Catching the Drift – Indexing Implicit Knowledge in Chemical Digital Libraries

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

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

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

Abstract. In the domain of chemistry the information gathering process is highly focused on chemical entities. But due to synonyms and different entity representations the indexing of chemical documents is a challenging process. Considering the field of drug design, the task is even more complex. Domain experts from this field are usually not interested in any chemical entity itself, but in representatives of some chemical class showing a specific reaction behavior. For describing such a reaction behavior of chemical entities the most interesting parts are their functional groups. The restriction of each chemical class is somehow also related to the entities' reaction behavior, but further based on the chemist's implicit knowledge. In this paper we present an approach dealing with this implicit knowledge by clustering chemical entities based on their functional groups. However, since such clusters are generally too unspecific, containing chemical entities from different chemical classes, we further divide them into sub-clusters using fingerprint based similarity measures. We analyze several uncorrelated fingerprint/similarity measure combinations and show that the most similar entities with respect to a query entity can be found in the respective sub-cluster. Furthermore, we use our approach for document retrieval introducing a new similarity measure based on Wikipedia categories. Our evaluation shows that the sub-clustering leads to suitable results enabling sophisticated document retrieval in chemical digital libraries.

Keywords: chemical digital collections, document ranking, clustering

1 Introduction

During the last years, the information gathering process and the way of searching for literature has changed. Beside text-based Web searches many data providers extended their services by offering user-centered, personalizable Web portals enabling searches over heterogeneous document collections and topical databases. However, the way of information seeking and offered data strongly depends on the scientific domain and the task at hand.

adfa, p. 1, 2011.

In the domain of chemistry information seeking is essentially focused on chemical entities. Whereas domain experts can easily identify chemical entities and classify them in the context of the document, it is currently a hard task to extract this information automatically. Chemical documents refer to entities usually using trivial names relying on the reader to figure out the contextual information. But also this does not help indexing: each chemical structure may have several different trivial names, often chosen with respect to the paper's context, e.g., pharmaceutical names, brand names, or terms from natural product chemistry. As always, the challenge for search engines using entity names is to discover all related synonyms and disambiguate terms based on the document context. In particular, practitioners as well as academic researchers are usually interested in finding all related documents to individual chemical entities. For both the search is basically recall-oriented because especially for synthesis procedures missing information about for instance existing patents or expected yields may lead to considerable financial losses.

Facing these problems, chemical information service providers offer specialized indexes. These indexes are built up by *manually* identifying all chemical structures from a document collection in structure databases. The resulting structure databases are accessed through graphical interfaces. By drawing a chemical structure a domain expert can thus formulate a query, which in turn will be parsed by the chemical query parser and matched against entities' fingerprints stored inside the structure database. The amount of manual work required for building up and maintaining such indexes results in high costs. Today, the most important provider is the Chemical Abstract Service (CAS) offering high quality data 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.

Focusing on the important field of drug design, the information gathering and indexing process is even more complex. A chemist from the area of synthetic chemistry is not only interested in a specific chemical entity, but in a representative of a chemical class adopting a specific role. Especially he is interested in entities having the same or similar characteristic chemical reactions. To assess if a chemical entity is relevant for his task in mind he uses his implicit knowledge about chemical classes and reaction behaviors. The characteristic reaction behavior of a chemical entity is defined by its functional groups. Functional groups are specific groups of atoms that will undergo the same or similar chemical reactions independent of the molecule they are part of. However, currently there is no knowledge base available allowing for this kind of automatic classification of chemical entities.

The goal of this paper is to open up the chemical knowledge stored in the Web for the field of drug design, respectively chemical synthesis. We will introduce an approach allowing a chemist from the field of drug design to search for documents containing chemical entities reflecting his implicit similarity perception. The first part in the introduced workflow is responsible for the automatic extraction of chemical entities from document collections and the index creation with synonym mark-up and disambiguation. We rely on our approach described in [1] to create for each document in the collection an extended index page including synonyms and different representations of all included chemical entities. In a next step, we use the entity's structure information included in the index pages to extract all functional groups. Entities having the same functional groups are grouped together into one cluster. Since for most entities this segmentation still does not fit to the chemist's implicit perception of chemical classes, the respective clusters are further refined into sub-clusters using fingerprint based similarity measures. We use the resulting clusters in a document retrieval scenario showing that the computed sub-clusters decrease the number of compared entities a lot while still including almost all relevant documents. Furthermore, we introduce a similarity measure dealing with the specific requirements of not only taking the simple occurrence of the query entity into account but reflecting the implicit similarity perception of the chemist.

