appendix B for a complete list) that consist mainly of environmental information. With this *a priori* knowledge about domain structure we were able to verify our trained algorithms.

We define that C_1 and C_2 are subsets of C , where C_1 contains the pages from all subdomains listed in Appendix B, and C_2 consists of web pages from all other subdomains. C_1 can be expected to consist mainly of relevant information; nevertheless it cannot be assumed that C_2 does not also contain some relevant information. This is because environmental affairs affect all areas of administration in one way or another. This will have some consequences when validating the results.

The quality of the classification can be quantified using these sets. On the one hand, positively classified documents in C_1 are counted, called *true-positives* (tp). On the other hand, unclassified documents in C_2 (called *true-negatives* (tn)) give another measure. Moreover, there are *false-positives* (classified pages in C_2 , fp) and *false-negatives* (unclassified pages in C_1 , fn).

It is necessary, however, to keep in mind that it is not possible in our experiment to tell the precision of fn and fp . The decision to define relevance of a page depending on the membership to one of the listed subdomains does not allow for an absolute measure. Some of the documents in fp will have environmental relevance despite not belonging to one of the relevant subdomains. For example, a page found at the Web site of Lower Saxonies Social Ministry, which gives detailed information about Tchernobyl¹³, has been classified as being highly relevant. Nevertheless this document is one of the *false-positives* in our experiment. This means that this page has been wrongly classified in the context of the setup of our experiment, but with respect to the content of this document the classification was correct. Moreover, there are pages in fn that have not been classified for a good reason and should belong to tn . For example, there are many navigational pages that have no real content but contain only links. Still, they are counted as *false-negatives* because they belong to one of the relevant subdomain. If only their content is considered they would be regarded as *true-negatives*. All these cases are, however, hard to quantify and can only be described by control samples.

In our tests we used 141813 Web pages of the Web site *niedersachsen.de*. From these pages, 35570 belonged to one of the relevant subdomains (see Appendix B for the subdomain names). The results for *NB* and *SVM* are given in Table 4.

Although we expected the *SVM* algorithm to perform better than *NB*, the results proved to the contrary. The number of *true-positives* is nearly the same as in the experiments with *NB*, but the number of *false-positives* is much higher. In all experiments, *NB* produces better *F1* values.

One of the surprising findings was that *SVM* profited much more from the enhanced dictionaries than *NB*. Although results for *NB* also improved when using the document set B_2 for training, the difference between the respective *F1* values is only 0.058. In contrast, the *F1* values in the *SVM* experiments differed by 0.363.

	Naïve Bayes		Support Vector Machines	
	$C(A+B_1)$	$C(A+B_2)$	$C(A+B_1)$	$C(A+B_2)$

¹³ http://www.ms.niedersachsen.de/live/live.php?navigation_id=5097&article_id=12913&psmand=17

Precision	0.674	0.812	0.272	0.628
Recall	0.588	0.608	0.165	0.521
F1	0.628	0.696	0.206	0.569
TP	20917	18666	5883	18516
FP	10098	4315	15715	10953
FN	14653	16904	29687	17054
TN	96145	101929	90528	95290

Table 4: Evaluation Results

A close look on the data reveals that *SVM* found many more *true-positives*, when the enlarged set B_2 was used, while *NB* came up with less *true-positives* in the same set-up. Both algorithms could reduce the numbers of *false-positives* significantly (note that the values for *fn* and *tn* depend directly on the values for *tp* and *fp* and need no further examination).

We have shown how the algorithms *Naïve Bayes* and *Support Vector Machines* performed on in our experiment. In the next section, we discuss the results and outline in which direction we intend to continue our investigations.

4 Conclusion and Outlook

The results on the classification of web pages with respect to their environmental relevance were ambiguous. Although the pages from the aforementioned relevant subdomains (see again Appendix B) could be classified to a high degree, a very large number of additional sites have been classified in the same manner. This is a well-known problem in document classification. It is not possible to maximize precision and recall at the same time, only a trade-off can be found, for which the weighted harmonic mean is a good indicator.

We found that the critical aspects in our experiments were on the one hand the creation of realistic and representative training sets, and on the other hand the testing of the trained algorithms. We discuss these problems in this section, followed by a brief outlook.

4.1 Effects of Dictionaries and Training Sets

In our experiments we were able to show that the document sets that are used in training both have a strong influence on the quality of the classification. This observation emphasizes the importance of a well-kept dictionary. In our experiments we found that the automatic generation of dictionaries gives reasonable results but the quality of the results increases with manual training.

