Automatic multi-label subject indexing in a multilingual environment

Boris Lauser[1],Andreas Hotho[2]

[1] FAO of the UN, Library & Documentation Systems Division,
00100 Rome, Italy
boris.lauser@fao.org
http://www.fao.org
[2] University of Karlsruhe, Institute AIFB,
76131 Karlsruhe, Germany
hotho@aifb.uni-karlsruhe.de

Abstract. This paper presents an approach to automatically subject index full-text documents with multiple labels based on binary support vector machines (SVM). The aim was to test the applicability of SVMs with a real world dataset. We have also explored the feasibility of incorporating multilingual background knowledge, as represented in thesauri or ontologies, into our text document representation for indexing purposes. The test set for our evaluations has been compiled from an extensive document base maintained by the Food and Agriculture Organization (FAO) of the United Nations (UN). Empirical results show that SVMs are a good method for automatic multi- label classification of documents in multiple languages.

1 Introduction

The management of large amounts of information and knowledge is of ever increasing importance in today's large organizations. With the ease of making information available online, especially in corporate intranets and knowledge bases, organizing information for later retrieval becomes an increasingly difficult task. Subject indexing is the act of describing a document in terms of its subject content. The purpose of subject indexing is to make it possible to easily retrieve references on a particular subject. It is the process of extracting the main concepts of a document, representing those concepts by keywords in the chosen language and associating these keywords with the document. In order to be unambiguous and carry out this process in a more standardized way, keywords should be chosen from a controlled vocabulary.
The AGROVOC[1] thesaurus, developed and maintained by the Food and Agricultural Organization[2] (FAO) of the United Nations (UN), is a controlled vocabulary developed for the agricultural domain. The FAO manages a vast amount of documents and

[1] [http://www.fao.org/agrovoc].
[2] [http://www.fao.org].

information related to agriculture. Professional librarians and indexers use the AGROVOC thesaurus as a controlled vocabulary to manually index all documents and resources managed by FAO's information management system. They are allowed to assign as many labels as necessary to index a document. In the following we call the automatic assignment process of suitable keywords to the documents the *multi-label* and *multi-class*[3] classification problem. This process is applied to resources in all the official FAO languages and herewith constitutes a *multilingual* problem. The cost of labour for professional indexers and the increase in growth in available electronic resources has resulted in a backlog of resources that are not indexed. Automatic document indexing could be particularly useful in digital libraries such as the ones maintained at the FAO to make more resources available through the system.

This paper presents an approach to use binary support vector machines (SVM) for automatic subject indexing of full-text documents with multiple labels. An extensive test document set has been compiled from FAO's large quantity of resources in which *multi-label* and *multilingual* indexing have been evaluated. Motivated by our text clustering results with background knowledge (cf. [7]), we have further analyzed the integration of domain specific background knowledge in the form of the multilingual AGROVOC thesaurus for performance improvement. With the evaluated results we will reason the integration of background knowledge with SVMs to be a promising approach towards (semi-) automatic, multilingual, multi-label subject document indexing.

The paper is outlined as follows: The next section introduces the reader to automatic text categorization, in particular support vector machines and the multi-label classification problem. Section 3 gives a brief introduction to ontologies and their representation. In Section 4, we explain in detail the compilation of the used test document set and the evaluation settings followed by a discussion of the results. We conclude by suggesting promising future possibilities for subject indexing of multilingual documents.

2 Automatic Text categorization

Text categorization is the process of algorithmically analyzing a document to assign a set of categories (or index terms) that succinctly describe the content of the document [11]. Various methods from different communities have been applied in automatic text categorization approaches, such as classical IR based classifiers, statistical learning classifiers, decision trees, inductive rule learning, expert systems or support vector machines (SVM). More comprehensive surveys of algorithms used for automatic classification can be found in [11], [1], and [12]. One application of text categorization is document indexing, in which several keywords taken from a controlled vocabulary such as a thesaurus or an ontology are assigned to a document in order to describe its subject. Support vector machines have been shown to outperform other approaches [1]. In this research, we therefore use an SVM-based approach to be applied to the

[3] In the following we only use the term multi-label.

multi-label classification problem as described in the following sections in accordance with the definitions given in [12]:

2.1 The classification problem

Multi-Label Classification Problem. In a multi-label classification problem, each document can be assigned an arbitrary number *m* (multiple labels) of *n* (multiple) possible classes. We have a set of training documents X and a set of classes $C = \{c_1, \ldots, c_n\}$. Each document $x_i \in X$ is associated with a subset $C_i \subseteq C$ ($|C_i| = m$) of relevant classes. The task is to find the most coinciding approximation of the unknown target function $\overline{\Phi} : X \times C \rightarrow \{true, false\}$ by using a function $\Phi : X \times C \rightarrow \{true, false\}$, typically called a classifier or learned model. $\overline{\Phi}$ reflects the unknown but "ideal" assignment of documents to classes.

