Automatic classification of scientific records using the German Subject Heading Authority File (SWD)

Christian Wartena and Maike Sommer*

Hochschule Hannover - University of Applied Sciences and Arts
Department of Information and Communication
Expo Plaza 12, 30539 Hannover, Germany
Christian.Wartena@fh-hannover.de
Maike.Sommer@stud.fh-hannover.de

Abstract. The following paper deals with an automatic text classification method which does not require training documents. For this method the German Subject Heading Authority File (SWD), provided by the linked data service of the German National Library is used. Recently the SWD was enriched with notations of the Dewey Decimal Classification (DDC). In consequence it became possible to utilize the subject headings as textual representations for the notations of the DDC. Basically, we we derive the classification of a text from the classification of the words in the text given by the thesaurus. The method was tested by classifying 3826 OAI-Records from 7 different repositories. Mean reciprocal rank and recall were chosen as evaluation measure. Direct comparison to a machine learning method has shown that this method is definitely competitive. Thus we can conclude that the enriched version of the SWD provides high quality information with a broad coverage for classification of German scientific articles.

1 Introduction

Subject classification is one of the major pillars to guarantee accessibility of records in large digital libraries. One of the worldwide most common classification systems is the Dewey Decimal Classification (DDC). The DDC is a universal classification system aiming at representing the entire knowledge of the world. It is used in more than 135 countries and translated into over 30 languages. More than 60 countries use the DDC even for their national bibliography. Apart from the worldwide use the DDC has a second strength: It is administrated by the Decimal Classification Editorial Policy Committee at the Library of Congress. Thus it is updated and developed continuously ([12]).

Classifying records according to DDC is a task that requires carefully reading and understanding of the abstracts and other available meta data as well as a

* This work is partially based on the Bachelor thesis of Maike Sommer ([17]).

detailed knowledge of the DDC class hierarchy. In a number of projects classifiers were built using machine learning techniques ([20], [19]). These approaches are problematic because the DDC-classes are very fine grained. Even in very large repositories, for most classes there are not sufficient training data. Thus a classification can only be made on the highest levels of the DDC hierarchy and even then the sparsity of data poses still a problem for some of the classes ([19]). A second problem is the dependency on training data. Especially, the data that are classified have to be comparable to the data that were used for training. E.g. if the collection to be classifies contains other text types than those used for training, the results might be worse than expected.

The basic principle of text classification based on machine learning is as follows. In the training phase words are given weights indicating how strong they characterize a certain class. During classification these weights are used to guess the most likely class for a text. Instead of determining weights in a training phase we can use a dictionary or thesaurus, if it contains information on the relation between words and the target classes, in our case the classes from the DDC. Recently, a large number of relations between subject headings of the German Subject Heading Authority File (*Schlagwortnormdatei*, SWD) and DDC-classes have been published ([4], [9]). Since most subject headings consist of just a single word or a very short phrase, we can use the SWD as a large lexical resource with a very broad coverage. Now, basically by counting the links of the subject headings found in a text to the DDC-classes we can predict the DDC-class for the text. The disadvantages of this approach are manifest: Weights are just 0 or 1, without any information how indicative a word is for a certain class. Furthermore, only the weights of the words are used and no dependencies between words can be modeled. The method has, on contrary, also the advantage that we need no training data and we directly can classify documents in domains that we did not see before. The success of this approach depends crucially on the quality of the thesaurus used. The main contribution of this paper is, that we show that the SWD is a vary valuable source of information in this respect.

In the following we will describe our approach in more detail and present results on the classification of the German language records from 7 repositories of different German universities. We compare our results with the results given by the Automatic Classification Toolbox for Digital Libraries (ACT-DL) from the University of Bielefeld (`http://clfapi.base-search.net/doc/index.html`), that uses a state-of.the-art machine learning approach. We show that the results are comparable in terms of mean reciprocal rank and in most cases better in terms of recall. The first measure is important with regard to fully automatic classification, the second measure is especially important in an interactive scenario in which the algorithm provides suggestions to a librarian.

