Taking Chemistry to the Task – Personalized Queries for Chemical Digital Libraries

Sascha Tönnies
L3S Research Center
Appelstrasse 9a
30167 Hannover
Germany
toennies@l3s.de

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

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

ABSTRACT

Nowadays, the information access is conducted almost exclusively using the Web. Simple keyword based Web search engines, e.g. Google or Yahoo!, offer suitable retrieval and ranking features. In contrast, for highly specialized domains, represented by digital libraries, these features are insufficient. Considering the domain of chemistry, where searching for relevant literature is essentially centered on chemical entities. Beside commercial information providers such as Chemical Abstract Service (CAS) numerous groups are working on building free chemical search engines to overcome the expensive access to chemical literature. However, due to the nature of chemical queries these are often overspecialized. Often we need meaningful similarity measures for chemical entities for query relaxation. In chemistry, the similarity measures are vast; more than 40 similarity measures are available and focus on different aspects of chemical entities. This vast number of similarity measures is obvious, because the desired search results highly depend on the working field of the chemist. In this paper we present a personalized retrieval system for chemical documents taking into account the background knowledge of the individual chemist. This is done by a query relaxation for chemical entities using similar substances. We evaluate our approach extensively by analyzing the correlation of commonly used chemical similarity measures and fingerprint representations. All uncorrelated measures are finally used by our feedback engine to learn preferred similarity measures for each user. We also conducted a user study with domain experts showing that our system can assign a unique similarity measure for 75% of the users after only 10 feedback cycles.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Storage and Retrieval – *Information Search and Retrieval*.

General Terms

Measurement, Experimentation, Human Factors.

JCDL'11, June 13–17, 2011, Ottawa, Ontario, Canada.

Keywords

Chemical Digital Libraries, Personalization, Query Relaxation

1. INTRODUCTION

Today, a keyword based Web search is the starting point for almost all information gathering processes. However, in some highly specialized domains a simple keyword based search is not sufficient. For example, the information gathering process in chemistry is entity centered. As a major information provider in the domain of chemistry, CAS subsidiary of the American Chemical Society (ACS) offers a specialized digital library indexing a variety of chemical document collections. Since digital libraries promise high quality information access, the ACS is maintaining their entity database, the CAS Registry, by manually indexing all chemical entities occurring in chemical literature. Further, they annotate the documents in order to build their CAS search index for chemical literature, resulting in a high quality digital library. This quality is prohibitively, gained at the expense of high costs for the manual indexing process. Moreover, search engine access is very expensive and strictly restricted to subscribers.

In an attempt to overcome the costly access to chemical literature, several groups are currently working on building free chemical search engines. Prime examples are the substance database PubChem[1] combining several chemical entity data sources and the document search engine ChemXSeer[2]. ChemXSeer relies on a highly complex process extracting chemical formulas in an automated way out of 150000 RSC publications and links them to the documents [1, 2]. Numerous publishers are also improving their information gathering process by adding chemical annotations to their documents. The prime example of this is RSC Publishing[3] utilizing the Oscar3 framework [3] to identify chemical entity names inside the document full text. These names are transformed into structural information, stored inside a structure database and linked to the document. However, these approaches still need special databases to handle chemical information.

In our previous work in [4] we have shown how structural data can be used for building up index pages for chemical documents. These index pages are indexed by Google and linked to the original documents. Since synonyms and different entity

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

[2] http://chemxseer.ist.psu.edu:8080/chemxseer

[3] http://pubs.rsc.org/

representations, like e.g. SMILES or InCHI code, are included in the index page a Boolean string based entity search retrieves almost all desired documents. However, chemists are usually searching for a very specific chemical entity occurring in a specific task. Besides the chemical entity this task is also necessary as query term to enhance the precision. Therefore, the search process is based on a combined Boolean query including both the chemical entity and a specific task.

In most cases the combination of chemical entity and specified task occur very seldom, resulting in a low number of retrieved documents (overspecialization). The task formulated as second query term imposes a hard constraint and cannot be relaxed. Considering the chemical entity, there are several other substances having the same functional properties. Therefore, it is inevitable to relax the first query term and search for entities with the same or similar properties.

In this paper we build a personalized retrieval system to overcome the problem of overspecialization. We relax the query term by computing similar entities. For computing similarity between chemical entities, the first necessary step is to convert the entities to a fingerprint representation. There are numerous fingerprints available, all of them emphasizing different attributes of a chemical entity, e.g. structural information, functional groups or number of atoms. Beside the different fingerprint representations, more than 40 similarity measures for chemical entities are available. In order to better understand the background story we first examined to which degree the similarity measures are correlated. The uncorrelated measures are further used in a feedback step in our system to learn which measure is most appreciated by the individual user. We evaluate which combination of fingerprint and similarity measure is useful for personalized query relaxation.

The rest of the paper is organized as follows: in section 2 we will give an overview of the related work. A typically use case for the daily work of a practitioner from the field of chemistry is introduced in section 3, followed by an analysis of different similarity measures used in chemistry in section 4. In section 5 we conducted a user study with domain experts. The outcome of the study leads to an architecture of a personalized retrieval system introduced in section 6. Finally we will conclude and give an outlook to future work.

2. RELATED WORK

Searching for chemical documents is essentially centered on the search for chemical entities. The problem faced is that of uniquely naming chemical structures in the text, while avoiding the ambiguity of systematic, IUPAC, trivial or brand names. Inspired by the work of Jacob H. van't Hoff and August Kekulé in the nineteenth century, drawings of chemical structures became the common way of communicating chemical information about substances and their reactions [5]. The chemical structure is easily interpretable by humans as a way to uniquely describe a chemical entity. Therefore, graphical representations of chemical entities are commonly used as query terms in searching for chemical information.

