An Evaluation of Text Retrieval Methods for Similarity Search of multi-dimensional NMR-Spectra

Alexander Hinneburg[1], **Andrea Porzel**[2], **Karina Wolfram**[2]
[1] Institute of Computer Science
Martin-Luther-University of Halle-Wittenberg, Germany
hinneburg@informatik.uni-halle.de
[2] Leibniz Institute of Plant Biochemistry (IPB), Germany
{aporzel,kwolfram}@ipb-halle.de

Abstract

Searching and mining nuclear magnetic resonance (NMR)-spectra of naturally occurring substances is an important task to investigate new potentially useful chemical compounds. Multi-dimensional NMR-spectra are relational objects like documents, but consists of continuous multi-dimensional points called peaks instead of words. We develop several mappings from continuous NMR-spectra to discrete text-like data. With the help of those mappings any text retrieval method can be applied. We evaluate the performance of two retrieval methods, namely the standard vector space model and probabilistic latent semantic indexing (PLSI). PLSI learns hidden topics in the data, which is in case of 2D-NMR data interesting in its owns rights. Additionally, we develop and evaluate a simple direct similarity function, which can detect duplicates of NMR-spectra. Our experiments show that the vector space model as well as PLSI, which are both designed for text data created by humans, can effectively handle the mapped NMR-data originating from natural products. Additionally, PLSI is able to find meaningful "topics" in the NMR-data.

1 Introduction

Nuclear magnetic resonance (NMR)-spectra are an important finger printing method to investigate the chemical structure of organic compounds from plants or other tissues. Two-dimensional-NMR spectroscopy is able to capture the influences of two different atom types at the same time (e.g. ^{1}H, hydrogen and ^{13}C carbon). The result of an 2D-NMR experiment can be seen as an intensity function measured over two variables[1]. Regions of high intensity are called peaks, which contain the real information about the underlying molecular structure. The usual visualizations of 2D-NMR spectra are contour plots as shown in figure 1. An ideal peak would register as a small dot, however, due to the limited resolution available (dependent on the strength of the magnetic field) multiple peaks may appear as a single merged object with non-convex shape. In the literature peaks are noted by their two-dimensional positions without any information about the shapes of the peaks. Content-based similarity search of 2D-NMR spectra would be a valuable tool for structure investigation by comparing spectra of unknown compounds with a set of spectra, for which the structures are known. While the principle is already in use for 1D-NMR spectra [7; 1; 11; 6; 2], to the best of our knowledge, no effective similarity search method is known for 2D-NMR-spectra.

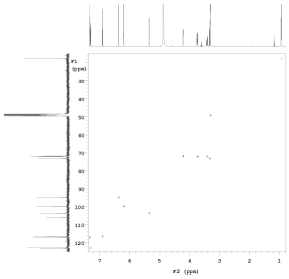

Figure 1: 2D-NMR spectrum of quercetrin. The plots at the axes are the corresponding 1D-NMR spectra.

Simplified, a 2D-NMR spectrum is a set of two-dimensional points. There is an analogy to text retrieval, where documents are usually represented as sets of words. Latent space models [5; 9; 3] were successfully used to model documents and thus improved the quality of text retrieval. Recently, a diversity of text mining approaches for different problems [4; 12; 8] have been proposed, which make use of probabilistic latent space models. The goal of this work is to show by example how to apply text retrieval and mining methods to biological data originating from experiments.

The contribution of this paper are methods to map 2D-NMR spectra to discrete text-like data, which can be analyzed and searched by any text retrieval method. We evaluate on real data the performance of two text retrieval methods, namely the standard vector space model [10] and PLSI [5] in combination our mapping methods for 2D-NMR spectra. Additionally, we propose a simple similarity function, which operates directly on the peaks of the spectra and serves as bottom line benchmark in the experimental evaluation. Our results indicate at a larger scope that text retrieval and mining methods, designed for text data created by humans, in combination with appropriate mapping functions may yield the potential to be also successful for experimental data from naturally occurring objects. In this paper we consider exemplarily ^{1}H, ^{13}C one-bond heteronuclear shift correlation 2D-NMR spectra.

The paper is structured as follows: first, in section 2, we introduce briefly the used text modeling methods while in section 3, we propose the mapping functions for 2D-NMR spectra. In section 4, we propose a simple similarity function as bottom line benchmark and define fuzzy duplicates.

[1] The measurements are in parts per million (ppm).

In section 5, we describe our experimental evaluation and section 6 concludes the paper.

2 Models for Text Retrieval

Like a 2D-NMR spectrum consists of a set of peaks, a document consists of many words, which typically are modeled as a set. So assuming a 2D-NMR spectrum can be transformed into a text-like object by mapping the continuous 2D peaks to discrete variables, a variety of text retrieval models can be applied. However, it is an open question, whether models designed for quite different data, namely texts created by humans, are effective on data which comes for naturally occurring compounds and thus do not include human design patterns.

