Context-Sensitive Ranking Using Cross-Domain Knowledge for Chemical Digital Libraries

Benjamin Köhncke[1], Wolf-Tilo Balke[2]

[1] L3S Research Center; Hannover, Germany
[2] TU Braunschweig, Germany

koehncke@L3S.de, balke@ifis.cs.tu-bs.de

Abstract. Today, entity-centric searches are common tasks for information gathering. But, due to the huge amount of available information the entity itself is often not sufficient for finding suitable results. Users are usually searching for entities in a specific search context which is important for their relevance assessment. Therefore, for digital library providers it is inevitable to also consider this search context to allow for high quality retrieval. In this paper we present an approach enabling context searches for chemical entities. Chemical entities play a major role in many specific domains, ranging from biomedical over biology to material science. Since most of the domain specific documents lack of suitable context annotations, we present a similarity measure using cross-domain knowledge gathered from Wikipedia. We show that structure-based similarity measures are not suitable for chemical context searches and introduce a similarity measure combining entity- and context similarity. Our experiments show that our measure outperforms structure-based similarity measures for chemical entities. We compare against two baseline approaches: a Boolean retrieval model and a model using statistical query expansion for the context term. We compared the measures computing mean average precision (MAP) using a set of queries and manual relevance assessments from domain experts. We were able to get a total increase of the MAP of 30% (from 31% to 61%). Furthermore, we show a personalized retrieval system which leads to another increase of around 10%.

Keywords: Chemical Digital Libraries, Personalization, Context Search

1 INTRODUCTION

Throughout the sciences the information gathering process is to a large degree based on Web sources today. Considering the exponentially growing amount of information on the Web it is thus essential for large-scale information providers, such as digital libraries, to build effective systems allowing for easy and flexible access to information relevant for specific user needs. Especially entity-centric searches have become common tasks for many researchers across almost all scientific domains. Considering for instance the domain of chemistry information gathering is prominently

focused on chemical entities. However, the actual search for chemical entities is by no means restricted to the chemical domain: in medicine it is important to find active ingredients of drugs, e.g., against infectious diseases. In biology chemical substances are important to understand complex metabolism processes. And in materials science chemical entities play a major role in developing novel materials like polymers or nanomaterials.

Thus throughout this paper we focus on chemical entities as an important example of entity-based searches. The problem for such entity-based searches is usually two-fold: one central complex is markup and disambiguation (which also included detecting synonyms), the other complex deals with similarity-based searches to find in some respect similar entities. The first complex for chemical entities is quite well researched and already actively used in portals, e.g., SMILES or InCHI code for representation, the OSCAR framework [1] for chemical markup, chemical search engine for formulae [2], or building enriched index pages for synonymy in chemical documents [3]. In contrast, the second complex has yet to make the step from laboratory usage to a widespread use in digital portals. The main problem is how to actually compute similarity between chemical entities?

Generally speaking, chemical entities are transformed into so-called fingerprint representations based on their chemical structures. A fingerprint is a sequence of bits where each bit represents the occurrence of a special chemical feature. Of course, there are different possibilities to encode chemical properties in a bit-sequence leading to different fingerprint representations. Also for the subsequent similarity computations different well-known measures are used, like e.g. Cosine, or Russell-Rao. [4] analyzed these different measures and found that they often produce entirely uncorrelated result lists. Thus, it seems that larger contexts or specific tasks may strongly influence the individual perception of the entities' similarity and relevance.

The problem is that none of the structural measures takes context information into account. But this is very important, because the similarity of two chemical substances is actually heavily related to the search context. Consider for instance the chemical entities *Zanamivir* and *Ibuprofen*. Both are used in the treatment of flu and are therefore similar regarding this pharmacological activity context. *Ibuprofen* is also used to treat inflammatory diseases such as rheumatoid arthritis. But, regarding this context both entities are very dissimilar: *Zanamivir* is a neuraminidase inhibitor and thus not at all useful for the treatment of rheumatoid arthritis. It is therefore necessary to personalize measures for entity similarity to the task or search context a user is currently engaged in. In brief, context used to disambiguate the user's explicit query can be expected to lead to focused and relevant retrieval results.