The remainder of this paper is organized as follows. In section 2 we discuss related work. In section 3 we present our approach. In section 4 we describe the data we have used in our experiment, the results of which are given in section 5 We conclude with a discussion of results (section 6) and outlook to future work (section 7).

2 Related Work

Waltinger et al. ([19]) treat exactly the same problem that we discuss in this paper, namely classifying English and German scientific abstracts into high-level DDC classes. They use a state-of-the art machine learning approach for text classification. Below we will compare our results of the ontology driven approach directly with the results obtained by their classifier, that is publicly available as a web service.

Various studies consider document labeling or classification with the labels of an ontology, using lexical and structural information from that ontology. Basically, occurrences of ontology concepts in the text are counted and in some manner the information is aggregated to determine the most central or important concepts. Usually these approaches require enrichment of the ontology with additional lexical information, in many cases obtained from WordNet. Examples of such approaches are [18], [14] and [7].

Another approach to using ontologies for text classification is to enrich the representation of the text with features derived from the ontology, like hypernyms or concept labels before applying the classification algorithms. E.g., Scott and Matwin ([16]) add WordNet hypernyms and Bloehdorn and Hotho ([3]) add hypernyms from Wordnet and other ontologies to the representation of the text. The latter authors also try out various disambiguation strategies for words that potentially represent more than one ontology concept. Improvements over the baseline using only the words from the text are in both cases not very convincing.

Addis et al. ([2]) consider text classification rather as a two step process. In the first step WordNet concepts (*synsets*) are extracted. In the second phase an existing mapping from synsets to DDC-categories is used to compute a DDC-classification for the text. However, the authors do not consider this process as their final classifier, but use it only to create text collections to train statistical classifiers on. The two step approach is treated more systematically by Chenthamarakshan et al.([5]), who explicitly distinguish between the process of finding representative concepts on the one hand side and learning a mapping from concepts to document classes on the other hand side. These approaches differ not only in the perspective on the task from those mentioned in the previous paragraph. They are also different because they do not add features to the simple word vector model, but replace the original representation.

In our approach we consider classification as a two step approach as well. Thus the main contribution of this paper is not the method presented, but rather the investigation in the potentials of the German Subject Heading Authority File, that was, to the best of our knowledge, not used for automatic classification before. Since the results turn out to be very competitive, the proposed method might also have practical value for application in libraries.

3 Approach

In our thesaurus based approach, the most relevant Dewey class for a text is determined by the Dewey classification of the words in this text according to the

thesaurus. In our case the thesaurus is the German Subject Heading Authority File (SWD) for which all terms have been related to Dewey classes.

In order to find all relevant words we stem all words in the text. To be sure that only relevant words are found we restrict our search for thesaurus terms to nouns only. The text analysis is implemented as a GATE pipeline ([6]). For stemming and part-of-speech tagging we use the TreeTagger ([15]). Search for thesaurus terms is implemented by Apolda ([21]).

In the next phase we can determine the class of the text on the basis of the identified occurrences of thesaurus terms. For this phase we keep only unambiguous words, since only these terms give a clear indication of the topic of the text. Usually enough unambiguous terms remain to determine the topic of the text. We consider a word as unambiguous if the word occurs as the label of only one subject in the thesaurus, or if the word is the preferred label of exactly one subject. E.g. the word *Student* occurs 9 times as an alternative label for subjects like *Studentenwohnheim* (student accommodation) or *Auslandsstudium* (study abroad). However, there is one subject that has *Student* as its preferred label. Thus we treat *Student* as a non-ambiguous term representing that subject. The word *Untersuchung* (investigation) in contrast is found 2 times as an alternative label but never as a preferred label. Thus this word is not considered in the following steps. In this way many very general terms are filtered out.