In the single-label case, only one class is assigned. The binary classification problem is a special case of the single-label problem and can be described as follows:

Binary Classification Problem. Each of the documents $x_i \in X$ is assigned to only one of two possible classes c_i or its complement $\hat{c}_i$.

There are different alternatives towards multi-label document indexing as carried out in the FAO. In this research we adopted the approach of transforming a multi-label classification problem into $|C|$ independent problems of binary classification. This requires that categories be stochastically independent, that is, for any c_i, c_k the value of $\Phi(x_i, c_i)$ does not depend on the value of $\Phi(x_i, c_k)$ and vice versa. In the case of document indexing at the FAO, this is a reasonable assumption.

2.2 Binary Support Vector Machines.

Vapnik first introduced support vector machines (SVM) in 1995 [5]. They have been applied to the area of text classification first by Joachims in 1998 [8]. In support vector machines, documents are represented using the vector space model:

Vector Space Model. A document x_i is transformed into a *d*-dimensional feature space $\mathbb{R}^d$. Each dimension corresponds to a term (word, also referred to as feature). The values are the frequencies of the terms in the document. A document is represented by its word-vector of term frequencies,

$$\vec{x}_i = (tf(x_i, t_1), \ldots, tf(x_i, t_{|T|})),$$

where T is the set of terms that occur at least once in at least one document in the whole set ($|T| = d$) and the $tf(x_i, t)$ represent the term frequency of term $t \in T$ in document x_i.

There are a wide variety of weights and ways to choose a term/feature. A more detailed discussion can be found in [12]. In this case, terms (or later concepts from the ontology) are chosen as features and the standard *tfidf* (Term Frequency Inverse Document Frequency) measure is used as term weight calculated as

$$tfidf(x_i,t) = \log(tf(x_i,t)+1) * \log\left(\frac{N}{df(t)}\right),$$

where $tf(x_i,t)$ is the frequency of term t in document x_i and N is the total number of documents (|X|) and $df(t)$ (document frequency) is the number of documents, a term t occurred in.
A binary SVM tries to separate all the word vectors of the training document examples into two classes by a hyper plane, maximizing the distance of the nearest training examples. Therefore, it is also referred to as the maximum margin hyper plane. A test document is then predicted by the SVM by determining, on which side of the hyper plane its word vector is. A very good and detailed introduction to SVM and document representations is provided in [14].

2.3 Related approaches

A different approach described in [10] uses a Bayesian classifier together with a document mixture model to predict multiple classes for each document. This approach takes into consideration all classes at the same time as opposed to splitting the whole problem into a number of binary classifiers.
A recently taken similar approach towards multi-label classification using binary classifiers is discussed in [6]. The difference to our approach is that these algorithms can be applied in online settings, where the examples are presented one at a time, as opposed to the batch setting used with support vector machines.

3 Background knowledge in form of ontologies

Apart from solving the multi-label problem, the additional incorporation of background knowledge as provided by domain specific ontologies is the second focus of this work. Since ontologies have been defined many times, we will abstain from giving a formal definition of domain ontologies in favour of introducing the main aspects in a short example. The underlining formal definition, used representation and notions in this work are in accordance with [3]. This is also the basis of our implementation in the KAON Framework[4]. Figure 1 shows a very small extract of the AGROVOC thesaurus, represented as an ontology. Refer to [9] for a detailed discussion of converting the AGROVOC thesaurus into an ontology. An ontology is basically a tree-ordered hierarchy structure of concepts as shown in Figure 1.

[4] [http://kaon.semanticweb.org/].

