Enriching Documents with Context Terms from Cross-Domain Ontologies

Benjamin KÖHNCKE,
Wolf-Tilo BALKE
L3S Research Center, Hannover, Germany

Entity-centric search has become a demanding problem for many domains on the Web. In particular, the suitable contextualization of result documents poses challenges in terms of selecting most adequate indexing terms for later retrieval. This holds even more, if no generally recognized ontologies for the respective domain are available. In this paper, we show that cross-domain ontology terms are actually more useful for indexing, than salient keywords taken from the documents. Moreover, learning typical contexts for groups of entities from collections indexed by strong cross-domain ontologies can considerably improve retrieval effectiveness. Our extensive experiments prove these results on real world document collections from the area of chemistry and computer science. In fact, our evaluation in different document retrieval scenarios show a vital increase of retrieval precision of up to 87% using documents annotated with cross-domain ontology terms as compared to 53% for BM25 searches and 43% for documents annotated with Wikipedia categories.

1. Introduction

Today's information searches are largely based on Web search. But due to the huge amount of available information the list of possibly relevant results is large for all kind of queries. To retrieve more focused results the users usually refine their queries by adding additional search terms in case the results have not been satisfying [1]. Due to these terms the context of the search request is defined. The problem is that these context terms must match the documents vocabulary. Otherwise, a lot of relevant results cannot be retrieved leading to a low Recall. This problem is widely known as the vocabulary problem [2].

One often used solution is to do a query expansion with meaningful, context-dependent terms relevant for the domain (often called salient terms). On the other hand, selecting too many, too few, or simply inadequate context terms risks missing relevant documents since, of course, only documents will be retrieved containing these terms. In our previous work, we analyzed different approaches, e.g. query expansion and Latent Semantic Analysis (LSA), and showed that they are not useful to solve this problem [3]. We further presented an approach in [4] computing the context of the documents and query terms on-the-fly using a semantic similarity measure based on Wikipedia. Other approaches focus on using knowledgebases with a fixed terminology for describing the search context [5].

In general, information contained in terminologies (or more general: ontologies) forms very useful background knowledge for classifying documents in a context-aware fashion. A prime example can be found in the biomedical domain where the National Library of Medicine (NLM) uses the MeSH (Medical Subject Headings) ontology to annotate and index documents from biomedical journals [6]. MeSH defines a set of controlled vocabulary thesaurus including a set of description terms that are hierarchically organized. All these annotated documents are included in MEDLINE which is currently the largest biomedical literature database. Web based interfaces have been developed to search over MEDLINE and other related collections. The most commonly used Web interface is PubMed comprising more than 21 million items of biomedical literature. However, MEDLINE indexed by MeSH is a rare case and is actually curated manually with expensive efforts. Most domains miss such overarching ontologies to annotate documents with suitable context information.

The goal of this paper is to overcome the lack of context annotations for domains that do not offer general ontologies. The idea is to use cross-domain knowledge from different, but related domains. The contribution of this paper therefore is to show that cross-domain knowledge is indeed useful to improve the general retrieval quality. We extract named entities from documents annotated with ontology terms and train classifiers to predict these ontology terms based on the extracted named entities. Documents from related domains are annotated with ontology terms based on these classification models. To ensure that the annotated terms are semantically related to the documents' context a semantic processor is introduced. The semantic processor computes the semantic similarity between the associated ontology terms and the document's named entities to filter unrelated terms. This computation is based on a general knowledgebase that acts as some kind of "glue" between the ontology terms from the source domain and the named entities used in the target domain.

For evaluating our approach we choose chemistry as an example domain, since here the search is entirely entity-centered and chemical documents still lack suitable context annotations on a large scale. To prove the generalization of our approach we also enrich documents from the domain of computer science with terms from the related domain of mathematics. In mathematics, documents are annotated with terms from the Mathematics Subject Classification (MSC) that also contains a whole sub-tree dealing with computer science. Our results prove that by annotating documents with terms from a controlled vocabulary the retrieval precision in context dependent searches can be dramatically increased from 53% using a BM25 ranking model to up to 87% with semantic, cross-domain annotations. We further show that simple query expansion with domain-specific terms is no suitable option in entity-centered searches. The results of our evaluation show that cross-domain annotations allow for high quality context dependent searches.

The rest of the paper is organized as follows: In section 2 we will give an overview of the related work. Section 3 shows a motivating example in a use case scenario from the domain of chemistry. In section 4 we describe our approach followed by a detailed evaluation in section 5. Finally we close with a short summary and give an outlook of our future work.

2. Related Work

An important aspect for information providers is how information is provided to the individual user. Due to the massive availability of documents in digital form it is necessary to annotate and classify them to assure a satisfying search experience for each user. The area of automated text categorization is a wide field dating back to the early '60s. Central approaches in the '80s were usually based on knowledge engineering, where a human expert defined a set of rules to classify documents under the given categories. Due to the machine learning paradigm and more powerful hardware devices the knowledge engineering approach lost popularity in the research community in the '90s. In machine learning a general inductive process automatically builds a classifier by learning the interesting characteristics from a set of pre-classified documents. Nowadays, text categorization plays a major role in information systems and is applied in many contexts, like e.g. document indexing or filtering, automated metadata generation or word sense disambiguation. An overview of machine learning in text categorization is given in [7].

The idea of classifying documents to enable context-sensitive document retrieval is currently discussed in several papers. In [8] an approach is presented to optimize Web search results using individual user preferences including preferred search contexts. The search contexts are discovered from raw query logs. It was shown that in terms of top-k search quality a system using context information outperforms existing personalization approaches without context information. In contrast, the approach in [5] does not use user profiles, but defines a query model extending conventional keyword queries and allowing users to specify their search contexts. The search context is defined by a subset of documents that the user is interested in. The idea is to use keyword statistics based on the user-defined context instead of using global, collection wide statistics, like e.g. TF*IDF, to rank the documents. Since these statistics cannot be computed at indexing time the challenge is to efficiently compute them at query time. The authors reduce this problem by evaluating aggregation queries and leverage materialized views to improve query performance. A more advanced approach described in [9] uses semantic information extracted from texts and some domain ontology to approximate concepts associated with documents. Since for document classification, respectively context annotation, it is necessary to know the set of possible classes in advance, using the controlled vocabulary and semantic relations of an ontology is beneficial.