The rest of the paper is organized as follows: In section 2 we will give an overview of the related work. Section 3 introduces our workflow for clustering chemical documents based on the chemist's implicit knowledge. Our evaluation is described in section 4. We will conclude with a summary and outlook to future work in section 5.

2 Related Work

Considering the domain of chemistry, information seeking is focused on chemical entities. In most cases a substructure search is performed using a graphical query interface. However, when searching for entities showing a specific reaction behavior, which is characterized by their functional groups, a substructure search is not sufficient since not all relevant entities will be found [2]. The approach described in [2] shows how chemical entities can be indexed to perform a functional groups search in a fast way. But, as we will show later it is still necessary to further decompose the resulting entity sets to meet the chemist's implicit knowledge. Another interesting approach is introduced in [3] where a small ontology for functional groups is built.

For document retrieval the first necessary step is the extraction of all chemical entities from the documents. For an automatic extraction the only open source chemical entity recognition tool currently available is the OSCAR framework [4], which can identify and extract multiple name variations of chemical entities. In combination with name-to-structure algorithms these entity names can be transformed into chemical structure information [5]. Besides structural information also several text-based formats have been developed for the internal digital representation and exchange of structures. Morgan [6] and Gluck [7] have developed algorithms to store two-dimensional atom-bond structural representations of chemical entities in a tabular form, so-called connection tables. In addition, linear notations have found widespread use. The early Wiswesser Line Notation (WLN) [8] or the later SMILES [9] are representations of chemical structures in the form of a linear string of alphanumeric symbols. The latest development is the InChI Code, an open standard for chemical structure description, by the IUPAC [10].

The idea of clustering data into groups of similar objects is widely used through almost all domains. [11] gives a comprehensible review of different clustering techniques used in data mining. Also the need for clustering chemical entities is discussed in several papers. An early approach described in [12] uses hierarchical clustering based on ring substituents derived from the WLN. The authors compare different clustering techniques based on two small datasets. In [13] an algorithm is presented, called *HierS*, which is also based on a hierarchical clustering method. The algorithm is unsupervised and uses explicit topological chemical graphs to construct hierarchical relationships between ring features of chemical compounds. An overview of different clustering methods used in computational chemistry can be found in [14].

In our scenario clusters are built using specific properties of chemical entities. All entities in one cluster have the same functional groups and have to be representatives of the same chemical class. For building sub-clusters we use a fingerprint-based similarity measure computed between each pair of entities. However, for similarity computation between chemical entities many different measures are available. The first necessary step for computing similarity is the transformation of a chemical substance into a fingerprint. Of course, the idea of measuring the similarity between two objects, each defined by a set of common attributes, is also discussed in many other domains, like e.g. biology [15]. The important point is that the used similarity coefficients are almost the same across all different application areas. Researchers from all domains have worked on finding the most meaningful measure. For chemical information systems the work done by Willett et.al, [16] and [17], gives overviews of the coefficients that have found widespread use. However, only a few comparative studies are available. Hubalek collected 43 similarity measures for the field of biology. After evaluating similarities, correlations, transformations of the value range and symmetry, 23 were excluded. The remaining measures were used for clustering fungi data resulting in five clusters of related coefficients [15]. For the domain of chemistry, Willet evaluated 13 similarity measures for binary fingerprint code [18]. Our approach described in [19] uses these measures and combine them with different fingerprint representations to identify correlation between them. Since many of them are uncorrelated we presented a personalized retrieval system for chemical documents using a feedback engine to find the best similarity measure for each individual chemist.