Most users perform the *contextualization* of searches manually by adding additional terms to their actual query, if the retrieval results have not yet been satisfying (query refinement) [5]. There are also first approaches *automatically enriching* a user's query with terms related to user's context, see e.g. [6]. However, for using context terms in document retrieval most approaches require documents to be annotated or classified with the related context terms using a fixed (controlled) vocabulary. For example, in the biomedical domain documents are annotated by terms from the well-known MeSH ontology. Since it is maintained manually, the offered terminology is of

high quality. But the almost completely MeSH-indexed MEDLINE digital library is a rare case and its manual curation is expensive, while automatic classification is still error-prone. Moreover, most document collections miss both, suitable annotations and the funds to add them. Considering for instance the linked open data community, hardly any collection dealing with chemical entities is properly annotated. Examples are *Linking Open Drug Data*, a task force within the World Wide Web Consortium's Health Care and Life Sciences Interest Group, or *clinical trials* describing relationships between active ingredients and diseases tested in clinical studies around the world.

In this paper we present an approach enabling *context searches for chemical entities using cross-domain knowledge* harvested from Wikipedia as a major knowledge base. One advantage of our approach is that every term occurring in Wikipedia can be used as context term. Instead of using a fixed vocabulary of predefined classes, we thus use the 'wisdom of the crowd' which is dynamic and ever-growing. The derived similarity measure is therefore not purely based on structural information of chemical entities, but extracts different features of chemical entities using common knowledge in the community. All features are combined in enriched profiles of chemical entities. These profiles are then used for similarity computations resulting in a personalized ranking function considering both, context as well as entity similarity. Our experiments show that it is indeed sensible to combine cross-domain features: the average precision is increased from 31% when using a Lucene fulltext filter for contextualization to up to 71% for personalized queries using our measure.

The rest of the paper is organized as follows: the next section gives an overview of related work. In section 3 we introduce our novel similarity measure based on chemical profiles incorporating cross-domain knowledge followed by a detailed evaluation in section 4. Finally, section 5 concludes the paper with an outlook on future work.

2 RELATED WORK

Today there are different groups of approaches using context information. The area of contextual search tries to proactively capture the information need of a user by automatically extending the user's query with information from the user's search. An approach using information from raw query search logs to discover context terms is described in [7]. The detected terms are included in user preferences used to optimize search results. It was shown that in terms of top-k search quality a system using context information outperforms existing personalization approaches without context information. In [6] three different algorithms are compared considering contextual search for the Web, i.e. query rewriting, rank-biasing and iterative filtering metasearch (IFM). The experimental results have shown that the query rewriting approach performs surprisingly well. Therefore, we will compare against a quite similar approach using query expansion for the context term in our evaluation.

Another famous ranking algorithm considering context information for Web searches is the topic-sensitive PageRank [8]. For each Webpage multiple importance scores with respect to various topics are computed. These scores are combined at

query time dependent on the topics stated in the query. Afterwards they can be combined with different IR measures to produce a suitable ranking. In [9] it was shown that context-sensitive ranking improves the retrieval quality for domain experts remarkably, compared with conventional ranking models. The proposed ranking model uses keyword statistics collected from the specified contexts to rank the documents. In comparison to our approach, here, it is still necessary to pre-classify the documents to their respective context terms. Since they are working on the MEDLINE corpus and all given documents are annotated with MeSH terms this classification is given.

But, such a corpus as MEDLINE where each document is indexed with several MeSH terms is a rare case and is actually curated manually with expensive efforts. In the domain of chemistry for example, no such ontology for annotating chemical documents with context information is available. Indeed, in the chemical domain only a few highly specialized controlled vocabularies are openly available, e.g. *Chemical Entities of Biological Interest* (ChEBI [10]). But our experiments with domain experts in [11] have shown that Wikipedia categories are more useful to describe the documents' context. 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. The approach presented in [12] also propose to use Wikipedia to enable cross-domain search. But their main focus is on analyzing tags used in Web 2.0 systems like Flickr and connect them to concepts in Wikipedia.

Beside the query context, of course, it is also necessary to consider the actual query term for retrieving suitable search results. In the chemical domain similarity search is centered on chemical entities. In previous work we have shown how to use structural information to create enriched index pages [3]. Indexing different unambiguous representations we were able to reach the retrieval quality of a chemical structure search using a common Google text search. Based on these index pages we analyzed how similarity between chemical entities is computed [4]. We analyzed the different possible combinations of fingerprints and similarity measures computing the k-tau correlation coefficient. We figured out that there are many uncorrelated measures. As a straight forward idea, we assumed that the uncorrelated combinations can be assigned to different chemical search tasks. But our experiments have shown that this is not possible and structure-based similarity measures are not useful for context searches.

