

# **Constructing an Associative Concept Space for Literature-based Discovery**

**This is a preprint of an article published in Journal of the American Society of Information**

**Science and Technology, 55(5):436-444**

**<http://www3.interscience.wiley.com/cgi-bin/jhome/76501873>**

**C. Christiaan van der Eijk, Erik M. van Mulligen, Jan A. Kors and Barend Mons**

Dept. of Medical Informatics, Erasmus MC - University Medical Center Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands

**Jan van den Berg**

Dept. of Computer Science, Faculty of Economics, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands

Correspondence to:

Christiaan van der Eijk

Dept. of Medical Informatics, Erasmus MC - University Medical Center Rotterdam

P.O. Box 1738, 3000 DR Rotterdam, The Netherlands

Tel: +31 10 4088151; fax: +31 10 408 9447

E-mail: [c.vandereijk@erasmusmc.nl](mailto:c.vandereijk@erasmusmc.nl)

**Abstract**

Scientific literature is often fragmented, which implies that certain scientific questions can only be answered by combining information from various articles. In this paper, a new algorithm is proposed for finding associations between related concepts present in literature. To this end, concepts are mapped to a multi-dimensional space by a Hebbian type of learning algorithm using co-occurrence data as input. The resulting concept space allows exploration of the neighborhood of a concept and finding potentially novel relationships between concepts. The obtained information retrieval system is useful for finding literature supporting hypotheses and for discovering hitherto unknown relationships between concepts. Tests on artificial data show the potential of the proposed methodology. In addition, preliminary tests on a set of Medline abstracts yield promising results.

## 1. Introduction

The exponential growth of scientific literature prevents many scientists from reading all articles appearing in their field. Information Retrieval (IR) has helped scientists to deal with this information explosion. Common IR techniques result in documents that are relevant to the user in a very direct way, focusing on the object of the search as defined by a set of keywords. However, indirectly related articles might be relevant to the user as well, because often no single article completely answers a scientific question. Only by combining elements from several articles on different topics, the answers to such questions can be found. For example, in genetics research, the results of chip array experiments need to be interpreted by combining information on several different genes and diseases scattered throughout biomedical literature (Hearst, 1999; Jenssen & Vinterbo, 2000). Unfortunately, finding such indirect relationships by comparing documents is a daunting task, because of the sheer number of possible combinations of articles. For this reason, much knowledge implicitly present in scientific literature remains unused. Computer-assisted searches can provide part of the solution, because computers can be used to represent and order combinations of articles or elements of articles in an informative fashion. This article focuses on combining concepts, where a concept is a standardized thesaurus form of a term or phrase. When two different terms are synonymous, they are normalized to the same concept. Using concepts instead of terms reduces the noise caused by natural language variation.

One way to find associations between concepts in literature is based on the notion of co-occurrence. The underlying assumption is that related concepts co-occur more frequently in articles than non-related concepts. A well-ordered representation of co-occurrence relations can be used to explore relations between concepts and to search for paths between concepts. Here a path is defined as a chain of co-occurring concepts. Each step in the path consists of two concepts that co-occur, and are therefore assumed to be related. Paths that connect concepts through more than one step are indicative of an indirect relationship between two concepts. Exploration of co-occurring concepts and searches for paths between concepts is expected to assist scientists in the search for indirectly related scientific papers.

Graphs are among the most intuitive ways to represent concepts and their relationships. However, searching for paths in large graphs is computationally expensive. More importantly, in a graph the positioning of concepts relative to each other is arbitrary, in particular for concepts that do not co-occur. However, when exploring literature for novel relationships, appropriate positioning may reveal important information. For example, two concepts that never appear in the same article may have many co-occurring concepts in common and on the basis of this information an interesting relationship might be suspected.

Several authors have already used the notion of co-occurrence as a basis for knowledge representation and discovery. Swanson and Smalheiser (1997) discovered valuable knowledge hidden in medical literature. They searched for paths between two sets of related terms allowing for one intermediary term to connect terms from the two sets. Others have built on Swanson's approach (Gordon & Dumais, 1998; Lindsay & Gordon 1999; Weeber, 1997). Stapley and Benoît (2000) and Jenssen, Laegreid, Komorowski, and Hovig (2001) identified gene symbols in biomedical articles and organized them in a graph, connecting co-occurring genes. Using this graph, they were able to group related gene symbols. They did not use the graphs to find paths between genes, but merely for exploratory purposes. Shatkay, Edwards, Wilbur, and Boguski (2000) predicted relationships among genes using similarity-based search algorithms. Outside the biomedical field, Kopcsa and Schiebel (1998) created maps of keywords on the basis of co-occurrence, which enabled them to see structure and trends in various research areas.

These approaches have several limitations. Firstly, both Swanson and Kopsca did not deal with the problem of synonymy, because their work was word-based and not concept-based. Stapley and Jenssen did use a thesaurus to identify concepts, but only for gene symbols. Shatkay countered the vocabulary problem by using similarity between documents as a way to identify documents about genes. However, her algorithm does require a human to select a good 'kernel' document for every gene in the analysis. Secondly, the handling capacity of existing systems appears limited. Swanson used only words from the titles of articles, and both Stapley and Jenssen looked only for gene symbols. Kopcsa's iterative learning algorithm was restricted to 200 keywords, which is too limited for our purposes. Thirdly, only in

Swanson's approach paths between concepts are constructed. His system generated all possible two-step paths between two concepts. Because the number of possible paths grows exponentially with the length of the path, the computational burden of extending the system to search beyond two-step connections is prohibitive (Lindsay & Gordon, 1999).

