Improving Domain Ontologies by Mining Semantics from Text

Michael Dittenbach[1] **Helmut Berger**[1] **Dieter Merkl**[1,2]

[1]E-Commerce Competence Center – EC3,
Donau-City-Straße 1, A–1220 Wien, Austria

[2]Research Group for Industrial Software Engineering,
Vienna University of Technology
Favoritenstraße 9–11/188, A–1040 Wien, Austria

{michael.dittenbach, dieter.merkl, helmut.berger}@ec3.at

Abstract

The creation and maintenance of domain ontologies is a costly and time-consuming task. With the advent of ontologies being used in many different fields of computer science, developing appropriate algorithms and methods to support or automatize ontology engineering have become an increasingly important goal. Hence, we present a connectionist approach to visualize semantic relations inherent in free-form text documents related to a specific domain. In particular, we exploit word co-occurrences to capture relatedness of words in order to generate numeric representations of the words' contexts. We use the self-organizing map, a well-known neural network model with unsupervised learning function, to map the high-dimensional data onto a two-dimensional representation for convenient browsing. This intuitive view on the domain vocabulary supports the construction and enrichment of domain ontologies by making relevant concepts and their relations evident. We underline this approach with an example from the tourism domain.

Keywords: Tourism Information System, Ontology Enhancement, Text Mining, Clustering, Self-Organizing Map

1 Introduction

Ontologies gained importance in many fields of computer science. Especially for information retrieval systems, ontologies can be a valuable means for representing and modeling domain knowledge to deliver search results of a higher quality. However, a crucial problem is an ontology's increasing complexity with growing size of the application domain. In this paper, we present an approach based on a neural network to assist domain engineers in creating or enhancing ontologies for information retrieval systems.

We show an example from the tourism domain, where free-form text descriptions of accommodations are used as a basis to enrich the ontology of a tourism information retrieval system with highly specialized terms that are hardly found in general purpose thesauri or dictionaries. This information retrieval system allows for accommodation queries being posed in natural language. In order to resolve the relevant concepts that define the constraints for matching accommodations as requested by users, we make use of a lightweight ontology where knowledge about the application domain, i.e. accommodations in tourism, is stored. The ontology covers concepts and the respective words representing them, which can be mapped onto the database schema.

In order to improve the quality of the retrieval results, we exploit information inherent in textual descriptions that are accessible but separated from the structured information the search engine operates on. The vector representations of the terms are created by generating statistics about local contexts of the words occurring in natural language descriptions of accommodations. These descriptions have in common that words belonging together with respect to their semantics are found spatially close together regarding their position in the text. This happens even though the descriptions are written by different authors, i.e. the accommodation providers themselves in case of our application. Therefore, we think that the approach presented in this paper can be applied to a variety of domains, since, for instance, product descriptions generally have similarly structured content. Consider for example, typical computer hardware descriptions where information about, say, storage devices are normally grouped together rather than being intertwined with input and display devices.

More specifically, we use the *self-organizing map (SOM)* to cluster terms relevant to the application domain to provide an intuitive representation of their semantic relations. With this kind of representation at hand, finding synonyms, adding new relations between concepts or detecting new concepts, which would be important to be added to the ontology, is facilitated.

The motivation for the work presented in this paper was, first, that creating the ontologies from scratch and refining them for the tourism information systems presented in Section 2, turned out to be more complex than anticipated. Second, the results of the field trial have shown that we have overlooked quite a number of synonyms for concepts that were therefore not detected by the natural language processing of our system. As already mentioned, regional characteristics, subjective and fuzzy criteria also played an important role in the users' queries. Third, comparing the number of attributes of accommodations at the time we have developed the first system (October 2001) with the current number, it can be seen that an ontology describing such a dynamic area like tourism is exposed to constant change, in this particular case, growth and reorganization. At the time of writing, accommodations can be queried by 159 features compared to 82 at the end of the year 2001. Some features like *massage* or *steam bath* that were then seen as part of *recreation*, are now subsumed under separate topics, i.e. *health* and *vitality*, with a number of related features. In particular, *health*

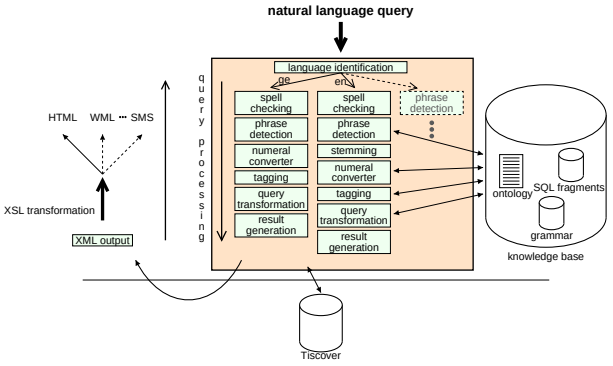


Figure 1: Pipelined system architecture.

includes 27 facilities and services like *hay baths*, *kinesiotherapy*, *herbal bath* or *Swedish massage*, to name but a few. Despite this increased number of wellness-related characteristics, also the variety of facilities offered for conference hosting, e.g. *microphones*, *video projector* or *laser pointer*, has grown to 20 items.

The remainder of the paper is structured as follows. In Section 2 we provide a brief review of our natural language tourism information retrieval system along with some results of a field trial in which the interface has been made publicly accessible. A number of related research projects are presented in Section 3. Section 4 gives an overview of the *SOM* and how it can be used to create a word category map. Following a description of our experiments in Section 5, we provide some concluding remarks in Section 6.

2 A Tourism Information Retrieval System

2.1 System Architecture

We have developed a natural language interface for the largest Austrian web-based tourism platform *Tiscover* (`http://www.tiscover.com`) (Pröll, Retschitzegger, Wagner & Ebner 1998). *Tiscover* is a well-known tourism information system and booking service in Europe that covers already more than 50,000 accommodations in Austria, Germany, Liechtenstein, Switzerland and Italy. Contrary to the original form-based *Tiscover* interface, our natural language interface allows users to search for accommodations throughout Austria by formulating the query in natural language sentences either in German or English. The language of the query is automatically detected and the result is presented accordingly. For the task of natural language query analysis we followed the assumption that shallow natural language processing is sufficient in restricted and well-defined domains. In particular, our approach relies on the selection of query concepts, which are modeled in a domain ontology, followed by syntactic and semantic analysis of the parts of the query where the concepts appear.

