An Adaptive Information Retrieval System based on Associative Networks

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

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

[2]Research Group of Industrial Software Engineering (RISE),
Favoritenstraße 9–11/188, A–1040 Wien, Austria

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

Abstract

In this paper we present a multilingual information retrieval system that provides access to Tourism information by exploiting the intuitiveness of natural language. In particular, we describe the knowledge representation model underlying the information retrieval system. This knowledge representation approach is based on associative networks and allows the definition of semantic relationships between domain-intrinsic information items. The network structure is used to define weighted associations between information items and augments the system with a fuzzy search strategy. This particular search strategy is performed by a constrained spreading activation algorithm that implements information retrieval on associative networks. Strictly speaking, we take the relatedness of terms into account and show, how this fuzzy search strategy yields beneficial results and, moreover, determines highly associated matches to users' queries. Thus, the combination of the associative network and the constrained spreading activation approach constitutes a search algorithm that evaluates the relatedness of terms and, therefore, provides a means for implicit query expansion.

Keywords: knowledge representation, associative networks, constrained spreading activation, natural language information retrieval.

1 Introduction

Providing easy and intuitive access to information still remains a challenge in the area of information system research and development. Moreover, as Van Rijsbergen (1979) points out, the amount of available information is increasing rapidly and offering accurate and speedy access to this information is becoming ever more difficult. This quote, although about 20 years old, is still valid nowadays if you consider the amount of information offered on the Internet. But how to address these problems? How to overcome the limitations associated with conventional search interfaces? Furthermore, users of information retrieval systems are often computer illiterate and not familiar with the required logic for formulating appropriate queries, e.g. the burdens associated with Boolean logic. This goes hand in hand with the urge to understand what users really want to know from information retrieval systems.

Standard information retrieval interfaces consist of check boxes, predefined option sets or selection lists forcing users to express her or his needs in a very restricted manner. Therefore, an approach leaving the means of expression in users' hands, narrows the gap between users' needs and interfaces used to express these needs. An approach addressing this particular problem is to allow query formulation in natural language. Natural language interfaces offer easy and intuitive access to information sources and users can express their information needs in their own words.

Hence, we present a multilingual information retrieval system allowing for query formulation in natural language. To reduce word sense ambiguities the system operates on a restricted domain. In particular, the system provides access to tourism information, like accommodations and their amenities throughout Austria.

However, the core element of the information retrieval system remains the underlying knowledge representation model. In order to provide a knowledge representation model allowing to define relations among information items, an approach based on a network structure, namely an associative network, is used. More precisely, this associative network incorporates a means for knowledge representation allowing for the definition of semantic relationships of domain-intrinsic information. Processing the network and, therefore, result determination is accomplished by a technique refereed to as spreading activation. Some nodes of the network act as sources of activation and, subsequently, activation is propagated to adjacent nodes via weighted links. These newly activated nodes, in turn, transmit activation to associated nodes, and so on.

We introduce a knowledge representation approach based on an associative network consisting of three layers. Moreover, a constrained spreading activation algorithm implements a processing technique that operates on the network. Due to the network structure of the knowledge representation model and the processing technique, implicit query expansion enriches the result set with additional matches. Hence, a fuzzy search strategy is implemented.

The remainder of the paper is organized as follows. In Section 2 we review the architecture of the information retrieval system that acts as a basis for the redeveloped approach presented herein. Moreover, Section 3 gives an overview about associative networks and we present an algorithm for processing such networks, i.e. spreading activation. In Section 4 we describe our approach based on associative networks and finally, some conclusions are given in Section 5.

2 AD.M.IN. – A Natural Language Information Retrieval System

Crestani (1997) points out that information retrieval is a science that aims to store and allow fast access to a large amount of data. In contrast to conventional database systems, an information retrieval system does not provide an exact answer to a query but tries to produce a ranking that reflects the intention of the user. More precisely, documents are ranked according to statistical similarities based on the occurrence frequency of terms in queries and documents. The occurrence frequency of a term provides an indicator of the significance of this term. Moreover, in order to get a measure for determining the significance of a sentence, the position of terms within a sentence is taken into account and evaluated. For comprehensive reports about information retrieval see Salton & McGill (1983), Salton (1989) and Baeza-Yates & Ribeiro-Neto (1999).

In order to adapt information retrieval systems to the multilingual demands of users, great efforts have been made in the field of multilingual information retrieval. Hull & Grafenstette (1996) subsume several attempts to define multilingual information retrieval, where Harman (1995) formulates the most concise one: *"multilingual information retrieval is information retrieval in any language other than English"*.

Multilingual information retrieval systems have to be augmented by mechanisms for query or document translation to support query formulation in multiple languages. Information retrieval is such an inexact discipline that it is not clear whether or not query translation is necessary or even optimal for identifying relevant documents and, therefore, to determine appropriate matches to the user query. Strictly speaking, the process of translating documents or queries represents one of the main barriers in multilingual information retrieval.

Due to the shortness of user queries, query translation introduces ambiguities that are hard to overcome. Contrarily, resolving ambiguity in document translation is easier to handle because of the quantity of text available. Nevertheless, state-of-the-art machine translation systems provide only an insufficient means for translating documents. Therefore, resolving ambiguities associated with translations remains a crucial task in the field of multilingual information retrieval. Ballesteros & Croft (1998), for instance, present a technique based on co-occurrence statistics from unlinked text corpora which can be used to reduce the ambiguity associated with translations. Furthermore, a quite straightforward approach in reducing ambiguities is to restrict the domain a multilingual information retrieval system operates on.