Once all subjects have been identified we count the Dewey classes they are related to. In the (enriched) SWD each word is related to one or more Dewey classes via an anonymous node. For each relation a confidence of correctness is given by an integer between 1 and 4. For our purposes we ignore all links with a confidence level of 1. Given a (non-ambiguous) term occurrence t we let $ddc(t)$ be the set of all DDC-classes that t is related to with a confidence level greater than 1. Most words are related to very specific class in DDC. In order to aggregate occurrence information on a higher level in the DDC-hierarchy we denote for each class c in the DDC-System the broader class at the n-th level as c^n. Since the DDC-system is a strict hierarchy c^n is uniquely defined for each class with a depth smaller than n. E.g. if c is the class 342.0684 then c^2 is 340. Now we can define the contribution of a term t to each DDC-class c as

$$w(t,c) = \frac{|\{c_i \in ddc(t) \mid c_i^n = c\}|}{|\{c_i \in ddc(t)\}|} \tag{1}$$

where n is the hierarchy level of c. Considering a text T as a set of term occurrences we define the weight of a class c for T as

$$w(T,c) = \sum_{t \in T} w(t,c). \tag{2}$$

This gives us almost a ranking of DDC-classes for a text T. Only in case two classes have the same weight we need to specify their ranking. In these cases we order the classes by the order of their first occurrences in the text, where an earlier occurrence implies a higher rank.

4 Data and experimental setup

The experiment we present here was enabled by the results of the CrissCross project conducted by the Cologne University of Applied Sciences (Fachhochschule Köln) in collaboration with the German National Library (Deutsche Nationalbibliothek). In this project a concordance between the German Subject Heading Authority File (Schlagwortnormdatei, SWD) and the DDC was constructed. In other words the subject headings were mapped to notations of the DDC. The SWD is a universal indexing language based on rules, namely the rules for the subject catalog (Regeln für den Schlagwortkatalog, RSWK) and the practice rules for the RSWK and the SWD. In contrast to the DDC there are not that many relations between the subject headings. In accordance with an unpublished study from 2004 almost 87 % of the subject headings do not have associative relations. Furthermore 34% have neither associative nor hierarchical relations. The enrichment of the SWD with DDC notations is helpful in structuring the SWD because it generates hierarchical, equivalence and associative relations through similar DDC notations. We already mentioned that there were not were not many relations between the subject headings before the Criss-Cross project ([11]). Thus a subject heading can be interpreted differently. The project group mostly mapped one subject heading to several DDC notations ([10]). Furthermore, the meaning of a subject heading is often very specific. Therefore the mapped DDC notations are also very specific, which means mappings to a deep hierarchy level. Hence this is called deep level mapping ([11]). In our experiment we only wanted to gain notations up to the second hierarchy level, that can easily be obtained through the DDC hierarchy as explained above. Furthermore, there is much variance to what extent a subject heading fits into a DDC class. To express this distinction, the project group invented four confidence levels (degrees of determinacy) with 1 for the lowest and 4 for the highest congruency. As aforementioned we disregarded all first level relations, because these mappings point to DDC notations with only a small thematic intersection ([1]).

The released version of the enriched SWD (`https://wiki.d-nb.de/download/attachments/34963694/SWD_s_rdf.zip`) has about 188,000 concepts linked to 51,748 DDC-classes. The concepts have preferred and alternative labels. These labels are however labels of subject headings and not intended to be used as a lexical resource for analyzing texts. Some concepts have labels that are very unlikely to appear in running texts. However, in many cases the terms are single words or small phrases that will appear in normal texts.

More problematic are however concepts that have labels that will occur in many texts for which the concept is not relevant. This can be the case with words that have a meaning that is related to some subject area but that also can be used in a more general way. E.g. the word *Zusammenhang* can be used in a general way, meaning *context* or *connection*, but it is also the alternative label of *Zussamenhang in einer Mannigfaltigkeit* (Connectedness in a manifold) that is mapped correctly onto the Dewey class 516.35 (Algebraic geometry). Another class of words causing problems in a similar way are the homographs and homonyms. E.g. the abbreviation *ALS* (for the disease amyotrophic lateral