3 Computing Context Similarity in Chemistry

In this section we introduce a similarity measure using external knowledge sources independent of chemical structures. Our measure considers both, entity- as well as context similarity. Finally, we are interested in documents including the query entity (or similar entities) in the sense of the specified context.

In our system a document, further denoted by d, is represented as the bag of words of its included chemical entities $E_d \subseteq E$, where E is the set of all chemical entities in the collection. Let D denote the collection of documents. A query for a context search is composed of two parts: $q = e_q \mid q_c$, where e_q is a chemical entity and q_c is the de-

sired context specification. q_c specifies a sub-collection $D_c \subseteq D$ such that $\forall d \in D_c$, d satisfies q_c. A chemical entity is defined as the trivial name of a chemical structure. The first necessary step is to extract all chemical entities from the documents. We use the OSCAR framework for an automatic extraction [1]. Next the similarity between these entities is computed.

3.1 Entity Similarity

To find a suitable similarity measure we use external knowledge from different information sources, create profiles of chemical entities containing different features, and finally compute the similarity based on these profiles. Since it was shown in [11] that Wikipedia is a reliable source for representing chemical documents we also used it here as main information source. For each chemical entity $e \in E_d$ we analyzed its corresponding Wikipedia page and extracted suitable features used in the chemical profiles. From each page we extracted a set of the assigned Wikipedia categories, a set of all other entities that are cited in the Wikipedia page (outgoing links), and a set of all other entities pointing to the respective page (incoming links).

Beside Wikipedia we also use another tool to automatically detect important entities in text, named OpenCalais. OpenCalais is a free Web service from Thomson-Reuters that does named entity recognition to extract events and relationships from text. It uses natural language processing and machine learning techniques to recognize instances of named entities. Since OpenCalais uses surface features, like e.g. capitalization, and is not based on handcrafted databases of entities it can detect new entities that may not be included in any knowledge base like Wikipedia.

For each chemical entity we analyze its Wikipedia page using OpenCalais and add the retrieved information to its chemical profile. In detail, we use the detected Calais entities, topics and tags. The Calais entities are further divided into several different types, ranging from types like e.g. medical treatment or medical condition, to types like e.g. person or operating system. The social tags are not really semantic features, but emulate how a person would tag a specific piece of content. The topics describe a category that the input content is about. They are based on the Calais categorization taxonomy. But, it is also possible that no topic is assigned to the input content.

To summarize, each chemical profile contains six different features. Each feature is used to compute the similarity between the query entity e_q and the entity $e_a \in E$.

Calais entity similarity: Let ts_q be the type set for e_q and ts_a the type set for e_a. Each type $t \in ts_x$ where $x \in \{q, a\}$ is associated with a set of related Calais entities, $t_n es_q$ and $t_n es_a$, where $1 \leq n \leq |ts_x|$. The similarity is computed using the Jaccard coefficient.

$$ts = \frac{ts_q \cap ts_a}{ts_q \cup ts_a} \quad (1)$$

The *ts* coefficient describes how many types the given chemical entities have in common. For each type they have in common the entity similarity is computed and normalized by the number of types e_q and e_a have in common.

$$es = \frac{\sum_{t \in ts_q \cap ts_a} \frac{{}_t es_q \cap {}_t es_a}{{}_t es_q \cup {}_t es_a}}{|ts_q \cap ts_a|} \quad (2)$$

The Calais entity similarity is computed as follows:

$$ces = (\gamma * ts) + ((1 - \gamma) * es) \quad (3)$$

where γ is a weighting factor and $0 \leq \gamma \leq 1$.

Calais tag and topic similarity: For tag and topic similarity the same measure is used. For each detected term (tag or topic term) a relevance score in the range of 0 to 1, further denoted as *rs*, is computed, describing the importance of each unique term.