The system architecture is depicted in Figure 1. To improve retrieval performance, we used a phonetic algorithm to find and correct orthographic errors and misspellings. It is furthermore an important issue to automatically identify proper names consisting of more than one word, e.g. *"Gries am Brenner"*, without having the user to enclose it with quotes. This also applies to phrases and multi-word denominations like *"city center"* or *"youth hostel"*, to name but a few. In the next query processing step, the relevant concepts and modifiers are tagged. For this purpose, we have developed an XML-based ontology covering the semantics of domain specific concepts and modifiers and describing linguistic concepts like synonymy. Additionally, a lightweight grammar describes how particular concepts may be modified by prepositions and adverbial or adjectival structures that are also specified in the ontology. Finally, the query is transformed into an SQL statement to retrieve information from the database. The tagged concepts and modifiers together with the rule set and parameterized SQL fragments, also defined in the knowledge base, are used to create the complete SQL statement reflecting the natural language query. A generic XML description of the matching accommodations is transformed into device-dependent output, customized to fit screen size and bandwidth.

Our information retrieval system covers a part of the *Tiscover* database, that, as of October 2001, provides access to information about 13,117 Austrian accommodations. These are described by a large number of characteristics including the respective numbers of various room types, different facilities and services provided in the accommodation, or the type of food. The accommodations are located in 1,923 towns and cities that are again described by various features, mainly information about sports activities offered, e.g. mountain biking or skiing, but also the number of inhabitants or the sea level. The federal states of Austria are the higher-level geographical units. For a more detailed report on the system we refer to (Dittenbach, Merkl & Berger 2003).

2.2 Implications of a Field Trial

The field trial was carried out during ten days in March 2002. During this time our natural language interface was promoted on and linked from the main *Tiscover* page. We obtained 1,425 unique queries through our interface, i.e. equal queries from the same client host have been reduced to one entry in the query log to eliminate a possible bias for our evaluation of the query complexity.

In more than a half of the queries, users formulated complete, grammatically correct sentences, about one fifth were partial sentences and the remaining set were keyword-type queries. Several of the queries consisted of more than one sentence. This confirms our assumption that users accept the natural language interface and are willing to type more than just a few keywords to search for information. More than this, a substantial portion of users is typing complete sentences to express their information needs.

To inspect the complexity of the queries, we considered the number of concepts and the usage of modifiers like *and*, *or*, *not*, *near* and some combinations of those as quantitative measures. We found out that the level of sentence complexity is not very high. This confirms our assumption that shallow text parsing is sufficient to analyze the queries emerging in a limited domain like tourism. We found out that information about regions and local attractions is inevitable and has to be integrated in such systems. We also noticed that users' queries contained vague or highly subjective criteria like *romantic*, *cheap* or *within walking distance to*. Even *wellness*, a term broadly used in tourism nowadays, is far from being exactly defined. A more detailed evaluation of the results of the field trial can be found in (Dittenbach, Merkl & Berger 2002).

Even more important for the research described in this paper, it turned out that a deficiency of our ontology was the lack of diversity of the terminology. This issue and the lessons learned during the creation of the initial ontology have led to the step of exploiting textual descriptions, i.e. web pages, of the hotels.

Besides the structured information about the accommodations, the web pages describing the accommodations offer a lot of additional information in form of natural language descriptions. Hence, the words occurring in these texts constitute a very specialized vocabulary for this domain. The next obvious step is to exploit this information to enhance the domain ontology for the information retrieval system. Due to the size of this vocabulary, some intelligent form of representation is necessary to express semantic relations between the words.

2.3 Alternative Knowledge Representation

The analysis of real-world queries that were received during the field trial described in the previous section has shown some deficiencies regarding the ontology of the original system. First, modeling fine-grained similarity relations between concepts is rather difficult with the ontology of the original system since no degree of similarity between two concepts could be defined. Additionally, with the rather fixed structure defining how the concepts are organized, some of the concepts the users asked for could be implemented either only in a cumbersome way or not at all. Especially highly subjective search criteria such as *romantic* posed a difficult task. Hence, we have developed a more flexible way of representing domain knowledge that suits the needs of our natural language information retrieval system.

This approach is based on *associative networks*. An associative network is a generic network consisting of pieces of information represented by nodes that are connected with either unlabeled or labeled links that can also be weighted to express a certain strength of the relations. Associative networks have quite a tradition in information retrieval and were first used to model relations between terms and terms, between terms and documents and between documents and documents.

A processing framework for associative networks is *spreading activation*, which emerged from the field of cognitive sciences. The basic idea is to distribute activation potentials expressed by numerical values along the directed, weighted links connecting the nodes of the network. This, usually, is an iterative process, where during one iteration a *pulse* is triggered and a termination criterion is checked. In other words, the activation is transferred from currently activated nodes along directed connections to the immediately adjacent nodes until, e.g., a certain number of iterations is reached.

In information retrieval, spreading activation has been used to process associative networks in order to retrieve a ranking of relevant information with regard to a search request (Cohen & Kjeldsen 1987, Salton & Buckley 1988, Crestani 1997). Pragmatically speaking, initial activations in the network are assigned to nodes that represent the terms in the query. Then, for a certain number of iterations the activation is spread across the network and the documents are ranked according to the final activation potentials.