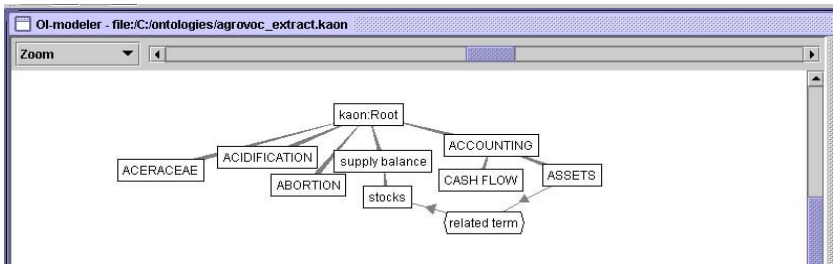

Figure 1: Small ontology extract

Each concept in the picture (drawn as a rectangle) has lexical entries (labels, synonyms) attached to it. The picture only shows the English labels of the concepts. The important fact for our purposes – explained in more detail in section 4 – is that a concept itself is actually language independent and internally represented by a URI[5] (Uniform resource identifier). Every concept has a super concept, e.g. "supply balance" is the super concept of "stocks". The highest concept is "root". In addition to the tree structure, an ontology can have arbitrary lateral relationships, as shown here with a 'related term' relationship. As opposed to simple thesauri, ontologies allow for many types of relationships, making it a more expressive instrument of abstract domain modelling.

4 Evaluation

4.1 The test document set

To evaluate this research, a set of training and test documents has been compiled from the agricultural resources of the FAO. Journals, proceedings, single articles and many other resources constitute an extremely heterogeneous test set, differing substantially in size and descriptive content of its documents. The metadata elements of the resources contain a subject element in addition to others such as title, URL, etc. Subject indexing is carried out using keywords from the AGROVOC thesaurus which provides over 16607 potential document descriptors to choose from. A maximum of 6 primary descriptors, describing the most important concepts of a resource, can be used to index a document. Additionally, an indefinite number of secondary descriptors can be chosen, as well as geographic descriptors (for example country information). Only the primary descriptor associations have been considered in this evaluation. Metadata information about FAO documents is stored in any of the three languages English, French and Spanish. The test sets have been compiled with the requirement of having at least 50 documents for each class. Table 1 shows an overview of the so compiled test sets in the 3 different languages.

[5] See also [http://www.w3.org/Addressing/].

Table 1: Compiled multi-label test document set in 3 languages

		English (en)	French (fr)	Spanish (es)
Total	# Documents	1016	698	563
	# Classes	7	9	7
Class Level	Max ($\frac{\#documents}{class}$)	315	214	179
	Min ($\frac{\#documents}{class}$)	108	58	58
	Avg ($\frac{\#documents}{class}$)	145,14	77,56	80,43
Document level	Max ($\frac{\#labels}{document}$)	3	3	3
	Min ($\frac{\#labels}{document}$)	1	1	1
	Avg ($\frac{\#labels}{document}$)	1,25	1,40	1,42

4.2 Performance measures

The common notions of precision and recall from the Information Retrieval (IR) community have been applied to measure performance of the conducted tests [12]. The initial document set X (pre-classified by human indexers) is split into a training document set X_{Tr} and a test document set X_{Te}, so that $X = X_{Te} \cup X_{Tr}$. The corpus of documents is pre-classified, i.e. the values of the function $\overline{\Phi}$ are known for every pair (x_i, c_i). The model is built with the training set and evaluated with the test set.

Precision and recall are measured for each class and calculated from four different numbers according to Table 2.

Table 2: Contingency table for class c_i

Class c_i		Expert judgements	
		YES	NO
Classifier judgements	YES	TP_i	FP_i
	NO	FN_i	TN_i

TP_i (true positives) is the number of documents correctly assigned to class c_i. FP_i (false positives), FN_i (false negatives) and TN_i (true negatives) are defined accordingly.

Overall performance is measured by summing up the values over all classes, and calculates precision and recall according to the micro-averaging approach [12] to:

$$\text{Precision}_{micro} = \frac{\sum_{i=1}^{|C|} TP_i}{\sum_{i=1}^{|C|} (TP_i + FP_i)} \qquad \text{Recall}_{micro} = \frac{\sum_{i=1}^{|C|} TP_i}{\sum_{i=1}^{|C|} (TP_i + FN_i)}$$

Precision is an indicator of how many of the predictions were actually correct. On the other hand, recall is an indicator of how many of the pre-classified documents have also been predicted, i.e. it provides an indication of how exhaustive the predictions were. Clearly, both figures must be measured to be able to draw a conclusion on performance.