Let tsm_q be the term set for e_q, and tsm_a the term set for e_a. The tag and topic similarity is computed using the following equation:

$$tsm = \beta * \frac{tsm_q \cap tsm_a}{tsm_q \cup tsm_a} \quad (4)$$

β is called the regulation factor which is computed as follows:

$$\beta = \frac{\sum_{t \in tsm_q \cap tsm_a} \frac{rsa_t + rsq_t}{2}}{|ts_q \cap ts_a|} \quad (5)$$

where rsa_t is the relevance score of term *t* for e_a and rsq_t the relevance score of *t* for e_q. The relevance scores are in the range of 0 to 1 and are assigned by OpenCalais. The regulation factor is used to give lower similarity scores to entities that indeed have many terms in common, but which have low relevance scores for the entity itself.

Wikipedia category similarity: For the Wikipedia category similarity we defined a quite similar formula as for the Calais tag and topic similarity. Let wc_q be the categories set for e_q and wc_a the categories set for e_a. For each Wikipedia category also a weighting factor (*wf*) is assigned describing how general the respective category is regarding the Wikipedia category graph. We use this factor to give more specific categories a higher score. The category similarity is computed using the following formula:

$$wc = wf * \frac{wc_q \cap wc_a}{wc_q \cup wc_a} \quad (6)$$

The weighting factor *wf* is defined as

$$wf = \frac{\sum_{wc \in wc_q \cap wc_a} dt_{wc}}{|wc_q \cap wc_a|} \quad (7)$$

where *dt* is the length of the shortest path from the respective Wikipedia category to the root category.

Wikipedia related entities similarity: Furthermore, we use the Jaccard coefficient to compute the similarity based on the related entities. For related entities we distinguish between entities linking to the Wikipedia page of e_a and e_q (further denoted as res_{in}) and entities that are linked from the Wikipedia pages of e_a and e_q (further denot-

ed as res_{out}). Let res_q be the set of related entities for e_q and res_a the set of related entities for e_a. The similarity is computed as follows:

$$res_{in/out} = \frac{res_q \cap res_a}{res_q \cup res_a} \tag{8}$$

Entity similarity: To compute the entity similarity of e_a and e_q we combine the different feature similarities in a linear fashion.

$$entSim = \omega * ces + \vartheta * tsm_{tag} + \sigma * tsm_{topic} + \vartheta * wf + \rho * res_{in} + \tau * res_{out} \tag{9}$$

Each feature is multiplied with a Boolean variable, i.e. $\omega, \vartheta, \sigma, \vartheta, \rho, \tau$, having the value 0 or 1. These variables are used for personalizing the entity similarity measure by switching features on and off. As we will see in the experiments it depends on the user preferences which combination of features leads to best retrieval results.

3.2 Context Similarity

The context similarity is also based on the knowledge covered by Wikipedia. We use the Wikipedia Miner [13] to access the Wikipedia corpus and compute the semantic similarity between the context term and all chemical entities in our corpus using the relatedness measure described in [14]:

$$contextSim(c, e) = \frac{\log(\max(|A|,|B|)) - \log(|A \cap B|)}{\log(|W|) - \log(\min(|A|,|B|))} \tag{10}$$

where c and e are the Wikipedia pages for the context term c and the entity e, C and E are the sets of pages that link to c, respectively e, and W is the set of all pages in Wikipedia.

A drawback of this measure is that we need to compute the semantic similarity between the context term and all other chemical entities in our collection. After computation the scores are stored in a database meaning that we only need to compute the similarity once for every context term. In case a new context term is entered in the system this computation has to be performed. The next time the context term is entered no computation is necessary and the scores can be directly retrieved from the database.

3.3 Combined Similarity

Our goal is to find the most similar entities for the query entity e_q in the given context q_c. The entity similarity computes the most similar entities for e_q and the context similarity finds the most related entities to the context term. The total similarity for query q is computed as follows:

$$totalSim = (\alpha * contextSim)^{\frac{+((1-\alpha) * entSim)}{|EF|}} \tag{11}$$

where EF is the set of features used for entity similarity computation and α is a weighting factor with $0 \leq \alpha \leq 1$.

4 Evaluation

For our experiments we used a data set of 44660 clinical studies[1]. We choose 10 different context terms which are all diseases, i.e. Malaria, Tuberculosis, Mumps, Tinnitus, Hypertension, Hepatitis A and C, Influenza, Dengue and Cancer. We automatically extracted all chemical entities using the OSCAR framework [1]. In total 1.573.264 entities have been annotated in the documents, 79223 of them are distinct.