sclerosis) is a homograph for the very frequent conjunction *als* (as). A number of these homographs can be filtered out, because the different meanings correspond to different parts of speech. As mentioned before we only consider words from the text that were tagged as noun. Another example is constituted by the word *IM* that is an alternative label of the term *Spitzel* (spy), since it is the abbreviation of *Informeller Mitarbeiter* (informal staff), especially for the intelligence department of the GDR. Its homograph *im* is a highly frequent word that is the contraction of the words *in dem* (in the). Furthermore, many auxiliaries and function words are included in the category linguistics. In order to avoid problems with these words we removed all concepts from the class 435 (German grammar) except for the subject headings *rational*, *irrational* and *Gloria* because they are mapped into a second DDC class apart from 435. Also the subject heading Grammis was not removed because it is not a stop word but the abbreviation for grammatical information system (Grammatisches Informationssystem des IDS (*Institut für Deutsche Sprache*, Institute for German language)) Additionally we removed all concepts with a question mark ("?") as preferred label and the following alternative labels: *im* and *in* (as abbreviation of *intelligentes Netz* (intelligent net) as an extension of a telephone network). After that we could use the subject headings as textual representations for the DDC-classes. In sum 314,287 preferred and alternative labels could be used as textual term representations.

Table 1. OAI-Metadata repositories used in this paper. From each repository all records of publications in German language with an abstract available at the date of retrieval were used.

URL	University	#records	date of retrieval
`http://opus.bsz-bw.de/fhhv/oai2/oai2.php`	Hanover UAS	271	2012-05-27
`http://opus.bibl.fh-koeln.de/oai2/oai2.php`	Cologne UAS	254	2012-06-13
`http://opus.bsz-bw.de/fhff/oai2/oai2.php`	Frankfurt am Main UAS	120	2012-06-13
`http://opus.kobv.de/tuberlin/oai2/oai2.php`	TU Berlin	2036	2012-06-13
`http://opus.bsz-bw.de/ubhi/oai2/oai2.php`	Univ. Hildesheim	97	2012-06-13
`http://www.opus-bayern.de/uni-regensburg/oai2/oai2.php`	Univ. Regensburg	790	2012-06-15
`http://opus.bsz-bw.de/phfr/oai2/oai2.php`	Freiburg Univ. of Education	258	2012-06-14

For testing the effectiveness of the proposed classification strategy we have used in the first place the repository of the Hochschule Hannover - University

of Applied Sciences and Arts. This repository supports the Open Archives Initiative Protocol for Metadata Harvesting ([13]). We have classified metadata records of this repository using different fields, like title, abstract and keywords at the first and second level of the DDC-hierarchy. In most realistic scenarios one will have the title and the abstract of a publication that has to be classified, but not keywords. Thus we concentrated on classification using title and abstract. Besides the repository of the Hochschule Hannover, we used 6 more repositories. The repositories were chosen on the basis of the presence of an OAI-PMH interface, the size of the repository and the availability of the required metadata and classification. We selected three repositories from universities (among which one technical university), three universities of applied sciences and one university of education. Details of the repositories are given in Table 1.

Since the SWD is a German resource, we are only interested in publications in German. Thus we have selected from the repositories only those publications that are marked explicitly as written in German. However, most German publications have German and English abstracts. We did not include a language detection but simply assumed that the first abstract is the German one. We did not find any counterexample. The universities have an emphasis on fundamental research and are internationally oriented. Hence they publish mainly in English. The majority of their German publications are PhD-theses. In contrast the universities of applied sciences (UAS) are regional oriented and have an emphasis on knowledge transfer. Moreover, they usually don't have PhD-Students. Thus, they have a lot of publications in German that are intended to inform professionals in industry about new research and developments. The Universities of Education have a position in between with PhD-theses but also a lot of other German publications.

Each of the repositories mentions a subject area of the publication. For all repositories this is a DDC class at the second hierarchy level. However, some of the repositories use the class 004 for computer science. We did not take this exception into account. All records with this label consequently will have a recall and mean reciprocal rank of 0 for every classification method.

