Learning a Lightweight Ontology for Semantic Retrieval in Patient-Centered Information Systems*

Ulrich Reimer[1], Edith Maier[1], Stephan Streit[1],
Thomas Diggelmann[2], and Manfred Hoffleisch[2]

[1] Institute for Information & Process Management
University of Applied Sciences St. Gallen
{firstname.lastname}@fhsg.ch
[2] semantic system ag
{td|mh}@ai-one.com

Abstract. The paper introduces a web-based eHealth platform currently being developed that will assist patients with certain chronic diseases. The ultimate aim is behavioral change. This is supported by online assessment and feedback which visualizes actual behavior in relation to target behavior. Disease-specific information is provided through an information portal that utilizes lightweight ontologies (associative networks) in combination with text mining. Emotional support is provided via virtual communities. The paper argues that classical word-based information retrieval is often not sufficient for providing patients with relevant information, but that their information needs are better addressed by concept-based retrieval. The focus of the paper is on the semantic retrieval component and the learning of a lightweight ontology from text documents, which is achieved by using a biologically inspired neural network. The paper concludes with preliminary results of the evaluation of our approach in comparison with traditional approaches.

Keywords: concept-based retrieval, ontology learning, neural network, lightweight ontology, term associations, associative net

1 Introduction

A growing share of the burden of disease, i.e. the direct and indirect health costs, is accounted for by chronic conditions. At the same time, health authorities across Europe have come to realize the tremendous costs involved in chronic care. It is therefore not surprising that countries are shifting in health policy towards more self-management and patient-centered care. However, self-management of disease is a skill that cannot be taken for granted but has to be learned and most people need assistance for this task. Motivating people to change their behavior

* The work reported in this paper is funded by the Swiss Confederation's innovation promotion agency (CTI) under grant number 9937.2.

and lifestyle has emerged as the central challenge in this respect. Tailoring the information to the needs and requirements of the individual user has found to be another prerequisite for success. In fact, user orientation plays an important role in SEMPER because setbacks in self-management initiatives in the past have also been attributed to the lack of personalization in the provision of information [6]. Concept-based retrieval is expected to address this shortcoming.

Before discussing this approach in detail, we give a brief overview of the SEMPER project which provides the framework for this endeavor (Sec.2). Subsequently, Sec.3 outlines the need for semantic retrieval. Sec.4 describes how a lightweight ontology is learned using a biologically inspired neural network. Sec.5 compares our approach with learning association nets by classical co-occurrence measures and provides first evaluation results. Sec.6 outlines future work.

2 SEMPER: A Support System for Patient Self-Care

The SEMPER[1] project develops an interactive, web-based platform that provides patients with ongoing assistance and encouragement for dealing with problems such as alcohol dependency and work-related disorders, especially those related to office work (e.g. stress, eye strain, repetitive strain injury). This will be realized through online assessment, disease-specific information, personalized monitoring and feedback as well as social and emotional support via virtual communities. The inclusion of new fields of application and/or target groups will be possible due to the open architecture of the platform.

The online components are not meant to replace consulting a doctor or other health professionals. Rather, we want to use the advantages of interactive technologies to lessen the burden of health professionals and complement face-to-face treatment. Figure 1 illustrates the main components of the SEMPER platform:

- *Motivation & monitoring support:* The online self-assessment questionnaire allows the user to specify the measures for changing behavior, such as daily exercise, or a certain maximum amount of alcohol intake per day (in the case of controlled drinking). This results in a personal action plan. Intended and actual behavior are then compared and the progress visualized.
- *Information portal:* The information portal provides health information and self-care training. This module focuses on increasing users' health literacy and improving their self-management skills. The patients can learn about symptoms, conditions, implications or consequences of their health conditions from a variety of information sources brought together on a single platform. They can learn about how their problems are related to their lifestyles and habits and how eventual behavioral changes can alleviate them.
  The information portal also allows access to relevant online communities which are included in the search. These represent a valuable social lifeline for those homebound due to illness, age or handicap, or those isolated in rural settings. Besides, in the case of alcohol-related problems, some may prefer

[1] see http://www.semper-net.ch/index.php?lang=en

the anonymous exchange online because of the social stigma attached to alcohol dependency. Besides, the knowledge to be found in online communities represents a resource into which health care professionals and researchers may be interested in tapping into so as to supplement their more structured research and to gain additional insights.