3 Clustering of Chemical Documents

The following scenario showcases the daily tasks of a researcher in the chemical domain. Assume our scientist is interested in anti-tuberculosis drugs, their pharmacological activities and synthesis. He may start by looking for information about *Isoniazid* and related drugs. *Isoniazid* is an organic compound and the treatment of choice for tuberculosis. Naturally our researcher is looking for experimental procedures for the synthesis of *Isoniazid*-like structures as he would like to minimize the side effects and the risk of resistance. In a first step, the chemist analyzes the structure of *Isoniazid* and identifies the parts of the molecule responsible for the specific reaction behavior. Furthermore, he implicitly knows that *Isoniazid* belongs to the chemical class of *hydrazines*. In particular, he is interested in chemical substances having the same reaction behavior and chemical class as *Isoniazid*. As starting point, our chemist will use *4-cynaopyridine* as it is already available in his laboratory. The question to solve is how to synthesize *4-cynaopyridine* to get a substance having the same functional properties as *Isoniazid*. Therefore, our chemist is searching for literature where the synthesis of *Isoniazid*-like structures is described. Furthermore, also the chemical entity *4-cynopyridine* should be included as educt. The first step is the search for entities with the same functional groups as *Isoniazid*: *hydrazine derivative, aromatic compound, carboxylic acid hydrazide, heterocyclic compound*. Furthermore, relevant entities must also belong to the class of *hydrazines*. Finally the result set is filtered for papers including the chemical entity *4-cynopyridine* as an educt. As final step, the chemist can now examine, if the reaction described in the papers can also be used for his chemical entity.

To allow for these types of searches in a document retrieval system it is necessary to detect the role of a chemical entity in the respective document. In addition, the even more important part and the main focus of this paper is the question of how to reflect the chemist's perception of chemical entities belonging to the same chemical class.

Our idea is to build clusters of chemical entities based on their functional groups. Each cluster describes a class of entities with similar reaction characteristics. Our approach comprises the following four steps, which are illustrated in Fig. 1.

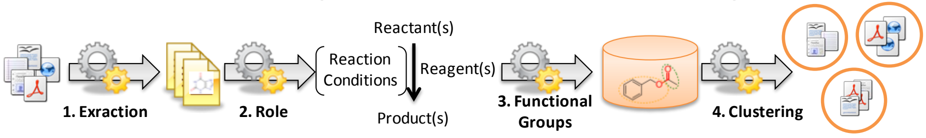


Fig. 1. Workflow Overview

Extraction of Chemical Entities: The first necessary step is to extract all chemical entities from the documents. We used our approach described in [1] to create an enriched index page for each document in the collection. The resulting index pages include further metadata, like synonyms, SMILES and InChI codes for all extracted chemical entities. The corresponding metadata has been extracted from a dumb of the PubChem database that includes all required information.

Role Identification: We have to determine the role of an entity within the document. Therefore, we analyzed our chemical document collection and identified 36 often used lexico-syntactic patterns. For a complete overview of the patterns see our website[1]. These patterns are in pseudo code, meaning that *[CHEMICAL]* is a placeholder for a recognized named entity, marked up by OSCAR. The placeholders are numbered consecutively and replaced by the respective role.

Calculation of Functional Groups: To determine the functional groups of a chemical entity the chemical structure of this entity is analyzed. There are different representations of chemical structures in the form of linear string notations. Usually the SMILES notation is used for functional groups analysis. We rely on the command line utility *checkmol*[2] to determine the functional groups of a chemical entity. Checkmol analyzes the input molecule for the presence of approximately 200 functional groups. We analyzed the output in a first short experiment with a group of domain experts and find out that checkmol simply recognize the presence of an *aromatic ring* but does not further investigate the dimension of contained aromatic rings. To

[1] http://www.l3s.de/vifachem/pattern.pdf

[2] http://merian.pch.univie.ac.at/~nhaider/cheminf/cmmm.html

enhance the quality of the resulting clusters, we added an extra parsing step to checkmol's output, to *determine* the dimension of an aromatic ring, resulting in n/m-aromatic rings, where n stands for the number of contained aromatic rings and m for the number of connected ring groups.