In this article we describe a mapping from a co-occurrence graph to an Associative Concept Space (ACS), in which concepts are assigned a position in space in such a way that the stronger the relationship between concepts, the closer they lie in the ACS. These spaces are different from the concept spaces created by Chen, Ng, Martinez, and Schatz (1997), who introduced an asymmetric cluster function to associate terms in order to generate and integrate thesauri. We will use the distance between concepts in the ACS as weights for the discovery of indirect relationships and for fast heuristic searching for paths in large sets of documents.

In the following, we first define co-occurrence and concept graphs more formally. We then describe the algorithm to construct the ACS, followed by a series of tests on artificial data that show the feasibility of the approach. Finally, the potential of the method is illustrated by two ACS examples derived from a set of Medline<sup>1</sup> abstracts.

## 2. Co-occurrence and graphs

A list of the thesaurus concepts identified in a document is called a concept fingerprint of that document. Fingerprints are created with an indexing algorithm (van Mulligen, Diwersy, Schmidt, Buurman & Mons, 2000) which is based on Salton's approach (1989). The indexing algorithm first detects sentences and eliminates stop-words from the sentences. After this it normalizes the remaining words, which means that nouns are reduced to the singular and verbs to the first person singular form. These normalized terms or phrases are then identified using a thesaurus. For this study the Medical Subject Headings 2002 (MeSH)

---

<sup>1</sup> Medline is maintained by the National Library of Medicine and can be searched using PubMed/Entrez, <http://www.ncbi.nih.gov/entrez>

thesaurus<sup>2</sup> was used. Because synonyms map to the same concept, the indexing algorithm deals with the synonym problem. For each identified concept a unique concept identifier is added to the fingerprint. This concept identifier is assigned a relevance score, based on term frequency and the specificity of the term in the thesaurus (which is the depth in the hierarchy). The fingerprints form compact representations of documents, because they are lists of concept identifiers; their use, instead of the original phrases and terms, improves speed and precision.

In this article we use co-occurrence of concepts in fingerprints. Fingerprint  $f_k$  of document  $k$  is the set of all  $m$  concepts  $c_i$  occurring in document  $k$ :

$$f_k = \{c_1, \dots, c_m\}$$

The occurrence  $o_k(c_i)$  of concept  $c_i$  in fingerprint  $f_k$  is defined as:

$$o_k(c_i) \Leftrightarrow c_i \in f_k$$

The size of the window within which co-occurrence is counted should be set as well. Here, windows equal entire fingerprints, so all combinations of two concepts in one fingerprint are counted. The co-occurrence  $\kappa_k$  of concepts  $c_i$  and  $c_j$  in fingerprint  $f_k$  is therefore defined as:

$$\kappa_k(c_i, c_j) \Leftrightarrow c_i \in f_k \wedge c_j \in f_k$$

Fingerprint  $f_k$  is said to support the co-occurrence of concepts  $c_i$  and  $c_j$  if and only if  $\kappa_k(c_i, c_j)$ .

The co-occurrence of concepts  $c_i$  and  $c_j$  in a set  $L$  of fingerprints is defined as:

$$\kappa_L(c_i, c_j) \Leftrightarrow \exists k : \kappa_k(c_i, c_j) \wedge f_k \in L$$

A straightforward representation of concepts and their co-occurrences is a graph. In this graph nodes denote concepts. In scientific literature, every association made in an article can be important, especially for domains that have received little attention from researchers. Therefore, we have decided to include a relation between two concepts, if they co-occur in at least one article. Nodes are connected by edges if and only if the corresponding concepts co-occur.

---

<sup>2</sup> <http://www.nlm.nih.gov/mesh/meshhome.html>

### 3. ACS construction

The ACS is a multidimensional Euclidian space in which concepts are positioned. Concepts that are connected by frequent co-occurrence paths, either directly or indirectly, should have a small distance in the ACS, while concepts with few or no paths between them should be far apart. To determine an appropriate position for the concepts a Hebbian learning algorithm based on the work of Van den Berg and Schuemie (1999) is used.

Each concept  $\underline{c}_i$  in the fingerprint set  $\underline{L}$  is associated with an  $\underline{n}$ -dimensional location vector  $\mathbf{x}_i$  in the ACS:

$$\mathbf{x}_i^T = (x_{i,1} \quad x_{i,2} \quad \dots \quad x_{i,n})$$

The ACS algorithm first assigns a random location vector in the  $\underline{n}$ -dimensional space to each concept. Then, for each fingerprint, all concepts in  $\underline{f}_k$  are moved to the centroid  $\mathbf{p}_k$  of  $\underline{f}_k$ , which is defined as the average of the vectors of the concepts in the fingerprint:

$$\mathbf{p}_k = \begin{bmatrix} p_{k,1} \\ \vdots \\ p_{k,n} \end{bmatrix} = \begin{bmatrix} \frac{\sum_{h=0}^m x_{h,1}}{m} \\ \vdots \\ \frac{\sum_{h=0}^m x_{h,n}}{m} \end{bmatrix}$$

where  $\underline{m}$  is the total number of concepts in fingerprint  $\underline{f}_k$ .