In our framework, the nodes represent domain-relevant concepts such as *hotel*, *sauna*, *vegetarian cuisine* or *playground*. A list of words, i.e. synonyms, is assigned to each concept. Furthermore, the concepts are assigned to the respective accommodations that *provide* them. Additionally, we have integrated information on geographical places such as cities and federal states into the network structure. Hence, cities are also represented by nodes and are connected to nearby cities with the weights of the links expressing closeness. Using this approach relieves us from the burden of treating geographical distances separately

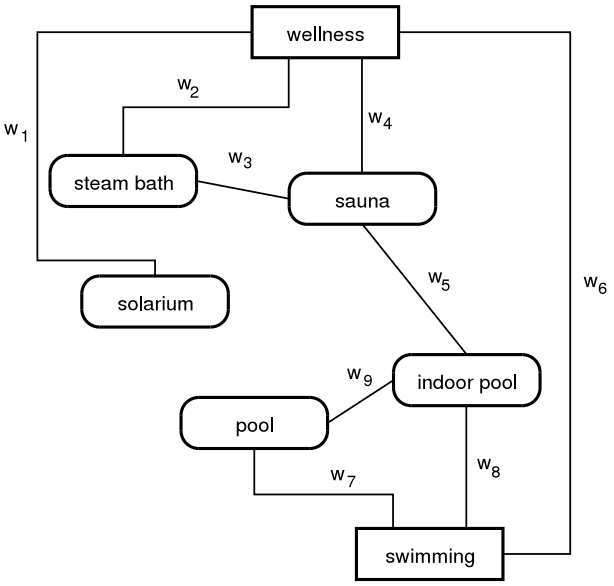


Figure 2: Detail of alternative knowledge representation using an associative network. The nodes (concepts) are connected by unlabeled, weighted links.

when analyzing queries including geographical references. As the second building block of our network, we use unlabeled, weighted links connecting the concepts with the weights defining the degree of similarity between two concepts.

An example of the network structure is given in Figure 2 where a small portion of the network related to the concept of *wellness* is shown. The nodes depicted by round-cornered boxes represent *concrete* concepts that are actually present in the database, i.e. services, facilities, accommodation types and the like. Boxes such as the one labeled *wellness* represent *abstract* concepts that are not directly found in the database but are related to one or more concrete concepts. This provides the possibility to answer a broader range of queries than compared to a form-based search interface where only the actual fields in the database can be selected.

After identifying of the relevant query terms an initial activation potential is set at the corresponding nodes. Then, the spreading activation process starts and iterates for a certain number of times. The final activation potentials in the network are summed up for each accommodation and the one best representing the requirements of the user's request is ranked highest. For an in-depth description of associative networks and spreading activation basics as well as our specific implementation we refer to (Berger, Dittenbach & Merkl 2003).

Even though this representation provides substantial advantages over the original approach regarding query processing complexity, the creation of the network is about as costly and time-consuming as the creation of a more customary type of ontology. Yet, the straightforward representation of nodes being connected by weighted links is especially suitable for the research presented in the remainder of the paper.

3 Related Work

Since ontologies became more frequently used in computer science, the number of approaches and algorithms for automatically or semi-automatically constructing and enriching ontologies has also increased.

In particular, the topic of mining concepts and relations from text corpora in the context of thesaurus and dictionary generation has been given a fresh impulse. We outline a few publications dealing with the topic of extracting knowledge from texts for thesaurus and ontology engineering.

Church & Hanks (1990) propose a measure to estimate word association norms based on mutual information. This *association ratio* can be used to estimate the level of associativity between two words based on the probability of occurring jointly in a fix-sized window of words and the probability of independent occurrence. This method of measurement can be used in lexicography to support the analysis of concordances for complex words.

A framework for maintaining domain-specific ontologies based on reasoning and hypothesis generation is reported in (Hahn & Schnattinger 1997) and (Hahn & Schnattinger 1998). The goal of this approach is to integrate new knowledge items occurring in a text into an already existing concept hierarchy. Based on a linguistic analysis of unknown lexical items, several hypotheses are generated and then ranked according to a plausibility measure derived from existing domain knowledge.

Grefenstette (1992) presents a system that syntactically analyzes texts to extract contexts for calculating a similarity measure between two terms. The user can specify a set of context relations that are extracted by the system according to the information gained by a linguistic analysis consisting of several processing steps including morphological analysis, grammatical disambiguation, noun and verb phrase detection and relation extraction. Then, the relations are measured using the Jaccard measure to compare terms relative to the contexts they share.

Velardi, Fabriani & Missikoff (2001), Missikoff, Navigli & Velardi (2002*a*) and Missikoff, Navigli & Velardi (2002*b*) describe a text mining tool for term extraction to support the ontology construction process. The system uses a rule set to detect named entities and extracts candidate concepts using shallow language processing techniques. Terms are then assessed according to some plausibility measure based on mutual information in order to rank them. Finally, the extracted terms are organized into a concept hierarchy using WordNet (Fellbaum 1998) and SemCor (Miller, Leacock, Tengi & Bunker 1993), a semantic concordance package where texts have been manually tagged with WordNet meanings.

A series of publications by Mädche & Staab (2000*a*), Mädche & Staab (2000*b*) and Mädche & Staab (2000*c*) deals with the topic of mining ontologies from text in the context of the *Semantic Net* (Berners-Lee, Hendler & Lassila 2001). Contrary to some of the work mentioned above, emphasis is put on extraction of non-taxonomic relations, i.e. relations that do not describe hierarchical *is-a* relations but rather relations such as *part-of* or *is-located-in*,to give an example. An algorithm for discovering generalized association rules is used to detect relations between concepts and assign them confidence values. The authors present an example from the tourism domain where relations such as *area – accommodation* or *room – furnishing* are detected. A closely related system, i.e. GETESS, described by Staab, Braun, Düsterhöft, Heuer, Klettke, Neumann, Prager, Pretzel, Schnurr, Struder, Uszkoreit & Wrenger (1999), is a system that gathers information from multiple sources on the Internet, extracts semantic relations, and provides a uniform natural language interface to query this information.

Hearst presents a method for automatic extraction of hyponym relations from text corpora (Hearst 1992, Hearst 1998). Hyponyms, i.e. *is-a* relations, are detected through lexicosyntactic patterns defining the position of noun phrases and certain keywords such as *including*, *especially* or *such as*. A whole set of patterns has been defined that are typically representing *is-a* relations and yield good results in detecting them. Hearst argues that the WordNet lexical database lacks many relations and proper nouns that would be useful for certain applications but do not suit a general purpose lexical database. Therefore, it is often necessary to enhance the database for specific purposes which is a costly task as already mentioned. Consequently, the major goal of this work is to automatically suggest relevant hyponymic term relations from domain-relevant documents to be integrated into WordNet.

