Using Wikipedia Categories for Compact Representations of Chemical Documents

Benjamin Köhncke
L3S Research Center
Appelstraße 9a
30167 Hannover
koehncke@l3s.de

Wolf-Tilo Balke
IFIS TU Braunschweig
Mühlenpfordtstraße 23
38106 Braunschweig
balke@ifis.cs.tu-bs.de

ABSTRACT
Today, Web pages are usually accessed using text search engines, whereas documents stored in the deep Web are accessed through domain-specific Web portals. These portals rely on external knowledge bases, respectively ontologies, mapping documents to more general concepts allowing for suitable classifications and navigational browsing. Since automatically generated ontologies are still not satisfactory for advanced information retrieval tasks, most portals heavily rely on hand-crafted domain-specific ontologies. This, however, also leads to high creation and maintaining costs. On the other hand, a freely available community maintained, if somewhat general, knowledge base is offered by Wikipedia. During the last years the coverage of Wikipedia has reached a large pool of information including articles from almost all domains. In this paper, we investigate the use of Wikipedia categories to describe the content of chemical documents in a compact form. We compare the results to the domain-specific ChEBI ontology and the results show that Wikipedia categories indeed allow useful descriptions for chemical documents that are even better than descriptions from the ChEBI ontology.

Categories and Subject Descriptors
H.3.7 [**Information Storage and Retrieval**]: Digital Libraries

General Terms
Measurement, Design, Experimentation

Keywords
Wikipedia Categories, Ontology, Document Topic, Tag Cloud

1. INTRODUCTION AND BACKGROUND
In recent years, the access to information and the information gathering process have changed. More and more information is made available online. Therefore, the first option is always a Web search. In most cases a user query returns a large set of matching pages that are somehow related to the query term. However, it has often been shown that consumers usually only examine the first 10 to 20 results, respectively the first two sites of the result set. Of course, the result set is ordered following a complex ranking system which indeed is not really transparent for the user (integrating Page-Ranks, etc.). The question is, why are these pages really marked as relevant regarding the query term?

To give users a certain feeling, in most engines the result lists are accompanied by snippets where the query term is highlighted. However, it is usually still not possible to get a good overview of the general topics relevant for the query from these snippets. Especially for high recall searches it would be helpful to retrieve a better structured result set offering a suitable overview of the general Web page categories. Actually, there are already approaches that cluster the results and offer a set of general categories for filtering. An example is the search engine Clusty.com. However, the problem still is that categories that occur in more pages are considered to be more relevant. Therefore, we do not get a complete overview of the whole category dimensions.

Moreover, information stored in the so-called Deep Web is only partly accessible from search engines and directories anyway. Usually the Deep Web information is made available through special Web portals offering possibilities of navigational access and category overviews. These portals are focused on specific domains and rely on external knowledge bases mapping documents to more general concepts allowing for suitable classifications. For example, in the area of medicine the GoPubMed portal (www.gopubmed.com) relies on the MeSH ontology which offers a controlled vocabulary used for indexing articles. For practitioners in the field of medicine the MeSH ontology thus offers a comprehensible, easy to use, and well-structured knowledge base.

The domain we will be focusing on in this paper is the area of chemistry which offers some challenging problems for information access due to its high degree of fragmentation. Generally speaking it consists of as much as 223 hierarchically structured different working fields (based on the classification of the German Chemical Society). This large number of different working areas makes it extremely unrealistic to build an overarching taxonomy or ontology for the entire domain. Each chemist has his/her own interpretation of chemical relatedness dependent on the domain he/she is working in. In fact, the chemical domain offers just one single highly specialized controlled vocabulary openly available, called 'Chemical Entities of Biological Interest' (ChEBI [5]). ChEBI focuses on small molecules which are either natural products or synthetic products used to intervene in the process of living organisms. Today, the only possibility to get a broad access to chemical documents is

CIKM'10, October 26–30, 2010, Toronto, Ontario, Canada.

through the Chemical Abstract Service (CAS)[1]. The CAS document repository is manually maintained resulting in high quality data which can be accessed through a Web portal called SciFinder. However, the access is limited to subscribed users at a price of about 30,000 USD/year for a single user subscription.