Xu, Netter & Stenzhorn (2000) describe an information retrieval system that aims at providing uniform multilingual access to heterogeneous data sources on the web. The MIETTA (Multilingual Tourist Information on the World Wide Web) system has been applied to the tourism domain containing information about three European regions, namely Saarland, Turku, and Rome. The languages supported are English, Finnish, French, German, and Italian. Since some of the tourism information about the regions were available in only one language, machine translation was used to deal with these web documents. Due to the restricted domain, automatic translation should suffice to understand the basic meaning of the translated document without having knowledge of the source language. Users can query the system in various ways, such as free text queries, form-based queries, or browsing through the concept hierarchy employed in the system. MIETTA makes it transparent to the users whether they search in a database or a free-form document collection.

2.1 The Architecture of the Original System

The software architecture of the natural language information retrieval system is designed as a pipeline structure. Hence, successively activated pipeline elements apply transformations on natural language queries that are posed via arbitrary client devices, such as, for instance, web browsers, PDAs or mobile phones. Due to the flexibility of this approach, different pipeline layouts can be used to implement different processing strategies. Figure 1 depicts the layout of the software architecture and illustrates the way of interaction of the pipeline elements.

In a first step, the natural language query is evaluated by an automatic language identification module to determine the language of the query. Next, the system corrects typographic errors and misspellings to improve retrieval performance. Before adding grammar rules and semantic information to the query terms, a converter transforms numerals to their numeric equivalents. Depending on the rules assigned to the query terms, a mapping process associates these terms with SQL fragments that represent the query in a formal way. Due to the fact that the system uses a relational database as backend this mapping process is crucial. In a next step the SQL fragments are combined according to the modifiers (e.g. *"and"*, *"or"*, *"near"*, *"not"*) identified in the query and a single SQL statement that reflects the intention of the query is obtained. Then the system determines the appropriate result and generates an XML representation for further processing. Finally, the XML result set is adapted to fit the needs of the client device.

The remainder of this section gives a brief outline of the system.

2.1.1 The Knowledge Base

A major objective of the Ad.M.In.(Adaptive Multilingual Interfaces) system was to separate the program logic from domain dependent data. In particular, language, domain and device dependent portions are placed in the knowledge base. Thus, the knowledge base represents the backbone of the system and consists of a relational database and a set of ontologies. The database stores information about domain entities, as, for instance, amenities of accommodations. The ontologies store synonyms, define semantic relations and grammar rules.

Basically, the knowledge base consists of separate XML files, whereas the synonym ontology is used to associate terms having the same semantic meaning, i.e. describes linguistic relationships like synonymy. The synonym ontology is based on a flat structure, allowing to define synonymy. Taking a look at the tourism domain, *"playground"* represents a concept possessing several semantic equivalents, as, for instance, *"court"*.

Unfortunately, the synonym ontology provides no means to associate concepts. Consider, for example, the three concepts *"sauna"*, *"steam bath"* and *"vegetarian kitchen"*. Straightforward, someone might derive a stronger degree of relatedness between the concepts *"sauna"* and *"steam bath"* as between *"sauna"* and *"vegetarian kitchen"*.

The second component of the knowledge base stores a set of grammar rules. More precisely, a lightweight grammar describes how certain concepts may be modified by prepositions, adverbial or adjectival structures that are also specified in the synonym ontology. For a more detailed description we refer to Berger (2001).

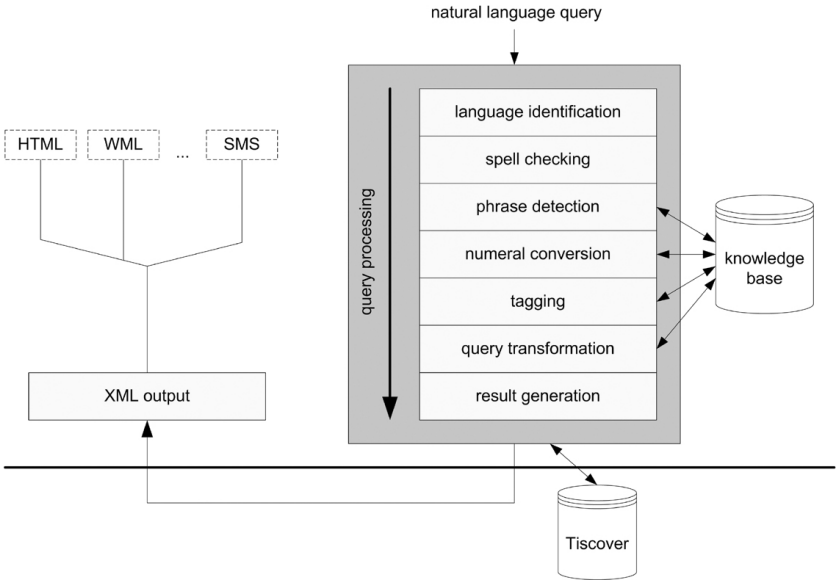

Figure 1: Software Architecture

2.1.2 Language Identification

To identify the language of a query, an n-gram-based text classification approach (cf. Cavnar & Trenkle (1994)) is used. An n-gram is an n-character slice of a longer character string. As an example, for $n = 3$, the *tri–grams* of the string *"language"* are: {*_la, lan, ang, ngu, gua, uag, age, ge_*}. Dealing with multiple words in a string, the blank character is usually replaced by an underscore *"_"* and is also taken into account for the construction of an n-gram document representation. This language classification approach using n-grams requires sample texts for each language to build statistical models, i.e. n-gram frequency profiles, of the languages. We used various tourism-related texts, e.g. hotel descriptions and holiday package descriptions, as well as news articles both in English and German language. The n-grams, with n ranging from 1...5, of these sample texts were analyzed and sorted in descending order according to their frequency, separately for each language. These sorted histograms are the n-gram frequency profiles for a given language. For a comprehensive description see Berger, Dittenbach & Merkl (2003).