Building (Sub-)Cluster: Each entity is located in one cluster and all other entities in that cluster have the same functional groups. These clusters were analyzed by a group of domain experts. The result was that some clusters are not specific because they contain entities from different chemical classes not fitting to the chemists' implicit perception. Therefore, we further decomposed such clusters into sub-clusters by computing fingerprint based similarity between the included entities. Fingerprints encode molecular structures in a series of binary digits (bits) where bits are set according to occurrences of particular structural features. For generating fingerprints, again the entity's unique SMILES representation is used. There are several types of chemical fingerprints focusing on different fragments of chemical entities [19].

As clustering algorithm we choose a partitioning method which constructs k partitions of the data. Each partition represents a cluster and satisfies the following requirements: each group must contain at least one object and each object must belong to exactly one group. One of the most famous algorithms from this group is the k-means clustering which we chose using the Weka33 API.

3.1 Cluster Based Document Retrieval

Each document is associated to the clusters based on its included chemical entities. To search for documents, the query entity is assigned to the respective sub-cluster and all related documents are retrieved. Instead of just delivering all documents, the result set is ordered according to a similarity measure. In our scenario a document is relevant for a query term if some chemical entity in the document has the same functional properties, respectively the same chemical class, as the query entity. Therefore, we need a specific similarity measure not only taking the simple occurrence of the query term into account. We developed a measure which is based on the Wikipedia category information. In [20] we have shown that Wikipedia categories can be used to describe the content of chemical documents. The Wikipedia categories are structured in a taxonomic tree based on the relationships between them. Here, the idea is to retrieve for each document the associated categories based on the included chemical entities. Since Wikipedia includes information from many different domains it is not sensible to use the whole category tree for describing chemical entities. The evaluation in [20] has shown that only categories that are up to two nodes away from the query node are useful. We retrieve the respective categories for each query term and each document in the query's sub-cluster. The documents are ranked according to the following similarity measure:

$$swc(q_i, d_j) = \frac{cq_i d_j}{cq_i} \times \frac{cd_j}{ed_j}$$

[3] http://www.cs.waikato.ac.nz/ml/weka

where q_i is the query term and d_j the respective document. The *swc* measure consists of two parts. The first quotient divides the number of categories found for query term *i* in the respective document *j* ($cq_i d_j$) by the total number of categories found for query term *i* (cq_i). The second quotient divides the total number of categories for the document (cd_j) by the total number of chemical entities found in that document (ed_j).

4 Experiments

For our evaluation we used a dump of the PubChem database[4] containing around 31.5 million chemical entities. For each entity we determined the functional groups and created an inverted index with name and entity allocation. In addition, we used a collection of 2588 chemical documents from the journal Archive for Organic Chemistry (ARKIVOC) which is one of the most renowned open access sources for organic chemistry. For each of these documents we extracted the chemical entities and their roles within a reaction.

4.1 Clustering Based on Functional Groups

In the first experiment we want to gain first insights about the entities contained in the PubChem dump. We took the entities SMILES codes and used our extended version of the command-line tool *checkmol* to determine the respective functional groups, resulting in a set of functional groups for each entity. All entities containing exactly the same set of functional groups are grouped into one cluster. The resulting distribution of all 31.5 million entities is shown in Fig. 2 (left).

# Contained Entities	# Clusters
1	773092
$1 < x \leq 10$	816817
$10 < x \leq 100$	226147
$100 < x \leq 1000$	36535
$1000 < x \leq 10000$	3615
$10000 < x \leq 100000$	143
$100000 < x$	0

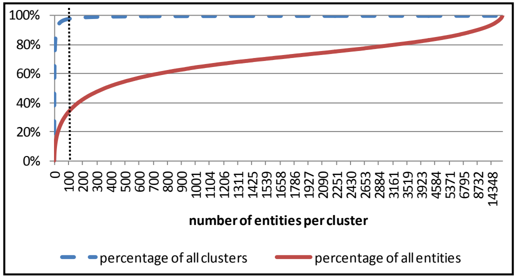


Fig. 2. Cluster sizes (left), Number of entities per cluster (right)