– *Maintenance & information cockpit:* This component allows to add or delete contents in the information portal. Moreover, since the ontologies used to enable semantic search are automatically extended by the system (see Sec.4 they need to be manually checked from time to time so that inadequate concepts and relationships can be removed.
– *User administration:* While the information portal of SEMPER can be used without a login, the motivation & monitoring support component requires a user to register and sign in so as to store the action plans and the data entered into the questionnaire. In the case of a registered user the user's personal data like age, date of diagnosis, etc. will be used to personalize the information provided via the information portal.

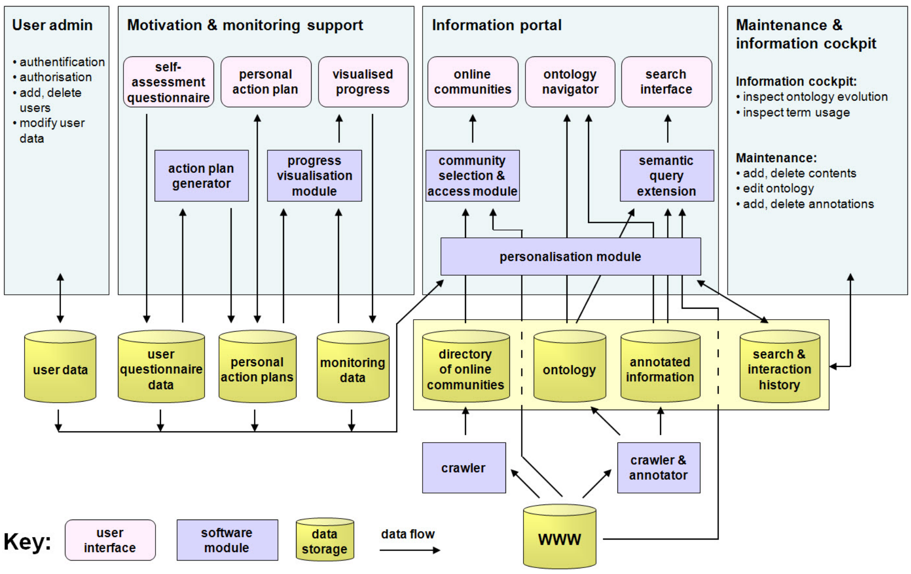

Fig. 1. System architecture of SEMPER

3 The Need for Semantic Retrieval

There is a huge gap between the information needs of a patient and their transformation into an appropriate query for obtaining the relevant information. Even for experienced users it can be quite cumbersome to find the information they are looking for because there may be many ways to refer to a particular concept (e.g. “MSD”, “Musculoskeletal disorder”, “lower back pain”). More importantly, some users may use a term from the subject-specific terminology while others use (and only know) popular terms. Even worse, patients often do not even know exactly what they are looking for and therefore have no clue which terms to use in their search. This is the reason why finding relevant information by means of standard search engines like Google can be very time-consuming and frustrating.

The solution lies in applying Semantic Web technology and moving from word-based information retrieval to concept-based retrieval. For this we need an ontology which relates semantically similar concepts with each other so that a search engine can extend a person’s query to include related concepts which are not referenced in the query [13]. For example, entering the search term “work-related disease” would also retrieve documents that contain the words “occupational disease” or “work-related disorders” if the underlying ontology contains the proper relations between these terms. Since SEMPER is specifically aimed at patients with work-related diseases and alcohol dependency, medical thesauri may be helpful but they only partially provide the query terms a user needs to get adequate search results. Therefore we have to combine them with additional, more application-specific ontologies.