2.1.3 Error Correction

To improve retrieval performance, potential orthographic errors have to be considered in the web-based interface. After identifying the language, a spell-checking module is used to determine the correctness of query terms. The efficiency of the spell checking process improves during the runtime of the system by learning from previous queries. The spell checker uses the *metaphone* algorithm (cf. Philips (1990)) to transform the words into their soundalikes. Because this algorithm has originally been developed for the English language, the rule set defining the mapping of words to the phonetic code has to be adapted for other languages. In addition to the base dictionary of the spell checker, domain-dependent words and proper names like names of cities, regions or states, have to be added to the dictionary. For every misspelled term of the query, a list of potentially correct words is returned. First, the misspelled word is mapped to its *metaphone* equivalent, then the words in the dictionary, whose *metaphone* translations have at most an edit distance (cf. Levenshtein (1966)) of two, are added to the list of suggested words. The suggestions are ranked according to the mean of first, the edit distance between the misspelled word and the suggested word, and second, the edit distance between the misspelled word's *metaphone* and the suggested word's. The smaller this value is for a suggestion, the more likely it is to be the correct substitution from the orthographic or phonetic point of view. However, this ranking does not take domain-specific information into account. Because of this deficiency, correctly spelled words in queries are stored and their respective number of occurrences is counted. The words in the suggestion list for a misspelled query term are looked up in this repository and the suggested word having the highest number of occurrences is chosen as the replacement of the erroneous original query term. In case of two or more words having the same number of occurrences the word that is ranked first is selected. If the query term is not present in the repository up to this moment, it is replaced by the first suggestion, i.e. the word being phonetically or orthographically closest. Therefore, suggested words that are very similar to the misspelled word, yet make no sense in the context of the application domain, might be rejected as replacements. Consequently, the word correction process described above is improved by dynamic adaptation to past knowledge.

Another important issue in interpreting the natural language query is to detect terms consisting of multiple words. Proper names like *"Bad Kleinkirchheim"* or nouns like *"parking garage"* have to be treated as one element of the query. Therefore, all multi-word denominations known to the system are stored in an efficient data structure allowing to identify such cases. More precisely, regular expressions are used to describe rules applied during the identification process.

2.1.4 SQL Mapping

With the underlying relational database management system PostgreSQL, the natural language query has to be transformed into a SQL statement to retrieve the requested information. As mentioned above the knowledge base describes parameterized SQL fragments that are used to build a single SQL statement representing the natural language query. The query terms are tagged with class information, i.e. the rel-

evant concepts of the domain (e.g. *"hotel"* as a type of accommodation or *"sauna"* as a facility provided by a hotel), numerals or modifying terms like *"not"*, *"at least"*, *"close to"* or *"in"*. If none of the classes specified in the ontology can be applied, the database tables containing proper names have to be searched. If a noun is found in one of these tables, it is tagged with the respective table's name, such that *"Tyrol"* will be marked as a federal state. In the next step, this class information is used by the grammar to select the appropriate SQL fragments. Finally, the SQL fragments have to be combined to a single SQL statement reflecting the natural language query of the user. The operators combining the SQL fragments are again chosen according to the definitions in the grammar.

3 Associative Networks

Quillian (1968) introduced the basic principle of a semantic network and it played, since then, a central role in knowledge representation. The building blocks of semantic networks are, first, nodes that express knowledge in terms of concepts, second, concept properties, and third,the hierarchical sub-super class relationship between these concepts.

Each concept in a semantic network represents a semantic entity. Associations between concepts describe the hierarchical relationship between these semantic entities via *is-a* or *instance-of* links. The higher a concept moves up in the hierarchy along *is-a* relations, the more abstract is its semantic meaning. Properties are attached to concepts and, therefore, properties are also represented by concepts and linked to nodes via labeled associations. Furthermore, a property that is linked to a high-level concept is inherited by all descendants of the concept. Hence, it is assumed that the property applies to all subsequent nodes. An example of a semantic network is depicted in Figure 2.

Semantic networks initially emerged in cognitive psychology and the term itself has been used in the field of knowledge representation in a far more general sense than described above. In particular, the term semantic network has been commonly used to refer to a conceptual approach known as *associative network*. An associative network defines a generic network which consists of nodes representing information items (semantic entities) and associations between nodes, that express, not necessarily defined or labeled, relations among nodes. Links between particular nodes might be weighted to determine the strength of connectivity.

3.1 Spreading Activation

A commonly used technique, which implements information retrieval on semantic or associative networks, is often referred to as *spreading activation*. The spreading activation processing paradigm is tight-knit with the supposed mode of operation of human memory. It was introduced to the field of artificial intelligence to obtain a means of processing semantic or associative networks. The algorithm, which underlies the spreading activation (SA) paradigm, is based on a quite simple approach and operates on a data structure that reflects the relationships between information items. Thus, nodes model real world entities and links between these nodes define the relatedness of entities. Furthermore, links might possess, first, a specific direction, second, a label and, third, a weight that reflects the degree of association. This conceptual approach allows for the definition of a more general, a more generic network than the basic structure of a semantic network demands. Nevertheless, it could be used to model a semantic network as well as a more generic one, for instance an associative network.