For example in the biomedical domain documents are annotated with one or more terms from the MeSH ontology. This ontology defines a controlled vocabulary specifying a variety of concepts in (biomedical) science. Each MeSH term represents a concept and a combination of these terms represents the context spanning the corresponding concepts. A researcher can use tools that visualize the MeSH ontology for specifying his/her search context by browsing through the ontology and selecting terms that are relevant for his/her context. An example is the GoPubMed portal (see http://www.gopubmed.com) where the user can do facetted searches by navigating through the MeSH ontology and filter the PubMed document corpus by choosing suitable ontology terms. In [10] it is also shown that the MeSH ontology is a valuable resource for representing MEDLINE documents at different abstraction levels. The authors evaluated the suitability of the ontology for classifying biomedical documents using a Bayesian Network classifier. Furthermore, it was shown that the classification accuracy can be improved by increasing the number of MeSH terms used for representing a document. Another approach trying to extend the ontology-based representation of biomedical documents is described in [11]. The initial MeSH annotations of biomedical documents have been extended with semantically similar concepts from the MeSH ontology. A simple edge-count similarity measure was used to evaluate the semantic proximity between different concepts.

In [12] an approach is presented focusing on the automatic annotation of MeSH terms to biomedical documents. Different classification systems are compared to reproduce manual MeSH annotations. Experiments also showed that the retrieval quality for biomedical documents can be improved by automatically annotating the user query with MeSH terms. A similar approach dealing with automatic query expansion in MEDLINE but using a pseudo-relevance feedback technique is described in [13].

Whereas for the biomedical domain a lot of work has been done to assist domain experts in searching for literature, other domains, like e.g. chemistry, still lack behind. Therefore, the most comprehensive database for chemical entities is still created manually by the Chemical Abstracts Services (CAS) as part of the American Chemical Society. CAS spends a tremendous amount of funding in the manual indexing of journal articles, conferences, patents and many other research publications in the chemical domain. Currently CAS registry comprises over 50 million of substances, but the access is strictly limited to subscribers at a price of about 30,000 USD/year for a single user subscription. Obviously for the growing open access movement this type of indexing documents is not a viable option.

Our approach shown in [14] describes a way of indexing chemical documents by building enriched index pages including different entity representations and synonyms. The entity recognition process was done automatically using the OSCAR framework [15]. We showed that using these pages also textual query interfaces, like e.g. Google or Yahoo!, can be used to search for chemical documents. Since indexing of chemical documents based on their contained entities can already be done automatically, again, the challenge is to enable context dependent searches to restrict the result set and enhance search precision. Although some Web portals for searching for chemical documents are freely available, like e.g. ChemXSeer or the ViFaChem portal , none of them allows for context-aware retrieval. The reason is that there is no suitable knowledgebase in chemistry offering a defined vocabulary comparable to the MeSH ontology in the biomedical domain. A possible approach might be the

automatic creation of ontologies. But, unfortunately, the quality of automatically generated ontologies for such complex domains as chemistry is not yet sufficient [15].

The approach in [16] discusses the problems of cross-domain knowledge transfer. The main focus lies on the problem that for classification training and test data have to follow the same distribution. Since for cross-domain classification this is usually not the case a two stage algorithm is presented based on semi-supervised classification. In [17] an approach enabling cross-domain search by exploiting Wikipeida is shown. The focus is on analyzing tags used in Web 2.0 systems like Flickr and connect them to concepts in Wikipedia. Other approaches use Wikipedia directly to improve document retrieval.

In [18] an approach is presented using machine learning techniques with Wikipeida to enrich document retrieval. The same authors presented a concept-based retrieval approach based on Explicit Semantic Analysis (ESA) in [19]. Their results show the usefulness of Wikipedia to compute semantic relatedness of natural language text. Another approach presented in [20] uses Wikipedia concept and category information for enriched document clustering. They argue that using ontology knowledge may lead to information loss or introduce noise.

In contrast to previous work, in our approach we enrich documents with cross-domain ontology terms and use Wikipeida to ensure that the associated terms are semantically related to the document to enable context-driven information retrieval.

3. Use Case

To show the importance of context dependent searches we present a use case from the domain of chemistry. Consider the typical work of two researchers working in different areas of chemistry: take for instance Frank, a synthetic chemist specialized in the synthesis of organic compounds of pharmalogical interests. Assume the second one to be Cathy, an analytical chemist specialized in forensic toxology. Search for chemical documents usually is induced by searching for substances, either by name or by graphical structure. Actually, as a starting point both researchers may be looking at the same class of compounds called synthetic cannabinoids, but of course from different angles and with a different focus.

As a synthetic chemist, Frank may be looking for compounds of the naphthoylindole family acting as analgesics. During his research he finds the substance *1-pentyl-3-(1-naphthoyl)indole*, a full agonist at both the CB_1 and CB_2 cannabinoid receptors, with some selectivity for CB_2. Now, to complete his work he is especially interested in documents describing synthetic methods for the preparation and isolation of these compounds as well as possible derivates, on lab scale with highest possible yield. Moreover, also older documents might be relevant as they often contain processes that are not covered by expensive to acquire patents.

In contrast, Cathy as an analytical chemist may be working on the analysis of a drug sample, which she believes to be Spice, a blend of synthetic cannabinoids. Besides herbal ingredients Spice contains a large and complex variety of synthetic cannabinoids, most often *cannabicyclohexanol* or *HU-210*. In fact, the trivial name *JWH-018* also represents the chemical substance *1-pentyl-3-(1-naphthoyl)indole*. But in this case Cathy is looking for documents containing information about analytical information of compounds contained in the drug spice, especially *JWH-018*. She is interested in instructions for sample preparation and workup, analytical methods, describing on how to perform both qualitative and quantitative analysis with spectral methods like mass spectrometry or NMR spectroscopy. Furthermore documents with reference of analytical data of *1-pentyl-3-(1-naphthoyl)indole* will be of special interests, because these data can be used as a reference for her own findings.

Please note that to search for relevant literature, both chemists use the same entity name, but in order to get focused information will have to filter the query results with context dependent terms. Let us assume that the chemists have access to a document repository containing documents from the Beilstein Journal of Organic Chemistry (BJOC) and the Eurasian Journal of Analytic Chemistry (EJAC). The BJOC contains more relevant literature for Frank, whereas EJAC includes mostly relevant documents for Cathy. For both fields we defined a set of general terms that should be used for query result filtering. Frank's set includes terms like e.g. synthesis, reduction, reaction, catalysis or oxidation. For Cathy the term set contains, among others, spectroscopy, separation, analysis, and chromatography. We created an Apache Lucene index, indexing 227 documents from both journals and performed a Boolean search for the substance with and without specific expansion terms. We used the following retrieval model: Let e be the query entity and $C=\{c_1,c_2,\ldots,c_n\}$ the set of all context terms. For the filtered query the queries are formulated as e *AND* $(c_1$ *OR* c_2 *OR* $\ldots$ *OR* $c_n)$, meaning all documents are returned containing the query entity and at least one context term.