Although easily recognized by the human eye, graphical representations of chemical entities still cannot be easily transferred into the digital world once published in a document. Over the last years several projects focused on developing a chemical optical recognition for the reconstruction of chemical structure information from digitized documents. However, recognition rates always have proven to be insufficient in a production environment [6, 7, 8, 9]. Thus, the graphical chemical structure in the full text cannot be used as an information source for the entity recognition and one has to rely on the textual representations.

There are a few approaches focusing on the special requirements of searching for chemical literature implementing own search engines. An implementation of such a chemical search engine is presented in [2]. The basic assumption of their work is that chemists search for literature using a chemical formula. Since chemical formulas are ambiguous this is not the case in reality. Nevertheless, the approach solves some challenging problems. The first necessary step is to extract chemical formula from text documents; casting the task as a classification problem. The authors used classification methods based on SVM and a probabilistic model based on CRF. To speed up the retrieval process the detected formulas are indexed in a next step. Finally ranking functions are presented enabling suitable document retrieval. This approach is further extended in [10] using multiple query semantics allowing partial and fuzzy searches. However, when focusing on the chemical structural information, one still needs to solve the challenging task of extracting entities and thus structural information from full texts in a fully automated way.

Even though, automatic entity extraction is currently considered for a variety of domains, in chemistry the only open source chemical entity recognition tool currently available is the OSCAR3 framework [3], 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 [11]. This automatic tagging of chemical entities still leads to recognition errors.

The problem is to uniquely naming chemical structures in text and for internal use. For a long time, chemists have developed different algorithms for converting a chemical structure to unique line notations. Based on the algorithms developed by Morgan [12] and Gluck [13] it is possible to store two-dimensional atom-bond structural representations of chemical entities in a tabular form, so-called connection tables; linear notations having widespread use. The early Wiswesser line notation (WLN) [14], or the later simplified molecular input line entry specification (SMILES) [15], ROSDAL [16] and SYBYL line notation [17] are representations of chemical structures in the form of a linear string of alphanumeric symbols. The latest development is the International Chemical Identifier (InChI), an open standard for chemical structure description, by the IUPAC [18].

Based on these unique line notations, we have, in previous work, introduced an approach enabling the search for chemical literature using enriched index pages [4]. We used OSCAR3 for extracting chemical terms from a document collection and created an index page for each document. These pages include the trivial name and many different representations of the extracted entities, e.g. InCHI and SMILES code. Indexing these unambiguous representations we were able to reach the retrieval quality of a chemical structure search using a normal Google text search. Nevertheless, the problem of query overspecialization is still not solved.

For similarity computation between chemical entities many different measures are available. As we will see later, some of them are uncorrelated because their result sets differ. The first

necessary step for computing similarity is the transformation of a chemical substance into a fingerprint. Since a lot of fingerprint transformations are available, the amount of possible combinations of fingerprints and similarity computations between them is really high. The idea of measuring the similarity of two objects, each defined by a set of common attributes, is discussed in many different domains, including e.g. biology [19] or chemistry [20]. Although these application areas are divers, the used similarity coefficients are almost the same. Since the performance always relies on the choice of an appropriate measure, many researchers have worked on finding the most meaningful measure. The work done by Willet et.al, [20] and [21] gives overviews of the coefficients that have found widespread use in chemical information systems.

Even though numerous binary similarity measures have been described in the literature by their properties and features ([22, 23, 24, 25]), only a few comparative studies are available. In the field of biology Hubalek collected 43 similarity measures and after evaluating similarities, correlations, transformations of the value range and symmetry, 23 were excluded. The remaining ones were used for cluster analysis on fungi data to produce five clusters of related coefficients [19]. In the domain of chemistry, Willet evaluated 13 similarity measures for binary fingerprint code [26]. Current work identified the most useful fingerprint based chemical similarity measures [21]. We use these measures and combine them with different fingerprint representations to identify correlation between them.

3. USE CASE

The following scenario is typical for the daily work of a practitioner in the chemical domain. Imagine a chemist from the field of drug design who is currently working on an improvement of Viagra®. In this scenario he is searching for related literature about the active ingredient, *Sildenafil* (see Figure 1).

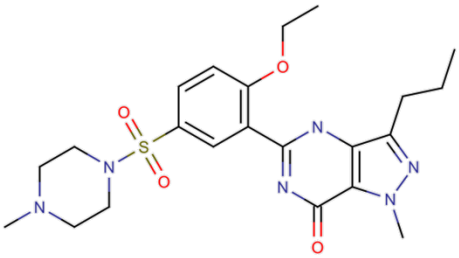


Figure 1 Structure of Sildenafil

Usually, the access to chemical literature is performed by a drawn query, using a specialized search interface. Today's chemical search engines are also able to relax the query entity by searching for entities with corresponding substructures or entities with similar properties. Problematically, all documents including these entities are returned. To overcome the obstacle of a specialized interface, we introduced an approach to open up the chemical domain for text based search engines [4]. The system generated an index page, containing all chemical entities included in the respective document, for each document in the digital library. Beside the entity name found in the document also all synonyms and different representations, like e.g. SMILES or InChI code are included. The evaluation has shown that the results were almost as good as a chemical structure search.

Chemistry is a wide field and chemical entities are usually used in many different contexts, e.g. drug design. Our chemist wants to overcome one specific side effect of *Sildenafil*, namely '*irregular heartbeat*' he is searching for documents describing this side effect. Since, to the best of our knowledge, current chemical search engines do not support the search for entities occurring only in a specific context, our chemist has to manually scan all retrieved documents.