Considering the spirit of open access journals it is, however, not desirable to rely on manually maintained, but therefore high priced commercial databases like Chemical Abstracts. Currently there are 111 chemistry journals listed in the Directory of Open Access Journals (www.doaj.org). However, to enable classification, respectively navigational browsing in an online portal, a high quality ontology is still essential. One opportunity is to use approaches for automatic ontology creation. These approaches are usually based on hierarchic clustering algorithms [1]. Another system, presented in [3], is based on an agglomerative clustering algorithm that exploits an external hypernym oracle to drive the clustering process. Unfortunately, the quality of automatically generated ontologies for such complex domains as chemistry is not yet sufficient [8].

On the other hand, a freely available, community maintained knowledge base is offered by Wikipedia. During the last years the coverage of Wikipedia has reached a large pool of information including articles from almost all areas. Each article is assigned to a number of categories which are hierarchically ordered and form a shallow ontology which is used by many users and constantly refined by Wikipedia editors [11]. Recent work proposes to use these categories to identify the topics of a document. In [7] an approach is presented using titles and categories of Wikipedia articles to characterize documents. After stopword removal and stemming a weight is assigned to each word of the source document. Afterwards all Wikipedia titles and related articles supported by words in the document are collected and weighted. Finally the assigned categories are retrieved and ranked. The top n categories are used to describe the document content. Interestingly in general scenarios Wikipedia categories seem more useful to describe documents than the respective full text.

Finally, in [6] a clustering method is introduced that uses Wikipedia concept and category information for document clustering. Beside an exact match strategy, also a relatedness-match is presented avoiding the synonym problem by not merely using Wikipedia article titles for matching, but also considering the content of the whole Wikipedia articles. The outcome of all these approaches is that Wikipedia is indeed useful for describing and summarizing the content of documents. However, all approaches were focused on general documents, respectively Web retrieval.

In this paper we investigated the usefulness of Wikipedia for document description in specialized domains, in particular in the area of chemistry. We take a document collection from the open access journal *Archive for Organic Chemistry* (ARKIVOC) and assign Wikipedia categories to each document. Furthermore, we also assign the terms of the domain-specific ChEBI ontology to each document. We then represent each document of the collection by a Wikipedia categories tag cloud and a corresponding ChEBI tag cloud. A survey with a team of domain-experts was used to evaluate the different representations and assess the degree to which each representation is useful.

[1] www.cas.org

2. WORKFLOW

In the area of chemistry searching for relevant literature is essentially centered on chemical entities. Therefore, the first necessary step is to extract all chemical terms from each document of the repository. We decided to use the OSCAR3 framework [4] which is currently the only open source project on the market focusing on the automatic extraction of chemical entities. It offers a wide range of functionalities for annotating chemical entities, reactions and concepts.

Usually a search for chemical literature is based on chemical structures which are drawn in special Web interfaces. But, recent work in [9] clearly shows that after proper indexing even specialized chemical document repositories like open access journals can be queried using a common Web search engine. The retrieval quality of the resulting enriched index pages is almost as good as chemical exact structure searches and significantly better compared to a full text search.

The framework presented in this paper was developed within the ViFaChem II project [10]. It produces compact document descriptions using external knowledge bases. These descriptions help practitioners to get an overview of the document's content. The first knowledge base we use is Wikipedia. Each Wikipedia article describes a single topic and is assigned to at least one Wikipedia category. These categories are organized in a hierarchically manner and form a kind of simple ontology [11]. The question is, if this ontology includes enough detailed information to describe also complex chemical documents?