For the substance only query a total of 23 documents is retrieved, 14 from the EJAC journal and 9 from the BJOC journal. The expanded search with Frank's synthetic chemistry term set retrieves 19 documents, 9 from BJOC and 10 from EJAC. For Cathy's query 13 docs are retrieved, 10 from EJAC and 3 from BJOC. Obviously the perfect result would have been that for Frank only the 9 BJOC journal documents are retrieved whereas for Cathy only the 14 EJAC documents are relevant. However, in this little example it is not possible to distinguish the documents from the different journals by filtering the results using context terms. It is also remarkable that for Cathy not all relevant documents have been retrieved leading to a lower recall for the filtered set. This shows that it is quite important to carefully choose all relevant context terms. We will later see in the experimental section that also statistical query expansion does not lead to better results. Thus, to enable context searches without losing important information it is necessary to enrich the documents with more general concept terms from some suitable controlled vocabulary.

4. Enriching Documents with Cross-Domain Knowledge

This section describes our approach for annotating documents with cross-domain context terms. The idea is to learn this context from collections that have already been indexed by strong ontologies although for slightly different domains. We define the search context as any set of terms from the source ontology. If the term is not contained in the source ontology it cannot be used as a context term.

Example: Imagine a user who is interested in documents relevant for a named entity in the context of 'computer science'. The term is searched in the source ontology and all sub-terms of the node 'computer science' are considered as relevant context terms. In this case, the search context is very general, resulting in many relevant context terms. Of course, if the user chooses a more specific context term the set of associated ontology nodes will be smaller.

Our system consists of three main parts. An overview of our proposed workflow is shown in Figure 1:

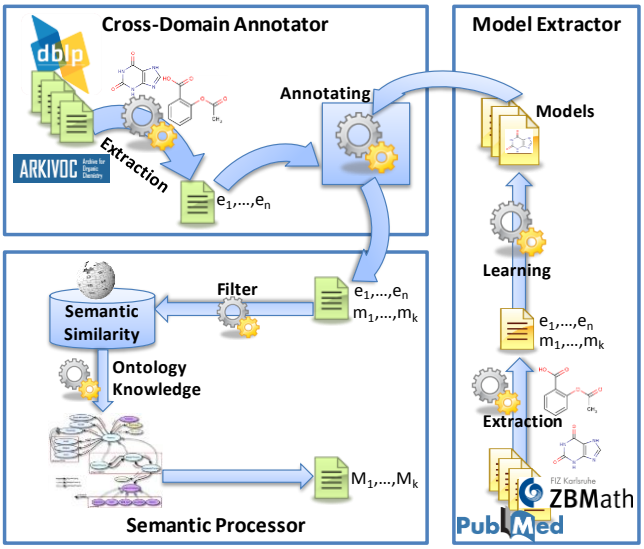

Figure 1. System overview

Model Extractor: First of all it is necessary to train classifiers to learn suitable models. Therefore, we take domain specific documents that have already been annotated with ontology terms and extract named entities. For example, we took MeSH annotated MEDLINE documents and extracted all chemical entities using the OSCAR framework. Afterwards for each document we have a list including named entities and a list with associated ontology terms. This information is used to learn a classification schema using the WEKA toolset [21]. We experimented with different classifiers. The results are shown in section 5.3.

Cross-Domain Annotator: Once the classifier has learned a model for each ontology term based on the set of named entities, these models are now used to annotate documents from related domains with ontology terms based on their contained named entities. To do this, the first step is to extract all named entities from the documents, e.g. by using the OSCAR framework for chemical documents. Afterwards the learned models are used to predict a set of adequate ontology terms for each document. For each assigned term a confidence value is given indicating the probability that the term was correctly assigned to the document.

Semantic Processor: The semantic processor takes the annotated documents from the annotator. The goal is to filter the set of associated terms and only keep the most relevant terms with respect to the entities included. For each entity e from the set of all entities E and for each ontology term m from the set of all terms M we compute the semantic similarity for each pair. The relevance of an ontology term for a document is the maximum of its semantic similarity values to any entity in the document. To compute this kind of similarity we need a knowledgebase containing both, named entities as well as ontology terms.

The most prominent general knowledgebase today is Wikipedia. Its usefulness for document retrieval compared to other knowledgebases, like e.g. WordNet or Open Directory Project (ODB), was shown in [22]. Furthermore, the provided knowledge is also useful for specialized domains like chemistry. In [23] we showed that Wikipedia category terms can be used to annotate chemical documents. The resulting document representations have been analyzed and voted by domain experts. The evaluation showed that the Wikipedia representation was considered to be very useful for domain experts.

Here, we use Wikipedia as "glue" to connect the domain-specific ontology terms and the vocabulary from the target domain. We compute the semantic similarity between any named entity and some ontology term in Wikipedia relying on the relatedness measure described in [24]:

$$relatedness(a,b) = \frac{\log(\max(|A|,|B|)) - \log(|A \cap B|)}{\log(|W|) - \log(\min(|A|,|B|))}$$

where a and b are two articles, A and B are the sets of articles that link to a, respectively b, and W is the set of all articles in Wikipedia. The relevance of an otology term m for a document d is then defined as:

$$relevance(d,m) = \max(relatedness(m,e_n))$$

where $e_n \in E$ and E is the set of all named entities occurring in d.

Finally, we have an ontology term vector assigned to each document where the ontology terms are ranked according to their Wikipedia relevance to the document's content. The extended documents are stored in our repository. When a new document is indexed the semantic similarity between each contained named entity to each term from the source ontology is computed. The results are stored in a relational database.

For performing a search, the query term is extended with suitable ontology terms by the semantic processor. For result set ranking the Dice similarity based on the sets of assigned ontology terms is computed as:

$$D_{sim} = \frac{2 * |D_m \cap Q_m|}{|D_m| + |Q_m|}$$

where Q_m is the set of assigned ontology terms for the query entity and D_m the set for the respective document. Finally the ranked document set is delivered to the user.

Example from chemistry: The chemical domain offers access to some highly specialized controlled vocabularies like for instance the *Chemical Entities of Biological Interest* (ChEBI [25]). However, experiments in [23] have shown that ChEBI terms are not too useful for annotating general chemical documents, in fact worse than general Wikipedia categories. The reason is that ChEBI focuses exclusively on a small subset of molecules, namely small molecules, which are either natural products or synthetic products used to intervene in processes of living organisms. Therefore, the key idea is to aggregate all the knowledge about chemical entities available in ontologies from other, but related domains. For instance, while the huge collection of MeSH-annotated MEDLINE documents mainly focuses on illnesses, it still relates them to drugs, i.e. chemical entities. Extensive discussions with domain specialists from different areas of chemistry showed that MeSH terms to some degree can be useful for describing properties of chemical entities. We thus use chemical entities occurring in MEDLINE documents to learn the associated MeSH terms.

Considering our chemist Frank from section 3 who was searching for literature on *1-pentyl-3-(1-naphthoyl)indole*. He submits the query q to our system. The query is handed on to the semantic processor which extends it with suitable MeSH terms. Please note, since Frank is a chemist only MeSH terms are used that are from the chemical sub-trees of the MeSH ontology. The extended query q_e is used for document retrieval. Here, a Boolean search is accomplished, meaning that all documents including the original query term q are retrieved. The extended query q_e is used to rank the documents according to the desired context (in this case chemistry).