Certainly, it is possible to build a search engine which has the ability to combine the entity and context as a query term. A simple architecture dealing with these combined queries is shown in Figure 2. Here, the combined user query *Q!* is sent to the search engine and split up into the chemical entity E_q and the specified context q_i. The documents including q_i can easily be computed using an inverted full text index. Searching for relevant documents regarding E_q is more difficult since we have to take all different entity representations (e.g. SMILES or InChI codes) and synonyms into account. To address ambiguity, we rely on chemical index pages (see [4]) to search for relevant documents. The intersection of both result sets shapes the final result set and is delivered to the user. Note that, due to the fact that in chemical documents the most relevant entity, i.e. the product of a synthesis, can occur only once, only Boolean queries are reasonable and traditional IR measure, e.g. TF*IDF, are not.

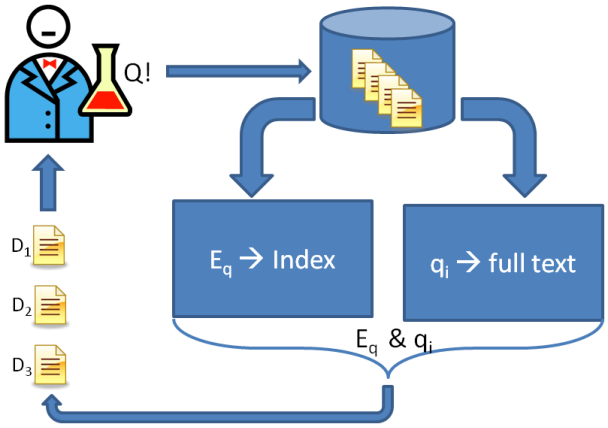


Figure 2: Simple architecture

Focusing on our chemist again, he can now search for documents about '*Sildenafil*' and '*irregular heartbeat*'. Unfortunately, he is still unable to fulfill his information need, because the active pharmaceutical ingredient '*Sildenafil*' is trademarked and cannot be used for other drugs. As a consequence, he must relax his query to find other chemical entities with similar properties within the same context. Indeed, this query relaxation should be done automatically by replacing the actual entity with similar entities. In the following sections we will extend the simple architecture to develop such a system.

4. FINGERPRINTS AND SIMILARITY MEASURES

In the context of chemical structure search a lot of work has been done in developing similarity measures for chemical entities resulting in a huge amount of available measures. All of these measures rely on a unique fingerprint representation of the chemical structure. In this section we will shortly describe the fingerprints and similarity measures widely used in the chemical domain.

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, the structure is converted into its unique SMILES representation [27]. There are several ways of creating fingerprints focusing on different fragments of chemical entities. Examples for typical fragments for generating fingerprints are:

- *Atom sequence*: A linear path of atoms and bonds through the molecule.
- *Ring composition*: An atom and bond sequence around a ring structure in the molecule.
- *Atom pairs*: A pair of atoms in the same molecule with number of bonds in the shortest path between them. The different atom pairs are usually further differentiated by, e.g., taking the number of attached hydrogens into account.

Sometimes fragments are too specific, leading to very low frequencies and sparse fingerprints, included atom and bond types can be generalized. We rely on the open source chemical development toolkit (CDK) [28] which includes the following fingerprints.

Standard Fingerprint This fingerprint examines the molecule and encodes the following:

- a pattern for each atom
- a pattern representing each atom and its nearest neighbors
- a pattern representing each group of atoms and bonds connected by paths up to 2 bonds long
- a pattern representing the atoms and bonds connected by paths up to 3 bonds long
- a pattern representing the atoms and bonds connected by paths up to 4, 5, 6, and 7 bonds long

Extended Fingerprint An Extended fingerprint includes in addition to the Standard fingerprint features for describing aromatic rings.

Graphonly Fingerprint This fingerprint is a specialized version of the Standard fingerprint that does not take the bond order into account.

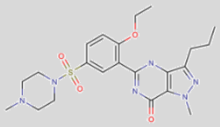
EState fingerprint generates 79 bit fingerprints using fragments describing the electronic and topological characterization of an atom, called electrotopological state (e-state) [29]. The fingerprint simply indicates if such a fragment is present in the structure or not.

Substructure Fingerprint currently supports 307 different substructures. A set bit indicates that the related substructure was found in the molecule.

Example: Substructure Fingerprint generation

Let us consider our chemist searching for *Sildenafil*. In a first step the name is converted to its unique SMILES representation: *CCCC1=NN(C2=C1NC(=NC2=O)C3=C(C=CC(=C3)S(=O)(=O)N4CCN(CC4)C)OCC)C*. This conversion is necessary, because SMILES codes include information about the molecular structure of a chemical substance. Now we want to create a Substructure fingerprint out of this SMILES code. For simplicity, let us consider that the Substructure fingerprint takes only 4 substructures into account. Each of the substructures is encoded in a SMARTS[4] pattern:

1. C=N-N-C: Pattern for an atomic arrangement taking bond orders into account.
2. C-S: Pattern for an atomic arrangement taking bond orders into account.



3. N-Br: Pattern testing the existence of a N-Br bond.
4. Oc1ccc(O)cc1: Pattern testing the presence of a specific substructure.

For each matching SMARTS pattern, we set the corresponding bit to 1. The resulting fingerprint for Sildenafil is 1100.

MACCS Fingerprint is the representation of the answer of 166 questions about a chemical structure [30].