As a baseline domain-specific knowledge base we also use the ChEBI ontology to annotate the documents from the collection. Compared to Wikipedia, the ChEBI ontology, however, is focused on chemical entities. It does not include information about reactions or other chemical terms. Beside the entities trivial names, also some SMILES and InCHI codes are available in ChEBI. However, ChEBI only includes small molecules which are of biological interest. Therefore, it is far away from being an all-embracing chemical ontology. Although it is focused on chemical molecules of biological interest, for practitioners from the field of biology it is often not useful due to the lack of biomolecular domain knowledge [2].

For each document we have a list of all extracted chemical terms. Each term is mapped to a Wikipedia page and a ChEBI ontology node. Since we are not interested in the Wikipedia pages themselves, we only extracted the associated category entries. Starting from this entry point all parent categories are extracted and appended to the chemical term. We did the same for the ChEBI ontology nodes. Each chemical term is described by a set of categories and a set of ontology nodes. Hence, the documents are described as the union of the category/ontology node sets of all included chemical entities. Finally, each document is represented by two different tag clouds, one containing the ChEBI nodes and the other containing Wikipedia categories.

3. EVALUATION

Our evaluation scenario is based on 2588 chemical documents from the journal *Archive for Organic Chemistry (ARKIVOC)* which is one of the most renowned open access sources for organic chemistry. It includes documents from bio-organic, organometallic and physical-organic chemistry. These documents

are processed with the OSCAR3 framework which is used to detect and extract chemical terms.

3.1 Where to find the most important entities?

Since document retrieval in the chemistry domain is centered on chemical entities we evaluated in the first experiment where to find the most important chemical terms regarding the document context. Therefore, a group of domain experts manually evaluated the importance of each chemical term occurring in a document.

We took a random set of documents from our collection and automatically extracted all chemical terms using the entity recognition module from OSCAR3. Beside chemical entities also reactions or other chemical concepts are annotated.

We then delivered the documents and the corresponding term lists to a team of domain experts who marked the most descriptive terms for each document. We observed that in most documents the relevant terms are already mentioned in the title and/or the abstract. Nevertheless, in some abstracts placeholders are used to link to a complex structure drawn in an image or to other complex text-fragments. For example, the 3 in *hexabromide 3* is linked to an image visualizing the complex structure of: *1,2,3,4,9,10-hexabromo-1,2,3,4-tetrahydroanthracene*. Especially the information contained in drawn representations is currently not automatically extractable.

3.2 Wikipedia Category Suitability

In the second experiment we took the important terms (title and abstract) found in our document collection and retrieved all associated Wikipedia categories. A team of chemists evaluated if these categories are useful for describing chemical documents.

In total our collection contains 11952 distinct chemical terms. Surprisingly, for 2163 we found an entry page in Wikipedia (18%) which include 745 distinct categories in total. Now, a team of domain experts analyzed the set of Wikipedia categories found for those pages. For each category our experts rated whether it is useful for classifying chemical documents or not. More than 25% are rated as good, but only two categories were considered very useful for classifying chemical documents: *addition reactions* and a*minoglycoside antibiotics*. However, more than 50% are at least not bad and can in principle be used for document classifying.

3.3 Mapping Traceability

In the next step we want to know whether the mapping of the chemical terms to ChEBI ontology nodes, respectively Wikipedia categories, is comprehensible. We analyzed the distribution of the important terms from our document collection and choose a random subset. For each term in this subset we retrieved the matching Wikipedia categories and ChEBI ontology nodes. Again, a team of domain experts evaluated the mapping quality of the associated nodes/categories. Furthermore, we evaluated this quality based on the level in the category/ChEBI tree.