We did a survey with domain experts to analyze the clusters. The result is that clusters containing up to 100 chemical entities are still reasonable for domain experts meaning they correspond to the chemist's implicit knowledge. Therefore, we can discover that 97.84 % of the resulting clusters can already be used. But these clusters only contain around 30% of all chemical entities (see Fig. 2 right). Most of the entities (around 21 million) are located in the remaining 2.16 % of the clusters. Therefore, it is necessary to split them up into more meaningful clusters.

[4] http://pubchem.ncbi.nlm.nih.gov

4.2 Building Meaningful Sub-Clusters

In this experiment we divide all clusters containing more than 100 chemical entities. For choosing a meaningful distance, respectively similarity function, we take the 5 different uncorrelated fingerprint/similarity measure combinations introduced in [19] into account. Each of these measures has different rankings with respect to the underlying fingerprint. To decide for a measure for the sub-cluster computation we evaluated for which of the measures the top-X ranked entities are in the same functional group cluster.

We randomly choose 100 clusters with more than 1000 entities per cluster. In addition, we choose 10 random queries and calculated the similarity between the query entity and all other entities from all clusters. For each entity we have six different fingerprint representations. The similarity is computed using the uncorrelated measures. Fig. 3 shows the average results based on the ten query entities.

The diagrams show that there are big differences between the different combinations. For the *substructure fingerprint* and the *Manhattan* distance always all top-100 entities are in the same functional groups cluster. For the top-1000 entities still around 800 are found in the same cluster. Since this combination retrieves the best results we decided to use it for sub-cluster computation.

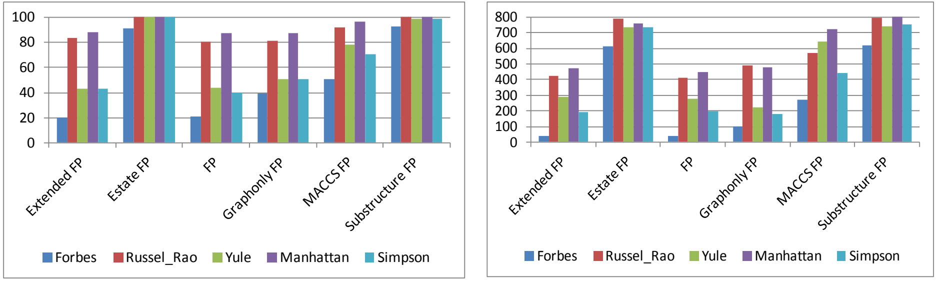


Fig. 3. Top-100 (left), Top-1000 (right)

Since we are using k-means clustering we also had to find a suitable k. The aim is that each entity in a cluster has the same chemical class. Therefore, we took a domain specific ontology including chemical classes as ground truth, the so-called ontology for chemical entities of biological interest (CheBI[5]). To the best of our knowledge CheBI is the only available ontology allowing for automatic classification of chemical entities. Since we are interested in decomposing the clusters with more than 100 entities (around 40000 clusters), we randomly took 2000 clusters (5%) out of this set. Since not all entities from our dataset are included in CheBI, we only choose clusters containing entities also included in the CheBI ontology. The idea is to take all entities from one cluster and assign the associated ontology classes to that cluster. Of course, it is not sensible to use all ontology nodes associated with one chemical entity. Nodes that are too general would lead to huge clusters that are again not meaningful. In [20] it was shown that only ontology nodes are meaningful that are at least three steps away from the entry node. Therefore, we only associated these classes with the respective cluster. We defined that the optimal segmentation is achieved, if all entities

[5] http://www.ebi.ac.uk/chebi

with different classes are in different sub-clusters. We manually built the respective sub-clusters and run the k-means algorithm varying the value for k. Our algorithm stops if k-means found the optimal solution, each entity is in one cluster for its own, or if no solution can be found. Evaluating the 2000 clusters we retrieved an optimal k for further splitting up the entities in the functional groups clusters in chemical classes of 4. Whereas CheBI includes for our dataset around 20000 chemical classes we were able to find more than 150000 classes for chemical entities. Please note, we cannot associate exact chemical class names to each cluster, but as we will see later, our results match the perception of the chemist's implicit knowledge of entities belonging to the same class.