The Hebbian learning rule is defined so that each concept is attracted to the centroid:

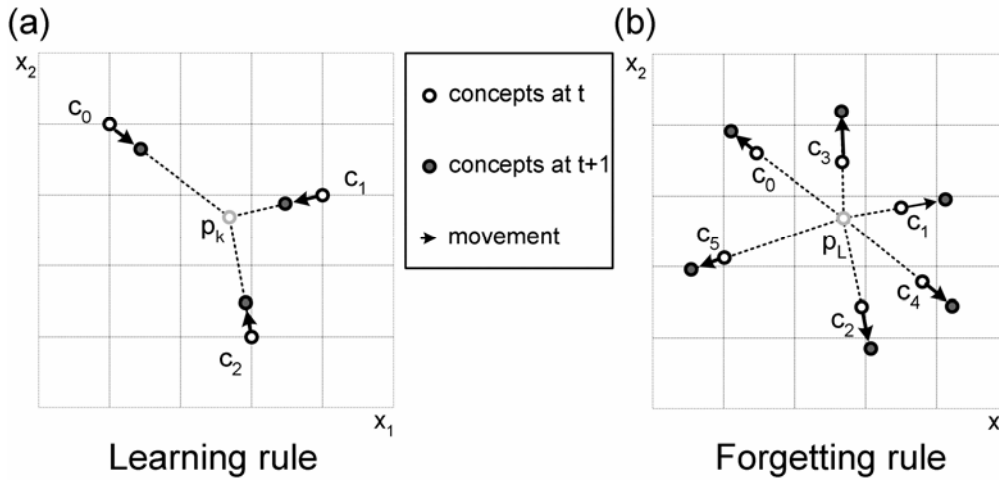
$$\mathbf{x}_i(t+1) = \mathbf{x}_i(t) + \eta(t) \frac{\mathbf{p}_k(t) - \mathbf{x}_i(t)}{\|\mathbf{p}_k(t) - \mathbf{x}_i(t)\|}$$

where  $\underline{t}$  is the learning cycle and  $\eta(t)$  ( $0 < \eta(t)$ ) is the learning rate. The learning rate is the ratio at which the distance between the concepts and  $\mathbf{p}_k$  is reduced and is set as in (van den Berg & Schuemie, 1999):

$$\eta(t) = \frac{2}{\min(t, u)}$$

where the constant  $\underline{u}$  is set by the user.

The greater  $\eta(t)$ , the greater the movement of the concepts in the space in case of co-occurrence. The movement is normalized to prevent far off points from moving great distances, because that would cancel the effects of previous learning. Figure 1a illustrates how concepts in the same fingerprint  $f_k$  move towards the centroid  $p_k$ . Note that one concept may occur in many different fingerprints, and the learning rule may change the positioning of the concept with each fingerprint in the ACS.



**Figure 1** The learning (a) and forgetting rule (b) applied in a two-dimensional space. Fingerprint  $f_k$  contains concepts  $c_0$ ,  $c_1$ , and  $c_2$  and  $p_k$  is its centroid. The set of all included concepts  $L$  contains  $c_0, c_1, \dots, c_5$  and its centroid is  $p_L$ .

After all fingerprints have been used for learning, the forgetting rule is applied. Concepts in fingerprint set  $\underline{L}$  are moved away from the overall centroid  $p_L$  which is calculated over all concepts in fingerprint set  $\underline{L}$ , to prevent congregation in one point and to separate points that do not co-occur. The rule increases the distance between concepts:

$$\mathbf{x}_i(t+1) = \mathbf{x}_i(t) - \lambda(\|\mathbf{p}_L(t) - \mathbf{x}_i(t)\|) \frac{\mathbf{p}_L(t) - \mathbf{x}_i(t)}{\|\mathbf{p}_L(t) - \mathbf{x}_i(t)\|}$$

where the repulsion function  $\lambda(y)$  is defined as in (van den Berg & Schuemie, 1999):

$$\lambda(y) = \begin{cases} 1 & \text{for } y < 1 \\ 1/y & \text{for } y \geq 1 \end{cases}$$

Figure 1b illustrates how concepts are pushed away from the centroid  $\underline{p}$  by the forgetting rule. The process of learning and forgetting can be repeated. In each repetition or cycle, the learning rule is applied using all fingerprints and the forgetting rule is applied once at the end of the cycle.

The distance between concepts in the resulting concept space reflects how often each concept and the concepts around it were drawn to each other because of co-occurrence in a fingerprint. In this way the Euclidian distance between two concepts is a measure of both co-occurrence and how many co-occurring concepts the two concepts have in common. The ACS system adds an edge between two concepts when they co-occur at least once, storing the identification codes of the fingerprints in which they co-occur. Thus, associations between concepts can be traced back to articles that support the association. The length of an edge is set to the Euclidian distance between the two concepts it connects.

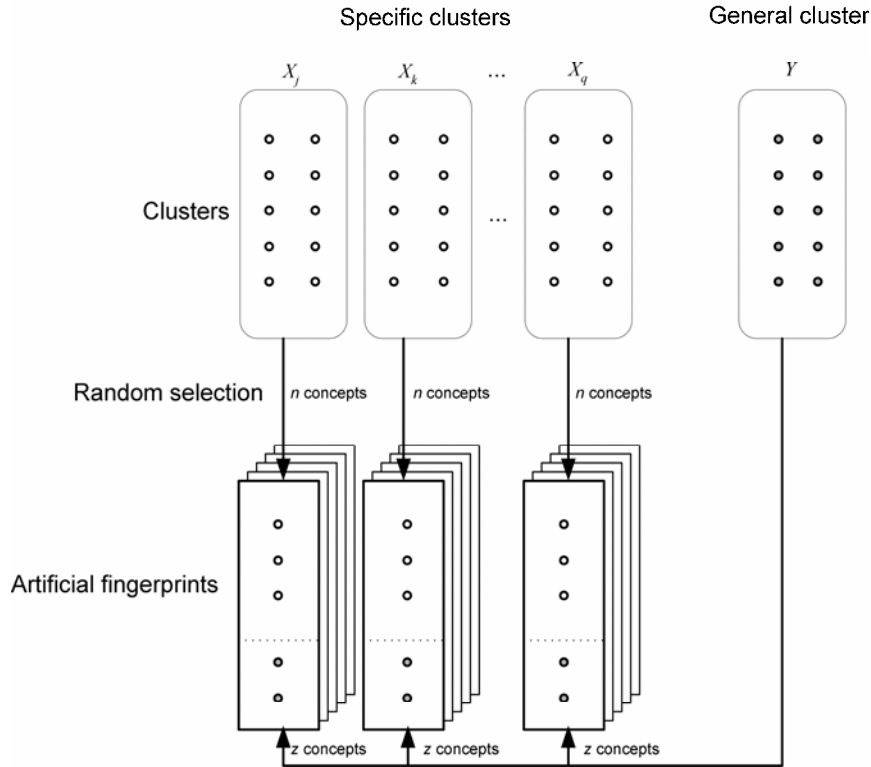
Combining the co-occurrence graph and the spatial representation allows for fast path searching in the graph, because at each step in the search algorithm the remaining distance to the target node from the neighboring node can be estimated using the Euclidian distance in the ACS. For example, the A\* algorithm uses this heuristic information to find a path between two given nodes efficiently (Weiss, 1999).