In total we found entries for 2163 disjunct terms occurring in title and abstract of the documents from our collection. Due to the automatic extraction process from OSCAR3 there are some recognition failures, e.g. the most often occurring term is *A* with 579 hits. For the query evaluation we therefore choose a representative subset. We took all terms occurring between 20 and 100 times resulting in a set of 129 chemical terms. From this subset we choose approximately 10% as query terms (resulting in 12 queries). For each query term we retrieved the matching ChEBI ontology node and all their parent nodes up to the ontology's root node. 121 of these terms were found in the ChEBI ontology (94%). Furthermore, we searched for the matching Wikipedia page and extracted their categories. Here, 79 of the 129 terms have a matching Wikipedia page (61%). Every Wikipedia page is assigned by at least one category. For each category we took all parent categories up to the root node. Finally we have for each query term a set of ontology terms and a set of Wikipedia categories. These sets are evaluated by a group of domain experts who rated for each category how close it is related to the query term. We also considered the level information for each node. Here, the level means the number of edges that needs to be passed to reach a node starting at the query node. The results visualized in Figure 1 show that for the domain-specific ontology the first three levels are almost equally important, whereas for the more general Wikipedia categories only terms of the first level are really relevant.

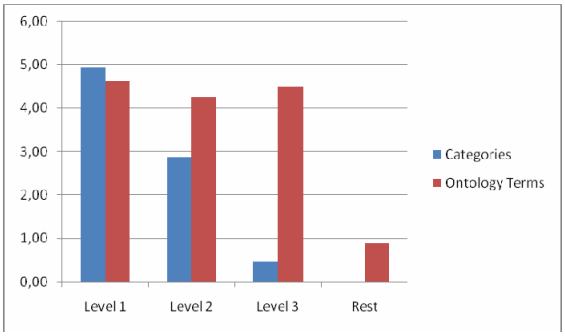

Figure 1. Level-based score distribution

The reason is that Wikipedia includes knowledge from many different domains. If we explore the categories graph we see that only a few steps away from the query term we reach categories that are not relevant for our domain. Comparing the categories and the ChEBI graph it is interesting that the different levels in the ChEBI graph include fewer nodes.

Table 1: Average number of associated terms for each chemical entity

	With Level Restriction	Without Level Restriction
Wikipedia	4	76
ChEBI	11	39

Table 1 gives an overview of the number of associated ontology/category terms for chemical entities. The values are averages and refer to the representative subset of the 129 chemical terms. Using the level information the number of associated terms has been highly decreased.

3.4 Comparing Wikipedia categories and ChEBI ontology terms

In the last experiment we created tag clouds based on the category information and the ChEBI ontology nodes associated with each document. Our group of domain experts evaluated the quality of

these clouds by stating how the content of the document is represented by the terms in the clouds. The rating values range is from 0 (bad) to 5 (very good). We randomly took a set of documents from our collection, extracted all chemical terms and searched for matching ChEBI terms and Wikipedia categories.

Since the really descriptive terms are occurring in the title or abstract and are usually rare, in our weighting scheme seldom occurring terms get higher weights. Figure 2 shows an example of a categories cloud. Please note, that the clouds only include the top 30 terms.

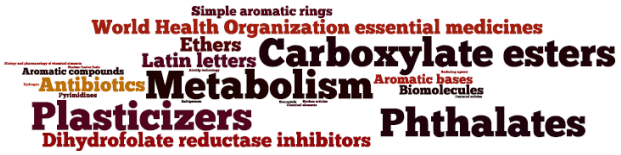

Figure 2. Example: category cloud

This cloud got an average score of 4, meaning to offer a good description of the corresponding document. Of course, some terms are also not useful, like for instance *latin letters*, but nevertheless, most terms give a good overview of the documents topic. Figure 3 shows the ChEBI ontology cloud for the same document.

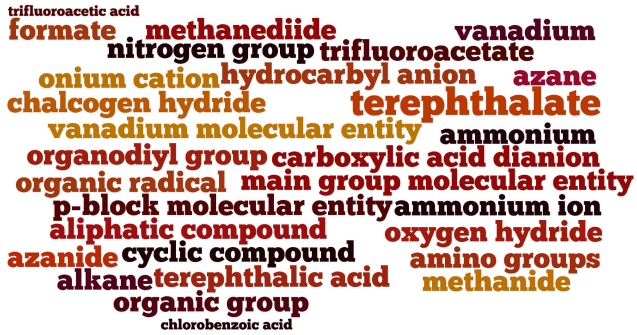

Figure 3. Example: ChEBI ontology cloud

The average score for this cloud is 1. To the means of our domain experts it still does not describe the document's content as well as the categories cloud. In total the average scores of the category clouds are 2.8 and of the ChEBI clouds 2.2.