4.3 Document Retrieval

During this experiment we evaluate the created clusters in a document retrieval scenario. Our document collection contains 2588 documents from the ARKIVOC journal. We generated enriched index pages using the approach described in [1] and extracted all contained chemical entities. Each document is associated to the functional groups clusters based on its contained entities. The last experiment has shown that the optimal segmentation of the entities in chemical classes is given for 4 sub-clusters. However, we will evaluate if this holds also for a document retrieval scenario.

First we have to randomly chose query entities and assess the relevance of each document for the respective query. Relevance can only be assessed manually by domain experts (in particular chemists), in what is a very expensive process. Therefore, we could not take the entire collection, but chose a subset of documents (still about 10% of the entire collection) for performing a precision/recall analysis. To choose a *representative* subset, we analyzed the number of occurrences of individual chemical entities in the document collection. It is not sensible to choose entities as query terms that either occur in almost every document or are extremely rare. We analyzed all entities occurring in less than 100 documents, but more than 20 documents. Furthermore, the entities should belong to functional groups clusters which have been further decomposed into sub-clusters using the fingerprint based similarity computations.

We retrieved all documents matching these queries and randomly chose a subset of 10%. From these documents we randomly selected a total of around 5% of the occurring entities resulting in 18 textual query terms. For the evaluation domain experts from the field of chemistry considered all retrieved documents with respect to each query and judged the relevance in a binary fashion. A document is marked as relevant if it contains entities having the same reaction behavior and belonging to the same chemical class as the query entity.

Now, we analyze if the sub-cluster decomposition is sensible meaning that all relevant documents for a query term are located in the same sub-cluster. Fig. 4 (left) shows the precision, recall and F-measure values for different values of k. The recall value is always around 93% meaning that some documents from other functional groups clusters were also marked as relevant. But if we are only using the functional groups clusters (k=1) we got a low precision value averaged over all queries of 42%. The precision value slightly increases up to 53% for k equals 12. According to the low precision values the classic F_1-Measure is on average only around 57%. The recall oriented F2-Measure results in an average of 68%.

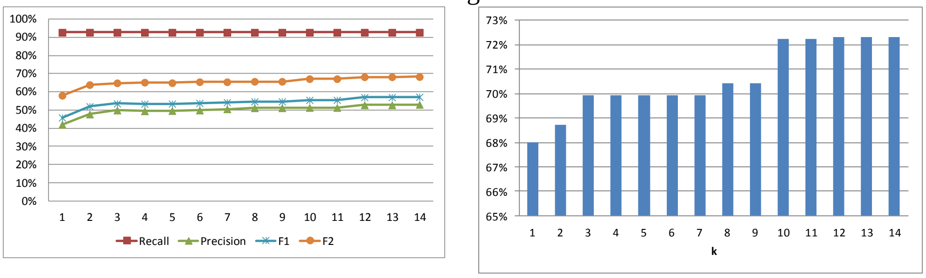
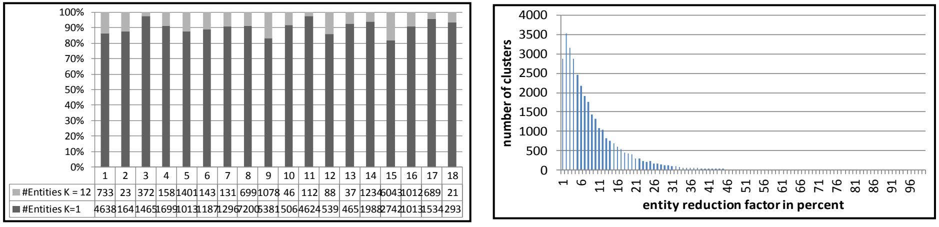


Fig. 4. Recall, Precision and F-Measures for varying k's (left), Mean Average Precision (MAP) for Wikipedia Categories Ranking and varying k's (right)

Please note we can further optimize the precision by using different k values for different queries, respectively different cluster sizes. For example, for query term 2 which is in a quite small cluster containing only 26 documents, already with k equals 2 we have a precision value of 60%. In contrast, for query term 15 the precision value for k equals 2 is 59% and for k equals 9 67%. Regarding all queries, the optimal value for k is varying between 1 and 12. But only 4 queries do not have their optimal precision value for k equals 12.