In our first experiments we were able to decrease the number of *fp* by nearly fifty percent only by removing meaningless words. The number of *fn* increased at the same time by approximately ten percent. The consequences are higher *precision* accompanied by lower *recall*.

But there are contradicting signs that seem to indicate that the overall effect of manual training of the dictionary has only little effect on the quality of classification. In the later phase of our experiments, we already had a sophisticated blacklist of words that is used to automatically exclude words from the dictionary. The dictionary for B_2 , which was created using that blacklist, contained 102 words: The resulting *F1*-value was 0.690. After our manual training the dictionary was reduced to 85 words, the *F1*-value raised to 0.695. This increase is so small that it hardly seems to justify the efforts that we made when controlling the vocabulary in the dictionary.

The conclusion from these observations is that it is justifiable to put some efforts into maintaining the black list. Those parts of a black list that cover general terms (e.g. *nächste*¹⁴, *vorherige*¹⁵) can be re-used in different classification experiments. But there are also specific terms in a black list that are not transferable from one application domain to another. Therefore, most of the work on the black list has to be done in the beginning of the training for a new application domain.

Moreover, random samples have shown that some of the documents have not been classified in the experiments because the relevant keywords were not in the dictionary. For example, the term *biosphärenreservat*¹⁶ is highly relevant yet not part of the dictionary. The document-frequency of this term for all pages of *uba.de* was 9, which is very low. This emphasizes the need for an extended training set of relevant documents. More documents from different sites could contribute to a better dictionary. Moreover, it is worth considering the inclusion of other document types (e.g. PDF) in the training sets. Until now we excluded these documents for reasons of performance.

Interestingly, the building of a dictionary for set B seemed to be more challenging than for set A. The algorithm was able to determine approximately one hundred words, less than 20 of them with a significance higher than 0.1. In set A, roughly 200 words have been collected, more than 80 with relevance >0.1. Obviously, it was more easy to find terms that indicate relevance.

Although the dictionary that was derived from set A contained more terms with higher relevance, the classification results for document class B were better. This indicates that the set of documents that was used in the training of the algorithms was not representative enough. It seems very probable that using a larger set of relevant documents in the training could enhance the quality of the testing results. Including only the Web site *uba.de* seemed to be a simple yet good idea in the beginning. After our first experiments, however, we think that the scope of this site is broad but not broad enough. There are still many missing or under-represented topics, for example words like *biber*¹⁷, *naturschutz*¹⁸, or *biosphärenreservat* are not found in the dictionary although their relevance is obvious. Adding more Web sites could help to improve this situation and are the next natural step for further experiments.

4.2 Optimizing Precision or Recall?

In the optimization of the performance of the algorithms it is important to decide what should be optimized. Higher *recall* values can be achieved if more relevant documents can be found. But, high recall is always connected with less *precision*, which means that also more irrelevant information is being found. In order to find high values for both recall and precision, not only the dictionary but also the algorithm is of importance.

Generally speaking a certain trade off between precision and recall has to be made. This is comparable to text classification problems in the field of E-Mail spam filters: When the algorithm is classifying too many mails as spam, it is possible to also have some non-spam mails classified as being spam.

¹⁴ next (engl. transl.)

¹⁵ last (engl. transl.)

¹⁶ biosphere reserve (engl. transl.)

¹⁷ beaver (engl. transl.)

¹⁸ nature conservation (engl. transl.)

4.3 Comparison Of Algorithms

The influence of different training sets on the quality of classification was much more significant with *SVM* as with *NB*. We have seen that the *SVM* algorithm performed much better on the enhanced set of documents B_2 . We would like to test in further experiments whether this improvement is due to the size of the dictionary or due to specific documents from the site *rlp.de*. Training the algorithms with a set of documents ($B_2 - B_1$) and again comparing the results could clarify this question. Furthermore different kernels will be inspected for the *SVM* algorithm as well as the usage of slack variables that allow the algorithm to ignore certain documents (support vectors) that can be identified as outliers. By this a better decision hyperplane in the higher dimensional vector space can be constructed. It is planned to apply these findings to the next sets of experiments. The

In addition, it would be interesting to know why *NB* performed better with document set B_2 although the number of *true-positives* declined (see Table 4). Further tests with even larger document sets could be made to confirm this observation. Examining several samples also revealed that *NB* is highly prone to wrong decisions if documents only contain a specific but strong feature. For instance one document was classified wrongly as environmentally relevant, because it contained the word water. The rest of the document was in dutch language and no other feature was found on the page to influence the classification. This also leads to the assumption that multi lingual documents have to be inspected in more detail.