#### **4. Test set-up**

A formal evaluation of the ACS on real data, i.e., existing scientific documents, is complicated because it is extremely difficult to establish a reference for the explicit and particularly the implicit associations present in a set of documents. For that reason we have used two other types of tests to evaluate the usefulness of the ACS. Firstly, the performance of the ACS algorithm was tested on simulated data sets for which the outcome is predictable. These controlled data are used to evaluate quantitatively the effects of several system parameters on stability and quality of the result. Secondly, two examples based on a real data from the Medline literature database are given to illustrate the search and discovery capabilities of the ACS in actual literature and to assess processing time.

The simulated data were automatically generated using a model of scientific literature. This model consists of a set of clusters; each cluster represents a field of interest, characterized by a set of specific concepts. A fingerprint contains concepts from one such cluster. Furthermore, it is assumed that each fingerprint contains a number of more general concepts, which occur throughout different fields of interest. These concepts can be thought of as elements from one cluster. Each fingerprint forms a mixture of specific and general concepts. This reflects a pattern often found in the Medline fingerprints we inspected, which usually contain a large number of infrequent, specific concepts and a few common, general concepts, like genes or alleles.

More formally, the simulation model consists of clusters  $X_j$ , which contain  $s_j$  specific concepts. For fingerprint  $f_k$  of length  $m$ ,  $z$  specific concepts are drawn at random from cluster  $X_j$ . An additional  $m-z$  concepts are added to the fingerprint to mimic the general concepts, and the ratio  $(m-z)/m$  is called the generality ratio. The general concepts are drawn at random from a special cluster  $Y$  of size  $s_Y$ . The process of fingerprint generation is illustrated in Figure 2.



**Figure 2** The fingerprint generation process. Fingerprints are generated by randomly selecting a number of concepts from one specific cluster  $X_i$  and adding a number of concepts from the general cluster  $Y$ .

Training the ACS with the simulated fingerprints should result in a space in which concepts from the same cluster are separated from other concepts. We have tested whether the ACS separates the clusters by using an invariant scattering criterion from cluster analysis (Duda & Hart, 1973). For this criterion the following definitions are used:

$p_j$  is the centroid for the concepts from cluster  $X_j$

$$p_j = \frac{1}{s_j} \sum_{x_i \in X_j} x_i$$

where  $x_i$  is the reference vector for a concept  $c_i$  and  $s_j$  is the number of concepts in cluster  $X_j$ .

The vector  $p_s$  is the centroid of all cluster centroids is:

$$p_s = \frac{1}{q} \sum_{j=1}^q s_j p_j$$

where  $q$  is the number of clusters and  $s_j$  the number of concepts in cluster  $X_j$ .

The scatter matrix  $M_j$  for cluster  $X_j$  is defined as:

$$M_j = \sum_{x_i \in X_j} (x_i - p_j)(x_i - p_j)^t$$

The within-cluster scatter matrix  $M_W$  is the sum of the cluster scatter matrices:

$$M_W = \sum_{j=1}^c M_j$$

The between-cluster scatter matrix  $M_B$  is defined:

$$M_B = \sum_{j=1}^c s_j (p_j - p_S)(p_j - p_S)^t$$

These are taken together in the invariant criterion:

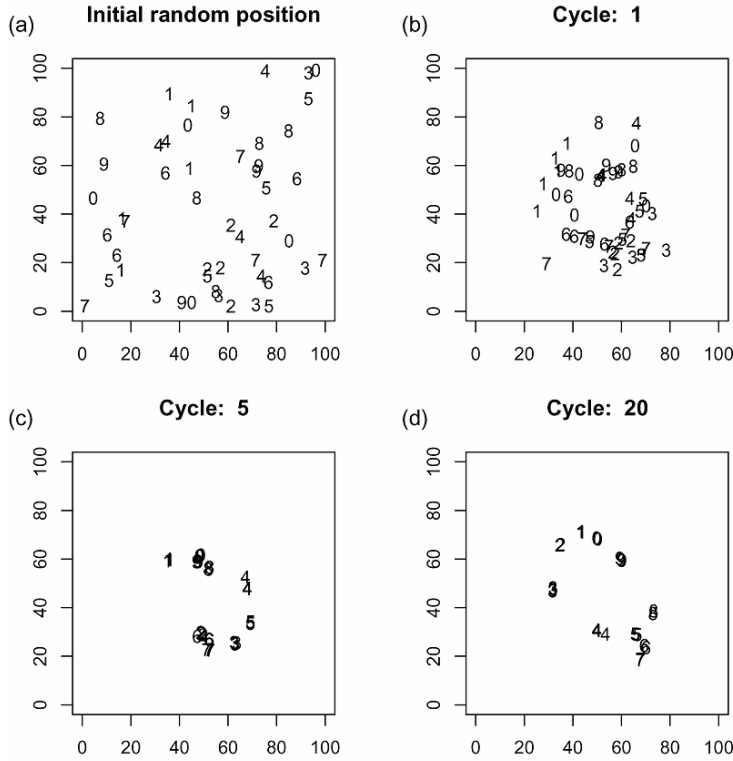
$$C = \text{trace } M_W^{-1} M_B$$

This criterion gives a ratio of the between-cluster scatter to the within-cluster scatter in the direction of the eigenvectors. A set of well separated clusters has a greater scatter between the clusters than within and yield a high value of  $C$ . The lower limit for  $C$  is 0, and no upper limit exists.