4 Learning a Lightweight Ontology Using a Biologically Inspired Neural Network

In SEMPER, we use a small subset of UMLS [5], namely synonyms, and various kinds of concept specialization. As said before, many search terms in our application setting are not included in a medical thesaurus, either because they are colloquial or because they relate to aspects simply not covered by a thesaurus. For example, a patient with work-related back pain might enter the query

```
lifestyle "back pain"
```

Relevant documents, however, may not contain the term “lifestyle” but related terms like “stress”, “nutrition”, “physical exercise”. If these terms are related with the term “lifestyle” the original query can be automatically expanded as shown in Figure 2. As can be seen we do not need a full-fledged ontology for this kind of query expansion but only an associative network (or lightweight ontology) where concepts are related with each other via an untyped relationship (or association) that expresses some kind of semantic nearness. An association between two concepts is labeled with an association strength between 0 and 1.

Given that many associative nets like the one shown in Figure 2 are needed and since it is not clear beforehand which terms they should contain, it is quite unrealistic to create them manually. We therefore adopt an unsupervised learning approach to acquire the associative nets automatically from text documents of the underlying domain. In contrast to full-fledged ontologies, association nets have the advantage that learning, except for regular manual pruning, happens mostly automatically (see [1] for an overview).

Figure 3 illustrates the overall learning setup. All the input documents are transformed into plain text documents (from original formats like pdf, doc, html). In the resulting documents stop words are eliminated and all words are converted into their uninflected, lemmatized form. After an initial learning step that uses a set of manually selected input documents the learning process proceeds incrementally on the basis of query results. The underlying rationale is that documents resulting from a user query are quite likely relevant for learning association nets. End-user queries therefore contribute to the continuous updating and extension of the associative nets.

In the following section we describe the neural net learner we use for learning the associations. Sec.4.2 discusses how the neural net learner is utilized in SEMPER and how the learning results are improved by giving the learner a notion of relevance. Sec.4.3 discusses the issue of directed vs. undirected, i.e. symmetric associations.

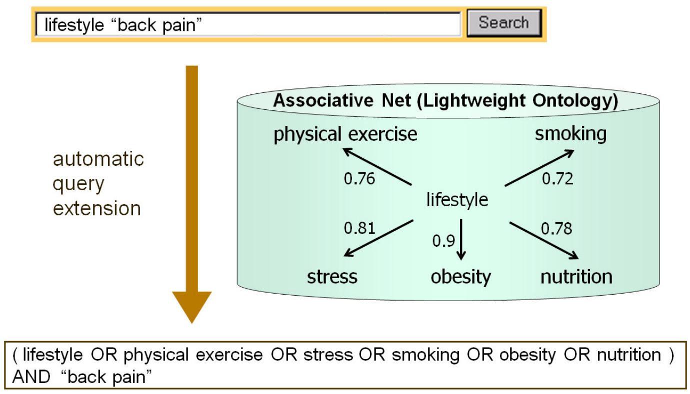

Fig. 2. Query expansion using an associative net (or lightweight ontology)

4.1 The Learning Algorithm: Biologically Inspired Neural Network

In SEMPER we use the neural net learner ai one$^{\mathrm{TM}}$provided by the project partner semantic system ag. Unlike any of the traditional neural nets [3], the neural networks based on ai one$^{\mathrm{TM}}$, the so-called "Hoffleisch neural networks" (because they were invented by Manfred Hoffleisch) or in short "HNN", are massively connected, asymmetrical graphs which are stimulated by binary spikes. HNN do not have any neural structures pre-defined by the user. Their building blocks resemble biological neural networks: a neuron has dendrites, on which the synapses from other neurons are placed, and an axon which ends in synapses at other neurons.

The connections between the neurons emerge in an unsupervised manner whilst the learning input is translated into the neural graph structure. The re-