5. Experiments

For the evaluation of our approach we used different document collections. For MeSH annotated biomedical documents, we took around 120,000 documents from PubMed Central (http://www.ncbi.nlm.nih.gov/pmc) which is a free full-text archive of biomedical and life sciences journal literature at the U.S. National Institutes of Health's National Library of Medicine (NIH/NLM).

Furthermore, for the chemical domain we used 2,700 documents from the journal Archive for Organic Chemistry (ARKIVOC) which is one of the most renowned open access sources for organic chemistry. To specifically focus on different contexts we took around 100 manually curated documents from the Beilstein Journal of Organic Chemistry (BJOC) which is an international, peer-reviewed open-access journal dealing with all aspects of *organic chemistry*. Furthermore, we curated around 130 documents from the Eurasian Journal of Analytic Chemistry (EJAC) which focuses on all aspects of *analytical chemistry* related with analytical methods, new instruments and reagents.

To prove the general usefulness of our approach we also did experiments with document collections from other domains. We took the Zentrallblatt Math (ZM) document repository containing 3 million documents. Each document is annotated with several terms from the MSC taxonomy. Furthermore, we took the DBLP computer science document repository containing 638,000 documents. Since these documents lack suitable annotations we use cross-domain knowledge from the ZM documents to improve the retrieval quality.

We performed the following experiments:

1. First, we evaluated whether a simple query expansion is already useful for entity centric search. We compared the term distributions of the EJAC and BJOC journal that are focused on different working fields: organic and analytical chemistry. Furthermore, we let domain experts define sets of context terms for both working fields. In addition, we also tried a statistical approach computing term-to-term co-occurrences for query expansion. Comparing the results using query expansion we can state that it is not a suitable choice for enabling context-driven retrieval in chemistry.
2. In the second experiment, we analyzed whether cross-domain knowledge can be useful for annotating chemical entities. We used the MeSH ontology to annotate chemical entities and discussed the results with domain experts. From the chemist's point of view the associated MeSH terms are comprehensible and quite useful to give insights on chemical properties as well as the applications scopes. Of course this experiment has more anecdotic character to give the reader an illustrative example of the annotated MeSH terms for a given chemical entity.
3. In the third experiment, we trained different classifiers to predict MeSH terms based on the chemical entities in a document. Our evaluation with a precision /recall analysis shows that it is indeed possible to predict MeSH terms using chemical entities.
4. In the fourth experiment, we use the learned classification models to annotate chemical documents with MeSH terms. Comparing different classifier confidence thresholds we present a semantic extension using Wikipedia semantic similarity to filter out irrelevant MeSH for chemistry.
5. Furthermore, we show in a document retrieval scenario that using the annotated documents context-driven searches are possible. We compare the results to a BM25 ranking and an enhanced baseline taking Wikipedia category information into account. The results indeed prove that our approach promise to dramatically increase the user's search experiences.

Finally, in the last experiment we prove the general usefulness of our approach to improve document retrieval. We enrich documents from the area of computer science with terms from the related domain of mathematics and evaluate the retrieval results.

5.1 Is it possible to Use Query Expansion?

The traditional way of searching for documents related to a specific context is to use query expansion. The user enters a query term and some context keywords, then all documents containing both terms are returned.

We did an experiment analyzing the word distribution of two chemical journals from different chemical working fields: organic chemistry (BJOC journal) and analytic chemistry (EJAC journal). If both collections use totally different terminology a query expansion should work to distinguish the documents.

We used Apache Lucene (http://lucene.apache.org/core) and the "Whitespace Analyzer" to index the documents. For the EJAC journal 55,350 and for BJOC 44,187 terms have been indexed. The overlap is indeed just 9,012 terms. Since the overlap between the two collections is quite small, it seems that query expansion should work fine. However, if we take a closer look at how often the different terms occur in the collections we immediately see that the terms occurring in only one collection are very rare (see Figure 1).

In contrast, the terms occurring in both collections are very frequent. The top-200 terms from EJAC and BJOC occur 12,787 times in the documents. Considering terms occurring in both collections the top-200 terms occur in more than 25,000 times in the documents. This leads to the assumption that query expansion is no suitable choice to distinguish documents from both collections.

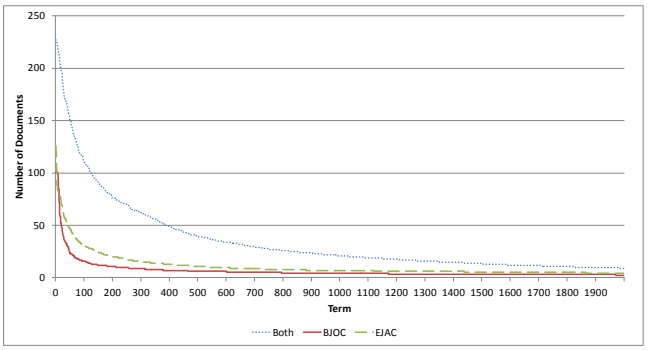

Figure 1. Comparing term distributions of different document collections

To prove this statement, we did a precision/recall analysis. As document collection we used the 2,700 documents from our ARKIVOC collection. Please note that these documents are from the same chemical sub-field as the BJOC collection: organic chemistry. For each of these documents we extracted all chemical entities using the OSCAR framework. Since relevance can only be assessed manually by domain experts (making it a very expensive process), we performed the precision/recall analysis only on a subset of documents (still about 10% of the entire collection). To choose a *representative* subset, we analyzed the number of occurrences of individual chemical entities in the document collection. Figure 2 shows the distribution of the 20,000 most often occurring chemical entities.

Since it is not sensible to choose entities for evaluation that occur either in almost all documents or are extremely rare, we chose our query entities for evaluation only from entities occurring in less than 100, but more than 20 documents (see the shaded area in Figure 2). We retrieved all documents matching the queries and randomly chose a subset of 10%. From these documents we randomly selected a total of 5% of the occurring entities resulting in 22 textual query terms.

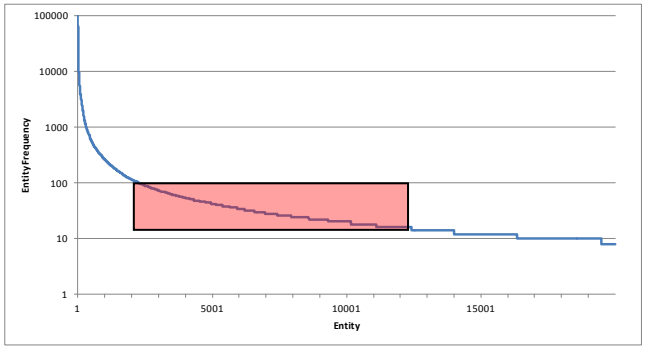

Figure 2. Entity distribution in collection

For a first experiment we also added the EJAC documents to our set and computed a Lucene index. Here we were interested in receiving all documents from the area of organic chemistry (ARKIVOC journal). All documents from the ARKIVOC journal containing the query term are marked as relevant. Of course, for a simple Boolean search without any context restriction all documents containing the query term have been found. But, there are also a lot of irrelevant documents leading to a low precision of only 31.6%.