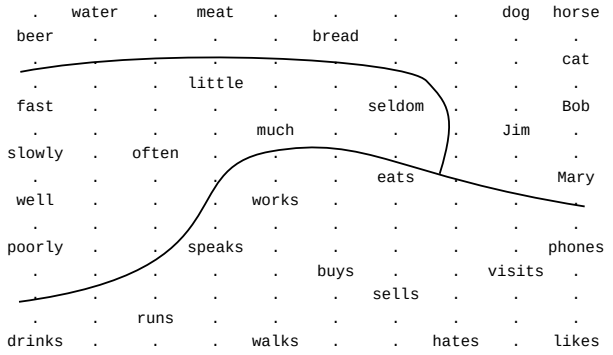


Figure 3: Semantic map of the original experiments by Ritter and Kohonen. The manually drawn cluster boundaries separate syntactic word classes.

Pike & Gahegan (2003) use the *self-organizing map* as a means for visualizing topics covered by discussions of experts in a web-based Delphi tool. Using this approach, major topics and issues that form in the course of the discussions can be identified. The *SOM* also displays relations between persons participating in a discourse, based on the content of their arguments.

For further reading we suggest the reviews by Ding & Foo (2002*a*) and Ding & Foo (2002*b*) who provide an extensive overview of ontology research and development covering ontology generation as well as mapping and evolving.

4 Semantic Word Clustering

4.1 Semantic Context Encoding

Ritter and Kohonen (Ritter & Kohonen 1989) have shown that it is possible to cluster terms according to their syntactic category by encoding word contexts of terms in an artificial data set of three-word sentences that consist of nouns, verbs and adverbs, such as, e.g. *"Jim speaks often"* and *"Mary buys meat"*. The resulting maps clearly showed three main clusters corresponding to the three word classes. It should furthermore be noted that within each cluster, the words of a class were arranged according to their semantic relation. For example, the adverbs *poorly* and *well* were located closer together on the map than *poorly* and *much*, the latter was located spatially close to *little* as shown in Figure 3. An example from a different cluster would be the verbs *likes* and *hates*.

Other experiments using a collection of fairy tales by the Grimm Brothers have shown that this method works well with real-world text documents (Honkela, Pulkki & Kohonen 1995). The terms on the *SOM* were divided into three clusters, namely nouns, verbs and all other word classes. Again, inside these clus-

ters, semantic similarities between words were mirrored. The results of these experiments have later been used as a basis to reduce the vector dimensionality for document clustering in the WEBSOM project (Kaski, Honkela, Lagus & Kohonen 1998). Here, a word category map has been trained with the terms occurring in the document collection to subsume words with similar context to one semantic category. These categories, obviously fewer than the number of all words of the document collection, have then been used to create document vectors for clustering. Since new methods of dimensionality reduction have been developed, the word category map has been dropped for this particular purpose (Kohonen, Kaski, Lagus, Salojärvi, Honkela, Paatero & Saarela 2000).

Nevertheless, since our objective is to disclose semantic relations between words we have decided to use word category maps. For training a *self-organizing map* in order to organize terms according to their semantic similarity, these terms have to be encoded as n-dimensional numerical vectors. The random values of the vector components are drawn from a uniform probability distribution, thus, being statistically independent. As shown by Honkela (1997), these random vectors are quasi-orthogonal in case of n being large enough. Consequently, unwanted geometrical dependence of the word representation can be avoided. This is a necessary condition, because otherwise the clustering result could be dominated by random effects overriding the semantic similarity of words.

In Figure 4, reproduced from Honkela (1997), the distributions of pairwise inner products of the random vectors for different dimensionalities are depicted. It can be seen that the vectors are not perfectly independent but bear enough statistical independence to be used for our experiments even with a dimensionality as low as 90. We have tested the quasi-orthogonality of the random vectors created for our experiments and came to the same results. The distribution of the pairwise inner products of 90-dimensional random vectors showed a standard deviation of 0.139, about the same as can be derived from the corresponding distribution in Figure 4.

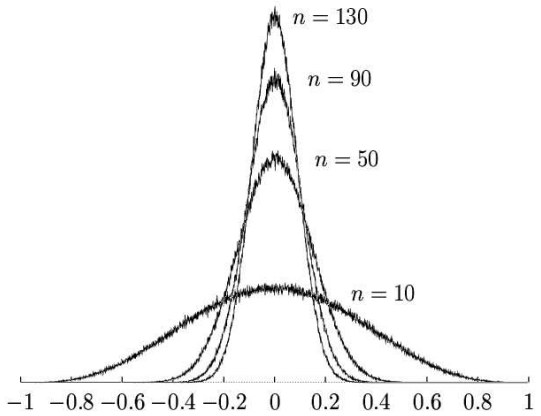


Figure 4: Distribution of pairwise inner products of the random vectors with n dimensions.

We assume that, in textual descriptions dominated by enumerations, semantic similarity is captured by contextual closeness within the description. For example, when arguing about the attractions offered for children, things like a *playground*, a *sandbox* or the availability of a *baby-sitter* will be mentioned close together regarding the word order. Analogously, the same is true for recreation equipment like a *sauna*, a *steam bath* or an *infrared cabin*. To capture this contextual closeness, we use word windows where a particular word i is described by the set of words that appear a fixed number of words before and after word i in the textual description. Given that every word is represented by a unique n-dimensional random vector, the context vector of a word i is built as the concatenation of the average of all words preceding as well as succeeding word i. Technically speaking, an $n \times N$-dimensional vector x_i representing word i is a concatenation of vectors $x_i^{(d_j)}$ denoting the mean vectors of terms occurring at the set of displacements $\{d_1, \ldots, d_N\}$ of the term as given in Equation 1. Consequently, the dimensionality of x_i is $n \times N$. This kind of representation has the effect that words appearing in similar contexts are represented by similar vectors in a high-dimensional space.

$$x_i = \begin{bmatrix} x_i^{(d_1)} \\ \vdots \\ x_i^{(d_N)} \end{bmatrix} \quad (1)$$

To illustrate this vector generation with an example, consider x_i in Equation 2 as the vector describing a term i by the average vectors of its immediate neighbors. In other words, the average contexts of words at displacements -1 and $+1$ constitute the contextual description.

$$x_i = \begin{bmatrix} x_i^{(-1)} \\ x_i^{(1)} \end{bmatrix} \quad (2)$$

Adhering to our tourism domain, consider, for example, the term *Skifahren* (skiing). The set of words occurring directly before the term at displacement -1 consists of words like *Langlaufen* (cross country skiing), *Rodeln* (toboggan), *Pulverschnee* (powder snow) or *Winter* to name but a few. By averaging the respective vectors representing these terms, a statistical model of word contexts is created.