The idea, underlying spreading activation, is to propagate activation starting from source nodes via weighted links over the network. More precisely, the process of propagating activation from one node to adjacent nodes is called a *pulse*. The SA algorithm is based on an iterative approach that is divided into two steps: first, one or more pulses are triggered and, second, a termination check determines if the process has to continue or to halt.

Furthermore, a single pulse consists of a *pre-adjustment phase*, the *spreading process* and a *post-adjustment phase*. The optional pre- and post-adjustment phases might incorporate a means of activation decay, or to avoid reactivation from previous pulses. Strictly speaking, these two phases are used to gain more control over the network. The spreading phase implements propagation of activation over the network. Spreading activation works according to the formula:

$$I_j(p) = \sum_{i}^{k} (O_i(p-1) \cdot w_{ij}) \tag{1}$$

Each node j determines the total input I_j at pulse p of all linked nodes. Therefore, the output $O_i(p-1)$ at the previous pulse $p-1$ of node i is multiplied with the associated weight w_{ij} of the link connecting node i to node j and the grand total for all k connected nodes is calculated. Inputs or weights can be expressed by binary values (0/1), inhibitory or reinforcing values (-1/+1), or real values defining the strength of the connection between nodes. More precisely, the first two options are used in the application of semantic networks, the latter one is commonly used for associative networks. This is due to the fact that the type of association does not necessarily have some exact semantic meaning. The weight rather describes the relationship between nodes. Furthermore, the output value of a node has to be determined. In most cases, no distinction is made between the input value and the activation level of a node, i.e. the input value of a node and its activation level are equal. Before firing the activation to adjacent nodes a function calculates the output depending on the activation level of the node:

$$O_i = f(I_i) \tag{2}$$

Various functions can be used to determine the output value of a node, for instance the sigmoid function, or a linear activation function, but most commonly used is the threshold function. The threshold function determines, if a node is considered to be active or not, i.e. the activation level of each node is compared to the threshold value. If the activation level exceeds the threshold, the state of the node is set to active. Subsequent to the calculation of the activation state, the output value is propagated to adjacent nodes. Normally, the same output value is sent to all adjacent nodes. The process described above is repeated, pulse after pulse, and activation spreads through the network and activates more and more nodes until a termination condition is met. Finally, the SA process halts and a final activation state is obtained. Depending on the application's task the activation levels are evaluated and interpreted accordingly.

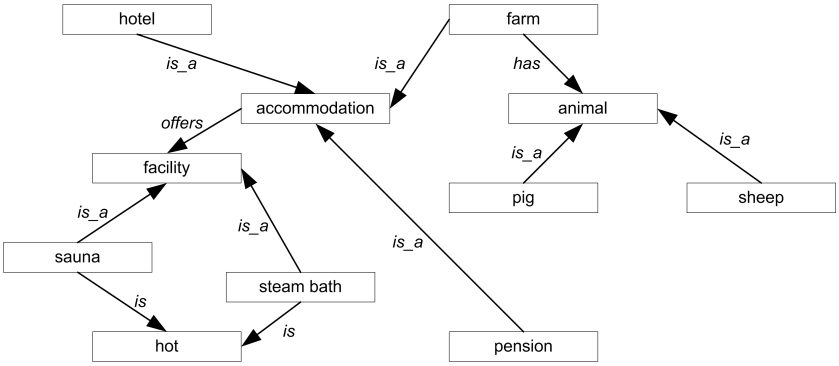

Figure 2: A semantic network example of tourism-related terms

3.2 Taming Spreading Activation

Unfortunately, the basic approach of spreading activation entails some major drawbacks. Strictly speaking, without appropriate control, activation might be propagated all over the network. Furthermore, the semantics of labeled associations are not incorporated in SA and it is quite difficult to integrate an inference mechanism based on the semantics of associations. To overcome these undesired side-effects the integration of constraints helps to tame the spreading process (cf. Crestani (1997)). Some constraints commonly used are described as follows.

- **Fan-out constraint:** Nodes with a broad semantic meaning possess a vast number of links to adjacent nodes. This circumstance implies that such nodes activate large areas of the network. Therefore, activation should diminish at nodes with a high degree of connectivity to avoid this unwanted effect.
- **Distance constraint:** The basic idea underlying this constraint is, that activation ceases when it reaches nodes far away from the activation source. Thus, the term *far* corresponds to the number of links over which activation was spread, i.e. the greater the distance between two nodes, the weaker is their semantic relationship. According to the distance of two nodes their relation can be classified. Directly connected nodes share a first order relation. Two nodes connected via an intermediate node are associated by a second order relation, and so on.
- **Activation constraint:** Threshold values are assigned to nodes (it is not necessary to apply the same value to all nodes) and are interpreted by the threshold function. Moreover, threshold values can be adapted during the spreading process in relation to the total amount of activity in the network.
- **Path constraint:** Usually, activation spreads over the network using all available links. The integration of preferred paths allows to direct activation according to application-dependent rules.

Another enhancement of the spreading activation model is the integration of a feedback process. The activation level of some nodes or the entire network is evaluated by, for instance, another process or by a user. More precisely, a user checks the activation level of some nodes and adapts them according to her or his needs. Subsequently, activation spreads depending on the user refinement. Additionally, users may indicate preferred paths for spreading activation and, therefore, are able to adapt the spreading process to their own needs.

4 Recommendation via Spreading Activation

One of the first information retrieval systems using constrained spreading activation was GRANT. Kjeldsen & Cohen (1987) developed a system that handles information about research proposals and potential funding agencies. GRANT's domain knowledge is stored in a highly associated semantic network. The search process is carried out by constrained spreading activation over the network. In particular, the system extensively uses path constraints in the form of *path endorsement*. GRANT can be considered as an inference system applying repeatedly the same inference schema:

$$\text{IF } x \text{ AND } R(x, y) \rightarrow y \tag{3}$$

$R(x, y)$ represents a path between two nodes x and y. This inference rule can be interpreted as follows: "if a founding agency is interested in topic x and there is a relation between topic x and topic y then the founding agency might be interested in the related topic y."