The last experiment has shown that we have an optimal segmentation for entities for k equals 4. For documents, k equals 4 is already good, but the precision value is slightly higher for k equals 12. We also tested higher values for k, but the precision did not increase anymore. Instead of delivering all documents in the sub-cluster randomly to the user, we also developed a similarity measure to rank the documents according to the query term. Fig. 4 (right) shows the mean average precision values for varying k's. Using the ranking we were able to reach a MAP value for k equals 12 of 72%. However, another interesting point is that even if the clusters include fewer documents, the recall value did not decrease. That means if a user is searching for documents with respect to a chemical entity with some characteristic reaction behavior and implicit knowledge of its chemical class, it is sufficient to find the cluster for this query entity and retrieve all included documents. If the query entity is located in a sub-cluster, it is not necessary to take chemical substances from other sub-clusters into account, even though they have the same functional groups. Our experiment has shown that almost all relevant documents are in the same sub-cluster as the query.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
▨ #Entities K = 12	733	23	372	158	1401	143	131	699	1078	46	112	88	37	1234	6043	1012	689	21
■ #Entities K=1	4638	164	1465	1699	1013	1187	1296	7200	5381	506	4624	539	465	1988	2742	1013	1534	293

Fig. 5. Number of entities for k=1 and k=12 (left), Number of clusters including x percent of the entities for k=12 compared to k=1 (right)

Fig. 5 (left) compares the number of entities in the functional groups cluster (k=1) to the number of entities for 12 sub-clusters for each query. The figure shows that the cluster sizes decrease on around 90% on average over all queries. Since the recall does not decrease the conclusion is that the cluster quality is high and only irrelevant entities are located in other sub-clusters.

Fig. 5 (right) shows the number of clusters including a certain percentage of entities for k equals 12. For example, there are 3500 sub-clusters where the number of entities has been reduced to 3% of the number of entities for k equals 1. This observation is quite important considering a facetted-search scenario. For example, in the ViFaChem2 portal[6] the user has the possibility to decide for relevant chemical entities after submitting a query. Our evaluation has shown that this set of offered chemical entities can be highly decreased by only considering entities from the same sub-cluster leading to a more sophisticated search experience.

5 Conclusion and Future Work

We presented an approach, clustering chemical entities based on their functional groups to reveal the chemists' implicit knowledge. A manual analysis of the resulting clusters by domain experts has shown that a simple functional groups clustering is not enough since most clusters are too unspecific and do not fit to the chemists' perception of chemical classes. Therefore, we used fingerprint based similarity measures to further divide these clusters into sub-clusters. Even though, we cannot assign explicit class names to the resulting sub-clusters, our evaluation has shown that they reflect the chemists' perception of chemical classes.

Further on, we used the clusters for document retrieval. The documents are assigned to the clusters based on their contained chemical entities. We did a precision/recall analysis with a group of domain experts showing that almost all relevant documents (recall of 93%) are located in the respective sub-cluster of the query. Instead of just delivering all documents from the respective cluster, we also introduced a ranking measure based on Wikipedia categories to further enhance the precision.

Another important point is that, without losing any relevant documents, the number of entities in the sub-clusters is dramatically decreased about 90% compared to the original functional groups clusters. This is quite important considering chemical Web portals using, e.g., facetted search, like for example the ViFaChem portal. It is a huge difference if the user retrieves 1000 or only 100 possible entries for the facets. Our evaluation has shown that all relevant chemical entities are located in the same sub-cluster. Thus, the number of possible hits is highly decreased resulting in a high quality search experience for the user.