4.3 Evaluation criteria and setup

Within this research work, three different test settings have been carried out:

Single-label vs. multi-label classification. The first evaluation focused on the comparison of single-label classification vs. multi-label classification. For this purpose, a second set of documents has been compiled from the document set shown in Table 1. This time, however, only the first primary descriptor assigned to the document was used, assuming that this is the most important descriptor for the respective document.
One binary support vector machine is trained for each unordered pair of classes on the training document set resulting in (m*(m-1))/2 support vector machines. Each document of the test set is then evaluated against each SVM. A binary SVM votes for the better class amongst the two it can choose from. A score is associated with each class calculated based on the number of votes for the respective class. The score is > 0 if more than 50% of a class's SVMs have voted for this class. In the single-label case, the class with the best score is assigned to a document. In the multi-label case, we introduced a score-threshold. All classes with a score greater than the score threshold were assigned to a document. Obviously, the number of assigned labels varies with the chosen score threshold.
Because the English document sets provide the most extensive test sets., they have been chosen for this first evaluation, The number of training examples per class has been varied from 5 up to 100. The number of test examples has been held at a constant rate of 50 test documents per class. In case of the multi-label test set, the score threshold has been varied between 0 and 0.6.

Multilingual classification. The second evaluation has been motivated by the idea that support vector machines basically operate independently of languages and document representations. The simplest possible scenario is a classifier that, given an arbitrary document, decides for one of the 3 classes (English, French or Spanish). A very simple document set has been compiled out of the single-label document sets that have been compiled for the previous evaluation, each pre-classified to its corresponding language class (English, French, Spanish) respectively. Each class contains more than 500 documents. The classifier has been trained varying the number of training documents per class between 5 and 200, leaving the number of test documents at a constant rate of 100 test documents per class.

Integration of domain-specific background knowledge. The third and last evaluation tests the effect of integrating the domain specific background knowledge provided by the AGROVOC ontology. The integration of background knowledge is accomplished by extending the word vector of a document with related concepts, extracted

from the domain ontology by using word-concept mappings and exploring concept relationships. The necessary steps to integrate the background knowledge are more formally outlined in Hotho et. al. [7]. In our evaluation, we varied two parameters: the **concept integration depth**, i.e. the maximum depth up to which super concepts and related concepts of a term are included; and the **concept integration mode**, for which 3 possible options exist:

- Add all concepts found in the ontology to the word vector (**add**)
- Replace all words in the word vector with corresponding concepts (**replace**)
- Consider only the concepts found in the ontology, i.e. for each document, create a word-vector only consisting of the domain specific concepts (**only**)

The idea behind this integration is to move the word vector of a document towards the specific domain and topic it represents, therefore making it more distinguishable from other word vectors. Domain specific background knowledge bares a certain potential to accomplish this task, in a way that it only contains the concepts, which are descriptive for the domain.

In our test case, the AGROVOC thesaurus has been pruned to reflect the domain of the compiled document sets. Pruning in this case means the extraction of only the relevant concepts for a specific domain, thus resulting in an ontology/thesaurus significantly smaller in size. The algorithm used here has been applied in other domains [13] and adapted within the project at the FAO [9].

We evaluated the integration of the pruned AGROVOC on the English document set for the single-label case. Apart from variation of the number of training and test examples per class and all possible concept integration modes, the concept integration depth has been varied from 1 to 2, 1 meaning that only matching concepts have been considered.

4.4 Results

Single-label vs. multi-label classification. For each parameter variation, 15 independent test runs have been conducted. In each run the document set has been split into an independent training and test set. Performance measures have been averaged over all 15 runs respectively.

In the single-label case, precision and recall are always the same and the calculation of both values is not needed. The precision values ranged from 47% (5 training examples per class) to 67% (100 training examples per class). In case of multi-label classification, both precision and recall have been calculated, since here they differ from each other substantially. Keeping the score threshold low implies that many labels – assumingly too many – get assigned to each test document. This results in low precision, because many of the classifications are wrong. However, in that case recall is high because most of the test documents get assigned the labels of the classes they are preclassified to. Table 3 shows the development of precision and recall depending on the score threshold exemplary for the English set with 50 training examples per class. By raising the score threshold, fewer labels get assigned to each document. In our test cases, precision could go up to as much as 45% and recall plummeted to as low as 76%. In order to make these contradictory effects comparable with the single-label

classification, the so-called breakeven value has been computed as the average mean of precision and recall, assuming that both measures are rated equally important.

Table 3: Results of multi-label classification with the English language test set. Development of precision and recall depending on the score threshold.