5 Results

All analyzed records provide a subject area that is in fact a second level DDC notation. It has to be noted that in many cases there is more than one possible label that could be regarded as true and a more or less arbitrary choice had to be made by the annotators. In fact labels closely related to the ground truth could be considered as correct as well ([8]). Furthermore, on closer inspection of the results, it turns out that in some cases of mismatch, the predicted label is the correct one and the label given by the repository was wrong ([17]). In the following we will nevertheless use these labels as the ground truth and consider only exact matches as being correct.

Since we consider assignment of DDC Notations as a classification task, in which each record should be assigned to exactly one category we have to observe the results for each record and not for each category, like one would do

Table 2. Mean reciprokal rank (MRR) and recall at 5 at first and second DDC level for SWD based classification of OAI Metadata from the SerWiss repository of the Hochschule Hannover using different fields and two classification methods, sci. the SWD-based classification and the ACT-DL classification service.

	SWD Based		ACT-DL	
Fields	MRR	rec@5	MRR	rec@5
title + abstract (DDC level 1)	0.68	0.89	0.67	0.90
title + abstract (DDC level 2)	0.48	0.66	0.39	0.37
title + keywords (DDC level 2)	0.61	0.76	0.32	0.39
title + abstract + keywords (DDC level 2)	0.61	0.77	0.39	0.47

in a retrieval setting. Thus the evaluation presented here differs from the one used in [19] who use the retrieval perspective. Since the algorithm produces a ranked list of results, we use mean reciprocal rank as an evaluation measure. Automatic classification might be used in a setting where a subject librarian is given suggestions for manual classification. Here it would be important that the correct label is always among the top 5 or top 10 results. Thus we also consider the recall at the fifth position in the ranked list (recall@5). Note that for each individual record the recall@5 is always 0 or 1.

Table 2 gives the results for classification of records form the Hochschule Hannover using different fields for both the ACT-DL classification service and the method presented in this paper. Though ground truth labels are given at the second level of the DDC-hierarchy, we can of course also evaluate the results at the first level. These first level results are given on the first line of the table.

Table 3. Mean reciprocal rank (MRR) and recall at 5 at second DDC level for SWD based classification of OAI Metadata of records from 7 German OAI-repositories using two classification methods, sci. the SWD-based classification and the ACT-DL classification service.

	SWD Based		ACT-DL	
Repository	MRR	rec@5	MRR	rec@5
Hanover UAS	0.48	0.66	0.39	0.37
Cologne UAS	0.32	0.44	0.35	0.39
Frankfurt UAS	0.55	0.75	0.49	0.62
TU Berlin	0.41	0.61	0.59	0.66
Univ. Hildesheim	0.25	0.39	0.25	0.29
Univ. Regensburg	0.61	0.80	0.65	0.72
Freiburg UE	0.53	0.75	0.33	0.36

Of course the results using the keywords gives the best results but is of least practical relevance for a library that wants to speed up the process of metadata generation for new publications. In the realistic situation only the abstract is

available, or author provided keywords that might be of less quality than the subject headings assigned by a librarian with in-depth knowledge of the subject authority file. Thus in Table 3, where we compare results for 6 more repositories, we use only title and abstract for classification. All the differences between the two methods are significant at the level of 0.001 according to the Wilcoxon signed rank test.

Finally, we have compared the results for different publication types. These results are given in Table 4.

Table 4. Mean reciprocal rank (MRR) and recall at 5 at second DDC level for SWD based classification of OAI Metadata for 6 most frequent publication types from 7 repositories.