We briefly introduce the essentials of the vector space model and PLSI to make the paper self contained. In the vector space model, documents are represented by vectors which have as many dimensions as there are words in the used vocabulary of the document collection. Each component of a documents vector reflects the importance of the corresponding word for the document. The typical quantity used is the raw term frequency (tf) of that word for the document, say the number of occurrences of that word in a document d. In order to improve the retrieval quality, those vectors are reweighed by multiply tf with the inverse document frequency (ifd) of a word. The inverse document frequency measures is large, if a word is included in only a small percentage of the documents in the collection. Formally, we denote the set of documents by $D = \{d_1, ..., d_J\}$ and the vocabulary by $W = \{w_1, ..., w_I\}$. The term frequency of a word $w \in W$ in a document $d \in D$ is denoted as $n(d, w)$ and the reweighed quantity is $\hat{n}(d, w) = n(d, w) \cdot idf(w)$. The similarity between a query document q and a document d from the collection is

$$S(d, q) = \frac{\sum_{w \in W} \hat{n}(d, w) \cdot \hat{n}(q, w)}{\sqrt{\sum_{w \in W} \hat{n}(d, w)^2} \cdot \sqrt{\sum_{w \in W} \hat{n}(q, w)^2}}$$

This can be interpreted as the cosine of the angles between the two vectors.

Probabilistic latent semantic indexing (PLSI) introduced in [5] extends the vector space model by learning topics hidden in the data. The training data consists of a set of document-word pairs $(d^{(i)}, w^{(i)})_{i=1,...,N}$ with $w^{(i)} \in W$ and $d^{(i)} \in D$. The joint probability of such a pair is modeled according to the employed aspect model as $P(d, w) = \sum_{z \in Z} P(z) \cdot P(w|z) \cdot P(d|z)$. The z are hidden variables, which can take K different discrete values $z \in Z = \{z_1, ..., z_K\}$. In the context of text retrieval z is interpreted as an indicator for a topic. Two assumptions are made by the aspect model. First, it assumes pairs (d, w) to be statistically independent. Second, conditional independence between w and d is assumed for a given value for z.

The probabilities necessary for the joint probability $P(d, w)$, namely $P(z)$, $P(w|z)$ and $P(d|z)$, are derived by an expectation maximization (EM) learning procedure. The idea is to find values for unknown probabilities, which maximize the complete data likelihood

$$\begin{aligned} P(S, z) &= \prod_{(d^{(i)}, w^{(i)}) \in S} [P(z) \cdot P(w^{(i)}|z) \cdot P(d^{(i)}|z)] \\ &= \prod_{d \in D} \prod_{w \in W} [P(z) \cdot P(w|z) \cdot P(d|z)]^{n(d,w)} \end{aligned}$$

with $S = \{(d^{i)}, w^{i)})_{i=1,...,N}\}$ is the set of all document-word pairs in the training set. In the E-step the posteriors for z are computed.

$$P(z|d, w) = \frac{P(z) \cdot P(w|z) \cdot P(d|z)}{\sum_{z' \in Z} P(z') \cdot P(w|z') \cdot P(d|z')}$$

The subsequent M-step maximizes the expectation of the complete data likelihood respectively to the model parameters, namely $P(z)$, $P(w|z)$ and $P(d|z)$.

$$\begin{aligned} P(d|z) &= \frac{\sum_{w \in W} P(z|d, w) \cdot n(d, w)}{\sum_{w \in W} \sum_{d' \in D} P(z|d', w) \cdot n(d', w)} \\ P(w|z) &= \frac{\sum_{d \in D} P(z|d, w) \cdot n(d, w)}{\sum_{w' \in W} \sum_{d \in D} P(z|d, w') \cdot n(d, w')} \\ P(z) &= \frac{\sum_{w \in W} \sum_{d \in D} P(z|d, w) \cdot n(d, w)}{\sum_{w \in W} \sum_{d \in D} n(d, w)} \end{aligned}$$

The EM algorithm starts with random values for the model parameters and converges by alternating E- and M-step to a local maximum of the likelihood.

There are serval ways possible to answer similarity queries using the trained aspect model. Because of its simplicity, we adopt the PLSI-U variant from [5]. The idea is to extend the cosine similarity measure from the tf-idf vector space model. The extension by Hofmann treats the learned multinomials $P(w|d)$ as term frequencies (tf). Note that $P(w|d) = P(d, w)/P(d)$ with $P(d) = \sum_{w' in W} n(d, w')/N$. The multinomials $P(w|d)$ are smoothen variants of the original term frequencies $\tilde{P}(w|d) = n(d, w)/(\sum_{w' \in W} n(d, w'))$. The proposed tf-weights are linear combinations of the multinomials $P(w|d)$ and $\tilde{P}(w|d)$. Thus, the new tf-idf weights used for the documents within the similarity calculation are

$$\hat{n}(d, w) = (\lambda \cdot P(w|d) + (1 - \lambda) \cdot \tilde{P}(w|d)) \cdot idf(w)$$

with $\lambda \in [0, 1]$. Hofmann suggests in [5] to set $\lambda = 0.5$. The tf-idf weights for the query are determined as in the standard vector space model. The smoothen tf-weight for a word which actually does not appear in the document may be still non-zero if the word belongs to a topic which is covered by the particular document. In that way a more abstract similarity search becomes possible.