To enhance precision we used a statistical query expansion method to define context terms. Since we are interested in documents for organic chemistry we computed a term-to-term co-occurrence matrix based on the ARKIVOC document subset. The position of each term in a document is also taken into account, meaning two terms that are close together, will get a higher score. Furthermore, we used popularity thresholds defining a required minimum and maximum popularity. Terms not fulfilling these thresholds are also not used as context terms. Finally, the query is expanded with the top-10 co-occurring terms using the query model introduced in chapter 3. This expansion leads to a small increase of the precision to 34.1%, but to a high decrease of the recall to 50.57%.

We also did a second experiment were domain experts considered all retrieved documents with respect to each query and judged the relevance in a binary fashion. As in our use case, we chose the sub-domain of synthesis chemistry for context search. The search is performed using a Lucene index on the documents. The average precision for a search using only the query terms is 17.1% which is very low. To enhance the precision the experts defined a set of typical context terms which are used for query expansion, like e.g. synthesis, reduction, reaction, catalysis or oxidation. But, using the combination of query term and context terms the precision actually decreased to 14.42%. Also the recall decreased to 45.1% meaning we miss relevant documents due to the context restriction. To ensure that the reason for the bad results is not the manual selection of the context terms, we also used a statistical

approach for context term selection. Here, we computed the term-to-term co-occurrence matrix based on all relevant documents (133 in total). But, as before, we could not get satisfying results. The precision increases compared to the manually selected context terms up to 23.1%, but the recall decreases to 41.4%.

These results prove that a simple query expansion is not useful for context-driven searches in chemistry. Therefore, we can state an urgent need for additional document annotations to enable context-driven searches.

5.2 Are MeSH Terms Useful for Describing Chemical Entities?

Please note, that this experiment has more anecdotic character to give the reader an illustrative example of the annotated MeSH terms for a given chemical entity. While analyzing the MeSH vocabulary with domain experts we found out that many of the included terms are also useful for describing chemical documents. Whole sub-trees of the ontology deal with chemical substances and general terminology. For example, 2,964 nodes are listed in the sub-tree for 'Organic Chemicals'.

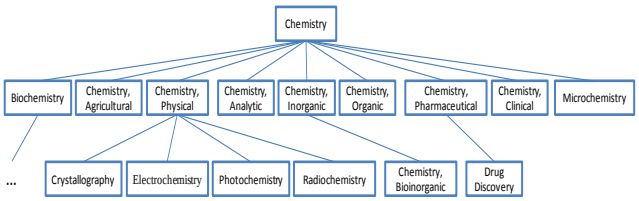

Figure 3. Extract of MeSH ontology for term 'chemistry'

Figure 4 shows an extract of the MeSH ontology dealing with chemical terminology for the node 'Chemistry'. The tree shows that there are different sub-nodes that represent different concepts from the chemical domain, like e.g. organic chemistry or analytic chemistry.

In a first experiment we tried to find out if the used terminology in MeSH is comprehensible for experts from the chemical domain. Therefore, we took the extracted chemical entities from our ARKIVOC collection and searched for them in our PubMed Central collection. In total we have 164,817 unique chemical entity names in the ARKIVOC collection. 151,287 (91.8 %) of them can also be found in PubMed Central.

To evaluate the MeSH vocabulary we annotated each chemical entity with a set of MeSH terms. We searched for the respective entity name in the titles and abstracts of the PubMed documents. If the name is found in the document, the document's MeSH terms are added to the entity's term set. We did not use the document's fulltext, because if the entity occurs in title or abstract, it should be more important for the document's context as if it occurs just somewhere in the fulltext. For each entity we created a tag cloud including all associated MeSH terms. As usual, the font size within the clouds is defined by the number of occurrences (i.e. the significance) of the respective term. We showed the tag clouds to domain experts and discussed, if they can associate the used terminology in the cloud with the chemical substance. From the experts' point of view the used terminology was comprehensive and while it contained some unrelated information, most of the terms were considered quite useful. To give an illustrative example, Figure 4 shows the MeSH term cloud for the chemical entity *Formaldehyde*.

Figure 4. MeSH term-cloud for Formaldehyde

For a long time, Formaldehyde was used in chipboards as agglutinant, respectively binding material. Due to its cancer-causing properties its evaporation leads to a contamination of the indoor air. Therefore, while not chemically relevant in a narrow sense terms like 'Air Pollution', 'Air Pollutants' or 'Indoor' occur prominently in the tag cloud.

Terms like 'Carcinoma', 'DNA', or 'Neoplasms' refer to the carcinogen effect that strongly confined the use of Formaldehyde. There are a lot of terms in the cloud dealing with the subject of cancer or biochemical processes. 'Receptors' indicates the cancer impact focused on biochemical aspects. Furthermore, the term 'Disinfectants' is one of its original fields of application, but still very useful for the individual chemists' context.

5.3 Predicting MeSH Terms Using Chemical Entities

In this experiment we aim at learning classification models to assign MeSH terms to documents based on their chemical entities. We tried different classifiers using the WEKA framework [21]. First of all, we needed to find out if chemical entities can be used to predict MeSH terms at all. For evaluation we took the 120,000 documents from the PubMed Central collection. Again, we used the OSCAR framework to automatically extract all chemical entities. From the set, around 114,000 documents include at least one chemical entity and could therefore be used for classifying. In total we found 151,287 unique chemical entities in the collection.

Of course, every document may have several MeSH terms. The problem is that WEKA does not support this kind of multi classes. Hence, it is necessary to train sev-

eral classifiers: One classifier for each MeSH term. Furthermore, it is important to get enough positive instances for each class to train the classifier. Therefore, we only used terms as classes that are included in at least 10 documents. Our goal is to predict the classes based on the chemical entities. Thus, we have for each MeSH term a file containing all chemical entities as attributes (around 150,000) and the respective MeSH term as class attribute. The instances are the documents represented in a sparse vector format where each dimension specifies the occurrence of the respective attribute. We did not choose *all* instances randomly, because then, due to our large dataset, the probability that most of the instances do not belong to the class is high. That would mean that during testing the probability is high that the classifier will not assign this class to an instance. Therefore, for each class we took all documents belonging to the class (positive examples) and randomly choose the same number of documents not belonging to the class (negative examples). Before training the classifier we used a filter to remove all irrelevant entities for the respective class. In total we trained 8,381 different classes.