Considering these fingerprints, we examined the most common useful measures (see Table 1) in the domain of chemistry collected in [21]. The variables of the formulas are defined as follows: If we consider two fingerprints of two chemical entities A and B, then:

- *a* is the count of bits set to 1 in entity A but not in entity B
- *b* is the count of bits set to 1 in entity B but not in entity A
- c is the count of the bits set to 1 in both entity A and entity B
- *d* is the count of the bits set to 0 in both entity A and entity B

Table 1: Reviewed similarity measures

Measure	Range	Formula
Cosine	[0, 1]	$\frac{c}{\sqrt{(a+c)*(b+c)}}$
Dice	[0, 1]	$\frac{2*c}{(a+c)*(b+c)}$
Euclidean	[0, 1]	$\sqrt{\frac{c+d}{a+b+c+d}}$
Forbes	[0, ∞]	$\frac{c*(a+b+c+d)}{(a+c)*(b+c)}$
Hamman	[-1, 1]	$\frac{(c+d)-(a+b)}{a+b+c+d}$
Jaccard / Tanimoto	[0, 1]	$\frac{c}{a+b+c}$
Kulczynski	[0, 1]	$0.5 * \left(\frac{c}{a+c}+\frac{c}{b+c}\right)$
Manhattan	[1, 0]	$\frac{a+b}{a+b+c+d}$
Matching	[0, 1]	$\frac{c+d}{a+b+c+d}$
Pearson	[-1, 1]	$\frac{(c*d)-(a*b)}{\sqrt{(a+c)*(b+c)*(a+d)*(b+d)}}$
Rogers-Tanimoto	[0, 1]	$\frac{c+d}{(a+b)+(a+b+c+d)}$
Russell-Rao	[0, 1]	$\frac{c}{a+b+c+d}$
Simpson	[0, 1]	$\frac{c}{\min\ ((a+c),(b+c))}$
Tversky	[0, 1]	$\frac{c}{\alpha*a+\beta*b+c}$
Yule	[-1, 1]	$\frac{(c*d)-(a*b)}{(c*d)+(a*b)}$

[4] http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html

5. EXPERIMENTS – IDENTIFING MEANINGFUL SIMILARITY MEASURES

In section 3, we demonstrated that it is inevitable to have a system with query relaxation to fulfill the information need of a chemist. The query relaxation engine should automatically determine similar entities and use them for query expansion. As shown in section 4, there are many different similarity measures available. The question that arises here is whether it is necessary to use all similarity measures in a retrieval system or if some of them are correlated meaning they deliver the same ranking.

5.1 Correlation Analysis

Since now, there is no work done in the literature, analyzing the correlation of the similarity measures applied on different fingerprints. Thus, our first goal was to explore if the underlying fingerprint has some influence on the similarity measures.

To do our first experiment, we took a random 1‰ sample of the PubChem database resulting in around 48.000 chemical entities. We downloaded their SDF files to have the structural information of all entities and converted them into their respective SMILES representations. These SMILES codes were necessary to generate the different fingerprint representations of each chemical entity using the CDK. In addition, we randomly choose 20 chemical entities as query entities. Since, in a later step, we want to use the similarity measures for a personalized retrieval system it seems reasonable to evaluate not only the complete result set of around 44000 entities but also smaller subsets. Thus, we decided to also evaluate the differences between the top-x results. Therefore, we computed for each combination of fingerprint, chemical entity and top-x the 16 fingerprint based similarity measures resulting in around 88 million similarity values.

As we can interpret the similarity value as a value in a ranking vector, we decided to use the Kendall rank correlation coefficient (KTau) [31] to determine the correlation of the different measures and fingerprints. We calculated the correlation coefficient for each ranking vector and the arithmetic mean over 20 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. Our experimental results have shown that the actual KTau values strongly differ over the fingerprints. For example the KTau value for the combination 'Euclidean / Russell-Rao / EState fingerprint' and 'Euclidean / Russell-Rao / Standard fingerprint' varies from 0.53 to -0.30 (see Table 2).

Due to the definition of the KTau, it is not straight forward to depict the uncorrelated similarity measures because *approximately zero* is not a well-defined threshold. To ensure a relatively high likelihood of correlation, we defined a threshold of 0.8. Based on this threshold, we evaluated how many uncorrelated similarity measures we have for each fingerprint. The results are shown in Figure 3. Interestingly, the EState fingerprint always has the minimum number of uncorrelated similarity measures.

Still, the concrete number differs from 5 to 3, which means that we have to take at least 3 different similarity measures (i.e. Yule, Russell-Rao and Forbes) into account. Given this result we notice that taking only the correlation coefficient into account is not discriminative enough; thus we consider additional discriminative properties.

Table 2: Similarity measures with highest variances over EState (1), Extended (2), Standard (3), Graphonly (4), MACCSS (5) and Substructure (6) fingerprint

Similarity Measure	1	2	3	4	5	6
Tanimoto / Euclidean	0,83	0,12	0,11	0,39	0,67	0,76
Cosine / Matching	0,82	0,05	0,04	0,40	0,67	0,76
Dice / Rogers Tanimoto	0,83	0,12	0,11	0,39	0,67	0,76
Euclidean / Russell-Rao	0,53	-0,29	-0,30	-0,09	0,38	0,33
Manhattan / Russell-Rao	-0,53	0,29	0,30	0,09	-0,38	-0,33
Tversky / Forbes	0,48	-0,11	-0,09	0,23	0,17	0,54
Forbes / Kulczynski	0,39	-0,40	-0,35	0,14	0,04	0,41
Hamman / Russell-Rao	0,53	-0,29	-0,30	-0,10	0,37	0,32
Jaccard / Rogers Tanimoto	0,83	0,12	0,11	0,39	0,67	0,76
Kulczynski / Euclidean	0,83	0,00	0,01	0,43	0,68	0,76
Matching / Russell-Rao	0,53	-0,29	-0,30	-0,09	0,38	0,33
Pearson / Russell-Rao	0,73	0,10	0,11	0,33	0,60	0,59
Rogers Tanimoto / Russell-Rao	0,53	-0,29	-0,30	-0,09	0,38	0,33
Russell-Rao / Rogers Tanimoto	0,53	-0,29	-0,30	-0,09	0,38	0,33
Simpson / Euclidean	0,66	-0,17	-0,11	0,32	0,48	0,55
Yule / Russell-Rao	0,67	0,01	0,02	0,19	0,50	0,49