For 2D-NMR spectra similarity search it is not clear, what is the best way to map the peaks of a spectrum to discrete words. We develop methods for this task in the next section. That will enable us to tackle the question, whether methods like the vector space model or PLSI, which is designed for text data, remains effective for experimental data from natural products.

3 Mapping of NMR Spectra

In this section we propose different methods to map the peaks of an NMR-spectrum from the continuous space of measurements to a discrete space of words. With the help of such a mapping, methods for text retrieval like PLSI can be directly applied. However, the quality of the similarity search depend on how the peaks are mapped to discrete words. A preliminary study of the proposed mappings appeared as poster in [13].

First we give a formal definition of 2D-NMR spectra. A two-dimensional NMR-spectrum of an organic compound captures many structural characteristics like rings and chains. Most important are the positions of the peaks.

As the shape of a peak and its height (intensity) strongly varies over different experiments with the same compound, the representation of a spectrum includes the peak positions only.

Definition 1 *A **2D NMR-spectrum** A is defined as a set of points $\{x_1, \ldots, x_n\} \subset \mathbb{R}^2$. The $|\cdot|$ function denotes the size of the spectrum $|A| = n$.*

The size of a spectrum is typically between 4 and 30 for small molecules found in naturally occurring products.

3.1 Grid-based Mapping

We introduce a simple grid-based method, on which we will build more sophisticated methods. A simple grid-based method is to partition each of the both axis of the two-dimensional peak space into intervals of same size. Thus, an equidistant grid is induced in the two-dimensional peak space and a peak is mapped to exactly one grid cell it belongs to. When a grid cell is identified by a discrete integer vector consisting of the cells coordinates the mapping of a peak $x \in \mathbb{R}^2$ is formalized as

$$g(x) = (g_c(x.c), g_h(x.h)) \text{ with } g_c(x.c) = \left\lfloor \frac{x.c}{w_c} \right\rfloor,\ g_h(x.h) = \left\lfloor \frac{x.h}{w_h} \right\rfloor$$

The quantities w_c and w_h are the extensions of a cell in the respective dimensions, which are parameters of the mapping. The grid is centered at the origin of the peak space. The cells of the grid act as words. The vocabulary generated by the mapped peaks consists of those grid cells which contain at least one peak. Empty grid cells are not included in the vocabulary. A word consists of a two-dimensional discrete integer vector.

Unfortunately the grid-based mapping has two disadvantages. First, close peaks may be mapped to different grid cells. This may lead to poor matching of related peaks in the discrete word space. Second, peaks of new query spectra are ignored when they are mapped to grid cells not included in the vocabulary. So some information from the query is not used for the similarity search which may weaken the performance.

3.2 Redundant Mappings

We propose three mappings which introduce certain redundancies by mapping a single peak to a set of grid cells. The redundancy in the new mappings shall compensate for the drawbacks of the simple grid-based mapping.

Shifted Grids

The first disadvantage of the simple grid-based method is that peaks which are very close in the peak space may be mapped to different grid cells, because a cell border is between them. So proximity of peaks does not guaranty that they are mapped to the same discrete cell.

Instead of mapping a peak to a single grid cell, we propose to map it to a set of overlapping grid cells. This is achieved by several shifted grids of the same granularity. In addition to the base grid some grids are shifted into the three directions $(1,0)(0,1)(1,1)$. An illustration of the idea is sketched in figure 2. In figure 2, one grid is shifted in each of the directions by half of the extent of a cell. In general, there may be $k-1$ grids shifted by fractions of $1/k, 2/k, \ldots, k-1/k$ of the extent of a cell in each direction respectively. For the mapping of the peaks to words which consist of cells from the different grids, two additional dimensions are needed to distinguish (a) the $k-1$ grids in each direction and (b) the directions themselves.

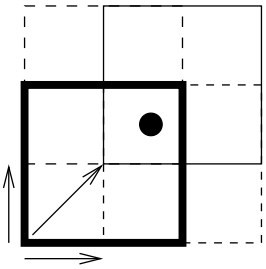

Figure 2: The four grids are marked as follows: base grid is bold, $(1,0)$, $(0,1)$ are dashed and $(1,1)$ is normal.