Score Threshold	Measure	50 Training Ex.
0.0	Precision	0.2727
	Recall	0.9329
	Breakeven	0.6028
0.1	Precision	0.2754
	Recall	0.9350
	Breakeven	0.6052
0.3	Precision	0.3412
	Recall	0.8721
	Breakeven	0.6066
0.5	Precision	0.4492
	Recall	0.7618
	Breakeven	0.6055
0.6	Precision	0.4539
	Recall	0.7702
	Breakeven	0.6121

Figure 2 shows all the results in one chart. The Spanish and French multi-label test sets have been additionally evaluated regarding language behaviour of the classifier. The breakeven values are shown depending on the training examples used for building the SVMs. Multi-label classification has shown overall worse performance than the single-label case. However, taking into account the higher complexity of the multi-label problem, the difference comparing the overall results between the two approaches is reasonable. Regarding performance of different languages, we can already infer from the multi-label results that languages different from English seem to perform equally well.

The breakeven values displayed here have been achieved with the overall superior configuration of a score threshold of 0.1. Raising the threshold further always resulted in similar breakeven values. No clear statement can be made on the use of varying the score threshold beyond that value. It depends on the intended goal of applying the classifier. If the classifier is used to help a human indexer by suggesting a large set of possible index terms from which the indexer can choose, then it is clearly advantageous to have a high recall, suggesting most of the 'good' terms amongst others. If, however, the automatic classifier is used without human support, it becomes more necessary to limit the risk of assigning wrong labels and aim for high precision. In the latter case, a higher score threshold should be chosen.

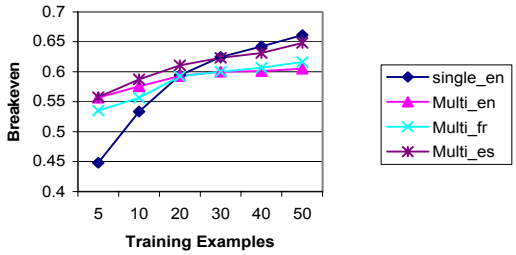

Figure 2: Results single-label vs. multi-label classification

Multilingual classification. The application of the scenario described in section 4.3 resulted in almost 100% precision in all test runs. This clearly shows that support vector machines are able to distinguish almost perfectly between languages.

Integration of domain-specific background knowledge. The integration of the pruned AGROVOC ontology was only able to show a slight increase in precision in the case of adding concepts to the word-vector and replacing words with their respective concepts. However, the performance gains did not show any significance. Figure 3 shows the results for the evaluation runs with 10 and 50 training examples per class. The leftmost values (ontology integration depth 0) display the results without ontology integration as reference. The remainder of the values belongs to each variation in integration mode (*add, replace, only*) and depth (*1* meaning that only the concepts which matched the label have been considered, whereas *2* means that also the direct super- sub- and related concepts have been considered).

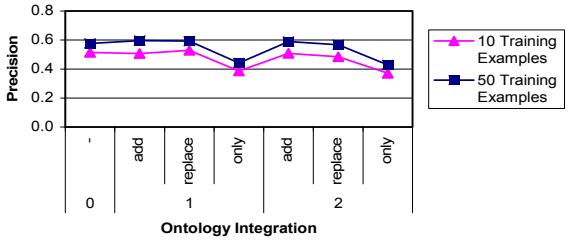

Figure 3: Results of ontology integration with the English single-label test document set.

In the case of totally replacing the word-vector of a document (concept integration mode *only*), the overall results even slightly decreased. This effect leads to the assumption that the used ontology misses domain specific vocabulary needed to unambiguously define the content of the domain documents. Considering our description of subject indexing made above, a document's content should be described by leaving out all non-domain-specific concepts.

5 Conclusion and outlook

Our results clearly show that SVMs behave robustly across different languages. The fact that no significant performance differences between the languages have been found in the multi-label case[6] indicates that SVMs can be applied to classify documents in different languages. SVMs seem to be especially applicable to the complex case of assigning multiple labels to a document. The inferior results of multi-label indexing compared to the single-label case are clearly explained by the increased complexity of the task. Among human classifiers, multi-label subject indexing is an inconsistent task; opinions vary from person to person and there is no single correct assignment of labels to a document regarding the type and number of chosen labels. Taking this phenomenon (also known as indexer-indexer inconsistency [4]) into consideration, the results found can even be interpreted as equally good. This is a rather optimistic hypothesis and since the two cases are not directly comparable, further research and evaluation are needed in order to confirm it. These results combined with the fact that the integration of background knowledge did not show any significant performance losses – except in the case of total replacement of a document's word-vector – leads us to an interesting conclusion for further research and evaluation. In the FAO (and most probably in many other environments), English resources heavily outweigh the availability of resources in other languages. As clearly shown in our results, the quality of SVMs strongly correlates with the number of used training examples. A desired scenario is therefore to be able to train the classifier with documents in one language only (i.e. English), and be able to use it to classify documents in other languages. This can be achieved by replacement of a document's word-vector by using only the concepts found in the multilingual domain specific background knowledge. AGROVOC is available online in 5 different languages and has been translated into many others. A document's word-vector thus becomes language independent and the resulting classification should be the same. With respect to the lower performance in case of replacing a document's word-vector with its domain-specific concepts only, future research should be applied towards testing the exhaustiveness of the AGROVOC ontology used here. On the other hand, the AGROVOC is a more generic thesaurus, used for the whole agricultural domain. Subsets of the documents used in this research are assumingly more specific to certain domains. It would therefore be especially of interest to re-evaluate the settings used in this test set by using a document set limited to a very specific domain and a suitable domain specific ontology.