Croft, Lucia, Crigean & Willet (1989) developed an information retrieval system initially intended to study the possibility of retrieving documents by *plausible inference*. In order to implement plausible inference constrained spreading activation was chosen accidently. The I^3R system acts as a search intermediary (cf. Croft & Thompson (1987)). To accomplish this task the system uses domain knowledge to refine user queries, determines the appropriate search strategy, assists the user in evaluating the output and reformulating the query. In its initial version, the domain knowledge was represented using a tree structure of concepts. The design was later refined to meet the requirements of a semantic network.

Belew (1989) investigated the use of connectionist techniques in an information retrieval system called Adaptive Information Retrieval (AIR). The system handles information about scientific publications, like the publication title and the author. AIR uses a weighted graph as knowledge representation paradigm. For each document, author and keyword (keywords are words found in publication titles) a node is created and associations between nodes are constructed from an initial representation of documents and attributes. A user's query causes initial activity to be placed on some nodes of the network. This activity is propagated to other nodes until certain conditions are met. Nodes with the highest level of activation represent the answer to the query by the AIR system. Furthermore, users are allowed to assign a degree of relevance to the results $(++, +, -, --)$. This causes new links to be created and the adaptation of weights between existing links. Moreover,

feedback is averaged across the judgments of many users.

A mentionable aspect of the AIR system is that no provision is made for the traditional Boolean operators like AND and OR. Rather, AIR emulates these logical operations because *"the point is that the difference between AND and OR is a matter of degree"*. This insight goes back to Von Neumann (as pointed out by Belew (1989)).

A system based on a combination of an ostensive approach with the associative retrieval approach is described in Crestani & Lee (2000). In the WebSCSA (Web Searching by Constrained Spreading Activation) approach a query does not consist of keywords. Instead, the system is based on an ostensive approach and assumes that the user has already identified relevant Web pages that act as a basis for the following retrieval process. Subsequently, relevant pages are parsed for links and they are followed to search for other relevant associated pages. The user does not explicitly refine the query. More precisely, users point to a number of relevant pages to initiate a query and the WebSCSA system combines the content of these pages to build a search profile. In contrast to conventional search engines WebSCSA does not make use of extensive indices during the search process. Strictly speaking, it retrieves relevant information only by navigating the Web at the time the user searches for information. The navigation is processed and controlled by means of a constrained spreading activation model. In order to unleash the power of WebSCSA the system should be used when users already have a point to start for her or his search. Pragmatically speaking, the intention of WebSCSA is to enhance conventional search engines, use these as starting points and not to compete with them.

Hartmann & Strothotte (2002) focus on a spreading activation approach to automatically find associations between text passages and multimedia material like illustrations, animations, sounds, and videos. Moreover, a media-independent formal representation of the underlying knowledge is used to automatically adapt illustrations to the contents of small text segments. The system contains a hierarchical representation of basic anatomic concepts such as bones, muscles, articulations, tendons, as well as their parts and regions.

Network structures provide a flexible model for adaptation and integration of additional information items. Nevertheless, Crestani (1997) points out that *"... the problem of building a network which effectively represents the useful relations (in terms of the IRs aims) has always been the critical point of many of the attempts to use SA in IR. These networks are very difficult to build, to maintain and keep up to date. Their construction requires in depth application domain knowledge that only experts in the application domain can provide."*

Dittenbach, Merkl & Berger (2003) present an approach based on neural networks for organizing words of a specific domain according to their semantic relations. A two-dimensional map is used to display semantically similar words in spatially regions. This representation can support the construction and enrichment of information stored in the associative network.

4.1 The Redeveloped System Architecture

To overcome the limitations of the knowledge base of the original system, the development of an alternative approach to model domain knowledge was necessary. Basically, the unassociated, non-hierarchic knowledge representation model inhibits the power of the system. Strictly speaking, the original system failed to retrieve results on a fuzzy basis, i.e. the results determined by the system provide exact matches only, without respect to first, interactions users made during past sessions, second, personal preferences of users, third, semantic relations of domain intrinsic information, and fourth, locational interdependencies.

In order to adapt the system architecture accordingly, an approach based on associative networks was developed. This associative network replaces the flat synonym ontology used in the original system. Moreover, both the grammar rules and the SQL fragments have been removed from the knowledge base. More precisely, the functionality and logic is now covered by newly developed pipeline elements or implicitly resolved by the associative network. In analogy to the original pipeline, the first three processing steps are accomplished. Next, a newly implemented *initialization*–element associates concepts extracted from the query with nodes of the associative network. These nodes act as activation sources. Subsequently, the newly designed *spreading*–element implements the process of activation propagation. Finally, the new *evaluation*–element analyzes the activation level of the associative network determined during the spreading phase and produces a ranking according to this activation level.

4.1.1 The Knowledge Representation Model

Basically, the knowledge base of the information retrieval system is composed of two major parts: first, a relational database that stores information about domain entities and, second, a data structure based on an associative network that models the relationships among terms relevant to the domain. Each domain entity is described by a freely definable set of attributes. To provide a flexible and extensible means for specifying entity attributes, these attributes are organized as <name, value> pairs. An example from the tourism domain is depicted in Table 1.

Hotel Wellnesshof	
category	4
facility	sauna
facility	solarium
facility	...