The third coordinate represents the fraction by which a cell is shifted and the fourth one represents the directions by the following coding: value 0 is (0,0), 1 is (1,0), 2 is (0,1) and 3 is (1,1). So each peak is mapped to a finite set of four-dimensional integer vectors. The mapping of a peak $x \in \mathbb{R}^2$ is

$$\begin{aligned} s(x) \quad = \quad & \{(g_c(x.c), g_h(x.h), 0, 0)\} \cup \\ & \bigcup_{i=1}^{k-1} \Big\{ (g_c(x.c + {}^{i}/_{k} \cdot w_c), g_h(x.h), i, 1), \\ & \quad (g_c(x.c), g_h(x.h + {}^{i}/_{k} \cdot w_h), i, 2), \\ & \quad (g_c(x.c + {}^{i}/_{k} \cdot w_c), g_h(x.h + {}^{i}/_{k} \cdot w_h), i, 3) \Big\} \end{aligned}$$

Thus, a single peak is mapped to $3(k-1)+1$ words. A nice property of the mapping is that there exists at least one grid cell for every pair of matching peaks both peaks are mapped to.

Different Resolutions

The second disadvantage of the simple grid-based mapping comes from the fact that empty grid cells (not occupied by at least one peak from the set of training spectra) do not contribute to the representation to be learned for similarity search. So peaks of new query spectra mapped to those empty cells are ignored. That effect can be diminished by making the grid cells larger. However, this is counterproductive for the precision of the similarity search due to the coarser resolution. Thus, there are two contradicting goals, namely (a) to have a fine resolution to handle subtle aspects in the data and (b) to cover at the same time the whole peak space by a coarse resolution grid so that no peaks of a new query spectrum have to be ignored.

Instead of finding a tradeoff for a single grid, both goals can be served by combining simple grids with different resolutions. Given l different resolutions $\{(w_c^{(1)}, w_h^{(1)}), \ldots, (w_c^{(l)}, w_h^{(l)})\}$ a peak is mapped to l grid cells of different sizes. In order to distinguish between the different grids an additional discrete dimension is needed. So the mapping function is

$$r(x) = \bigcup_{i=1}^{l} \{(g_c^{(i)}(x), g_h^{(i)}(x), i)\}$$

with $g_c^{(i)}$ and $g_h^{(i)}$ use $w_c^{(i)}$ and $w_h^{(i)}$ respectively. Note that a hierarchical, quad-tree like partitioning is a special case of the proposed mapping function with $w_c^{(i)} = 2^{i-1} w_c$ and $w_h^{(i)} = 2^{i-1} w_h$.

Combining shifted Grids with different Resolutions

Both methods are designed to compensate for different drawbacks of the simple grid mapping. So it is nat-

ural to combine both mappings. The parameters of such a mapping are the number of shifts k, the number of different grid cell sizes l and the actual sizes $\{(w_c^{(1)}, w_h^{(1)}), \ldots, (w_c^{(l)}, w_h^{(l)})\}$. Beside the two coordinates for the grid cells, additional discrete dimensions are needed for the shift, the direction and the grid resolution. Using the the definitions from above the mapping function of the combined mapping of a peak is

$$
\begin{aligned}
c(x) = \quad & \bigcup_{i=1}^{l} \left\{ \left(g_c^{(i)}(x.c), g_h^{(i)}(x.h), 0, 0, i\right) \right\} \cup \\
& \bigcup_{j=1}^{k-1} \Big\{ \left(g_c^{(i)}(x.c + {}^{j}/_{k} \cdot w_c^{(i)}), g_h^{(i)}(x.h), j, 1, i\right), \\
& \quad \left(g_c^{(i)}(x.c), g_h^{(i)}(x.h + {}^{j}/_{k} \cdot w_h^{(i)}), j, 2, i\right), \\
& \quad \left(g_c^{(i)}(x.c + {}^{j}/_{k} \cdot w_c^{(i)}\right), \\
& \quad g_h^{(i)}(x.h + {}^{j}/_{k} \cdot w_h^{(i)}), j, 3, i) \Big\}
\end{aligned}
$$

Thus a single peak is mapped to $l(3(k-1)+1)$ words. In the next section all mappings are compared with respect to the effectiveness for similarity search.

4 Directly Computing Similarity

In this section, we introduce a method to directly compute similarity between pairs of spectra. This method will be used in the experiments as a bottom line benchmark. We also propose on the basis of direct similarity a definition of fuzzy duplicates.

As a peak in a spectrum has two numeric attributes, which can vary continuously, we formalize the notion of matching peaks. A simple but effective approach is to require that a peak matches other peaks only within a certain spatial neighborhood. The neighborhood is defined by the ranges α and β.

Definition 2 *A peak x from spectrum A **matches** a peak y from spectrum B, iff $|x.c - y.c| < \alpha$ and $|x.h - y.h| < \beta$, where $.c$ and $.h$ denote the NMR measurements for carbon and hydrogen respectively.*

Note that a single peak of a spectrum can match several peaks from another spectrum. Given two spectra A and B, the subset of peaks from A which find matching partners in B is denoted as $matches(A, B) = \{x \colon x \in A, \exists y \in B \colon x \text{ matches } y\}$. The function $matches$ is not symmetric, but helps to define a symmetric similarity measure

Definition 3 *Let be A and B two given spectra and $A' = matches(A, B)$ and $B' = matches(B, A)$, so the **similarity** is defined as*

$$sim(A, B) = \frac{|A'| + |B'|}{|A| + |B|}$$

The measure is close to one if most peaks of both spectra are matching peaks. Otherwise the similarity drops towards zero.