4.2 Self-Organizing Map Principles

The *self-organizing map (SOM)* (Kohonen 1982, Kohonen 1995) is an unsupervised neural network providing a mapping from a high-dimensional input space to a usually two-dimensional output space while preserving topological relations as faithfully as possible. The *SOM* consists of a set of units arranged in a two-dimensional grid, with a weight vector $m_i \in \Re^n$ attached to each unit i. Data from the high-dimensional input space, referred to as input vectors $x \in \Re^n$, are presented to the *SOM* and the activation of each unit for the presented input vector is calculated using an activation function. Commonly, the Euclidean distance between the weight vector of the unit and the input vector serves as the activation function, i.e. the smaller the Euclidean distance, the higher the activation.

In the next step the weight vector of the unit showing the highest activation is selected as the winner and is modified as to more closely resemble the presented input vector. Pragmatically speaking, the weight vector of the winner is moved towards the presented input by a certain fraction of the Euclidean distance as indicated by a time-decreasing learning rate $\alpha(t)$ as shown in Equation 3.

$$m_i(t+1) = m_i(t) + \alpha(t) \cdot h_{ci}(t) \cdot [x(t) - m_i(t)] \quad (3)$$

Thus, this unit's activation will be even higher the next time the same input signal is presented. Furthermore, the weight vectors of units in the neighborhood of the winner are modified accordingly as described by

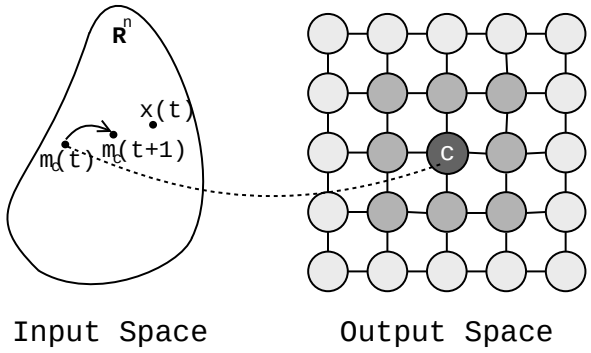


Figure 5: Adaption of weight vectors during SOM training

a neighborhood function $h_{ci}(t)$ (cf. Equation 4), yet to a less strong amount as compared to the winner. The strength of adaptation depends on the Euclidean distance $||r_c - r_i||$ between the winner c and a unit i regarding their respective locations $r_c, r_i \in \Re^2$ on the 2-dimensional map and a time-decreasing parameter σ.

$$h_{ci}(t) = \exp\left(-\frac{||r_c - r_i||^2}{2 \cdot \sigma^2(t)}\right) \quad (4)$$

Starting with a rather large neighborhood for a general organization of the weight vectors, this learning procedure finally leads to a fine-grained topologically ordered mapping of the presented input signals. Similar input data are mapped onto neighboring regions on the map.

A simple graphical representation of the *self-organizing map* architecture and its learning process is provided in Figure 5. In this figure the output space consists of a square of 25 units, depicted as circles, forming a grid of 5×5 units. One input vector $x(t)$ is randomly chosen and mapped onto the grid of output units. In the second step of the learning process, the winner c showing the highest activation is selected. Consider the winner being the unit depicted as the dark gray unit labeled c in the figure. The weight vector $m_c(t)$ of the winner c (the middle unit of the map as indicated by the dashed line) is now moved towards the current input vector. This movement is symbolized in the input space on the left-hand side. As a consequence of the adaptation, unit c will produce an even higher activation with respect to the input pattern x at the next learning iteration, $t + 1$, because the unit's weight vector, $m_c(t + 1)$, is now closer to the input pattern x in terms of the input space. Neighboring units that are also subject to adaptation are depicted as shaded units in the figure. The shading of the various units corresponds to the amount of adaptation, and thus, to the spatial width of the neighborhood-kernel. Generally, units in close vicinity of the winner are adapted more strongly, and consequently, they are depicted with a darker shade in the figure.

5 Experiments

5.1 Data

The data provided by *Tiscover* consist, on the one hand, of structured information as described in Section 2, and, on the other hand, of free-form texts describing the accommodations. Because accommodation providers themselves enter the data into the system, the descriptions vary in length and style and are are not uniform or even quality controlled regarding spelling. HTML tags, which are allowed to format the descriptions, had to be removed to have plain-text files for further processing. For the experiments presented hereafter, we used the German descriptions of the accommodations since they are more comprehensive than the English ones. Especially small and medium-sized accommodations provide only a very rudimentary English description, many being far from correctly spelled.

To give a practical example and to point out the different writing styles, consider Figure 6 showing two sample descriptions of a hotel and a private lodging provider respectively. On the left-hand side in Figure 6(b) we can see a comparatively extensive description of a sport hotel. It contains enumerations of sporting activities that can be undertaken on the hotel grounds, e.g. tennis, beach volleyball, cycling, horseback riding and many more. Furthermore, the description underlines the possibilities to relax by describing the different swimming facilities, different types of saunas and other wellness-related facilities and services. Please note the semantic grouping of the characteristics regarding the position in the text. The private accommodation in Figure 6(b) on the other hand, just mentions the cozy and family-friendly atmosphere and the location close to a ski lift tying the accommodation to a very large skiing area that can be conveniently reached with.