We tried three different classifiers and compared their results in a precision/recall analysis. For all classifiers we used the default options and 10 times cross-validation. The results are shown in Table 1. The labels 'class yes' and 'class no' mean that the classifier predicts that a document has, respectively has not, the given class. The best classifier is the SVM having average precision and recall values of around 80% for all cases. The SVM implementation in WEKA is named SMO and implements the sequential minimal optimization algorithm for training a support vector classifier, see [26] for details. The results show that it is possible to use chemical entities for assigning MeSH terms to documents.

Table 1. Avg. precision/recall of different classifiers (in %)

Classifier	Class yes		Class no	
	Precision	Recall	Precision	Recall
naïve Bayes	79.62	77.98	**79.67**	78.50
C4.5 (J48)	79.10	66.99	71.50	**81.49**
SVM (SMO)	**79.72**	**80.00**	76.91	78.87

5.4 Annotating Chemical Documents with MeSH Terms

In this experiment we assigned MeSH terms to chemical documents and assessed their usefulness. We used the SVM classifier to annotate each of the 2,700 documents of our ARKIVOC collection. The classifier takes all entities from each document and applies all learned models. In total we have around 8000 different classes. Figure 6 shows the number of associated MeSH terms for each document. In average 3,316 terms are assigned to a document. If the classifier decides to assign a term (class) to the document also a confidence value is computed.

To know which terms are more related to chemistry than others we analyzed the MeSH ontology with domain experts and figured out important parts of the ontology for the domain of chemistry. The MeSH ontology consists of 19 main categories ranging from 'Anatomy' to 'Geographical Location'. Of course, not all of them are relevant from the chemist's point of view. From the 19 main categories we identified the 'Chemicals and Drugs' category to be of special interest for chemists. This category contains 20,249 subcategories covering for example a lot of different organic and inorganic chemicals. Another interesting subtree containing more general terms, called 'Chemistry', can be found under the 'Natural Science Disciplines' node in the 'Disciplines and Occupations' category (Figure 3).

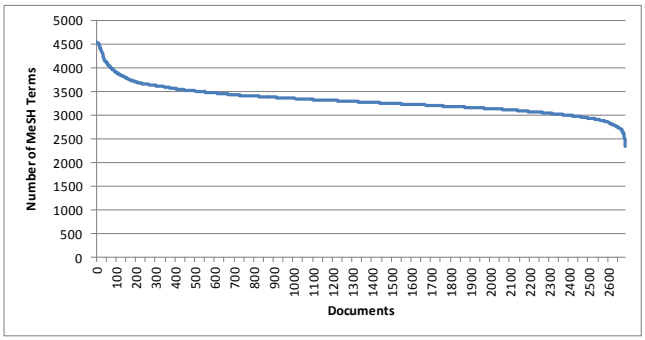

Figure 5. Number of assigned MeSH terms per document

To evaluate the usefulness of our approach we have to determine the quality of the assigned terms. Therefore, we defined that all terms from the chemical sub-trees are relevant. We took the assigned MeSH-terms from each ARKIVOC document and ranked them according to their confidence value. Then we took the top-k terms and computed the Mean Average Precision (MAP) for varying values of k.

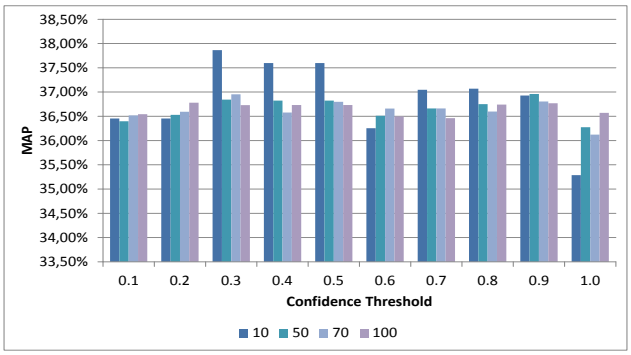

Figure 6. MAP for varying confidence thresholds for top-k MeSH terms

Figure 6 shows the MAP for varying values of k and different confidence thresholds. A confidence threshold of, e.g. 0.5 means that each assigned term has *at least* a confidence of 0.5. The MAP is quite low for almost all confidence thresholds (around 37%). The highest value is reached for a confidence threshold of 0.3 for the top-10 MeSH terms. Here, the MAP is 38% meaning that from 10 assigned terms only 4 are relevant for the area of chemistry. The problem is that the confidence value does not describe how the term is semantically related to the document. It only says to what percentage the classifier is sure that the term has to be assigned to the document. To

further enhance the quality of the assigned terms we need a semantic filter. Therefore, we used Wikipedia to compute the relevance of a MeSH-term for the respective document. The relevance is defined as the maximum semantic similarity of an assigned MeSH-term compared to each chemical entity occurring in the document. Again we measured MAP, this time varying the relevance threshold.

Figure 7 shows the results for varying Wikipedia relevance thresholds. The results show that the MAP is much better using the Wikipedia relevance. For the Top-10 assigned terms the best MAP (78%) is reached for a relevance threshold of 0.6. However, for the top-50 to top-100 terms the MAP drops to around 65%. Regarding all top-k terms, a threshold of 0.7 retrieves the best results with a MAP of always at least 74%.

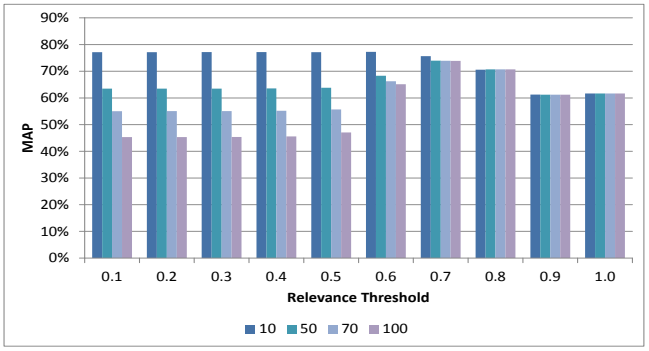

Figure 7. MAP for varying Wikipedia relevance thresholds for top-k MeSH terms

This experiment has shown that using the knowledge from Wikipedia or similar sources can dramatically increase the quality of the assigned MeSH-terms. The combination of MeSH-terms and Wikipedia seems to be quite useful to enrich chemical documents, supporting our results in [23].

5.5 Using MeSH for Chemical Document Retrieval

In this experiment we analyzed whether the assigned cross-domain MeSH terms really lead to suitable improvements for chemical document retrieval. As document sets we used our PubMed Central (PMC), ARKIVOC and BJOC collections. In contrast to PMC the documents from ARKIVOC and BJOC are *all* from the area of organic chemistry and are therefore closely related. In total we got 120,000 documents. We randomly chose 25 query terms out of all chemical entities from our collection. We are interested in documents containing the respective query entity in the context of organic chemistry. For each query we took 50 documents from the organic chemical journals and 50 documents from PMC. We only took documents where the respective query entity occurs in title or abstract. The relevance was assessed manually by domain experts. For each of these queries we computed the semantic similarity to each of our learned MeSH terms using Wikipedia. We assigned all MeSH terms with a relevance threshold of more than 0.7 to the respective query term. Since we are interested in retrieving all documents in the context of organic chemistry, we filtered the assigned MeSH terms to only use terms from the respective sub-tree of the MeSH ontology.