An important application of similarity search is the detection of duplicates to increase the data quality of a collection of 2D-NMR-spectra. Clearly a naive definition of duplicates does not work, like two duplicate spectra A and B need to have the same size and the peaks at the same positions. The reason is that the spectra are measured experimentally and so the peak positions differ even if the same probe is analyzed twice. So flexibility should be allowed for the peak positions. Another problem appears when two spectra of the same substance are measured with different resolutions. In case a spectrum is measured with low resolution it may happen that neighboring peaks are merged to a single one. A restriction to an one-to-one relationship between matching peaks can not handle such cases.

We propose a definition of fuzzy duplicates based on the direct similarity measure, which can deal with both of the mentioned problems.

Definition 4 *A pair of 2D-NMR-spectra A and B are **fuzzy duplicates**, iff $sim(A, B) = 1$.*

By that definition it is only required that every peak of a spectrum finds at least one matching peak in the other spectrum.

Group	#Spectra	#Peaks
Pregnans	11	17–26
Anthrquinones	8	3–6
Aconitanes	8	22–26
Triterpenes	17	24–31
Flavonoids	18	5–8
Isoflavonoids	16	5–7
Aflatoxins	8	8–10
Steroids	12	16–23
Cardenolides	15	18–25
Coumarins	19	3–8

Table 1: Groups with number of spectra and range of peaks

5 Evaluation and Results

In this section we present the results for duplicate detection, a comparison of the effectiveness of the mappings for similarity search, and mining aspects of 2D-NMR-data.

5.1 2D-NMR-Data

The substances included in the database are mostly secondary metabolites of plants and fungi. They cover a representative area of naturally occurring compounds and originate either from experiments or from simulations[2] based on the known structure of the compound. The database includes about 587 spectra, each has about 3 to 35 peaks. The total number of peaks is 7029. Ten small groups of chemically similar compounds are included in the database for controlled experiments. The groups with the number of spectra and number of peaks are listed in table 1 left. The peak space with all peaks in the database is shown in figure 3 right. Two groups, steroids and flavonoids, are selected as examples and shown with their peak distribution within figure 3 right.

Natural steroids occur in animals, plants and fungi. They are vitamins, hormones or cardioactive poisons like digitalis or oleander. The steroids in the database ar mostly hormones like androgens and estrogens. Flavonoids are aromatic substances (rings). Some flavonoids decrease vascular permeability or possess antioxidant activity which can have an anticarcinogenic effect.

5.2 Detection of Duplicates

We used the direct similarity function introduced in section 4 to detect duplicates in the database. With a setting

[2] ACD/2D NMR predictor, version 7.08, http://www.acdlabs.com/

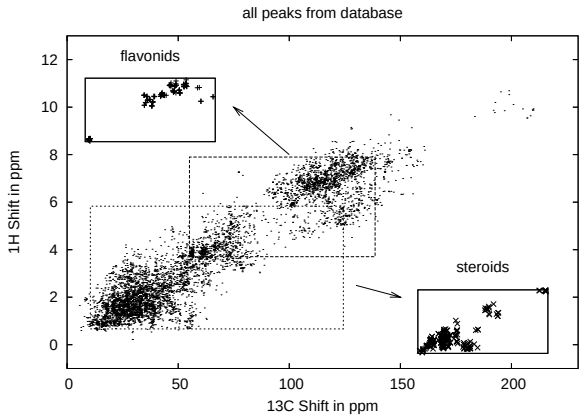

Figure 3: Distribution of the peaks of all spectra with the distribution within the groups of flavonoids and steroids.

of $a = 3$ppm and $b = 0.3$ppm, which are reasonable tolerances, 54 of 171991 possible pairs are reported as fuzzy duplicates. An inspection by hand revealed that 30 pairs are just very similar spectra, but 24 are candidates for real duplicates. Many of the found pairs come from the groups shown in table 1. Some pairs consist of an experimental and a simulated spectrum of the same substance which confirms the usefulness of the definition. There was also a surprise, namely the pair Thalictrifoline/Cavidine. Both structures differ only in the stereochemical orientation of one methyl group. Evidently, in this case the commercial software package used for the simulation is not able to reflect the different stereochemistry in calculated spectra. In the future, fuzzy duplicates will be used to improve the quality of collections of 2D-NMR spectra.

5.3 Performance Evaluation

The different methods for similarity search of 2D-NMR-spectra are compared using recall-precision curves. The search quality is high, when both – recall and precision – are high. So the upper curves are the best.