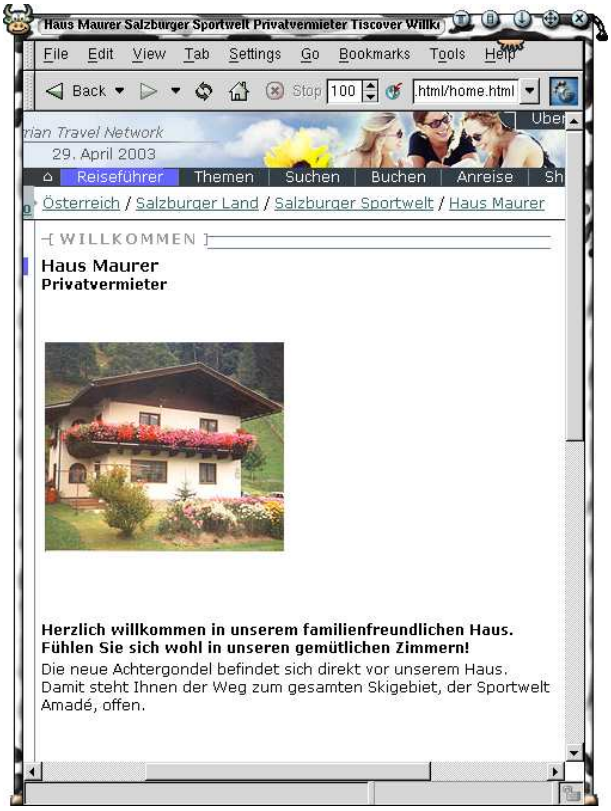
It has been shown with a text collection consisting of fairy tales that, with free-form text documents, the word categories dominate the cluster structure of such a map (Honkela et al. 1995). To create semantic maps primarily reflecting the semantic similarity of words rather than categorizing word classes, we removed words other than nouns and proper names.

Therefore, we used the characteristic, unique to the German language, of nouns starting with a capital letter to filter the nouns and proper names occurring in the texts. Obviously, using this method, some other words like adjectives, verbs or adverbs at the beginning of sentences or in improperly written documents are also filtered. Contrarily, some nouns can be missed, too. A different method of determining nouns or other relevant word classes, especially for languages other than German, would be part-of-speech (POS) taggers. But even state-of-the-art POS taggers do not reach an accuracy of 100% (Manning & Schütze 2000). For the rest of this section, the numbers and figures presented, refer to the already preprocessed documents, if not stated otherwise.

The collection consists of 12,471 documents with a total number of 481,580 words, i.e. on average, a description contains about 39 words. For the curious reader we shall note that not all of the 13,117 accommodations in the database provide a textual description. The vocabulary of the document collection comprises 35,873 unique terms, but for the sake of readability of the maps we reduced the number of terms by excluding those occurring less than ten times in the whole collection. Consequently, we used 3,662 terms for creating the semantic maps.

In Figure 7, a natural language description of a holiday flat in Vienna is shown. Beginning with the location of the flat, the accessibility by public transport is mentioned, followed by some terms describing the dining and living room together with enumerations of the respective furniture and fixtures. Other parts of the flat are the sleeping room, a single bedroom and the bathroom. In this particular example, the only words not being nouns or proper names are the determiner *Die* and the preposition *In* at the beginning of sentences. For the sake of convenience, we have provided an English translation in Figure 8.

(a) An opulently equipped hotel praising its sports-related facilities and the possibilities of relaxation.



(b) A rather short description of a private accommodation simply stating the cozy atmosphere and the proximity to the ski lift.

Figure 6: Two different accommodation descriptions.

Die Ferienwohnung Lage Stadtrand Wien Bezirk Mauer In Gehminuten Schnellbahn Fahrminuten Wien Mitte Stadt Die Wohnung Wohn Eßraum Kamin SAT TV Küche Geschirrspüler Schlafzimmer Einzelbetten Einbettzimmer Badezimmer Wanne Doppelwaschbecken Dusche Extra WC Terrasse Sitzgarnitur Ruhebetten Die Ferienwohnung Aufenthalt Wünsche

Figure 7: A sample description of a holiday flat in a suburb of Vienna after removing almost all words not being nouns or proper names.

the$_{(fem.)}$, holiday flat, location, outskirts, Vienna, district, Mauer, in, minutes to walk, urban railway, minutes to drive, Wien Mitte (station name), city, the$_{(fem.)}$, flat, living, dining room, fireplace, satellite tv, kitchen, dishwasher, sleeping room, single beds, single-bed room, bathroom, bathtub, double washbasin, shower, separate toilet, terrace, chairs and table, couches, the$_{(fem.)}$, holiday flat, stay, wishes

Figure 8: English translation of the description shown in Figure 7.

5.2 Results

For encoding the terms we have chosen 90-dimensional random vectors. The vectors used for training the semantic map depicted in Figure 9 were created by using a context window of length four, i.e. two words before and two words after a term. But instead of treating all four sets of context terms separately, we have put terms at displacements -2 and -1 as well as those at displacements $+1$ and $+2$ together. Then the average vectors of both sets were calculated and finally concatenated to create the 180-dimensional context vectors. Further experiments have shown that this setting yielded the best result. For example, using a context window of length four but considering all displacements separately, i.e. the final context vector has length 360, has led to a map where the clusters were not as coherent as on the map shown below. A smaller context window of length two, taking into account only the surrounding words at displacements -1 and $+1$, had a similar effect. This indicates that the amount of text available for creating such a statistical model is crucial for the quality of the resulting map. By subsuming the context words at displacements before as well as after the word, the disadvantage of having an insufficient amount of text can be alleviated, because having twice the number of contexts with displacements -1 and $+1$ is simulated. Due to the enumerative nature of the accommodation descriptions, the exact position of the context terms can be disregarded.

The *self-organizing map* depicted in Figure 9 consists of 20 × 20 units. Due to space considerations, only a few clusters can be detailed in this description and enumerations of terms in a cluster will only be exemplary. The semantic clusters shaded gray have been determined by manual inspection. They consist of very homogeneous sets of terms related to distinct aspects of the domain. The parts of the right half of the map that have not been shaded, mainly contain proper names of places, lakes, mountains, cities or accommodations. However, it shall be noted, that

e.g. names of lakes or mountains are homogeneously grouped in separate clusters.

In the upper left corner, mostly verbs, adverbs, adjectives or conjunctions are located. These are terms that have been inadvertently included in the set of relevant terms as described in the previous subsection. In the upper part of the map, a cluster containing terms related to pricing, fees and reductions can be found. Other clusters in this area predominantly deal with words describing types of accommodation and, in the top-right corner a strong cluster of accommodation names can be found. On the right-hand border of the map, geographical locations, such as *central*, *outskirts*, or *close to a forest* have been mapped, and a cluster containing skiing- and mountaineering-related terms is also located there.

A dominant cluster containing words that describe room types, furnishing and fixtures can be found in the lower left corner of the map. The cluster labeled *advertising terms* in the bottom-right corner of the map, predominately contains words that are found at the beginning of the documents where the pleasures awaiting the potential customer are described.