OSCAR also uses a name-to-structure algorithm which associates chemical structures to the found entities. Since we want to compare against the fingerprint-based similarity measures we filtered out all found entities that do not have structural information (in this case a SMILES code). This leads to a total of 721 distinct chemical entities independent of the documents' context. Since our measure relies on Wikipedia we analyzed how many of the chemical entities can be found. We used the WikipediaMiner [13] to search for the chemical entities in Wikipedia. For 92.6% (668) we found a matching Wikipedia page.

4.1 Correlation Analysis

In this experiment we analyzed if we need all cross-domain features in the chemical profile for similarity computation or if some of them are correlated. We randomly chose around 10% of all chemical entities as query terms, resulting in 72 queries in total. Using these terms we computed the rankings to all other chemical entities in our set based on the six feature similarities introduced in section 3.

Since we can interpret the similarity value as a value in a ranking vector, we used the Kendall rank correlation coefficient (KTau) [15] to determine the correlation of the different measures. We calculated the correlation coefficient for each ranking vector and the arithmetic mean over 72 queries. A KTau of 1 means that the agreement of two rankings is perfect, -1 indicates a perfect disagreement and for independent rankings one would expect the coefficient to be *approximately* 0. For each pairwise comparison of two rankings we averaged the Ktau values over all queries. We only considered those queries which are significant meaning having a p-Value less than 0.05. The highest correlation is found between the Wikipedia in-links and the Wikipedia categories, followed by the Open Calais topic ranking and the Wikipedia categories. However the values are still very small (< 0.45) so that we consider the rankings as uncorrelated. Therefore, all features deliver different rankings and are used in our similarity measure.

4.2 Comparing Different Rankings

In this experiment we compare the rankings of the different similarity measures. As stated earlier, a query is defined as follows: A query for a context search is composed of two parts: $q = e_q \mid q_c$, where e_q is a chemical entity and q_c is the desired context specification. Basically, we compared the feature similarity against the fingerprint-

[1] http://clinicaltrials.gov/ct2/home

based similarity measures. Since the relevance ratings for two entities differ between different context terms it is not sensible to evaluate the entity ranking without considering the search context. For considering the context in the fingerprint-based measures we used the following procedure. The documents in our collection, further denoted by D, are filtered and only those related to q_c are retrieved. From this document set, denoted by D_c, the chemical entities are extracted and ranked using the different similarity measures. We evaluated different possibilities for building D_c. First, we use a Boolean approach where D_c contains all documents including the context term q_c. Second, we use an approach using statistical query expansion where q_c is expanded using the most co-occurring terms.

For building a ground truth to compare the different rankings against, we randomly choose a set of 10 chemical entities and related context terms as queries. In order to make manual relevance assessment feasible, we pooled together the top-20 entities retrieved for each query and similarity metric. The relevance assessment was done manually by domain experts. The experts marked for each query all chemical entities from the sampling sets that are relevant for the query in a Boolean fashion. To evaluate the rankings we computed the mean average precision (MAP) based on the relevance assessments.

First, we analyze the results of the Boolean retrieval model. The document set is filtered using q_c, meaning only documents are included containing q_c in the fulltext. The filtering was done using a Lucene fulltext index. The highest MAP of 31% is reached using the Forbes similarity measure based on the Substructure fingerprint. The average recall using the Boolean approach is 82.7%. That means some relevant entities are filtered out. The reason is that not all relevant documents contain the context term in the fulltext.

For the second baseline approach we use a retrieval model including statistical query expansion. We computed a term-to-term co-occurrence matrix based on our document set. We also considered the position of the term in the document, 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 not used as context terms. Finally, the context term q_c is expanded with the top-10 co-occurring terms. We used the following retrieval model: Let $C=\{q_c,c_1,...,c_n\}$ be the set including q_c and all expanded terms. The expanded context query is formulated as q_c *OR* c_1 *OR ... OR* c_n, meaning all documents are returned containing q_c or any of the expanded terms. The highest MAP of 23% is reached using the Yule similarity measure based on the Extended fingerprint. The MAP is even lower than for the Boolean approach. The reason is that using query expansion the set of entities is getting bigger. This is also proved if we take a look at the recall. It has increased up to 89.5%. These results confirmed the experiments in [4] showing that fingerprint-based measures are not suitable for context searches.