		SWD Based		ACT-DL	
Publ. type	#records	MRR	rec@5	MRR	rec@5
PhD Thesis	2503	0.47	0.66	0.63	0.70
Master thesis	277	0.32	0.44	0.37	0.41
Essay	195	0.59	0.75	0.29	0.36
Monograph	176	0.46	0.62	0.43	0.51
Festschrift	106	0.43	0.76	0.38	0.40
Lecture	84	0.40	0.96	0.03	0.06

6 Discussion

The results of the SWD based approach are similar to those given by ACT-DL, which is rather surprising given the simplicity of our approach. Especially the recall@5 is very good for the SWD based approach as compared to the machine learning method: For 6 out of 7 repositories the recall@5 was even better. The mean reciprocal rank is 3 out of 7 cases better, in 1 case the same and in 3 cases worse. This shows that our method is rather successful in getting the correct label among the best 5 candidates but has difficulties to decide which one to put on top. A detailed analysis of a small subset shows that in many cases the first and the second result have the same weight and the ordering is arbitrary. Here machine learning techniques considering statistical relations between words of different categories or at least using a priori probabilities for the categories could improve the results.

The results split up for different publication types are probably most interesting. Especially, the SWD-based approach is able to outperform the machine learning approach for the less typical publication types. It is likely that there were not many examples of these publication types in the training data for the ACT-DL classifier, while at the same time there is a considerable difference in vocabulary between the publication types. Thus, it should be fairly easy to adjust the classifier by including additional training documents to get better results for

these publication types as well. Nevertheless it shows the weakness of the machine learning approach: it is extremely dependent on the proper composition of the training data. The thesaurus based approach on the other hand, might not reach the best possible results, but is independent of training.

The quality of results that can be achieved with the thesaurus based approach of course depends on the coverage and quality of the thesaurus. In the work presented here we could show that the enriched version of the German Subject Heading Authority File (SWD) is a high quality resource for classifying German scientific records into DDC-classes.

7 Conclusion and future work

We have shown that the SWD with the mapping of SWD-subject headings to DDC classes provides a very valuable resource that can be used for classification of scientific records. With basic methods we could already achieve results that are comparable to results from state-of-the-art machine learning algorithms. There are various possibilities for improvement. In the first place, the SWD is a file of subject headings. Especially for complex or ambiguous concepts subject headings are often formulated in a way that might never be found in a running text. Thus, lexical enrichment might improve the results. Furthermore, we have simply counted the links to DDC classes. This works only, to some extend, if the number of terms per class is well balanced. In general, it does not have to be that case that class from which the most terms are found, is also the most likely class for the text. Here a machine learning approach as proposed by [5] could be used.

Another issue for further research is the reason why for some repositories the SWD-based approach is better, while for others the trained classifier is superior. The difference can partly be explained by the different distribution of text types. The reason might also be hidden in some properties of the repository but also in the policies and habits of the libraries that assign the labels that we have used as ground truth.

References

1. Leitfaden zur Vergabe von DDC-Notationen an SWD-Schlagwrtern (September 2010), `\url{http://linux2.fbi.fh-koeln.de/crisscross/CrissCross_Endg_Grundlagenpapier_Sept2010.pdf}`
2. Addis, A., Angioni, M., Armano, G., Demontis, R., Tuveri, F., Vargiu, E.: A novel semantic approach to document collections. In: Isaías, P., Paprzycki, M. (eds.) IADIS Multi Conference on Computer Science and Information Systems. vol. 2008, pp. 73–85 (2008)
3. Bloehdorn, S., Hotho, A.: Text classification by boosting weak learners based on terms and concepts. In: Rastogi, R., Morik, K., Bramer, M., Wu, X. (eds.) Proceedings of the 4th IEEE International Conference on Data Mining (ICDM 2004), 1-4 November 2004, Brighton, UK. pp. 331–334. IEEE Computer Society (2004)
4. Boteram, F., Hubrich, J.: Specifying intersystem mapping relations: Requirements, strategies and issues. Knowledge Organization 37(3), 216–222 (2010)