Interesting inter-cluster relations showing the semantic ordering of the terms can be found in the bottom part of the map. The cluster labeled *farm* contains terms describing, amongst other things, typical goods produced on farms like, *organic products*, *jam*, *grape juice* or *schnaps*. In the upper left corner of the cluster, names of farm animals (e.g. *pig*, *cow*, *chicken*) as well as animals usually found in a petting zoo (e.g. *donkey*, *dwarf goats*, *cats*, *calves*) are located. This cluster describing animals adjoins a cluster primarily containing terms related to children, toys and games. Some terms are *playroom*, *tabletop soccer*, *sandbox* and *volleyball*, to name but a few.

This representation of a domain vocabulary supports the construction and enrichment of domain ontologies by making relevant concepts and their relations evident. To provide an example, we found a wealth of terms describing sauna-like recreational facilities having in common that the vacationer sojourns in a closed room with well-tempered atmosphere, e.g. *sauna*, *tepidarium*, *bio sauna*, *herbal sauna*, *Finnish sauna*, *steam sauna*, *thermarium* or *infrared cabin*. On the one hand, major semantic categories identified by inspecting and evaluating the semantic map can be used as a basis for a top-down ontology engineering approach. On the other hand, the clustered terms, extracted from domain-relevant documents, can be used for bottom-up engineering an existing ontology.

6 Conclusions

In this paper, we have presented a method, based on the *self-organizing map*, to support the construction and enrichment of domain ontologies. The words occurring in free-form text documents from the application domain are clustered according to their semantic similarity based on statistical context analysis. More precisely, we have shown that when a word is described by words that appear within a fix-sized context window, semantic relations of words unfold in the *self-organizing map*. Thus, words that refer to similar objects can be found in neighboring parts of the map. The two-dimensional map representation provides an intuitive interface for browsing through the vocabulary to discover new concepts or relations between concepts that are still missing in the ontology. This approach is especially suitable for finding new concepts and relations to be added to the associative network, because types of relations are not made visible by the *self-organizing map*. We have illustrated this approach with an example from the tourism domain. The clustering results revealed a number of relevant tourism-related terms that can now be integrated into the ontology to provide better retrieval results when searching for accommodations. Especially in a commercial setting such as the one of *Tiscover*, the quality of the knowledge base is crucial. Hence, we think that an approach supporting the manual construction and enhancement of an ontology in such a system is very important and provides a useful tool in addition to (semi-)automatic ontology learning approaches. We achieved this by analysis of self-descriptions written by accommodation providers, thus assisting substantially the costly and time-consuming process of ontology engineering.

References

Berger, H., Dittenbach, M. & Merkl, D. (2003), Activation on the move: Querying tourism information via spreading activation, *in* 'Proceedings of the 14th International Conference on Database and Expert Systems Applications (DEXA 2003)', Prague, Czech Republic. accepted for publication.

Berners-Lee, T., Hendler, J. & Lassila, O. (2001), 'The semantic web', *Scientific American* .

Church, K. W. & Hanks, P. (1990), 'Word assiciation norms, mutual information, and lexicography', *Computational Linguistics* **16**(1), 22–29.

Cohen, P. R. & Kjeldsen, R. (1987), 'Information retrieval by constrained spreading activation in semantic networks', *Information Processing and Management* **23**(4), 255–268.

Crestani, F. (1997), 'Application of spreading activation techniques in information retrieval', *Artificial Intelligence Review* **11**(6), 453–582.

Ding, Y. & Foo, S. (2002*a*), 'Ontology research and development. Part 1: A review of ontology generation', *Journal of Information Science* **28**(2), 123–136.

Ding, Y. & Foo, S. (2002*b*), 'Ontology research and development. Part 2: A review of ontology mapping and evolving', *Journal of Information Science* **28**(2), 375–388.

Dittenbach, M., Merkl, D. & Berger, H. (2002), What customers really want from tourism information systems but never dared to ask, *in* 'Proc. of the 5th Int'l Conference on Electronic Commerce Research (ICECR-5)', Montreal, Canada.

Dittenbach, M., Merkl, D. & Berger, H. (2003), A natural language query interface for tourism information, *in* A. J. Frew, M. Hitz & P. O'Connor, eds, 'Proceedings of the 10th International Conference on Information Technologies in Tourism (ENTER 2003)', Springer-Verlag, Helsinki, Finland, pp. 152–162.

Fellbaum, C., ed. (1998), *WordNet: An Electronic Lexical Database*, MIT Press.

Grefenstette, G. (1992), Sextant: Exploring unexplored contexts for semantic extraction from syntactic analysis, *in* 'Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics (ACL 1992)', Newark, DE, pp. 324–326.