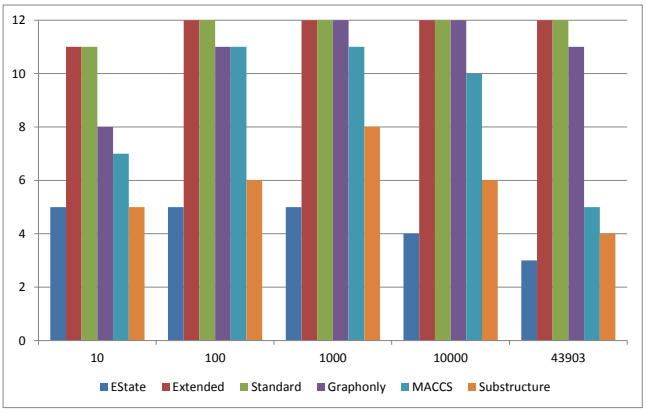


Figure 3 Number of minimal independent rankings for top-x and a threshold of 0.8

5.2 Task Based Analysis

This huge variety of uncorrelated similarity measures is eligible because chemical similarity differs according to the task a chemist is working on. Intuitively, we consider that each measure might be useful for a specific task and therefore conducted experiments with example tasks using synthesis and drug design.

For drug design we took, among others, *Sildenafil* as query entity. The idea is to retrieve alternative substances with similar chemical properties. Let us consider there are two scientists from the area of drug design Peter and Bob. Both are searching for *Sildenafil*, but with different additional conditions. Peter is interested in *pyrazolopyrimidinones* with a piperazine ring system connected to the sulfonyl group. In contrast to Sildenafil Peter is looking for a free N-side at the piperazine to examine further reactions at this position. A good hit for this query scenario is *Demethylsildenafil* (see Figure 4).

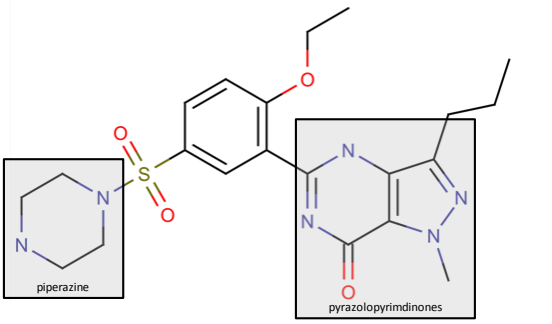


Figure 4: Demethylsildenafil

Bob is interested in *pyrazolopyrimidinones* with a secondary amine connected to the sulfonyl group as he is interested to perform alkylation reactions at his position. *Udenalfil* with its N-alkylated secondary amine side chain represents a top candidate for this kind of query (see Figure 5).

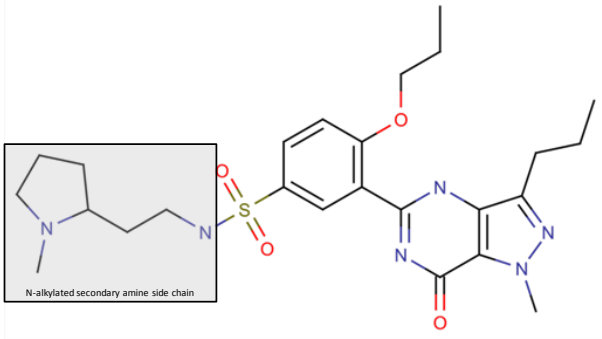


Figure 5: Udenafil

To evaluate the ranking results of the different similarity measures, we took all chemical entities that were retrieved by a similarity search in PubChem for the query term *Sildenafil*. We also assured that the entities of interest defined by the domain experts, *Demethylsildenafil* and *Udenafil*, are included in this set. We computed similarity values for *Sildenafil* and each entity in this set using all uncorrelated similarity measures. The domain experts analyzed all result sets and evaluated which similarity measure retrieves the best. The output of the experiment is that there is no suitable measure delivering both as relevant defined entities under the top-10. For Peter who expected *Demethylsildenafil* as relevant hit the combination of EState fingerprint and Tanimoto measure delivers the best results, ranking *Demethylsildenafil* on rank 9 and *Udenafil* on rank 335. For Bob expecting *Udenafil* as most relevant entity the combination of Substructure fingerprint and Tanimoto measure gives the best result, ranking *Udenafil* on rank 2 and *Demethylsildenafil* on rank 228. Although both chemists are from the field of drug design, they expect different ranking results for the same query term. Therefore, it is not possible to use one fixed similarity measure for one specific task. Of course, we also tried queries for the other tasks but with the same result: it is not possible to assign one similarity measure to a specific task.

To better judge the impact of the task, we interviewed a group of domain experts to find reasons for this behavior. We figured out that each individual chemist has some kind of special background knowledge or experiences that he implies, like e.g. costs for synthesis or which substances are already in the fund of the company. This background knowledge cannot be expressed by the query term resulting in insufficient result sets.

5.3 Feedback Analysis

The task based experiment has shown that there is a need for personalized retrieval systems. The idea is to build a system where each individual user trains the system and the system will learn the similarity measure which fits best to his needs. Consequently, we conducted a user study with domain experts from the area of drug design and synthesis, to discover if already a simple feedback step would result in an explicit combination of similarity measure and fingerprint. Furthermore, we are interested in the number of feedback cycles that are necessary until such a system is stable.