For our future work we plan to investigate if our approach is also useful in other domains. Considering the field of biology, the search for information is also focused on entities, like e.g. genes or proteins. Just as in chemistry, a huge amount of these entities is available having specific characteristics, like e.g. phosphorylation sites that may be used in a clustering approach. We will see if this huge amount of data can also be confined without losing any relevant information in a document retrieval system.

[6] www.chem.de

6 REFERENCES

1. 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.
2. N. Haider, "Functionality Pattern Matching as an Efficient Complementary Structure/Reaction Search Tool: An Open-Source Approach.," *Molecules*, vol. 15, no. 8, 2010.
3. H. J. Feldman, et al., "CO: A Chemical Ontology for Identification of Functional Groups and Semantic Comparison of Small Molecules.," *FEBS letters*, vol. 579, no. 21, 2005.
4. 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.
5. J. A. Townsend, et al., "Chemical Documents: Machine Understanding and Automated Information Extraction," *Journal of Organic & Biomolecular Chemistry*, vol. 2, 2004.
6. H. L. Morgan, "The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service," *Journal of Chemical Documentation*, vol. 5, no. 2, 1965.
7. D. J. Gluck, "A Chemical Structure Storage and Search System Developed at Du Pont.," *Journal of Chemical Documentation*, vol. 5, no. 1, 1965.
8. E. Smith, P. Baker, and W. Wiswesser, *The Wiswesser Line-Formula Chemical Notation (WLN)*, vol. 102, no. 2. Chemical Information Management (Cherry Hill, N.J.), 1975.
9. D. Weininger, "SMILES, A Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules," *Journal of Chemical Information and Modeling*, vol. 28, no. 1, 1988.
10. S. E. Stein, S. R. Heller, and D. Tchekhovskoi, "An Open Standard For Chemical Structure Representation: The IUPAC Chemical Identifier," *In Proc. of the International Chemical Information Conference*, 2003.
11. P. Berkhin, "A Survey of Clustering Data Mining Techniques," *Journal of Grouping Multidimensional Data*, no. c, 2006.
12. G. W. Adamson and D. Bawden, "Comparison of Hierarchical Cluster Analysis Techniques for Automatic Classification of Chemical Structures," *Journal of Chemical Information and Modeling*, vol. 21, no. 4, 1981.
13. S. J. Wilkens, J. Janes, and A. I. Su, "HierS: Hierarchical Scaffold Clustering Using Topological Chemical Graphs," *Journal of Medicinal Chemistry*, vol. 48, no. 9, 2005.
14. G. M. Downs and J. M. Barnard, "Clustering Methods and their Uses in Computational Chemistry," in *Reviews in ComputationalCchemistry*, vol. 18, 2002.
15. Z. Hubálek, "Coefficients of Association and Similarity, Based on Binary (Presence-Absence) Data: An Evaluation," *Journal of Biological Reviews*, vol. 57, no. 4, 1982.
16. P. Willett, J. M. Barnard, and G. M. Downs, "Chemical Similarity Searching," *Journal of Chemical Information and Modeling*, vol. 38, no. 6, 1998.
17. J. Holliday, C. Hu, and P. Willett, "Grouping of Coefficients for the Calculation of Intermolecular Similarity and Dissimilarity Using 2D Fragment Bit-Strings," *Journal of Combinatorial Chemistry; High Throughput Screening*, vol. 5, no. 2, 2002.
18. P. Willett, "Similarity-based Approaches to Virtual Screening," *Journal of Biochemical Society Transactions*, vol. 31, 2003.
19. S. Tönnies, B. Köhncke, and W.-T. Balke, "Taking Chemistry to the Task – Personalized Queries for Chemical Digital Libraries," *In Proc. of the Joint Conf. on Digital Libraries (JCDL)*, 2011.
20. B. Köhncke and W.-T. Balke, "Using Wikipedia Categories for Compact Representations of Chemical Documents," *In Proc. of the Int. Conf. of Information and Knowledge Management (CIKM)*, 2010.