All documents in our set are already annotated with MeSH terms. The PMC documents in our collection have on average about 10 MeSH terms. Therefore, we also used the top-10 terms for our chemical documents. Terms are ordered by Wikipedia's relevance score. For performing a search all documents containing the respective query term are retrieved. For result set ranking we computed Dice similarities on the sets of assigned MeSH terms.

To evaluate if the annotation of MeSH terms leads to better retrieval results we compared the results to two different baselines. The first baseline uses the BM25 ranking model with standard parameters. We searched for the 25 query terms using a Lucene fulltext index without additional MeSH terms for the chemical documents. Secondly, we compared our approach to a Wikipedia category baseline to evaluate the retrieval improvement of the semantic processor. As described in [23] we annotated each document with Wikipedia categories based on the chemical entities contained. Also all query entities are annotated with Wikipedia categories. All documents containing the query are retrieved and ranked using Dice similarity based on annotated categories.

To compare the different rankings we computed the mean average precision (MAP) for the top-k documents over all queries (see Figure 8).

For the BM25 ranking precision values are around 35% with the highest value of 35.54% for the top-20 documents. The Wikipedia ranking has a low precision value of 23.5% for the top-5 which increases to 43.2% for the top-45 documents. But, using MeSH annotations average precision can be dramatically improved: For the MeSH ranking precision values are almost constant around 83% with the highest value of 87% for the top-30 documents.

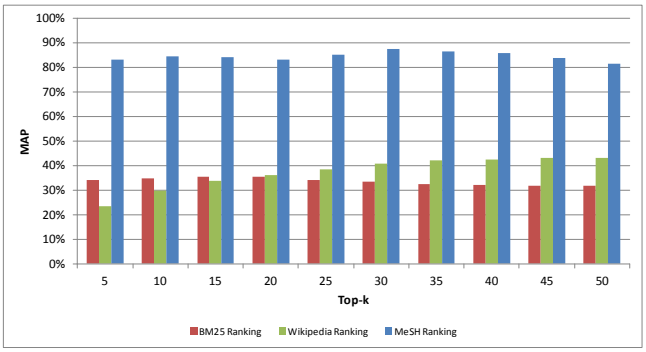

Figure 8. MAP for top-k documents

The results show that using knowledge about chemical entities from other domains for extending chemical documents promises a high increase of the retrieval quality for domain experts. Without additional annotations the top-k result sets include more than 60% of irrelevant hits. With Wikipedia annotations this number can still be decreased to 50%. Using semantically enriched documents only 15% of the retrieved results are irrelevant.

5.6 Enriched Retrieval for Computer Science

In this experiment we prove the general usefulness of our approach. We took documents from the ZM repository, where each document is annotated with several terms from the MSC taxonomy. While analyzing the taxonomy we found out that a whole sub-tree is relevant for the related domain of computer science. Therefore, we took the DBLP document repository containing 638,000 documents from computer science. Since these documents lack suitable annotations we use cross-domain knowledge from the ZM documents to improve the retrieval quality.

We extracted named entities from the ZM documents and trained a SVM classifier to learn the MSC classes. For entity extraction we used the Wikipedia Miner which annotated all entities matching Wikipedia articles. We also extracted named entities from the DBLP documents and associated MSC classes based on the learned classification models. The assigned MSC classes are filtered using our semantic processor. Finally, the usefulness of the annotations is evaluated in a document retrieval experiment.

We randomly choose 30 query entities and took 150 documents containing these entities from DBLP and 150 documents from ZM. The relevance of each document for each query was manually judged by a group of 10 domain experts. All experts are Ph.D. students or postdoctoral researchers from the field of computer science. The goal is to find documents containing the query term which are relevant for the context computer science. As described for the MeSH experiments the query term is associated with terms from the MSC taxonomy. Since the context is computer science the terms are filtered to those from the respective sub-tree. Again we compared against the BM25 and the Wikipedia categories baseline. Figure 9 shows the results for the top-k retrieved documents using mean average precision (MAP).

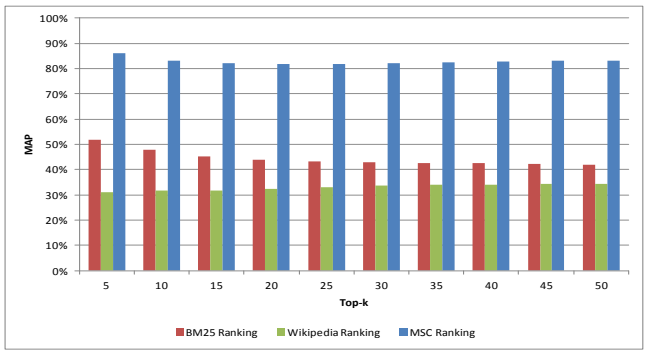

Figure 9. MAP for top-k documents in computer science

Interestingly the results for the BM25 ranking are better than in the chemical domain. The reason is that query terms in computer science are more general than chemical entities leading to better retrieval precision even for fulltext searches. Nevertheless, a cross-domain ranking using MSC classes outperforms both baselines. The highest MAP of 85.9% is reached for the top-5 documents.

This experiment proved that cross-domain knowledge from related domains is very useful to improve the retrieval quality. However, it is important to also filter the annotated cross-domain terms to ensure that they are semantically related to the document's context. Therefore, it is important to use a general knowledgebase, like e.g. Wikipedia, as 'glue' to connect the domain-specific ontology terms to the vocabulary used in the other domain.

6. Conclusions and Future Work

In this paper, we show that the usage of cross-domain knowledge may dramatically improve the retrieval quality of context dependent, entity-centered searches. We proposed an approach using cross-domain knowledge to learn models for annotating documents from domains lacking suitable ontologies. To assure that annotated terms are semantically related to the documents' context we used Wikipedia as general knowledgebase to filter out all unrelated terms by computing the semantic similarity between each term and a document's named entities.

As main use case, we choose the domain of chemistry, since here searches are almost entirely focused on chemical entities. However, no suitable controlled vocabulary for annotating all documents with context information is available. Nevertheless, there is a strong need for context dependent searches to enable high precision retrieval. We annotated chemical documents with ontology terms from the related domain of biomedicine where documents are annotated with terms from the MeSH ontology. We also proved the general usefulness of our approach by enriching documents from computer science with ontology terms from the related domain of mathematics.