Table 1: <name,value>-pair example for entity *"Hotel Wellnesshof"*

The associative network consists of a set of nodes and each node represents an information item. Moreover, each node is member of one of three logical layers defined as follows:

- **Abstraction layer:** One objective of the redevelopment of the knowledge base was to integrate information items with abstract semantic meaning. More precisely, in contrast to the knowledge base used in the original system which only supported modeling of entity attributes, the new approach allows the integration of a broader set of terms, e.g. terms like *"wellness"* or *"summer activities"* that virtually combine several information items.

- **Conceptual layer:** The second layer is used to associate entity attributes according to their semantic relationship. Thus, each entity attribute has a representation at the conceptual layer. Furthermore, the strengths of the relationships between information items are expressed by a real value associated with the link.

- **Entity layer:** Finally, the entity layer associates entities with information items (entity attributes) of the conceptual layer, e.g. an entity possessing the attribute *"sauna"* is associated with the *sauna*–node of the conceptual layer.

The building blocks of the network are concepts. A concept represents an information item possessing several semantically equivalent terms, i.e. synonyms, in different languages. Each concept possesses one of three different roles:

- **Concrete** concepts are used to represent information items at the conceptual layer. More precisely, concrete concepts refer to entity attributes.
- Concepts with an **abstract** role refer to terms at the abstraction layer.
- Finally, the **modifier** role is used to categorize concepts that alter the processing rules for abstract or concrete concepts. A modifier like, for instance, *"not"* allows the exclusion of concepts by negation of the assigned initialization value.

Moreover, concepts provide, depending on their role, a method for expressing relationships among them. The *connectedTo* relation defines a bidirectional weighted link between two concrete concepts, e.g. the concrete concept *"sauna"* is linked to *"steam bath"*. The second relation used to link information items is the *parentOf* association. It is used to express the sub-super class relationship between abstract concepts or concrete and abstract concepts.

A set of concepts representing a particular domain is described in a single XML file and acts as input source for the information retrieval system. During initialization, the application parses the XML file, instantiates all concepts, generates a list of synonyms pointing at corresponding concepts, associates concepts according to their relations and, finally, links the entities to concrete concepts. Currently, the associative network consists of about 2,200 concepts, 10,000 links and more than 13,000 entities. The concept network includes terms that describe the tourism domain as well as towns, cities and federal states throughout Austria.

To get a better picture of the interdependencies associated with the layers introduced above see Figure 3. Each layer holds a specific set of concepts. Abstract concepts associate concepts at the same or at the conceptual layer. Concepts at the conceptual layer define links between entity attributes and associate these attributes with entities at the entity layer. Finally, entities are placed at the lowest layer, the entity layer. Concepts at the entity layer are not associated with items at the same layer. Consider, for example, the abstract concept *"indoor sports"* and the concept *"sauna"* as concepts from which activation originates from. First, activation is propagated between the abstraction layer to the conceptual layer via the dashed line from *"indoor sports"* to *"table tennis"*. We shall note, that dashed lines indicate links between concepts of different layers. Thus, *"sauna"* and *"table tennis"* act as source concepts and, moreover, activation is spread through the network along links at the conceptual layer. Activation received by concepts at the conceptual layer is propagated to the entities at the entity layer stimulating, in this particular case, the entities *"Hotel Stams", "Hotel Thaya"* as well as *"Wachauerhof"*. Moreover, a fraction of activation is propagated to adjacent concept nodes at the conceptual layer, i.e. *"solarium"*, *"whirlpool"* as well as *"tennis"*, and to entities, i.e. *"Hotel Wiental"* and *"Forellenhof"*, respectively.

4.1.2 Processing the Associative Network

Due to the flexibility and adaptivity of the original system, the integration of the redesigned parts has been accomplished with relatively little effort. In particular, the existing knowledge base has been replaced by the associative network and additional pipeline elements to implement spreading activation have been incorporated.

Figure 4 depicts the redeveloped knowledge base on which the processing algorithm operates. The conceptual layer stores concrete concepts and the weighted links among them. Associating abstract concepts with concrete concepts is done at the abstraction layer. Each entity has a unique identifier that is equivalent to the entity identifier stored in the relational database. Furthermore, entities are connected to concepts at the conceptual layer. More precisely, an entity is connected to all attributes it possesses. As an example consider the entity *"Hotel Stams"* as depicted in Figure 4. This hotel offers a *"sauna"*, a *"steam bath"* and a *"solarium"* and is, therefore, linked to the corresponding concepts at the conceptual layer.

First, a user's query, received by the information retrieval system, is decomposed into single terms. After applying an error correction mechanism and a phrase detection algorithm to the query, terms found in the synonym lexicon are linked to their corresponding concept at the abstraction or conceptual layer. These concepts act as activation sources and, subsequently, the activation process is initiated and activation spreads according to the algorithm outlined below.

At the beginning, the role of each concept is evaluated. Depending on its role, different initialization strategies are applied:

- **Modifier role:** In case of the *"not"* modifier, the initialization value of the subsequent concept is multiplied with a negative number. Due to the fact that the *"and"* and *"or"* modifiers are implicitly resolved by the associative network, they receive no special treatment. More precisely, if, for instance, somebody is searching for an accommodation with a sauna or solarium, those accommodations offering both facilities will be ranked higher than others, providing only one of the desired facilities. Furthermore, the *"near"* modifier reflecting geographic dependencies, is automatically resolved by associating cities or towns within a circumference of 15km. Depending on the distance, the weights are adapted accordingly, i.e. the closer they are together, the higher is the weight of the link in the associative network.
- **Abstract role:** If a source concept is abstract, the set of source concepts is expanded by resolving the *parentOf* relation between parent and child concepts. This process is repeated until all abstract concepts are resolved, i.e. the set of source concepts contains members of the conceptual layer only. The initial activation value is propagated to all child concepts, with respect to the weighted links.
- **Concrete role:** The initial activation level of concrete concepts is set to initialization value defined in the XML source file. The spreading activation process takes place at the conceptual layer, i.e. the *connectedTo* relations between adjacent concepts are used to propagate activation through the network.