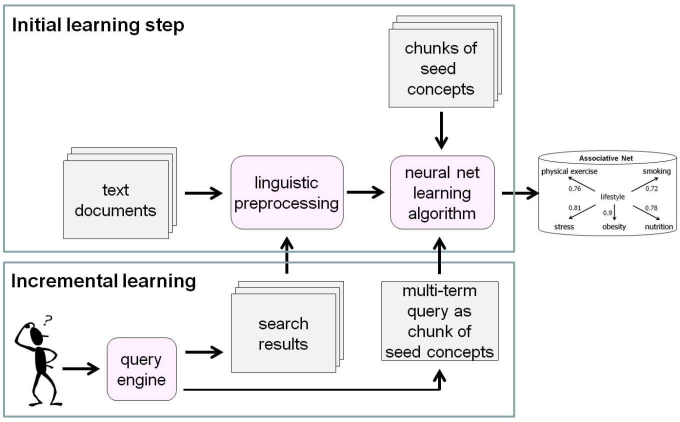

Fig. 3. Scenario for unsupervised learning of associative nets

sulting graph can be queried by means of specific stimulations of neurons. In traditional neural systems it is necessary to set up the appropriate network structure at the beginning according to what is to be learned. Moreover, the supervised learning employed by neural nets such as the perceptron [14, 2] requires that a teacher be present who answers specific questions. Even neural nets that employ unsupervised learning (like those of Hopfield [4] and Kohonen [7]) require a neighborhood function adapted to the learning issue. In contrast, HNN require neither a teacher nor a predefined structure or neighborhood function. In the following we characterize HNN according to their most prominent features.

Exploitation of context
In the SEMPER project, HNN are used for the learning of associative networks. The learning input consists of documents from the application domains, which are broken down into segments rather than entered whole: all sentences are segmented into sub-sentences according to grammatical markers. By way of experimenting we have discovered that a segment should ideally consist of 7 to 8 words. This is in line with findings from cognitive psychology [12]. Breaking down text documents into sub-sentences is the closest possible approximation to the ideal segment size. The contexts given by the sub-sentence segments help the system to learn (see Sec.4.2 for more details): The transitivity of term co-occurrences over the various input contexts (i.e. sub-sentences) are a crucial contribution to creating appropriate associations. This can be compared with the higher-order co-occurrences explored in the context of latent semantic indexing [8].

Continuously evolving structure
The neural structure of a HNN is dynamic and changes constantly in line with neural operations. In the neural context, change means that new neurons are produced or destroyed and connections reinforced or inhibited. Connections that are not used in the processing of input into the net for some time will get gradually weaker. This effect can also be applied to querying, which then results

in the weakening of connections that are rarely traversed for answering a query. In SEMPER, however, the setting is such that the net is not changed by queries.

Asymmetric connections
The connections between the neurons need not be equally strong on both sides and it is not necessary that a connection should exist between all the neurons (cp. Hopfield's correlation matrix [4]).

Spiking neurons
The HNN is stimulated by spikes, i.e. binary signals which either fire or do not. Thresholds do not play a role in HNNs. The stimulus directed at a neuron is coded by the sequence of spikes that arrive at the dendrite.

Massive connectivity
Whenever a new input document is processed, new (groups of) neurons are created which in turn stimulate the network by sending out a spike. Some of the neurons reached by the stimulus react and develop new connections, whereas others, which are less strongly connected, do not. The latter nevertheless contribute to the overall connectivity because they make it possible to reach neurons which could otherwise not be reached. Given the high degree of connectivity a spike can pass through a neuron several times since it can be reached via several paths. The frequency and the chronological sequence in which this happens determine the information that is read from the net.

General purpose
There is no need to define a topology before starting the learning process because the neural structure of the HNN develops on its own. This is why it is possible to retrieve a wide range of information by means of different stimulation patterns. For example, direct associations or association chains between words can be found, the words most strongly associated with a particular word can be identified, etc.

4.2 HNN in SEMPER

Creating an HNN in SEMPER In SEMPER the learning input consists of a sequence of sub-sentences where each word is mapped to the corresponding neuron[2]. If no such neuron exists, it is created. The stimuli which are then sent across the system from each neuron

- build new connections to neurons from the previous input segments,
- reinforce connections that are traversed,
- weaken connections that are not traversed.

Querying an HNN in SEMPER To query associations, the neuron representing a particular search term sends a spike to all outgoing connections. The neurons that a spike passes through are the associated words. As mentioned before, each neuron can be stimulated several times since it can be reached via

[2] Normally, a word is represented by a group of neurons but for the sake of simplicity we assume here that one neuron corresponds to one word

several paths. The stimulation frequency and the distribution of the stimuli over time determine the strength of the association.