Moreover, especially in multinational organizations and environments like that provided at the FAO, more and more documents are actually multilingual, containing parts written in different languages. The integration of background knowledge as described above obviously has potential in showing robust behaviour towards those kinds of documents.

In conclusion, the results shown here are preliminary steps towards a promising option to use support vector machines for automatic subject indexing in a multilingual envi-

[6] This result could be confirmed with further test runs conducted on the document sets compiled for single-label classification.

ronment. Future research should exploit different other domains, in order to prove or confute the findings made here.

Acknowledgements. We express our gratitude to the FAO of the UN, Rome for the funding of our work. We especially thank all our colleagues there for their substantial contribution in requirements analysis and the compilation of the test document sets.

References

1. Aas, K.; Eikvil, L.: Text Categorization: a survey. Technical Report #941, Norwegian Computing Center. 1999.
2. Berners-Lee, T.; Fielding, R.; Irvine, U.C.; Masinter, L.: Uniform Resource Identifiers (URI): Generic Syntax. IETF Request for Comments: 2396. [Online: http://www.ietf.org/rfc/rfc2396.txt]. August 1998.
3. Bozsak, E.; Ehrig, M.; Handschuh, S.; Hotho, et al: KAON -- Towards a Large Scale Semantic Web. In: Bauknecht, K.; Min Tjoa, A.; Quirchmayr, G. (Eds.): Proc. of the 3rd Intl. Conf. on E-Commerce and Web Technologies (EC-Web 2002), 2002, 304-313.
4. Chung, Y.; Pottenger, W. M.; Schatz, B. R.: Automatic Subject Indexing Using an Associative Neural Network, in: Proceedings of the 3 rd ACM International Conference on Digital Libraries (DL'98), pp. 59-68. ACM Press, 1998.
5. Cortes, C.; Vapnik, V.: Support-Vector Networks. In Machine Learning, 20(3):273-297, September 1995.
6. Crammer, K.; Singer, Y.: A new family of online algorithms for category ranking. In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 151-158. Tampere, Finland. 2002.
7. Hotho, A.; Staab, S. and Stumme, G. (2003). Text clustering based on background knowledge (Technical Report 425), University of Karlsruhe, Institute AIFB. 36 pages.
8. Joachims, T.: Text categorization with Support Vector Machines: Learning with many relevant features. In Machine Learning: ECML-98, Tenth European Conference on Machine Learning, 1998. pp. 137--142. [http://citeseer.nj.nec.com/article/joachims98text.html].
9. Lauser, B.: Semi-automatic ontology engineering and ontology supported document indexing in a multilingual environment. Internal Report. University of Karlsruhe, Institute of Applied Informatics and Formal Description Methods. 2003.
10. McCallum, A.: Multi-label text classification with a mixture model trained by em. AAAI'99 Workshop on Text Learning. 1999.
11. Ruiz, M. E.; Srinivasan, P.: Combining Machine Learning and Hierarchical Structures for text categorization. In Proceedings of the 10th ASIS SIG/CR Classification Research Workshop, Advances in Classification Research–Vol. 10. November 1999.
12. Sebastiani, F.: Machine learning in automated text categorization. Tech. Rep. IEI-B4-31-1999, Consiglio Nazionale delle Ricerche, Pisa, Italy. 1999.
13. Volz, R.: Akquisition von Ontologien mit Text-Mining-Verfahren. Technical Report 27, Rentenanstalt/Swiss Life, CC/ITRD, CH-8022 Zürich, Switzerland, ISSN 1424-4691. 2000.
14. Witten, I.; Frank, E.: Data Mining, Practical Machine Learning Tools and techniques with Java implementations. Morgan Kaufmann. 1999.