To conclude, we want to state that it is really surprising that from the domain experts view the Wikipedia clouds are more descriptive for chemical documents than a domain-specific knowledge base, like the ChEBI ontology. We already mentioned that the domain of chemistry includes many different working areas. The chemists have a different understanding of the relations between chemical substances heavily dependending on the area they are working in. We did a final survey with a team of chemists to analyze if the relations between entities in the ChEBI-ontology are comprehensible. The results are that even for chemists from the area of organic chemistry not all relations in the ontology are comprehensible. Thus, for them, the mapping of the ontology terms to the chemical entities is also invalid. Indeed, the Wikipedia categories offer a suitable alternative to domain-specific ontologies. Since the quality of the document descriptions is good the next step is to use this knowledge in a document retrieval scenario.

4. CONCLUSIONS

In this paper we have shown an approach using Wikipedia categories to generate compact representations of chemical documents. As a baseline we used a domain-specific ontology (ChEBI) to represent the documents and compared the results. Each document from our repository is described by a Wikipedia categories cloud and a ChEBI ontology cloud. Our evaluation by a team of domain experts has shown that the Wikipedia categories are even more expressive for describing chemical documents than the handcrafted, domain-specific ChEBI ontology terms. Therefore, we have shown that the Wikipedia categories system can be used in domain-specific portals to overcome the problem of expensive, manually created ontology knowledge.

5. REFERENCES

[1] G. Bisson, C. Nédellec, and L. Canamero, "Designing clustering methods for ontology building-The Mo'K workbench," *In Proc. of the ECAI Ontology Learning Workshop*, Berlin, Germany, 2000.

[2] J. Choi, M.J. Davis, A.F. Newman, and M. Ragan, "A semantic web ontology for small molecules and their biological targets," *Journal of Chemical Information and Modeling*, vol. 50, 2010.

[3] P. Cimiano and S. Staab, "Learning concept hierarchies from text with a guided hierarchical clustering algorithm," *In Proc. of Workshop on Learning and Extending Lexical Ontologies by using Machine Learning Methods*, Bonn, Germany, 2005.

[4] P. Corbett and P. Murray-Rust, "High-throughput identification of chemistry in life science texts," *Computational Life Sciences II*, 2006.

[5] K. Degtyarenko, P. de Matos, et al., "ChEBI: a database and ontology for chemical entities of biological interest," *In Nucleic Acids Research*, vol. 36, 2008.

[6] X. Hu, X. Zhang, et al., "Exploiting Wikipedia as external knowledge for document clustering," *In Proc. of Int. Conf. on Knowledge Discovery and Data Mining (KDD)*, New York, USA, 2009.

[7] P. Schonhofen, "Identifying Document Topics Using the Wikipedia Category Network," *In Proc. of Int. Conf. on Web Intelligence (WI)*, Hong Kong, China, 2006.

[8] S. Sie and J. Yeh, "Automatic Ontology Generation Using Schema Information," *In Proc. of Int. Conf. on Web Intelligence (WI)*, Hong Kong, China, 2006.

[9] S. Tönnies, B. Köhncke, O. Koepler, and W. Balke, "Exposing the Hidden Web for Chemical Digital Libraries," *In Proc. of ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, Surfers Paradise, Australia, 2010.

[10] S. Tönnies , B. Köhncke , O. Koepler , and W. Balke, "Building Chemical Information Systems - the ViFaChem II Project," *13. GI-Fachtagung Datenbanksysteme für Business, Technologie und Web,* Münster, Germany, 2009.

[11] J. Yu, J. Thom, and A. Tam, "Ontology evaluation using Wikipedia categories for browsing," *In Proc. of Int. Conf. on Information and Knowledge Management (CIKM)*, Lisboa, Portugal, 2007.