For the user study, we have randomly chosen 10 query entities from PubChem, each of them representing one feedback cycle inside the system. Based on the results shown in section 5.1 we used the 5 uncorrelated measures Russell-Rao, Yule, Forbes, Simpson and Manhattan for calculating the similarity values. In a first step, we retrieved the top-10 entities for each similarity measure and put them in one set which did not include duplicates and was unranked. In a second step, the chemists marked all relevant entities resulting in their personalized ranking vector. For each query we took the respective ranking vector and compared it to the top-10 vector of the uncorrelated similarity measures by computing precision at 10.

Example: As an illustrating example we take the results of the domain expert introduced in our use case scenario searching for *Sildenafil* (Figure 6). One can see that there are perfect candidates for the personalized similarity measure, i.e. a combination of the Extended fingerprint and the Yule, Forbes or Simpson measure.

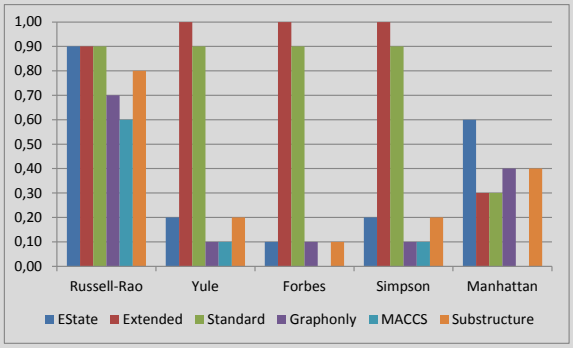


Figure 6: P@10 values for the query *Sindenafil*

However, of course one query is not enough to decide for a specific similarity measure. Figure 7 shows the average precision at 10 values for the chemist regarding 10 different queries. Regarding all queries the personalized similarity measure has slightly changed. Finally, the best matching similarity measure is Russell-Rao based on the Graphonly fingerprint. Only six feedback cycles were necessary to find this ideal combination for this chemist, meaning the preferred similarity measure did not change again after 6 queries. The second best measure is the combination of Yule and the Extended fingerprint.

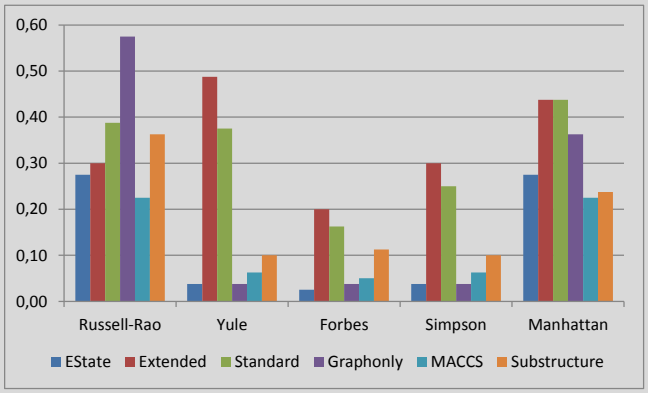


Figure 7: Average P@10-values for one chemist over all queries

The second question to evaluate was the number of needed feedback cycles until the system was stable for an individual user. For this purpose, we defined the system as stable, if a precision value did not change more than 2% over 3 queries. We can state, that for 75% of the domain experts, the system was able to determine an explicit combination of similarity measure and fingerprint within our ten feedback cycles. The particular number of needed feedback cycles varies between 3 and 8. For the remaining 25% we could not determine a combination after 10 feedback cycles.

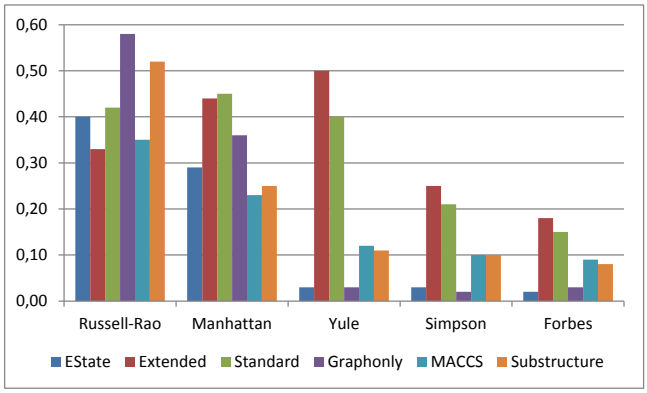


Figure 8: P@10 values for arithmetic mean over all experts and queries

Furthermore, we analyzed the arithmetic mean over all experts and queries (see Figure 8). One can see that the Russell-Rao measure outperforms all other measure applying it on the EState, Graphonly, MACCS and Substructure fingerprint. The best measure for the Extended fingerprint is Yule and for the Standard fingerprint it is Manhattan. Remember, these results cannot be applied out of the box to all users because the individual expectations can differ a lot. However, they are candidates for solving the well-known *new user problem*, if the user decides at least on a specific fingerprint or taking the overall best measure for a global starting point, i.e. the combination of Russell-Rao and the Graphonly fingerprint. In the next section we will describe how to integrate our findings into an architecture of a chemical search engine.

6. SYSTEM ARCHITECTURE WITH FEEDBACK COMPONENT

We now revisit the plight of our chemist posed in the use case scenario (section 3). His aim is to find relevant documents dealing with the chemical substance E_q. Since literature for *Sildenafil* covers a lot of different topics, he further restricts the query by entering the context he is interested in, namely the side effect of irregular heartbeat. Figure 9 shows our advanced architecture dealing with such kind of queries. In addition to the simple architecture, we add a query relaxation module to be able to relax the query part E_q with similar entities. The final result set only includes documents containing the context term q_i and the chemical substance E_q or q_i and at least one other similar entity for example $E_{q'}$. The document result set is ranked according to the similarity value of the included entities. As a result of the ranking function, the documents containing q_i and E_q are always top ranked followed by documents including q_i and the most similar entity $E_{q'}$.