Providing a notion of relevance A problem we have encountered when performing our first experiments with learning association nets is the lack of a notion of relevance. Even if the text documents for the initial learning input have been selected manually and are highly representative for the application domain they contain lots of terms that are not specific to the domain of discourse. Since the learning algorithm behaves like a "neutral text reader" and treats all words as equally relevant a large amount of irrelevant associations irrelevant for the application are created. In order to reduce the number of irrelevant associations we provide the learning system with a sense of relevance similar to human readers who always read texts with a particular bias and concentrate on the statements they are interested in.

This we achieve by providing an ontology as background knowledge. Rather than using a fully-fledged background ontology it is sufficient to provide a number of "seed concepts" from which associations to other concepts will grow. Additionally, we group several seed concepts into chunks (see Fig.4). Each chunk includes about 4 to 8 concepts that are semantically interrelated. These chunks typically overlap, i.e. they share concepts. They have the same effect on the neural net learning system as the (sub-)sentences of an input text, i.e. they allow transitive term occurrences to emerge, except that now we have manually predefined contexts that comprise highly interrelated concepts. Their impact on the learning output is considerable. More specifically, the chunks of seed concepts support the learning according to the following principles:

1. A learned association with one seed concept induces a weaker association with the seed concepts in the same chunk. See example shown in Figure 4: A newly learned association between "life philosophy" and "self esteem" triggers weaker associations between "life philosophy" and "obesity", "nutrition" and "physical exercise".
2. A learned association with a seed concept from a chunk C induces a (still) weaker association with all the seed concepts in those chunks with which chunk C shares a concept. See example in Figure 4: A newly learned association between "life philosophy" and "self esteem" (in $chunk_2$) triggers associations between "life philosophy" and all the concepts in $chunk_1$.

As can be seen from the example given, some of the associations induced in steps 1 and 2 are highly relevant (e.g. between "life philosophy" and "back pain", "posture", "stress relief") while others are less relevant (e.g. between "life philosophy" and "physical exercise") or not relevant at all (e.g. between "life philosophy" and "ergonomic workplace"). However, as more associations are added and more input text is processed, the relevant associations get stronger while the irrelevant ones remain at a low association strength.

A chunk of seed concepts can graph-theoretically be interpreted as an n-ary relation between the concepts in the chunk. Thus, the set of chunks forms a *hypergraph*.

The chunks are created by domain experts based on their understanding of which concepts belong together. The only guidelines are that a chunk should preferably consist of at least 4 and no more than 8 concepts, and that chunks should overlap to facilitate the creation of new associations. Overlapping chunks are quite natural because concepts tend to occur in different contexts within a domain of discourse (like "physical exercise" in Figure 4).

User queries in the semantic information portal that contain more than one query term are taken as chunks of seed concepts and added incrementally to the learning system (see Fig.3) – even when they do not meet the ideal criterion of consisting of 4 to 8 concepts.

$chunk_1$ = {physical exercise, back pain, ergonomic workplace, posture, stress relief}
$chunk_2$ = {obesity, nutrition, physical exercise, self esteem}

Fig. 4. Examples of chunks of seed concepts

4.3 Directed Associations

The associations in classical approaches are based on co-occurrence measures and are thus undirected. However, as Figure 2 shows the associations needed in SEMPER for expanding terms in a query are directed. This is important because while it makes sense to include "smoking" when the query contains the term "lifestyle" it does not make sense to expand the search term "smoking" with the term "lifestyle" (see also [10]). This has to be reflected in the association strength: The association from "lifestyle" to "smoking" would have a strong weight while the association from "smoking" to "lifestyle" might actually be present but with a much smaller weight. Therefore it is either not considered at all or filtered out because it is below a given cut-off value.

Due to the characteristics of the HNN-based learning system the learned associations are directed because the learning algorithm relies heavily on the context in which a certain term t occurs. Put differently, the learning system considers which other terms occur significantly frequently when the term t is present, which is by nature asymmetrical. In statistical terms this can be interpreted as a conditional probability, i.e. the probability that e.g. "smoking" occurs in a document when the seed concept "lifestyle" is present. A directed association has more semantics than an undirected association and resembles a semantic implication like hyponomy or partonomy.

5 Evaluation of the Association Net Learning

Measuring the success and effectiveness of SEMPER as a health promotion and disease self-management platform has been described elsewhere [11]. In this paper we focus on the evaluation of the learned association nets.