5. Chenthamarakshan, V., Melville, P., Sindhwani, V., Lawrence, R.D.: Concept labeling: Building text classifiers with minimal supervision. In: Walsh, T. (ed.) IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011. pp. 1225–1230. IJCAI/AAAI (2011)
6. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: A framework and graphical development environment for robust nlp tools and applications. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA. pp. 168–175. ACL (2002)
7. Gazendam, L., Wartena, C., Brussee, R.: Thesaurus based term ranking for keyword extraction. In: Tjoa, A.M., Wagner, R. (eds.) Database and Expert Systems Applications, DEXA, 10th International Workshop on Text-based Information Retrieval, TIR. pp. 49–53. IEEE (2010)
8. Gazendam, L., Wartena, C., Malaisé, V., Schreiber, G., De Jong, A., Brugman, H.: Automatic annotation suggestions for audiovisual archives: Evaluation aspects. Interdisciplinary Science Reviews, 34 2(3), 172–188 (2009)
9. Hubrich, J.: Crisscross: Swd-ddc-mapping. Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen & Bibliothekare 61(3), 50–58 (2008)
10. Hubrich, J.: Thematische suche in heterogenen informationsrume. In: Bergner, U., Gmpel, E. (eds.) The ne(x)t Generation, das Angebot der Bibliotheken: 30. sterreichischer Bibliothekartag Graz 15. - 18.09.2009. Schriften der Vereinigung Österreichischer Bibliothekarinnen und Bibliothekare, vol. 7, pp. 234–242. Neugebauer, Graz-Feldkirch (2009)
11. Jacobs, J.H., Mengel, T., Müller, K.: Benefits of the crisscross project for conceptual interoperability and retrieval. In: Gnoli, C., Mazzocchi, F. (eds.) Paradigms and conceptual systems in knowledge organization. Proceedings of the Eleventh International ISKO Conference. pp. 236–241. ERGON-Verlag (2010)
12. Joan, S. (ed.): Dewey Dezimalklassifikation und Register, Dt. Ausg., vol. 1. Sauer, München, 22 edn. (2005)
13. Lagoze, C., Van de Sompel, H., Nelson, M., Warner, S.: Open Archives Initiative-Protocol for Metadata Harvesting-v. 2.0. Open Archives Initiative (2002), `http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm`
14. Malaisé, V., Gazendam, L., Brugman, H.: Disambiguating automatic semantic annotation based on a thesaurus structure. In: Hathout, N., Muller, P. (eds.) Actes de la 14e conférence sur le Traitement Automatique des Langues Naturelles (communications orales). pp. 197–206. Association pour le Traitement Automatique des Langues, Toulouse (2007)
15. Schmid, H.: Improvements in part-of-speech tagging with an application to german. In: Proceedings of the ACL SIGDAT-Workshop (1995)
16. Scott, S., Matwin, S.: Text classification using wordnet hypernyms. In: Use of WordNet in natural language processing systems: Proceedings of the conference. pp. 45–51. Association for Computational Linguistics (1998)
17. Sommer, M.: Automatische Generierung von DDC Notationen für Hochschulveröffentlichungen (2012), Bachelor Thesis
18. Tiun, S., Abdullah, R., Kong, T.E.: Automatic topic identification using ontology hierarchy. In: Gelbukh, A.F. (ed.) Computational Linguistics and Intelligent Text Processing, Second International Conference, CICLing 2001, Mexico-City, Mexico, February 18-24, 2001, Proceedings. Lecture Notes in Computer Science, vol. 2004, pp. 444–453. Springer (2001)

19. Waltinger, U., Mehler, A., Lösch, M., Horstmann, W.: Hierarchical classification of oai metadata using the ddc taxonomy. In: Bernardi, R., Chambers, S., Gottfried, B., Segond, F., Zaihrayeu, I. (eds.) Advanced Language Technologies for Digital Libraries - International Workshops on NLP4DL 2009. Lecture Notes in Computer Science, vol. 6699, pp. 29–40. Springer (2011)
20. Wang, J.: An extensive study on automated dewey decimal classification. Journal of the American Society for Information Science and Technology 60(11), 2269–2286 (2009)
21. Wartena, C., Brussee, R., Gazendam, L., Huijsen, W.: Apolda: A practical tool for semantic annotation. In: Database and Expert Systems Applications, DEXA, 7th International Workshop on Text-based Information Retrieval, TIR. pp. 288–292. IEEE (2007)