For our feature-based approach we combine context- and entity similarity in one single measure. Since our measure computes the similarities for all chemical entities the recall is always 100%. To regulate the weighting between context- and entity similarity a variable *alpha* is used (see 3.3). If alpha is 0 no context similarity is used

and if it is 1 no entity similarity is used. Fig.1 shows the MAP results for the cross-domain similarity measure for varying alpha values.

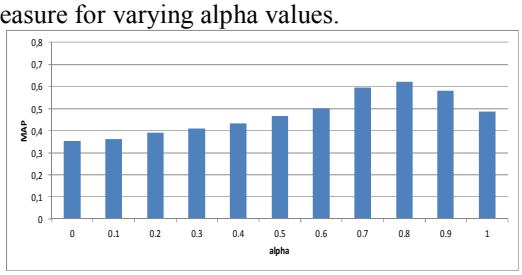

Fig. 1. MAP values dependent on alpha

The best result of a MAP of 61% is reached for alpha equals 0.8. That means the context similarity is slightly higher weighted. Using this measure we were able to increase the MAP from 31% for the Boolean approach to 61%. Since this result is an average over all chemists and all queries we tried to further increase it by personalizing the similarity measure.

4.3 Personalized Ranking

The idea is to build a personalized retrieval system where each individual user trains the system and the system will learn the best similarity measure for the user. The system includes a simple feedback step where the user marks the chemical entities most relevant for him. Therefore, we conducted a user study with domain experts from the area of drug design and synthesis. For the user study, we have randomly chosen ten queries consisting of chemical entity and context. Each query represents a feedback cycle in the system.

Since the measure for computing the entity similarity is composed of six different features, we analyzed which feature combination is the best for the individual chemist. The goal is to find a suitable feature combination for computing the entity similarity within the feedback cycles. Thus, we need to compute all possible combinations and analyze which leads to the best results. Let us consider we have a finite set *EF* containing *n* features. The number of different subsets we need to combine is computed using the power set, $|P(EF)| = 2^n$. Since we have 6 different features we can combine them in $2^6 - 1 = 63$ different ways. We need to subtract 1 since we do not need to compute the empty set which is also contained in the power set.

For each chemist and each query we computed the 63 different rankings and compared them to the manual relevance judgments by computing the average precision. For each query we analyzed which feature combinations lead to the best result. Unfortunately it was not possible to find the optimal solution for each chemist. But we found out that in average 4 different feature combinations are enough to always find the most suitable ranking. These combinations have been found after 7 feedback cycles in average. That means that we only need to compute 4 different rankings instead of 63 and have a high probability that the most suitable solution is found.

Fig.2 (left) shows the number of top rankings for the different feature combinations over all chemists. It is interesting to see that more than half of the combinations never lead to the best ranking. Of course, this statistic will change over time depending on the different users submitting queries to the system, but it is useful to overcome the well known *new user problem*. For new users it seems to be a good choice to use the overall best measure as global starting point, i.e. tsm_{tag} and tsm_{topic} or res_{in} (see 3.1).

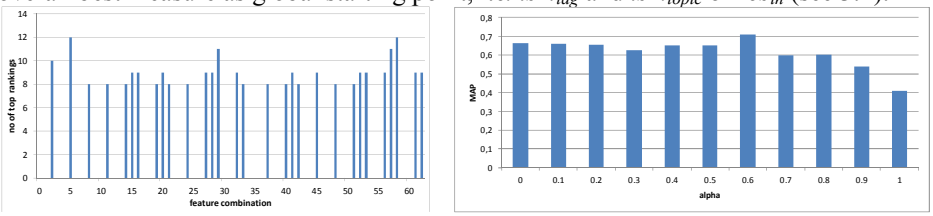

Fig. 2. Number of top rankings for different feature combinations (left)
Example: MAP values for varying alpha for one chemist over 10 queries (right)

Now, that we found the best feature combinations we use them to analyze which weighting between entitiy- and context similarity is the best by varying the alpha value. For each chemist and each query we took the best feature combination and compute the average precision using the chemist's relevance vector. Fig.2 (right) shows the MAP results for one chemist for varying alpha over 10 queries. For this chemist, the best results are retrieved using an alpha of 0.6. Compared to the impersonalized measure the mean average precision is increased of up to 71%. In average over all users the mean average precision increases about 9% using personalization.