## 5. Results

### 5.1 Simulated Data

Figure 3 illustrates the effect of learning in a 2-dimensional ACS. The position of a concept in the space is marked with the number of its cluster. The different ACS's in Figure 3 were trained using 100 artificial fingerprints of size 10. The fingerprints were generated from 10 clusters of 10 concepts with 4 general concepts per fingerprint.



**Figure 3** The initial distribution of concepts in a 2-dimensional space (a) and distribution after 1 (b), 5 (c) and 20 (d) learning cycles.

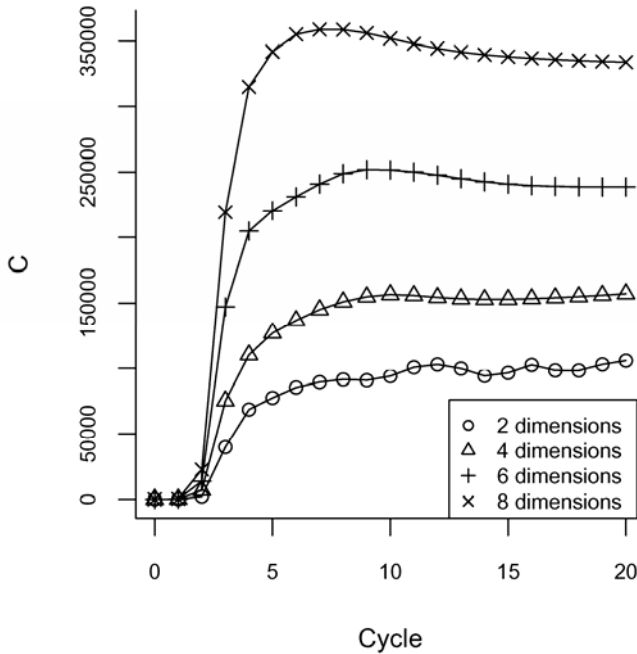
Figure 3a shows an initial random distribution, where concepts are randomly scattered in space. One learning cycle (Fig. 3b) changes the distribution drastically. After 5 cycles (Fig. 3c) concepts from the same cluster have already been grouped. After 20 cycles (Fig. 3d) concepts from different clusters are well separated. Other runs resulted in different positions of the clusters, because the random initialization was different each time. However, the relative positions of the clusters after 20 cycles were very similar, the circle having been rotated differently.

Using the scatter criterion  $C$  as a measure of success, two series of tests were conducted. In our tests two other algorithm parameters had little effect on the resulting space: the  $\underline{u}$  in the learning parameter

$$\eta(t) = \frac{2}{\min(t, u)}$$

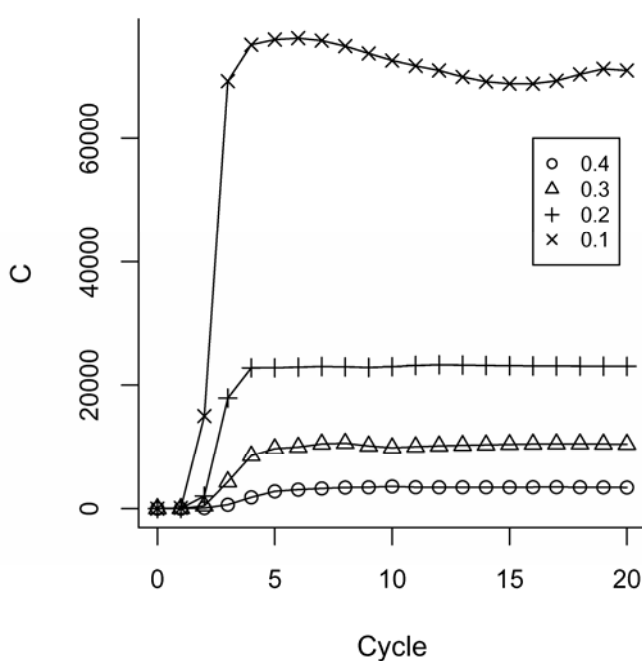
and the range from which the initial random numbers were drawn. For that reason, in all

tests,  $\underline{u} = 10$ , and the random numbers were generated from a uniform distribution in the range  $[0,100]$ . In all tests, each fingerprint contained 10 concepts and for each cluster 10 fingerprints were included.



**Figure 4** Scatter criterion  $C$  against the number of learning cycles for an ACS of varying dimensions.

Figure 4 demonstrates the effect of increasing the number of dimensions. The generality ratio was fixed at 0.4 and the number of concepts per cluster was set at 10. Figure 4 shows that using more dimensions results in higher  $C$ -values. The number of dimensions determines the number of degrees of freedom and a larger number of degrees of freedom will lead to better clustering. Figure 4 also shows that for these parameters the quality of the mapping stabilizes after around 7 learning cycles. Repeated experiments show some variation in the resulting  $C$ -values, again due to the changing random initialization. The number of cycles at which stabilization occurred depended on the size of the clusters.