After the initialization phase has completed, the iterative spreading process is activated. During a single

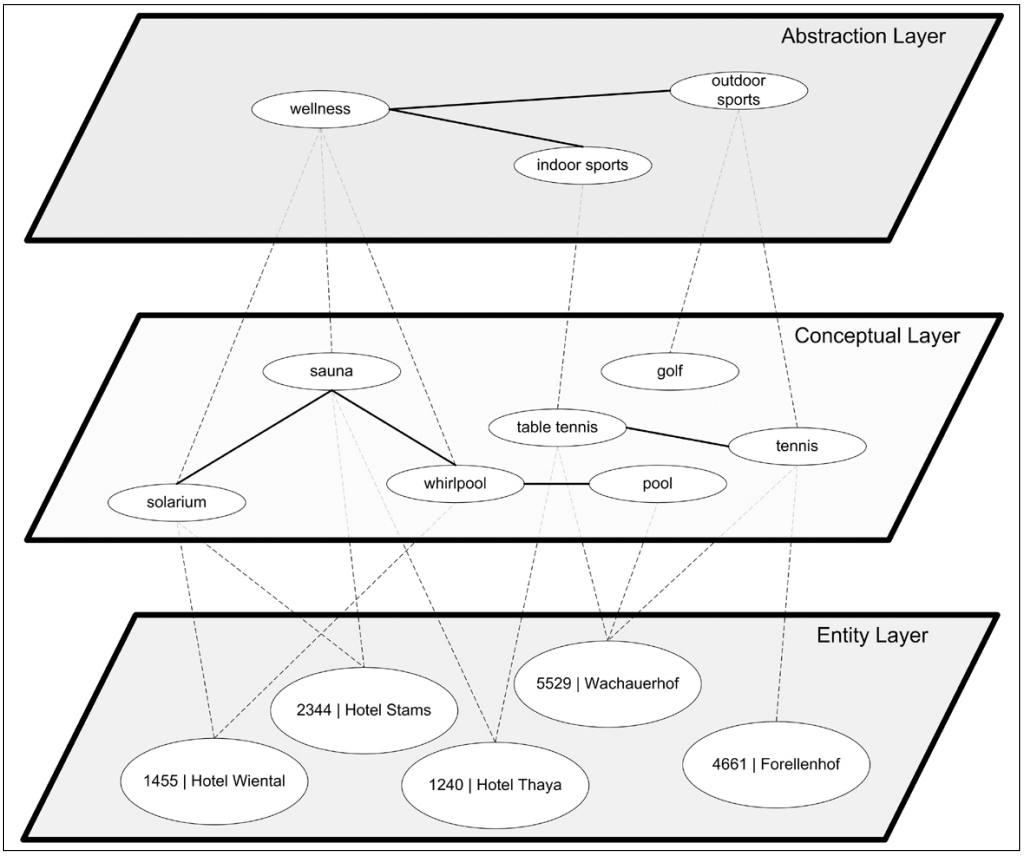

Figure 3: Network layer interdependencies

iteration one pulse is performed, i.e. the number of iterations equals the number of pulses. Starting from the set of source concepts determined during initialization, in the current implementation activation is spread to adjacent nodes according to the following formula:

$$O_i(p) = \begin{cases} 0 & \text{if } I_i(p) < \tau, \\ \frac{F_i}{p+1} \cdot I_i(p) & \text{else, with } F_i = (1 - \frac{C_i}{C_T}) \end{cases} \quad (4)$$

The output, $O_i(p)$, sent from node i at pulse p, is calculated as the fraction of F_i, which limits the propagation according to the degree of connectivity of node i (i.e. fan-out constraint, cf. Section 3.2), and $p + 1$, expressing the diminishing semantic relationship according to the distance of node i to activation source nodes (i.e. distance constraint, cf. Section 3.2). Moreover, F_i is calculated by dividing the number of concepts C_i directly connected to node i by the total number of nodes C_T building the associative network. Note, τ represents a threshold value.

Simultaneous to calculating the output value for all connected nodes, the activation level $I_i(p)$ of node i is added to all associated entities. More precisely, each entity connected to node i receives the same value and adds it to an internal variable representing the total activation of the entity. As an example, if the concept node *"sauna"* is activated, the activation potential is propagated to the entities *"Hotel Stams"* and *"Hotel Thaya"* (cf. Figure 4). Next, all newly activated nodes are used in the subsequent iteration as activation sources and the spreading process continues until the maximum number of iterations is reached.

After the spreading process has terminated, the system inspects all entities and ranks them according to their activation. Figure 5 depicts the results determined for the example query

Ich und meine Kinder möchten in einem Hotel in Kitzbühel Urlaub machen. Es sollte ein Dampfbad haben.[1]

In this particular case, the entities *"Schwarzer Adler Kitzbühel"* and *"Hotel Schloss Lebenberg – Kitzbühel"* located in *"Kitzbühel"* are suggested to be the best matching answers to the query. Moreover, the result set includes matches that are closely related to the user's query. Thus, depending on the relations stored in the associative network, entities offering related concepts are activated accordingly. More precisely, not only the attributes *"hotel"*, *"steam bath"* and *"kids"* are taken into account, but also all other related entity attributes (e.g. *"sauna"*, *"whirlpool"*, *"solarium"*, etc.) have some influence on the ranking position. Furthermore, accommodations in cities in the vicinity of *"Kitzbühel"* providing the same or even better offers are also included in the result set. Thus, the associative network provides a means for exact information retrieval and incorporates a fuzzy search strategy that determines closely related matches to the user's query.