Our evaluation has shown that the traditional way of using query expansion with domain specific terms is not useful to restrict the search results to the desired context. We further showed that context dependent searches using cross-domain annotations are possible. We evaluated our annotations in a document retrieval scenario comparing our approach to a BM25 ranking model based on Lucene and an enhanced baseline using Wikipedia categories. While with BM25 the mean average precision is around 53% for the computer science domain and with Wikipedia categories around 43% with our approach it is increased up to 87%. It is remarkable that by using cross-domain knowledge combined with general knowledge provided by Wikipedia the retrieval quality can be increased up to 30% for context driven searches.

For our future work we plan to integrate our promising results into a Web portal, e.g. the ViFaChem Portal (www.chem.de), to enable context-driven searches. Furthermore, we plan to build a general framework allowing for cross-domain context annotations. In this framework the different components, i.e. the domain-specific ontology, documents from a related domain and the general knowledgebase, needs to be easily replaceable. Thus, it is necessary to define interfaces for the different components of our approach on a protocol layer.

References

[1] R. Kraft and J. Zien, "Mining anchor text for query refinement," in *Proceedings of the 13th International Conference on World Wide Web (WWW)*, 2004.

[2] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais, "The vocabulary problem in human-system communication," *Communication of the ACM (CACM)*, vol. 30, 1987.
[3] B. Köhncke, P. Siehndel, and W. Balke, "Bridging the Gap–Using External Knowledge Bases for Context-Aware Document Retrieval," in *Proceedings of the International Conference on Asia-Pacific Digital Libraries (ICADL)*, 2013.
[4] B. Köhncke and W. Balke, "Context-Sensitive Ranking Using Cross-Domain Knowledge for Chemical Digital Libraries," in *Theory and Practice of Digital Libraries*, 2013.
[5] L. Chen, et al., "Context-sensitive ranking for document retrieval," in *Proceedings of ACM SIGMOD*, 2011.
[6] S. J. Nelson, D. Johnston, and B. L. Humphreys, "Relationships in Medical Subject Headings," in *Relationships in the Organization of Knowledge*, 2001.
[7] F. Sebastiani, "Machine learning in automated text categorization," *ACM Comput. Surv.*, vol. 34, no. 1, 2002.
[8] D. Jiang, et al., "Context-aware search personalization with concept preference," in *Proceedings of ACM Conference on Information and Knowledge Management (CIKM)*, 2011.
[9] S. H. Nguyen and W. Swieboda, "Extended Document Representation for Search Result Clustering," *Stud. Comput. Intell.*, 2012.
[10] R. Laza, et al., "Evaluating the effect of unbalanced data in biomedical document classification.," *J. Integr. Bioinform.*, vol. 8, no. 3, Jan. 2011.
[11] F. Camous, et al., "Ontology-based MEDLINE document classification," in *Proceedings of the 1st International Conference on Bioinformatics Research and Development*, 2007.
[12] D. Trieschnigg, et al., "MeSH Up: effective MeSH text classification for improved document retrieval.," *Bioinformatics*, vol. 25, no. 11, Jun. 2009.
[13] S. Yoo and J. Choi, "Improving MEDLINE document retrieval using automatic query expansion," in *Proceedings of the 10th International Conference on Asian Digital Libraries (ICADL)*, 2007.
[14] S. Tönnies, B. Köhncke, O. Koepler, and W.-T. Balke, "Exposing the Hidden Web for Chemical Digital Libraries," in *Proceedings of the 10th ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 2010.
[15] P. Corbett and P. Murray-Rust, "High-throughput identification of chemistry in life science texts," in *Proceedings of the 2nd International Symposium on Computational Life Sciences*, 2006, vol. 4216.
[16] Y. Zhen and C. Li, "Cross-domain knowledge transfer using semi-supervised classification," in *Proceedings of the 21st Australasian Joint Conference on Artificial Intelligence: Advances in Artificial Intelligence*, 2008.
[17] C. Liu, et al., "Cross Domain Search by Exploiting Wikipedia," in *Proceedings of the International Conference on Data Engineering (ICDE)*, 2012.
[18] E. Gabrilovich and S. Markovitch, "Computing semantic relatedness using wikipedia-based explicit semantic analysis," in *Procs. of the 20th Int. Joint Conference on Artificial Intelligence (IJCAI)*, 2007.
[19] O. Egozi, S. Markovitch, and E. Gabrilovich, "Concept-Based Information Retrieval Using Explicit Semantic Analysis," *ACM Trans. Inf. Syst.*, vol. 29, no. 2, 2011.
[20] X. Hu, X. Zhang, C. Lu, E. K. Park, and X. Zhou, "Exploiting Wikipedia as external knowledge for document clustering," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2009.
[21] M. Hall, E. Frank, and G. Holmes, "The WEKA data mining software: an update," *ACM SIGKDD Explor. Newsl.*, vol. 11, no. 1, 2009.
[22] P. Wang, et al. "Using Wikipedia knowledge to improve text classification," *Knowl. Inf. Syst.*, vol. 19, no. 3, 2008.
[23] B. Köhncke and W.-T. Balke, "Using Wikipedia categories for compact representations of chemical documents," in *Procs. of ACM Conference on Information and Knowledge Management (CIKM)*, 2010.
[24] D. Milne and I. Witten, "Learning to link with wikipedia," in *Proceedings of ACM Conference on Information and Knowledge Management (CIKM)*, 2008.
[25] K. Degtyarenko, et al., "ChEBI: a database and ontology for chemical entities of biological interest.," *Nucleic Acids Res.*, vol. 36, Database Issue, 2008.
[26] J. Platt, "Fast Training of Support Vector Machines using Sequential Minimal Optimization," in *Advances in Kernel Methods*, 1998.

Benjamin KÖHNCKE

Benjamin Köhncke received his PhD in the area of digital libraries from the Technische Universität Braunschweig, Germany, after working several years for L3S Research Center at Leibniz Universität Hannover, Germany. His main research is in the area of (semantic) metadata indexing in chemical digital libraries, however with the aim to generalize major findings also to other scientific fields like medicine or bio-informatics. He received his B.A and M.Sc degrees in computer science from Leibniz Universität Hannover, Germany.

Wolf-Tilo BALKE

Wolf-Tilo Balke currently heads the Institute for Information Systems (IfIS) at Technische Universität Braunschweig, Germany, and serves as a director of L3S Research Center, Hannover, Germany. Before, he was associate research director at L3S and a research fellow at the University of California at Berkeley, USA. His research is in the area of databases and information service provisioning, including personalized query processing, retrieval algorithms, preference-based retrieval and ontology-based discovery and selection of services. In 2013 Wolf-Tilo Balke has been elected as a member of the Academia Europaea. He is the recipient of two Emmy-Noether-Grants of Excellence by the German Research Foundation (DFG) and the Scientific Award of the University Foundation Augsburg, Germany. He has received his B.A and M.Sc degree in mathematics and a PhD in computer science from University of Augsburg, Germany.