5 CONCLUSION AND FUTURE WORK

For digital library providers it is important to allow for context searches to assure high quality retrieval. Our experiments showed that structure-based similarity measures cannot retrieve suitable results for context searches. Therefore, we presented an approach using cross-domain knowledge gathered from Wikipedia to enable context searches in the chemical domain. The beauty of the presented approach is that digital library providers can easily integrate it into their workflow of metadata enrichment. The necessary steps are the extraction of chemical entities, creation of enriched chemical profiles using Wikipedia and the similarity computation using the profiles. Each profile consists of six different features, i.e. Wikipedia categories, in- and out-links, and three additional features extracted using OpenCalais. The features are combined in a linear fashion and used to compute entity similarities. For context similarity we also relied on Wikipedia and computed the semantic similarity of each chemical entity for the specific query context. Finally entity- and context similarity are combined in one similarity measure.

Our experiments have shown that the cross-domain similarity measure outperforms the structure-based measures. We compared our measure against two baselines: a Boolean retrieval model and a model using statistical query expansion for the context terms. We computed the mean average precision (MAP) using a set of queries and

manual relevance assessments from domain experts. We were able to get a total incensement of the MAP of 30% (from 31% to 61%). To further increase the precision we introduced a personalized retrieval system based on user feedback by varying the features used for entity similarity and the weighting between context and entity similarity. Using the best feature combination for each query we were able to further increase the MAP up to 71%.

For our future work we plan to generalize our approach and use it in other domains. It will be interesting to see if cross-domain knowledge from Wikipedia is also useful in domains using different entities, like e.g. genes in biology. There, we can also compare against classification approaches, like e.g. SVM, since we can rely on a fixed set of context terms, like e.g. provided by the MeSH ontology. Furthermore, instead of using Boolean variables in the entity similarity measure, it might be interesting to learn the weighting parameters using a learning to rank framework.

6 REFERENCES

1. P. Corbett and P. Murray-Rust, "High-throughput identification of chemistry in life science texts," in *Proc. of the Int. Symp. on Computational Life Sciences*, vol. 4216, 2006.
2. B. Sun, et al., "Identifying, Indexing, and Ranking Chemical Formulae and Chemical Names in Digital Documents," *ACM Transactions on Information Systems*, vol. 29, 2011.
3. S. Tönnies, B. Köhncke, O. Koepler, and W.-T. Balke, "Exposing the Hidden Web for Chemical Digital Libraries," in *Proc. of the Joint Conf. on Digital Libraries (JCDL)*, 2010.
4. S. Tönnies, et al., "Taking Chemistry to the Task – Personalized Queries for Chemical Digital Libraries," in *Proc. of the Joint Conf. on Digital Libraries (JCDL)*, 2011.
5. R. Kraft and J. Zien, "Mining anchor text for query refinement," in *Proc. of the Int. Conf. on World Wide Web (WWW)*, 2004.
6. R. Kraft, C. C. Chang, F. Maghoul, and R. Kumar, "Searching with context," in *Proc. of the Int. Conf. on World Wide Web (WWW)*, 2006.
7. D. Jiang, et al., "Context-aware search personalization with concept preference," in *Proc. of Conf. on Information and Knowledge Management (CIKM)*, 2011.
8. T. Haveliwala, "Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search," *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, 2003.
9. L. Chen and Y. Papakonstantinou, "Context-sensitive ranking for document retrieval," in *Proc. of ACM SIGMOD Conf.*, 2011.
10. K. Degtyarenko, et al., "ChEBI: a database and ontology for chemical entities of biological interest.," *Nucleic acids research*, vol. 36, Database issue, 2008.
11. B. Köhncke and W.-T. Balke, "Using Wikipedia categories for compact representations of chemical documents," in *Proc. of Conf. on Information and Knowledge Management (CIKM)*, 2010.
12. C. Liu, S. Wu, S. Jiang, and A. K. H. Tung, "Cross Domain Search by Exploiting Wikipedia," in *Int. Conf. on Data Engineering (ICDE)*, 2012.
13. D. Milne and I. H. Witten, "An open-source toolkit for mining Wikipedia," *Artificial Intelligence*, vol. 194, 2012.
14. D. Milne and I. Witten, "Learning to link with wikipedia," in *Proc. of Conf. on Information and Knowledge Management (CIKM)*, 2008.
15. M. G. Kendall, "A New Measure of Rank Correlation," *Journal of Biometrika*, vol. 30, no. 1–2, 1938.