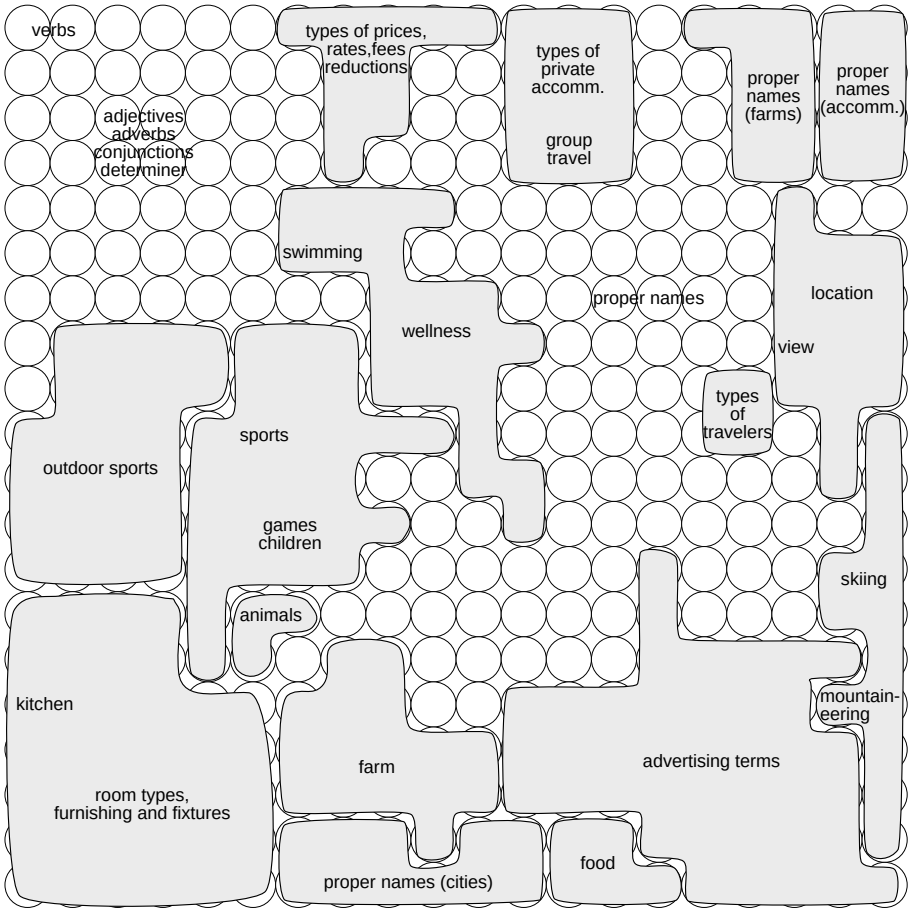


Figure 9: A self-organizing semantic map of terms in the tourism domain with labels denoting general semantic clusters. The cluster boundaries have been drawn manually.

Hahn, U. & Schnattinger, K. (1997), Knowledge mining from textual sources, *in* F. Golshani & K. Makki, eds, 'Proceedings of the 6th International Conference on Information and Knowledge Management (CIKM 1997)', ACM Press, Las Vegas, NV, pp. 83–90.

Hahn, U. & Schnattinger, K. (1998), Towards text knowledge engineering, *in* 'Proceedings of the 15th National Conference on Artificial Intelligence (AAAI 1998)', Madison, WI, pp. 524–531.

Hearst, M. (1992), Automatic acquisition of hyponyms from large text corpora, *in* 'Proceedings of the 14th International Conference on Computational Linguistics (COLING 1992)', Nantes, France.

Hearst, M. (1998), Automated discovery of wordnet relations, *in* C. Fellbaum, ed., 'WordNet: An Electronic Lexical Database', MIT Press.

Honkela, T. (1997), Self-Organizing Maps in Natural Language Processing, PhD thesis, Helsinki University of Technology.

Honkela, T., Pulkki, V. & Kohonen, T. (1995), Contextual relations of words in grimm tales, analyzed by self-organizing map, *in* F. Fogelman-Soulie & P. Gallinari, eds, 'Proceedings of the International Conference on Artificial Neural Networks (ICANN 1995)', EC2 et Cie, Paris, France, pp. 3–7.

Kaski, S., Honkela, T., Lagus, K. & Kohonen, T. (1998), 'WEBSOM–self-organizing maps of document collections', *Neurocomputing, Elsevier* **21**, 101–117.

Kohonen, T. (1982), 'Self-organized formation of topologically correct feature maps', *Biological Cybernetics* **43**.

Kohonen, T. (1995), *Self-organizing maps*, Springer-Verlag, Berlin.

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**(3), 574–585.

Mädche, A. & Staab, S. (2000*a*), Discovering conceptual relations from text, *in* W. Horn, ed., 'Proceedings of the 14th European Conference on Artificial Intelligence (ECAI 2000)', IOS Press, Berlin, Germany.

Mädche, A. & Staab, S. (2000*b*), Mining ontologies from text, *in* R. Dieng & O. Corby, eds, 'Proceedings of the 12th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2000)', number 1937 *in* 'LNAI', Springer-Verlag, Juan-les-Pins, France, pp. 189–202.

Mädche, A. & Staab, S. (2000*c*), Semi-automatic engineering of ontologies from text, *in* 'Proceedings of the 12th International Conference on Software Engineering and Knowledge Engineering (SEKE 2000)', Chicago, IL.

Manning, C. & Schütze, H. (2000), *Foundations of statistical natural language processing*, MIT Press.

Miller, G. A., Leacock, C., Tengi, R. & Bunker, R. T. (1993), A semantic concordance, *in* 'Proceedings of the 3rd DARPA Workshop on Human Language Technology', Morgan Kaufmann, Plainsboro, NJ, pp. 303–308.

Missikoff, M., Navigli, R. & Velardi, P. (2002*a*), 'Integrated approach to web ontology learning and engineering', *IEEE Computer* pp. 60–63.

Missikoff, M., Navigli, R. & Velardi, P. (2002*b*), The usable ontology: An environment for building and assessing a domain ontology, *in* 'Proceedings of the 1st International Semantic Web Conference (ISWC 2002)', number 2342 *in* 'LNCS', Springer-Verlag, Chia, Italy, pp. 39–53.

Pike, W. & Gahegan, M. (2003), Constructing semantically scalable cognitive spaces, *in* W. Kuhn, M. Worboys & S. Timpf, eds, 'Proceedings of the International Conference on Spatial Information Theory (COSIT 2003)', number 2825 *in* 'LNCS', Springer-Verlag, Ittingen, Switzerland, pp. 332–348.

Pröll, B., Retschitzegger, W., Wagner, R. & Ebner, A. (1998), 'Beyond traditional tourism information systems – TIScover', *Information Technology and Tourism* **1**.

Ritter, H. & Kohonen, T. (1989), 'Self-organizing semantic maps', *Biological Cybernetics* **61**(4), 241–254.

Salton, G. & Buckley, C. (1988), On the use of spreading activation methods in automatic information retrieval, *in* 'Proceedings of the 11th International Conference on Research and Development in Information Retrieval (SIGIR 1988)', Grenoble, France, pp. 147–160.

Staab, S., Braun, C., Düsterhöft, A., Heuer, A., Klettke, M., Neumann, G., Prager, B., Pretzel, J., Schnurr, H.-P., Struder, R., Uszkoreit, H. & Wrenger, B. (1999), GETESS–searching the web exploiting german texts, *in* M. Klusch, O. M. Shehory & G. Weiss, eds, 'Proceedings of the 3rd Workshop on Cooperative Information Agents (CIA 1999)', number 1652 *in* 'LNCS', Springer-Verlag, Uppsala, Sweden, pp. 113–124.

Velardi, P., Fabriani, P. & Missikoff, M. (2001), Using text processing techniques to automatically enrich a domain ontology, *in* 'Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS 2001)', ACM Press, Ogunquit, ME, pp. 270–284.