Also as already mentioned further improvements will be considered to increase the performance of the *NB* classifier (Rennie/Shih/Teevan/Karger, 2003, 616-623).

4.4 Validating the Results

We have found that the numbers for *tp*, *tn*, *fp*, and *fn* vary depending on the used training data and the used algorithm. It is hard, however, to interpret these variations. Many of the pages in the relevant subdomains do not have any relevant content. A common reason why a document does not become classified and is treated as a *false-negative* is that the document does not contain any useful information but only links to further pages or documents. At the same time, many pages from irrelevant subdomains do have relevant content. Random samples have shown that quite a large number of *false-positives* were in fact *true positives* because they clearly showed relevant content (see again Footnote 13).

Therefore, some of the pages in *fp* are in fact *true-negatives*, while some of the pages in *fn* are in fact *true-positives*. A more reliable validation is needed for more confidence in the significance of the results. Right now, the only available information about the relevance of a document is its membership to a subdomain. Additional information about the relevance on document level has to be collected in form of control samples. Random samples among *fp* and *fn* showed that some of them were indeed relevant in our sense, but a large number of samples were not relevant at all.

Instead of random samples, a pre-classified collection of test documents would be more efficient. A typical test scenario in the field of document classification is the Reuters-21578 corpus of test documents¹⁹. Here, the classification of a document is known beforehand. This collection cannot be used for our experiments because our approach needs a separate subset of documents with environmental relevance. This is not available with the Reuters corpus. As far as we know, an equivalent for the domain of environmental information does not exist. One strategy would be to compile such a corpus of documents where all the variety of environmental information should be covered. This, however, is a very ambitious project.

¹⁹ <http://www.daviddlewis.com/resources/testcollections/reuters21578/>

4.5 Future Directions

4.5.1 Improving training sets

As the URL sets that were used during the experiments were assumed to be either related to environmental topics or not, it is clear that this assumption is not perfect. For instance a lot of pages exist in the *umweltbundesamt.de* set that are not related to the environment at all. Nevertheless these were included in the training process of the experiments. Current work is done in the software tool to aid the user in adjusting the training sets. That will be done by exposing the classified URLs from the test or from real world testing with *niedersachsen.de* according to the score they got during classification: Web pages having very low scores for a given topic as well as web pages having very high scores will be presented the user for manual inspection. If the user decides, that the given document does not belong to the class it should be he will be offered the possibility to shift the document to another training class or remove it from the training set at all.

We call this method “hybrid training” and expect it to greatly improve the quality of the training sets to be used and finally enhance the quality of the following training process. Apart from this manual method another technique for improving training sets seems to be quite promising: The usage of the rough set theory coming originally from the field of document clustering allows cleaning training sets from outliers (Arco/Bello/Garcia, 2006).

4.5.2 Classification of Environmental Topics

Also a promising approach could be the classification of web pages with respect to specific environmental themes, i.e. pollution. Here, the classifier is trained using theme-specific subsets of Web pages from the same Web sites as in the aforementioned scenario. Based on this training a dictionary with a list of theme-specific keywords can be created. We expect better results from this setup as it should be easier to find more specific keywords for the dictionary. The techniques for multi-class classification are, however, different from our experiments. For example, *SVM* is inherently a two-class (binary) classifier and in the proposed case it is needed to create n binary classifiers when using n classes (Feldman/Sanger, 2008, 67).

For example when considering three different topics t_1 , t_2 and t_3 the *SVM* has to be trained for the classes t_1 vs. (t_2 and t_3), t_2 vs. (t_3 and t_1) and t_3 vs. (t_1 and t_2). In the setup given in this paper that has not been needed but we expect multi-label classification to furthermore improve the results.

5 Literature

- Arco, L., Bello, R., Garcia, M. M., 2006: On Clustering Validity Measures and the Rough Set Theory. In Proceeding: Fifth Mexican International Conference on Artificial Intelligence (MICAI '06)
- Chang, C.-C., Lin, C.-J., 2011; LIBSVM : a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1--27:27, 2011. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>
- Dean, J., Ghemawat, S., 2004: MapReduce: Simplified Data Processing on Large Clusters. In Proceedings: Symposium on Operating System Design and Implementation. San Francisco, CA. URL: <http://labs.google.com/papers/mapreduce-osdi04.pdf>
- Feldman, R., Sanger, J., 2008: *The Text Mining Handbook*. New York: Cambridge University Press
- Manning, C. D., Raghavan, P., Schütze, H., 2008: *Introduction to Information Retrieval*. New York: Cambridge University Press