In principle, we are not interested in the learned association nets by themselves but use them as a means to improve retrieval. Ultimately, the nets will have to be evaluated with regard to their impact on retrieval quality in terms of both the ease of finding relevant results and the recall and precision of the results. However, this presents a tremendous challenge. Therefore, we are currently evaluating the appropriateness of the learned association nets by presenting them to domain experts. Additionally, we are comparing them with association nets learned by a classical algorithm from the same set of input documents. That algorithm uses the cosine of two term vectors to determine the association weight between the two terms. The term vectors are obtained by arranging the input documents into a document-term matrix. We used 280 input documents resulting in about 13000 lemmatized words. The evaluation is being performed according to the following setup:

1. Two sets of terms are generated: T_S the set of all terms from the associations generated in SEMPER that are relevant to our domain of discourse and have an association to another term with at least the weight cw. Accordingly, T_{cos} is the set of all terms from the associations generated by the cosine measure that are relevant and have an association to another term with at least the weight cw. Experiments are done with different values for the cut-off value cw: 0.8, 0.9. Thus, T_S and T_{cos} contain the relevant and most strongly associated words in each of the association nets.
2. Let $weight(t_1, t_2)$ be the weight of the association from term t_1 to term t_2. For each $t_S \in T_S, t_{cos} \in T_{cos}$ where $t_S = t_{cos}$, and for $cw \in \{0.8, 0.9\}$:
   (a) Determine $AH_S = \{t \,|\, t \text{ is associated with } t_S, weight(t_S, t) > cw\}$
   (b) Determine $AH_{cos} = \{t \,|\, t \text{ is associated with } t_{cos}, weight(t_{cos}, t) > cw\}$
   (c) A domain expert determines the number of relevant terms in AH_S and AH_{cos}.
   (d) Let $assoc\text{-}relevant_a(S)$ = number of relevant terms in S and $assoc\text{-}relevant_r(S) = assoc\text{-}relevant_a(S)/|S|$.
   Compare $assoc\text{-}relevant_r(AH_S)$ with $assoc\text{-}relevant_r(AH_{cos})$ (greater values are better).

In brief, the evaluation only considers words that have the strongest association weights to another word and are relevant to the domain of discourse (Step 1). For the words that occur in both lists of most strongly associated words domain experts determine how many of the strongest associations found by SEMPER and the cosine measure, respectively, are relevant for the domain of discourse (Step 2d).

While the evaluation setup described above makes sense in principle, our setting required a minor adaptation. The problem is that the cosine measure only gives reasonable results if calculated on a much larger document collection as we did. The 280 documents we used for learning in SEMPER resulted in many association weights of 1 because many words occur together only in 2 documents and thus have exactly the same co-occurrence pattern. Therefore we slightly changed the computation of the set L_{cos} by excluding all associations with the weight of exactly 1.

We are in the midst of the evaluation (a final version of this paper would include the full evaluation results). Preliminary results are based on the comparison of single associations from AH_S and AH_{cos} by domain experts. The average values of *assoc-relevant*$_r$ for both sets are about the same (approximately 0.8). However, the cosine measure quite often delivers very few associations (between 1 and 4). This is not sufficient to implement semantic retrieval, e.g. via generating a tag cloud. In contrast, our approach based on HNN learned many more associations and is therefore better suited for semantic retrieval. The SEMPER approach has further benefits when compared to classical approaches to association learning:

- As has been discussed in Sec.4.3 classical co-occurrence measures deliver symmetric association weights while in fact this is not what we need for query extension. The HNN uses a different paradigm based on contexts so that the *directed associations* it delivers better meet our requirements.
- Another very important issue is the *computation time* needed for calculating the association measures. In SEMPER, using the HNN, the computationally complex part is the import of the documents. This takes an ordinary notebook about 30 minutes for the 280 documents. Querying association nets is instantaneous. Computation of the cosine measure requires going through all word combinations. This takes about a day for the 280 documents! Other association measures like latent semantic indexing [9] also suffer from a high computational complexity.
- Moreover, due to the nature of the HNN used in SEMPER we can give it further input documents to learn from as we encounter them (cf. Sec.4). Thus, in SEMPER we can apply *incremental learning* which is not possible with co-occurrence measures!
- Unlike other term association measures our approach already delivers reasonable associations from a rather *small number of input texts*.