First, a series of experiments is conducted using our proposed mapping functions in combination with the vector space model. Each spectrum from the ten groups is used as a query while the rest of the respective group should be found as answers. The plots in figure 4 and 5 show averages over all queries. The results for the simple grid-based mapping are shown in figure 4a. The sizes of the grid cells are varied over $w_c = 4, 6, 8, 10$ and $w_h = 0.4, 0.6, 0.8, 1.0$ respectively. Small sizes give the best results.

The use of shifted grids improves the performance substantially over simple grids, as shown in figure 4b,c. The plots show the experiments for $k = 2, 3$. The results for $k = 2$ and $k = 3$ are almost identical. However, the vocabulary for $k = 2$ is much smaller. In practise, the smaller model with $k = 2$ shifts is favored.

Also the mapping based on grids with different grid cell sizes are assessed. Due to lack of space, only the results from combinations of $w_c^{(1)} = 4, w_h^{(1)} = 0.4$ with other sizes are reported, because those performed best among all combinations. Figure 4d shows that also the mapping based on different grid cell sizes outperforms the simple grid-based mapping. But the improvement is not as much as for shifted grids. The set of resolutions $\{(w_c^{(1)} = 4, w_h^{(1)} = 0.4), (w_c^{(2)} = 10, w_h^{(2)} = 1.0)\}$ performs best.

Also, experiments are performed with the combination of the previous two mappings, namely a combination of shifted grids with those of different resolutions. The performance results are shown in figure 4e which indicates that the best combination, namely the resolution set $\{(w_c^{(1)} = 4, w_h^{(1)} = 0.4), (w_c^{(2)} = 10, w_h^{(2)} = 1.0)\}$ with $k = 2$ shifts, outperforms both previous mappings. This is more clearly seen in figure 4f which compares the best performing settings from the above experiments.

Next, a series of similar experiments is conducted using our proposed mapping functions in combination with PLSI. Random initialization is used for the EM training algorithm described in section 2. All curves are averages from cross validation over all groups. As PLSI is trained on the data beforehand, we used cross validation where the current query is not included in the taining data. As the groups are very small, the leave-one-out cross validation scheme is employed. The results for PLSI are shown in figure 5a-f. PLSI requires to chose the number of hidden aspects. For the experiments reported so far, the PLSI model is used with 20 hidden aspects. Also different numbers of aspects are tested using the best combination of mappings. Figure 5g shows that the performance with 10 aspects drops a bit The increase in the numbers of aspects from 20 to 32 is only marginally reflected in increase of search performance. So 20 is a reasonable number of aspects for the given data.

In summery, the experiments with both text retrieval methods show, that the mappings based on shifted grids and those with different resolutions perform significantly better than the simple grid-based mapping. In both cases, the combination of shifted grids and grids with different resolutions is even better than the individual mappings. The comparison between PLSI and the vector space model (figure 5h) shows that both have similar performance for small recall but for large recall PLSI has a better precision.

Last, the direct similarity function is tested (figure 5i). The size of the matching neighborhood is varied over $\alpha = 4, 6, 8, 10$ and $\beta = 0.4, 0.6, 0.8, 1.0$ respectively. The search quality is quite low. In fact on average, it fails to deliver a spectrum from the answer set in the top ranks which is indicated by the hill-like shape of the curves.

In conclusion, the results prove experimentally that the vector space model as well as the PLSI model, which are designed for text retrieval, are indeed effective for similarity search of 2D-NMR spectra from naturally occurring products.

5.4 Analysis of the latent Aspects

We analyzed the latent aspects learned by the PLSI model using the mapping based on the combination of shifted grids with different resolutions. The grid cells (words) with high probability for a given aspect are plotted together to describe the aspects meaning. Some aspects specialized on certain regions in the peak space which are typical for distinct molecule fragments like aromatic rings or alkane skeletons. However, also more subtle details of the data are captured by the aspect model. For example, the main aspect for the group of flavonoids specializes not only on the region for aromatic rings which are the main part of flavonoids. It also includes a smaller region which indicates oxygen substitution. A closer inspection of the database revealed that indeed many of the included flavonoids do have several oxygen substituents. The main aspect for flavonoids with the respective peak distribution of the flavonoid group