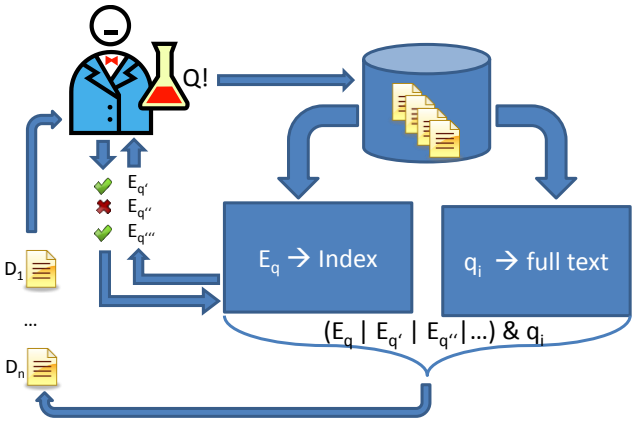


Figure 9: Advanced architecture

As described in the previous section, a lot of uncorrelated measures are available resulting in totally different rankings and it is not obvious which similarity measure / fingerprint combination is most applicable. For a new user, the system uses the *best* similarity measure by computing the arithmetic mean over all available user feedbacks (see section 5.3) and learns the best individual similarity measure in some feedback steps.

For each feedback step, the system is calculating the top-x results of all uncorrelated measures for a query. Out of this list, the user has to decide which chemical entities are relevant for him. In a next step, the system calculates the precision at 10 values for each measure and uses the best matching one. If the chosen measure does not change over a number of different queries it is accepted as default measure for this user and the feedback step is skipped for subsequent queries. Of course, if the user is not satisfied by the proposed ranking, he can force the system to learn or to use another measure.

7. CONCLUSION AND FUTURE WORK

Increasingly, the information gathering process relies upon the Web. The usage of keyword based search engines works fine in most cases, but for highly specialized domains a simple keyword based search is insufficient. In the chemistry domain users have additional requirements, in particular their information gathering needs focus on the search of chemical entities. Often, the desired search results highly depend on the background knowledge of a chemist and cannot be expressed in the query. To better assist the chemists in their information access needs, personalized retrieval systems using different similarity measures are essential. However, there is a huge variety of different measures available computing similarity between chemical entities. And not only do a huge variety of different similarity measures exist, different fingerprint representations of chemical entities are available, too. The result is lack of clarity in the usage of the measures due to the overwhelming possible combinations.

In this paper we presented a personalized retrieval system for chemical documents using a feedback engine for finding the best similarity measure for an individual chemist. In our experiments we took 16 widely used similarity measures for chemical entities and analyzed the correlation between them using Kendall's Tau. The results show that many of them are uncorrelated, meaning they deliver different rankings.

Chemistry is a wide field with many different subdomains. Therefore, chemists are focused on specific tasks when searching for literature, for example drug design or synthesis. We have analyzed whether the uncorrelated similarity measures, which are based on different fingerprint representations, fit to typical search tasks in chemistry. The different fingerprints represent different chemical aspects. For example, the Substructure fingerprint only considers the structure of a molecule, whereas the MACCS fingerprint uses a set of questions regarding more properties of a molecule than just the structure. We investigated if it is possible to assign one similarity measure to one specific task. We conducted a user study with domain experts and have shown that for the same task, e.g. drug design, different domain experts preferred different similarity measures. Hence, it is not possible to assign one similarity measure to one specific task, meaning there is no similarity measure always delivering the most suitable result set for that task. During discussions with domain experts we discovered that chemists usually have special background knowledge when searching for literature that cannot be expressed in the query, like e.g. costs for synthesis or which substances are already in the fund of the company.

These experiments have shown the need for personalized retrieval systems in chemistry. We have introduced one possible solution, including a feedback cycle analyzing which similarity measures retrieve the best results for each individual chemist. Our experiments have shown that we were able to stabilize the system for 75% of our participants within 10 feedback cycles. The impact of this work is that an exact assignment of a combination of similarity measure and fingerprint for the domain expert is possible. A valuable contribution of our work is in determining that chemists from the same subdomain, e.g. synthesis, have chosen different similarity measures. This observation sustains the need for personalized retrieval systems.

Since personalization is a fundamental aspect we will evaluate the usefulness of a relevance feedback system in our future work. The question is whether the retrieval quality is still increasable when learning a personalized similarity measure instead of using an established one.

8. REFERENCES

[1] P. Mitra, C. Giles, B. Sun, and Y. Liu, "Chemxseer: a digital library and data repository for chemical kinetics," *Proceedings of the International Conference on Information and Knowledge Management (CIKM)*, 2007, pp. 7-10.

[2] B. Sun, Q. Tan, P. Mitra, and C.L. Giles, "Extraction and search of chemical formulae in text documents on the web," *Proceeding of the 16h International Conference on World Wide Web (WWW)*, 2007, pp. 251-260.

[3] P. Corbett and P. Murray-Rust, "High-throughput identification of chemistry in life science texts," *Proceedings of the 2nd International Symposium on Computational Life Sciences*, Springer Berlin Heidelberg, 2006, pp. 107-118.

[4] S. Tönnies, B. Köhncke, O. Koepler, and W.-T. Balke, "Exposing the hidden web for chemical digital libraries," *Proceedings of the 10th joint Conference on Digital Libraries (JCDL)*, 2010, p. 235.

[5] R. Hoffmann and P. Laszlo, "Representation in Chemistry," *Angewandte Chemie International Edition in English*, vol. 30, 1991, pp. 1-16.