**Figure 5** Scatter criterion  $C$  against the number of learning cycles for an ACS of a varying generality ratio.

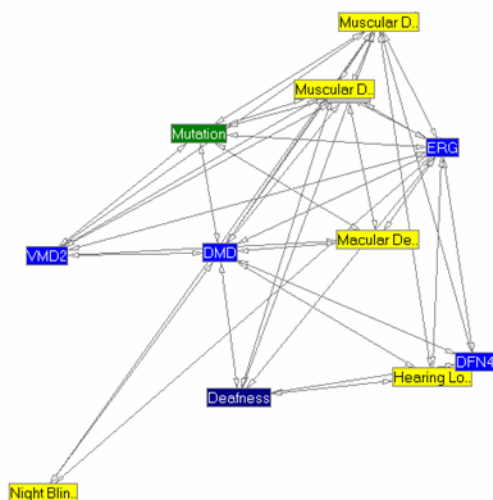
We have tested for variations in the simulated data by varying the generality ratio (defined in section 4). The size of the clusters determines the complexity of the generating model. The generality ratio determines how many concepts in a fingerprint are cluster-specific and therefore informative. Figure 5 shows the effect of different generality ratio levels on the quality of the mapping. For these tests, the number of concepts per cluster was set at 10. The generality ratio level greatly influenced the resulting  $C$ -values, with lower ratios resulting in better mappings. The generality ratio levels did not seem to have an effect on the number of learning cycles needed to reach a stable solution. We have also varied the size of the fingerprints and the number of fingerprints per cluster, but the shape of the curves remained essentially the same as those for the generality ratio (data not shown).

### 5.2 Medline data

To see whether the learning algorithm resulted in a potentially interesting depiction of the co-occurrence patterns in a set of literature, we have used a set of articles from Medline. This set was obtained by

querying PubMed with the query “Duchenne OR DMD OR dystrophy OR limb-girdle OR LGMD OR BMD” which resulted in 13,423 articles (February 9, 2003). This test set was indexed by the indexing algorithm using the Medical Subject Headings 2002 vocabulary (MeSH). The resulting fingerprints contained an average of 29 concepts per abstract. In total, the set of fingerprints contained 9,770 different concepts. In the fingerprints 5,378 different concepts that had a relevance score greater than 0.4. Between these concepts 109,430 edges were created. The ACS learning algorithm was applied on this test set and a graphical interface (Van Mulligen, Van der Eijk, Kors, Schijvenaars, & Mons, 2002) was used to inspect the resulting 8-dimensional ACS. In this interface some distances may be distorted, because an 8-dimensional ACS is projected onto a 2-dimensional plane by taking only the first two axes in account. Nevertheless, concepts close to each other in ACS will inevitably appear close to each other on the plane. Some concepts that appear near in the plane might be far-off in the actual concept space, but the visualization tool also shows the distances so these instances can be distinguished from truly near concepts.

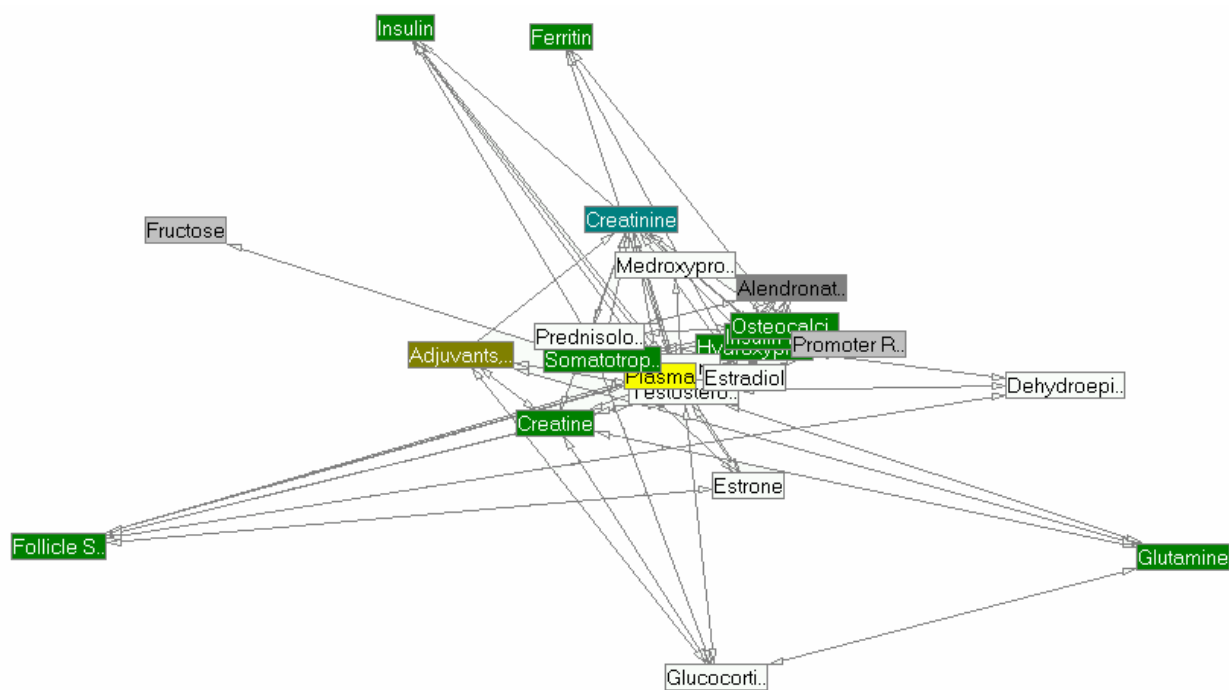
A small subset of the ACS after 10 cycles is shown in Figure 6. Different grey levels denote the semantic types in the UMLS ontology (National Library of Medicine, 1995) that are attached to concepts. For example, “Muscular Dystrophy” in figure 6 is of the semantic type “Disease or Syndrome”.



**Figure 6** Two-dimensional projection of part of the 8-dimensional ACS derived from a set of Medline abstracts on Muscular Dystrophy. The ACS suggests a relationship between “Macular Degeneration” and “Deafness”.