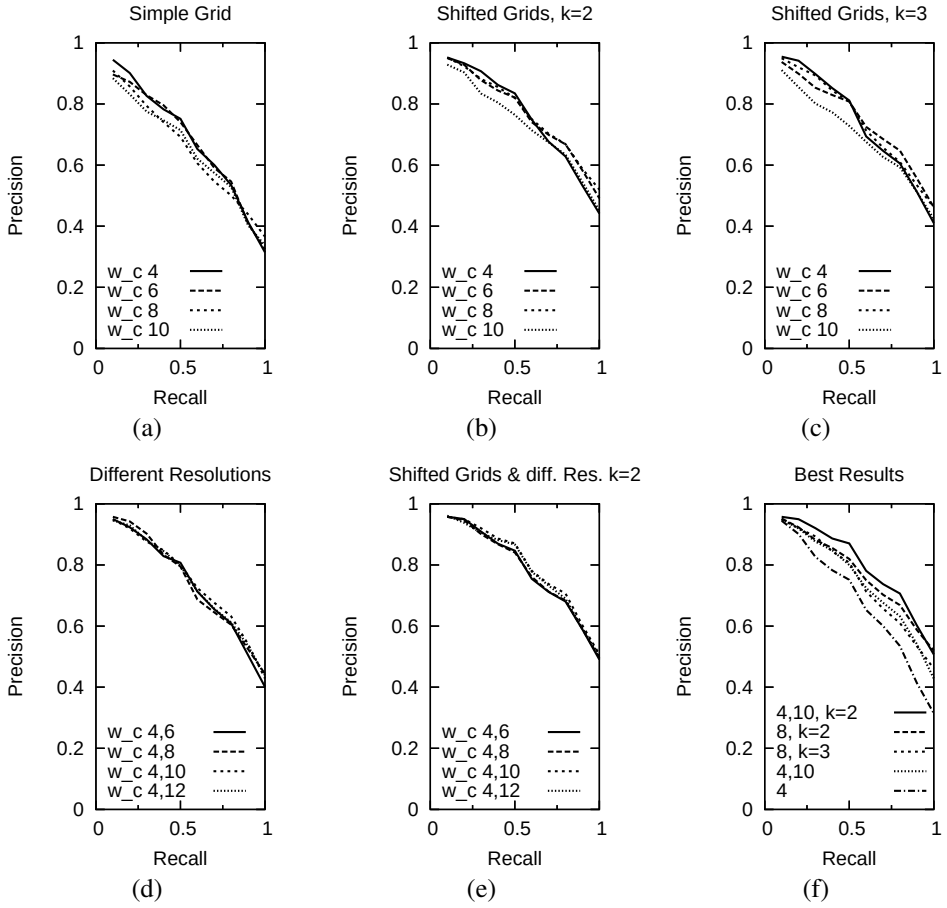

Figure 4: Average recall-precision curves using the vector space model

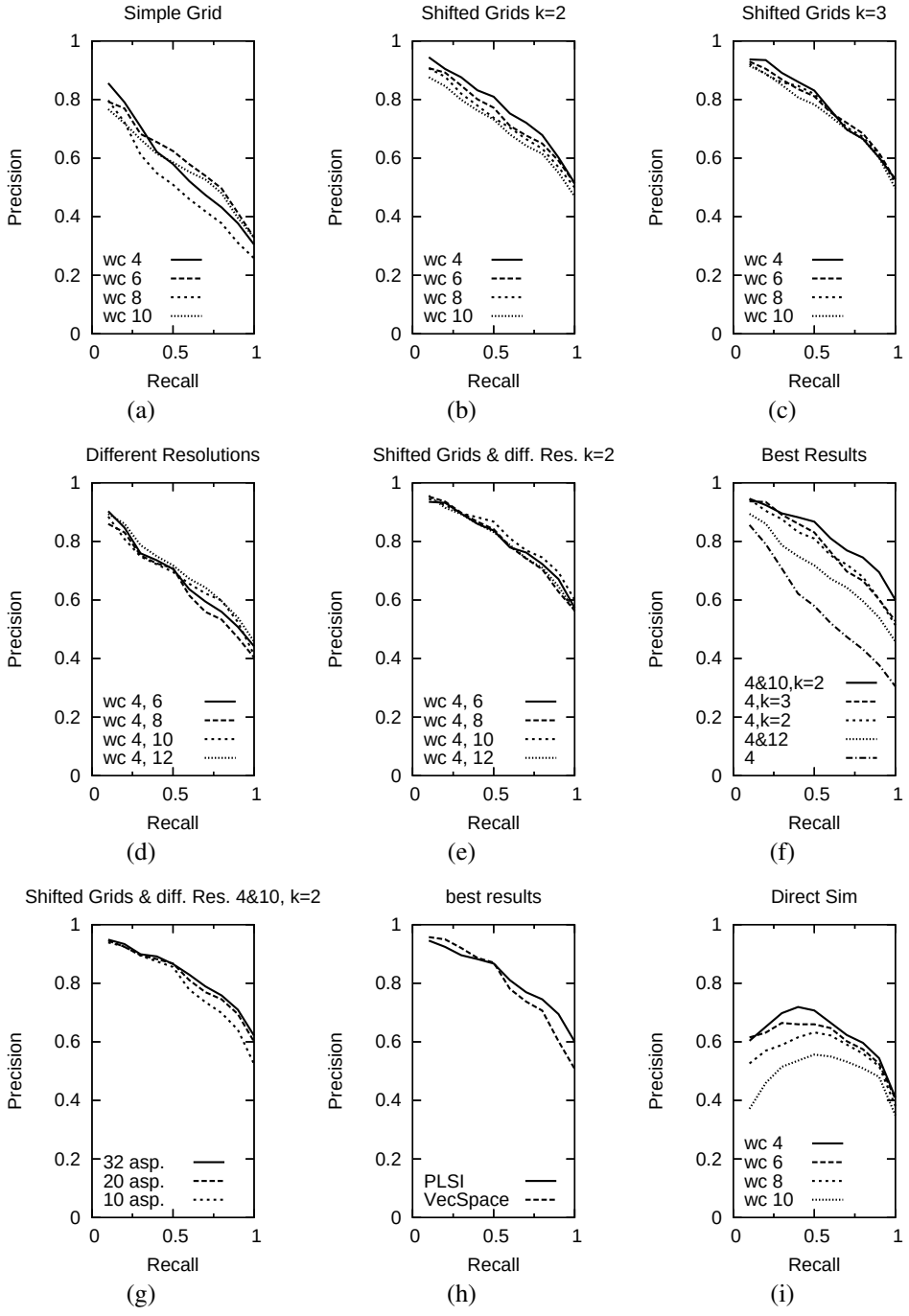

Figure 5: Average recall-precision curves from leave-one-out cross validation experiments with the PLSI model (a-g), best results of PLSI and vector space model (h) and results for the direct similarity (i).

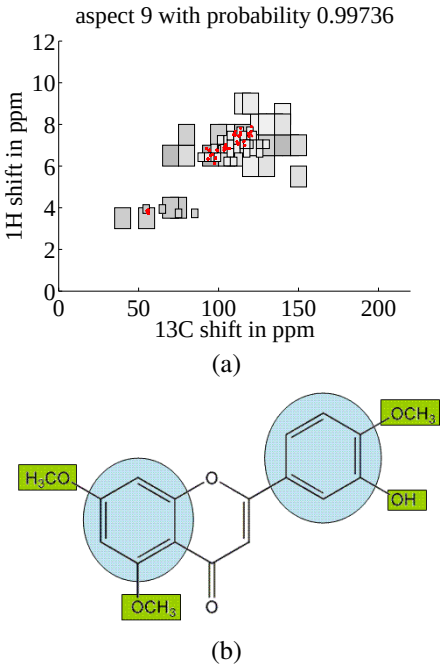

Figure 6: (a) Main aspect of the flavonoid group which includes the region of aromatic rings (upper right cluster) and the region for oxygen substituents (lower left cluster). The gray shades indicate the strength of the association between grid cell and aspect. (b) An example of an flavonoid (3'-Hydroxy-5,7,4'-trimethoxyflavone) where the aromatic rings and the oxygen substituents (methoxy groups in this case) are marked.

is shown in figure 6a. We believe a detailed analysis of the aspects found by the model may help to investigate unknown structures of new substances when their NMR-spectra are included in the training set.

6 Conclusion

We proposed redundant mappings from continuous 2D-NMR spectra to discrete text-like data which can be processed by any text retrieval method. We demonstrated experimentally the effectiveness of the our mappings in combination with the vector space model and PLSI. Further analysis revealed that the aspects found by PLSI are chemically relevant. In future research we will study more recent text models like LDA [3] in combination with our mapping methods.

References