6 Further work and outlook

So far we have been able to verify the benefits of using the learned association nets for improving retrieval both by automatically extending a user query and by showing the user in a tag cloud terms that are most strongly associated to the query terms. Using an HNN to learn the association nets from input texts in an unsupervised manner proved also highly practicable and valuable. Based on these encouraging results we are planning to continue improving the learning results by enhancing the learning setup. For example, we are looking for a methodology to determine the chunks of seed concepts in a more systematic way. Furthermore, segmenting an input text into sub-sentences for input into the HNN is certainly not ideal because words that succeed each other are not necessarily the ones which are syntactically most closely related. Instead we are considering performing a more in-depth syntactic analysis to obtain a more appropriate segmentation of an input text. Finally, we are experimenting to gain a better understanding of the effect the input texts have on the quality of the learned associations, in particular in terms of number of input texts, narrower or wider scope of contents being covered, and document length.

Finally, we intend to further enhance the quality of the semantic search by taking the individual user profiles into account (see personalization module in Fig.1). Each user will be characterized by a profile which is automatically built up in the background based on user-specific data (like date of diagnosis, education level, etc.) and the queries and interactions of the user as well as users with similar problems. The user profiles will allow interpreting a search term within the user's context. For example, a search for back pain will show (among others) documents about MSD if that is what the user is suffering from.

References

1. Drumond, L., Girardi, R.: A survey of ontology learning procedures. In: de Freitas, F.L.G., Stuckenschmidt, H., Pinto, H.S., Malucelli, A., Corcho, Ó. (eds.) WONTO: 3rd Workshop on Ontologies and their Applications. CEUR Workshop Proceedings, vol. 427. CEUR-WS.org (2008)
2. Grossberg, S.: Contour enhancement, short-term memory, and constancies in reverberating neural networks. Studies in Applied Mathematics 52, 213–257 (1973)
3. Gurney, K.: An Introduction to Neural Networks. CRC Press (1997)
4. Hopfield, J.: Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Sciences of the USA 79(8), 2554–2558 (1982)
5. Kashyap, V., Borgida, A.: Representing the UMLS semantic network using OWL: (Or "What's in a Semantic Web Link?"). In: Proc. 2nd International Semantic Web Conference (ISWC 2003). pp. 1–16. Springer (2003)
6. Kennedy, A., Reeves, D., Bower, P., Lee, V., Middleton, E., Richardson, G., Gardner, C., Gately, C., Rogers, A.: The effectiveness and cost effectiveness of a national lay-led self care support programme for patients with long-term conditions. Journal of Epidemiology Community Health 61(3), 254–261 (2007)
7. Kohonen, T.: Self-Organizing Maps. Springer (1995)
8. Kontostathis, A., Pottenger, W.M.: Detecting patterns in the lsi term-term matrix. In: Proc. ICDM02 Workshop on Foundations of Data Mining and Discovery (2002)
9. Landauer, T.K., Foltz, P.W., Laham, D.: An Introduction to Latent Semantic Analysis. Discourse Processes 25, 259–284 (1998)
10. Lemaire, B., Denhière, G.: Incremental construction of an associative network from a corpus. In: Proceedings of the 26th Annual Meeting of the Cognitive Science Society (CogSci'2004). pp. 825–830 (2004)
11. Maier, E., Reimer, U., Guttormsen, S., Zimmermann, P.: Semper: A web-based support system for patient self-management. In: Proc. 23rd Bled eConference (2010)
12. Miller, G.A.: The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. Psychological Review 63, 81–97 (1956)
13. Reimer, U., Brockhausen, P., Lau, T., Reich, J.: Ontology-based knowledge management at work: The swiss life case studies. In: Davies, J., Fensel, D., van Harmelen, F. (eds.) Towards the Semantic Web. Ontology-driven Knowledge Management, pp. 197–218. John Wiley (2002)
14. Rosenblatt, F.: The perceptron: a probabilistic model for information storage and organization in the brain. Psychological Review 65, 386–408 (1958)