In figure 6 one can discern “Deafness” and “Hearing Loss” in close proximity, as is to be expected, but both are also close to “Macular Degeneration”, which is loss of vision due to degeneration of the part of the eye called the macula. The relationship between these concepts is not immediately obvious and in our subset of Medline no abstract contains both terms, which is why figure 6 does not show a link between them. However, a query of the whole of Medline for articles containing both “Deafness” and “Macular Degeneration” yielded 28 results (June 13, 2003), some of which clearly link deafness and macular dystrophy, a condition that leads to degeneration of the macula. The association between deafness and macula degeneration which is explicit in several Medline abstracts was only implicitly present in the subset we used. The ACS algorithm revealed this implicit association.

Figure 7 shows another example from the same ACS.



**Figure 7** Two-dimensional projection of part of an 8-dimensional ACS of a set of Medline abstracts on Muscular Dystrophy. The ACS suggests a relationship between “Insulin” and “Ferritin”.

Here we find that “Insulin” and “Ferritin”, amongst others, are positioned closely together, without these concepts co-occurring in our set of abstracts. Again we might infer that they are related and a PubMed search yields 212 articles that contain both terms (June 27, 2003). As in the previous example, the ACS was able to reveal implicit information for a set of fingerprints.

We have also trained an ACS for a set of European public sector information from Trias Politica Online<sup>3</sup> (TP Online) consisting of 415,191 articles with a total of 37,976 different concepts and 480,985 edges. One training cycle for an ACS based on this set took about 4 minutes on a 2.4 GHz Pentium IV.

## 6. Discussion and outlook

Literature-based discovery requires computer assistance to present the enormous number of potentially interesting relations between scientific papers in an insightful way. Using co-occurrence of concepts, we built a concept space that presents views on the patterns of co-occurrence between concepts. Our experiments show that spatial information in the ACS reflects co-occurrence patterns, while it also allows co-occurrence between two concepts to be retrieved. An example ACS derived from a set of Medline abstracts yielded promising results. In the ACS, we were able to rediscover established knowledge; with a larger set of articles we expect not only to find established knowledge, but also to uncover novel associations. We are currently initiating large-scale experiments to further assess the usefulness of the ACS for the discovery process, in close collaboration with domain experts.

Several studies describe techniques to map relationships to a space or map, albeit for other purposes. For example, co-citation statistics have been displayed graphically by Chen and Carr (1999) and Small (1999). Others have mapped documents to a plane as a tool for retrieving documents and displaying relationships between documents (Benoit, 2002; Iliopoulos, Enright & Ouzounis, 2001; Kohonen, Kaski, Lagus, Salojärvi, Honkela, Paatero & Saarela, 2000; Lin, 1997).

---

<sup>3</sup> <http://www.triaspolitica.org>

Further improvements of the various algorithms will be investigated. The indexing algorithm calculates a relevance score for each concept based on several statistics, e.g. the term frequency and the inverse document frequency. The relevance scores and the location of concepts in articles could be used to add weights to the learning and forgetting rules. An additional source of input to be employed by the ACS learning algorithm is the relational information present in many thesauri and ontologies. For example, thesauri may contain concept hierarchies, which link concepts through is-a relationships. These could be used to improve the initial distribution, by placing concepts close in the hierarchy in each other's proximity in the ACS. The hierarchy can also be used to refine learning. The algorithm might add the co-occurrence of children to the co-occurrence of their parents, so that when e.g. a symptom co-occurs with a variant of a disease, the disease in general is also drawn to that symptom. Additionally concepts may have been assigned a semantic type, as in the UMLS ontology (National Library of Medicine, 1995). In this ontology, a network of potential relations between semantic types has been defined. This network could be used to present a list of possible relations between two concepts. A priori knowledge from ontologies or from previous runs could also be used to improve the initial positioning of concepts in the ACS, which is now done at random.

Our current system assesses the statistical relationships of concepts in documents. For a researcher the semantics of a relationship will often be of interest. Several Natural Language Processing (NLP) tools have been developed for mining both genes and proteins interactions with interesting results (Blaschke and Valencia, 2001; Craven & Kumlien, 1999; Rindflesch, Tanabe, Weinstein & Hunter, 2000; Sekimizu, Park & Tsujii, 1998; Stephens, Palakal, Mukhopadhyay, Raje & Mostafa 2001; Wong; 2001). However, the scope and scalability of these tools is problematic because NLP techniques are computationally expensive. Therefore, a combination of the ACS approach with NLP techniques to identify the nature of the relationship between co-occurring concepts may prove valuable.

In conclusion, the ACS presents a variety of ways to discover novel relationships between concepts. The user may inspect an ACS to discover possibly interesting associations using the graphical user interface

(Van Mulligen, Van der Eijk, Kors, Schijvenaars, & Mons, 2002). This allows the user to explore the set of potentially interesting concepts around a selected concept and the relationships in this set and to retrieve the literature that supports the relationships. By adding a path finding algorithm users will be assisted further in their search for associations. The ACS interface can also suggest for closer inspection pairs of concepts that lie close in space, but are not connected in any article, simply by generating lists of concepts ordered by distance to a selected concept, possibly filtering on semantic categories of concepts, if such information is provided by the thesaurus. We believe that the ACS together with the user interface is a powerful tool for knowledge discovery.

### **Acknowledgements**

We would like to thank Collexis B.V. for making available their software and concept fingerprints.