[1] A. Tsipouras, J. Ondeyka, C. Dufresne et al. Using similarity searches over databases of estimated c-13 nmr spectra for structure identification of natural products. *Analytica Chimica Acta*, 316:161–171, 1995.

[2] A. S. Barros and D. N. Rutledge. Segmented principal component transform-principal component analysis. *Chemometrics & Intelligent Laboratory Systems*, 78:125–137, 2005.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[4] L. Cai and T. Hofmann. Text categorization by boosting automatically extracted concepts. In *SIGIR '03*, 2003.

[5] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR '99*, 1999.

[6] P. Krishnan, N. J. Kruger, and R. G. Ratcliffe. Metabolite fingerprinting and profiling in plants using nmr. *Journal of Experimental Botany*, 56:255–265, 2005.

[7] M. Farkas, J. Bendl, D. H. Welti et al. Similarity search for a h-1 nmr spectroscopic data base. *Analytica Chimica Acta*, 206:173–187, 1988.

[8] Q. Mei and C. Zhai. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *KDD '05*.

[9] A. Popescul, L. H. Ungar, D. M. Pennock, and S. Lawrence. Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments. In *UAI '2001*.

[10] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, 1975.

[11] C. Steinbeck, S. Krause, and S. Kuhn. Nmrshiftdb-constructing a free chemical information system with open-source components. *J. chem. inf. & comp. sci.*, 43:1733 –1739, 2003.

[12] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths. Probabilistic author-topic models for information discovery. In *KDD '04*.

[13] K. Wolfram, A. Porzel, and A. Hinneburg. Similarity search for multi-dimensional nmr-spectra of natural products. In *PKDD '06*, 2006.