- Rennie, J. D. M., Shih, L., Teevan, J., Karger, D. R., 2003: Tackling the Poor Assumptions of Naïve Bayes Text Classifiers. In Proceedings of the Twentieth International Conference on Machine Learning, pp. 616-623
- Rogati, M., Yang, Y., 2002: High Performing Feature Selection for Text Classification. In Proceedings of the ACM Conference on Information and Knowledge Management. pp. 659-661

Appendix A: Dictionaries

Dictionary for document set A (in combination with document set B₁):

selben , fachbibliothek , what , grundlagen , energieeffizienz , umweltbundesamt , energy , system , gesamt katalog , press , schadstoffe , drinking , ausdrucken , background , health , luft , umwelthygiene , sachgebiete , provide , beiträge , order , policy , contaminated , stand , bewertung , data , development , texte , insbesondere , trinkwasser , beispiel , internationale , german , control , umweltrecht , umwelt , water , pollutants , register , adresse , international , stellenangebote , know , soil , letzte , agriculture , kennnummer , verkehr , anlagensicherheit , warenkorb , umweltökonomie , change , germany , nahrungsmittelindustrie , presse , protection , transport , climate , planning , production , agency , sustainable , umweltmanagement , maßnahmen , resource , ressourcenschonung , pollution , deutschland , sites , other , verbandsklage , toxic , umweltschutz , contact , safety , chemicals , entwicklung , umweltinformation , mediendatenbank , environment , internationaler , wasser , which , reihe , schutz , products , nachhaltige , klimaschutz , chemikalienpolitik , related , monitoring , altlasten , lärm , associational , neuerscheinungen , federal , grundwasser , ratgeber , raumbezogene , last , möchten , events , umweltbundesamtes , awareness , produktion , economics , landwirtschaft , technische , stoffe , claims , abfallwirtschaft , waste , changed , gewässerschutz , informiert , ergebnisse , free , papers , senden , nachhaltiger , plant , available , gesamtliste , luftreinhaltung , hygiene , verwandte , industry , wissen , jahres , produkte , umweltplanung , that , conservation , antarktis , pollutant , gesundheit , rahmen , materialien , from , transfer , home , teil , antarctic , services , anforderungen , release , langfassung , publikation , umweltbewusstsein , more , partnerverlage , umsetzung , spatial , kostenlos , umweltbeobachtung , consumption , this , environmental , bestellen , report , 02.05.2011 , foodstuffs , management , energie , issues , 10.08.2010 , reach , änderung , richtlinie , kinder , basis , konsum , boden , hintergrundpapiere , cyanocenter , noise

Dictionary for document set A (in combination with document set B₂):

available , insbesondere , texte , nachhaltiger , policy , adresse , development , pollutant , fachbibliothek , toxic , german , stellenangebote , luft , protection , resource , gesamt katalog , rahmen , luftreinhaltung , bereich , system , consumption , conservation , antarktis , kostenlos , spatial , water , materialien , altlasten , stoffe , plant , verwandte , partnerverlage , publikation , report , chemikalienpolitik , release , mediendatenbank , internationaler , raumbezogene , energie , noise , umweltplanung , umwelt , landwirtschaft , änderung , boden , entwicklung , umweltbundesamt , erhalten , climate , planning , control , nachhaltige , health , safety , anlagensicherheit , produktion , ziel , deutschland , lärm , anforderungen , associational , technische , grundwasser , sites , internationale , rechtliche , related , order , energy , hintergrundpapiere , reach , informationen , trinkwasser , antarctic , schadstoffe , foodstuffs , gesundheit , verkehr , umweltbeobachtung , maßnahmen , umweltökonomie , umweltmanagement , change , jugendliche , sustainable , pollution , reihe , umweltrecht , products , kennnummer , kinder , waste , pollutants , cyanocenter , ressourcenschonung , environment , schutz , nahrungsmittelindustrie , services , umweltbewusstsein , awareness , environmental , papers , federal , richtlinie , economics , hygiene , transport , contaminated , wasser , teil , bewertung , free , background , produkte , konsum , gewässerschutz , umwelthygiene , bestellen , energieeffizienz , abfallwirtschaft , daten , agriculture , umsetzung , verbandsklage , wissen , sachgebiete , beiträge , langfassung , fragen , management , international , aktuelle , chemicals , germany , ergebnisse , neuerscheinungen , know , industry , klimaschutz , data , grundlagen , umweltschutz , monitoring , agency , gesamtliste , umweltinformation , umweltbundesamtes , claims , soil , transfer , ratgeber

Words only present in the dictionary for document set B₁ (in combination with document set A) and not contained in the dictionary for set B₂:

übermitteln , polizeireviere , verschiedene , broschüre , beruf , steuererklärung , amt24 , vordrucke , organisation , regierung , fälle , finanzamt , ?esky , steuern , hinweis , elster , finanzämter , anreise , übersichtskarten , karriere ,

folgende , bestimmte , elektronisch , zentrale , straÙe , pressemitteilungen , steuerportal , beratungsstellen , benötigen , anschrift , ausbildung ,

Words only present in the dictionary for document set B₂ (in combination with document set A) and not contained in the dictionary for set B₁:

weiterbildung , schrift , wissenschaft , familie , mediathek , wirtschaft , zusammenarbeit , sport , demografie , frauen , bild , euro , jugend , förderung , direkt , infrastruktur , landesregierung , ministerium , ministerpräsident , integration , rund , ministerien , landes , kultur ,

Words only present in the dictionary for document set A (when combined with document set B₁):

selben , what , press , drinking , ausdrucken , provide , stand , beispiel , register , letzte , warenkorb , presse , production , other , contact , which , last , möchten , events , changed , informiert , senden , jahres , that , from , home , more , this , [02.05.2011](#) , issues , 10.08.2010 , basis ,

Dictionary for document set B₁ (in combination with document set A):

aktuelle , schüler , eltern , übermitteln , esignatur , postanschrift , ansprechpartner , finanzen , justiz , land , polizeireviere , verschiedene , bildung , elektronische , journalisten , landesamt , zuständigkeit , bereich , erhalten , broschüre , lehrer , beruf , steuererklärung , innern , europa , polizei , besucheradresse , amt24 , einrichtungen , vordrucke , telefon , rechtliche , staatskanzlei , datenschutz , geschichte , hinweise , government , organisation , regierung , telefax , arbeit , fälle , möglichkeit , fragen , jugendliche , angebote , aufgaben , finanzamt , ?esky , soziales , steuern , hinweis , elster , finanzämter , verfügung , anreise , verwaltung , übersichtskarten , karriere , archiv , nutzen , folgende , bestimmte , elektronisch , behördenwegweiser , ziel , nachgeordnete , zentrale , straÙe , staatsministerium , pressemitteilungen , medieninformationen , bürger , steuerportal , beratungsstellen , benötigen , Verbraucherschutz , behörden , daten , touristen , anschrift , menschen , ausbildung , informationen , investoren

Dictionary for document set B₂ (in combination with document set A):

esignatur , direkt , verwaltung , datenschutz , hinweise , wirtschaft , telefon , jugend , landes , bildung , frauen , aufgaben , familie , landesamt , einrichtungen , nutzen , schrift , besucheradresse , möglichkeit , angebote , government , förderung , integration , elektronische , postanschrift , menschen , medieninformationen , bürger , finanzen , eltern , weiterbildung , Verbraucherschutz , telefax , polizei , behörden , land , staatsministerium , justiz , journalisten , mediathek , ministerien , investoren , infrastruktur , sport , wissenschaft , ministerium , ansprechpartner , kultur , arbeit , demografie , bild , staatskanzlei , innern , verfügung , behördenwegweiser , ministerpräsident , zusammenarbeit , nachgeordnete , schüler , landesregierung , europa , zuständigkeit , euro , geschichte , archiv , touristen , lehrer , soziales , rund

Appendix B: List of environmentally relevant subdomains, listed by domain-names:

niedersachsen.de

umwelt , numis , mu , mu1 , lbeg , nlwkn , naturparke , weltnaturerbe , wattenmeerstiftung , nlga , nna-en , nna , elbtalae , life-duemmer , naturschutzgebiete , natur-erleben , umweltbericht , umweltskarten

sachsen.de

umwelt.sachsen.de , landwirtschaft.sachsen.de , smul.sachsen.de , forsten.sachsen.de , sachsen.de/geoinformationen , publikationen.sachsen.de , gdi.sachsen.de , portalu.sachsen.de , wasserbuch.sachsen.de , statistik.sachsen.de/html/360.htm.* , statistik.sachsen.de/html/832.htm.* , statistik.sachsen.de/html/834.htm.* , statistik.sachsen.de/html/835.htm.* , statistik.sachsen.de/html/836.htm.* , statistik.sachsen.de/html/837.htm.*

rlp.de

mulewf.rlp.de , mwkel.rlp.de/Klimaschutz , mwkel.rlp.de/Kreislaufwirtschaft , mwkel.rlp.de/Bodenschutz , mwkel.rlp.de/Landesplanung , mwkel.rlp.de/Strahlenschutz