## References

- Benoit, G. (2002). Data Discretization for Novel Relationship Discovery in Information Retrieval. *Journal of the American Society for Information Science* 53: 736-746.
- Blaschke, C. & Valencia, A. (2001). The potential use of SUISEKI as a protein interaction discovery tool. *Genome Informatics* 12: 123-34.
- Chen, C. & Carr, L. (1999). Visualizing the evolution of a subject domain: A case study. In *Proceedings of the IEEE Visualization '99 Conference*, (pp. 449-452), San Francisco, CA: IEEE Computer Society Press.
- Chen H, Ng TD, Martinez J, Schatz BR. (1997). A Concept Space Approach to Addressing the Vocabulary Problem in Scientific Information Retrieval: An Experiment on the Worm Community System. *Journal of the American Society for Information Science* 48(1):17-31.
- Craven, M. & Kumlien, J. (1999). Constructing biological knowledge bases by extracting information from text sources. In Lengauer, T., Schneider, R., Bork, P., Brutlag, D., Glasgow, J., Mewes, H-W., and Ralf Zimmer (eds.) *Proceedings of the 7<sup>th</sup> International Conference on Intelligent Systems in Molecular Biology (ISMB-99)* (pp. 77-86) Menlo Park, CA: AAAI Press.
- Duda, R. O. & Hart, P. E. (1973). *Pattern classification and scene analysis*. New York: Wiley-Interscience 1973.
- Gordon, M. D. & Dumais, S. (1998). Using Latent Semantic Indexing for Literature Based Discovery. *Journal of the American Society for Information Science and Technology* 49: 674-685.
- Hearst, M. A. (1999). Untangling Text Data Mining. In *Proceedings of the 37<sup>th</sup> Annual Meeting of the Association for Computational Linguistics (ACL'99)*, College Park, MD
- Iliopoulos, I., Enright, A. J. & Ouzounis, C. A. (2001). Textquest: document clustering of Medline abstracts for concept discovery in molecular biology. In *Proceedings of the 6<sup>th</sup> Pacific Symposium on Biocomputing (PSB 2001)* (pp. 384-95).

- Jenssen, T. K., Laegreid, A., Komorowski, J. & Hovig, E. (2001). A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics* 28: 21-28.
- Jenssen, T. K. & Vinterbo, S. (2000). A set-covering approach to specific search for literature about human genes. *Proceedings of the American Medical Informatics Association Symposium*: (pp. 384-8).
- Kohonen, T., Kaski, S., Lagus, K., Salojärvi, J., Honkela, J., Paatero, V. & Saarela A. (2000). Self organization of a massive document collection. *IEEE Transactions on Neural Networks* 11: 1-31.
- Kopcsa, A. & Schiebel, E. (1998). Science and Technology Mapping: A New Iteration Model for Representing Multidimensional Relationships. *Journal of the American Society for Information Science* 49: 7-17.
- Lin, X. (1997). Map displays for information retrieval. *Journal of the American Society for Information Science* 48: 40-54.
- Lindsay, R. K. & Gordon, M. D. (1999). Literature-Based Discovery by Lexical Statistics. *Journal of the American Society for Information Science* 50: 574-587.
- National Library of Medicine. (1995). UMLS knowledge sources; Documentation,
- Rindflesch, T. C., Tanabe, L., Weinstein, J. N. & Hunter, L. (2000). EDGAR: extraction of drugs, genes and relations from the biomedical literature. In *Proceedings of the 5<sup>th</sup> Pacific Symposium on Biocomputing (PSB 2000)*, pp. 517-28.
- Salton G. (1989). *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Reading, Mass.: Addison-Wesley.
- Sekimizu, T., Park, H. S. & Tsujii, J. (1998). Identifying the Interaction between Genes and Gene Products Based on Frequently Seen Verbs in Medline Abstracts. *Genome Informatics* 9: 62-71.
- Shatkay H, Edwards S, Wilbur WJ, Boguski M. (2000). Genes, themes and microarrays: using information retrieval for large-scale gene analysis. In *Proceedings of the International Conference on Intelligent Systems for Molecular Biology 2000*;8:317-28.

- Small, H. (1999). Visualizing Science by Citation Mapping. *Journal of the American Society for Information Science* 50: 799-813.
- Stapley, B. J. & Benoit, G. (2000). Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. In *Proceedings of the 5<sup>th</sup> Pacific Symposium on Biocomputing (PSB 2000)* (pp. 529-40).
- Stephens, M., Palakal, M., Mukhopadhyay, S., Raje, R. & Mostafa, J. (2001). Detecting gene relations from Medline abstracts. In *Proceedings of the 6<sup>th</sup> Pacific Symposium on Biocomputing (PSB 2001)* (pp. 483-95)
- Swanson, D. R. & Smalheiser, N. R. (1997). An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial Intelligence* 91: 183-203.
- Van den Berg, J. & Schuemie, M. (1999). Information Retrieval Systems using an Associative Conceptual Space. In *Proceedings 7th European Symposium on Artificial Neural Networks (ESANN'99)*, Bruges, Belgium.
- Van Mulligen, E. M., Diwersy, M., Schmidt, M., Buurman, H. & Mons, B. (2000). Facilitating networks of information. In *Proceedings of the American Medical Informatics Association Symposium* (pp. 868-72)
- Van Mulligen, E. M., Van der Eijk, C. C., Kors, J. A., Schijvenaars, B. J. A. & Mons, B. (2002). Research for Research: Tools for knowledge discovery and visualization. In *Proceedings of the American Medical Informatics Association Symposium 2002* (pp. 835-839).
- Weeber, M., Vos, R., Klein, H. & de Jong-van den Berg, L.T.W. (1997). Using Concepts in Literature-based Discovery: Simulating Swanson's Raynaud - Fish Oil and Migraine - Magnesium Discoveries. *Journal of the American Society for Information Science and Technology* 52: 548-557
- Weiss, M. A. (1999). *Data structures and algorithm analysis in C++*. Reading, Massachusetts, Addison Wesley Longman, Inc.

Wong, L. (2001). PIES, a protein interaction extraction system. Proceedings of the 6<sup>th</sup> Pacific Symposium on Biocomputing (PSB 2001) (pp. 520-31)