[6] J.R. McDaniel and J.R. Balmuth, "Kekule: OCR-optical chemical (structure) recognition," *Journal of Chemical Information and Modeling*, vol. 32, Jul. 1992, pp. 373-378.

[7] M. Zimmermann, L.T. Bui Thi, and M. Hofmann, "Combating Illiteracy in Chemistry: Towards Computer-Based Chemical Structure Reconstruction," *ERCIM News*, 2005, pp. 40-41.

[8] A.T. Valko and A.P. Johnson, "CLiDE Pro: the latest generation of CLiDE, a tool for optical chemical structure recognition.," *Journal of Chemical Information and Modeling*, vol. 49, Apr. 2009, pp. 780-7.

[9] I.V. Filippov and M.C. Nicklaus, "Optical Structure Recognition Software To Recover Chemical Information: OSRA, An Open Source Solution.," *Journal of Chemical Information and Modeling*, vol. 49, Mar. 2009, pp. 740-3.

[10] B. Sun, P. Mitra, and C.L. Giles, "Mining, indexing, and searching for textual chemical molecule information on the web," *Proceeding of the 17th International Conference on World Wide Web (WWW)*, 2008, pp. 735-744.

[11] J.A. Townsend, S.E. Adams, C.A. Waudby, V.K. de Souza, J.M. Goodman, and P. Murray-Rust, "Chemical documents: machine understanding and automated information extraction," *Journal of Organic & Biomolecular Chemistry*, vol. 2, 2004, p. 3294--3300.

[12] 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, 1965, pp. 107-113.

[13] D.J. Gluck, "A Chemical Structure Storage and Search System Developed at Du Pont.," *Journal of Chemical Documentation*, vol. 5, Feb. 1965, pp. 43-51.

[14] E.G. Smith and P.A. Baker, eds., *The Wiswesser Line-Formula Chemical Notation (WLN)*, Cherry Hill, N. J.: Chemical Information Management, 1976.

[15] D. Weininger, “SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules,” *Journal of Chemical Information and Modeling*, vol. 28, 1988, pp. 31-36.

[16] J. Barnard, C. Jochum, and S. Welford, “ROSDAL: A universal structure/substructure representation for PC-host communication,” *Chemical Structure Information Systems: Interfaces, Communication and Standards, ACS Symposium Series No. 400*, American Chemical Society, 1989, p. 76–81.

[17] S. Ash, M. a Cline, R.W. Homer, T. Hurst, and G.B. Smith, “SYBYL Line Notation (SLN): A Versatile Language for Chemical Structure Representation,” *Journal of Chemical Information and Modeling*, vol. 37, Jan. 1997, pp. 71-79.

[18] S.E. Stein, S.R. Heller, and D. Tchekhovskoi, “An Open Standard For Chemical Structure Representation: The IUPAC Chemical Identifier,” *Proceedings Of The International Chemical Information Conference*, Nimes: Infonortics, 2003, pp. 131-143.

[19] Z. Hubálek, “Coefficients of association and similarity, based on binary (presence-absence) data: An Evaluation,” *Journal of Biological Reviews*, vol. 57, Nov. 1982, pp. 669-689.

[20] P. Willett, J.M. Barnard, and G.M. Downs, “Chemical Similarity Searching,” *Journal of Chemical Information and Modeling*, vol. 38, Nov. 1998, pp. 983-996.

[21] J. Holliday, C. Hu, and P. Willett, “Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D fragment bit-strings,” *Journal of Combinatorial Chemistry; High Throughput Screening*, vol. 5, 2002, p. 155–166.

[22] R.M. Cormack, “A Review of Classification,” *Journal of the Royal Statistical Society. Series A (General)*, vol. 134, 1971, pp. 321-367.

[23] L.A. Goodman and W.H. Kruskal, “Measures of Association for Cross Classifications,” *Journal of the American Statistical Association*, vol. 49, 1954, pp. 732-764.

[24] L.A. Goodman and W.H. Kruskal, “Measures of Association for Cross Classifications. II: Further Discussion and References,” *Journal of the American Statistical Association*, vol. 54, 1959, pp. 123-163.

[25] L.A. Goodman and W.H. Kruskal, “Measures of Association for Cross Classifications III: Approximate Sampling Theory,” *Journal of the American Statistical Association*, vol. 58, 1963, pp. 310-364.

[26] P. Willett, “Similarity-based approaches to virtual screening,” *Journal of Biochemical Society Transactions*, vol. 31, Jun. 2003, pp. 603-606.

[27] E. Anderson, G.D. Veith, and D. Weininger, *SMILES, a Line Notation and Computerized Interpreter for Chemical Structures*, US Environmental Protection Agency, Environmental Research Laboratory, 1987.

[28] C. Steinbeck, C. Hoppe, S. Kuhn, M. Floris, R. Guha, and E.L. Willighagen, “Recent Developments of the Chemistry Development Kit (CDK) - An Open-Source Java Library for Chemo- and Bioinformatics,” *Journal of Current Pharmaceutical Design*, vol. 12, Jun. 2006, pp. 2111-2120.

[29] L.H. Hall and L.B. Kier, “Electrotopological state indices for atom types: A novel combination of electronic, topological, and valence state information,” *Journal of Chemical Information and Computer Sciences*, vol. 35, 1995, p. 1039–1045.

[30] J.L. Durant, B.A. Leland, D.R. Henry, and J.G. Nourse, “Reoptimization of MDL Keys for Use in Drug Discovery,” *Journal of Chemical Information and Modeling*, vol. 42, Nov. 2002, pp. 1273-1280.

[31] M.G. Kendall, “A New Measure of Rank Correlation,” *Journal of Biometrika*, vol. 30, 1938, pp. 81-93.