5 Conclusion

A natural language system based on an approach described in Berger (2001) and Berger, Dittenbach, Merkl & Winiwarter (2001) has been reviewed in this paper and, furthermore, provided the basis for the research presented herein. The reviewed system offers multilingual access to information on a restricted domain. In this particular case the system operates on

[1] Me and my kids would like to spend our holidays in a hotel in Kitzbhel. It should have a steam bath.

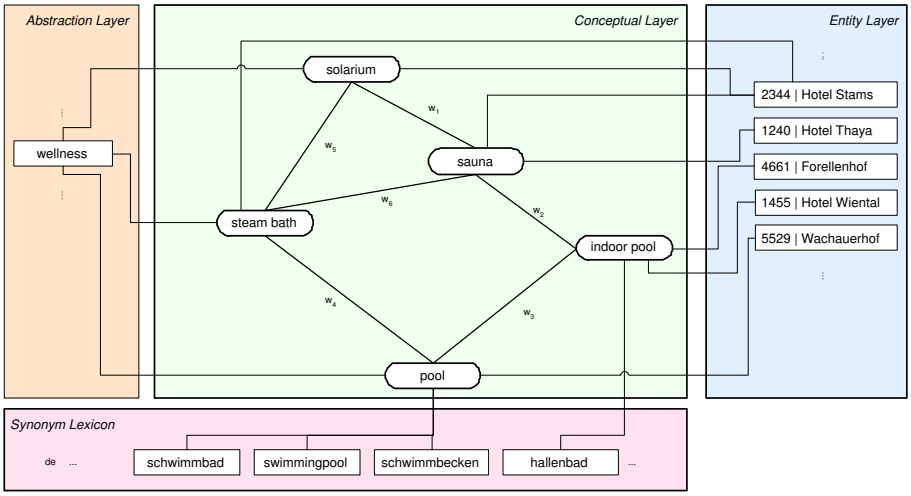

Figure 4: Knowledge base architecture

the tourism domain. Moreover, users of the search interface are encouraged to formulate queries in natural language, i.e. they are able to express their intentions in their own words.

We developed a knowledge representation model that facilitates the definition of semantic relations between information items exemplified by terms of the tourism domain. In particular, an associative network based on a three layered structure was introduced. First, the abstraction layer allows modelling of terms with a subjective or broader semantic meaning, second, the conceptual layer is used to define relations via weighted links between terms, and, finally, the entity layer provides a means to associate elements stored in a relational database with information items in the associative network. Moreover, a constrained spreading activation algorithm implements a processing technique operating on the network. Generally, the combination of the associative nature of the knowledge representation model and the constrained spreading activation approach constitutes a search algorithm that evaluates the relatedness of terms and, therefore, provides a means for implicit query expansion.

The flexible method of defining relationships between terms unleashes the ability to determine highly associated results as well as results that are predefined due to personal preferences. Moreover, especially designed associative networks can be used to model scenarios, as, for instance, a winter holiday scenario that favors accommodations offering winter sports activities by adapting the weights of links accordingly.

One important task for further enhancement is the possibility to express the relevance of query terms. Users should be able to assign a degree of significance to terms. Consider, for example, a user searching for an accommodation with several amenities in the capital city of Austria. Moreover, the user is a vegetarian. Therefore, a means for expressing the importance of vegetarian kitchen is needed. In order to accomplish this requirement, the system might be extended to *understand* words that emphasis terms, e.g. in analogy to modifiers like *"and"*, *"or"*, *"near"*, etc. the word *"important"* is handled like a modifier and influences the activation level of the following query term. Additionally, an interface providing a graphical instrument to express relevance by means of a slide controller might be considered.

Furthermore, an associative network might act as a kind of *short term memory*. More precisely, during a user session a particular network is used to store the activation level determined during past user interactions. A user, for instance, is searching for a hotel in Vienna. Thus, the associative network stores the activation level for further processing. Next, the user might restrict the results to accommodations offering a sauna. This spreading process is carried out using the associative network determined during the previous interaction.

References

Baeza-Yates, R. A. & Ribeiro-Neto, B. (1999), *Modern Information Retrieval*, Addison-Wesley, Reading, MA.

Ballesteros, L. & Croft, W. B. (1998), Resolving ambiguity for cross-language retrieval, *in* 'Research and Development in Information Retrieval', pp. 64–71.

Belew, R. K. (1989), Adaptive information retrieval: Using a connectionist representation to retrieve and learn about documents, *in* N. J. Nicholas J. Belkin & C. J. Van Rijsbergen, eds, 'Proceedings of the 12th International Conference on Research and Development in Information Retrieval (SIGIR'89)', ACM, pp. 11–20.

Berger, H. (2001), Adaptive multilingual interfaces, Master's thesis, Vienna University of Technology.

Berger, H., Dittenbach, M. & Merkl, D. (2003), Querying tourism information systems in natural language, *in* 'Proceedings of the 2nd International Conference on Information System Technology and its Applications (ISTA 2003)', Kharkiv, Ukraine.

Berger, H., Dittenbach, M., Merkl, D. & Winiwarter, W. (2001), Providing multilingual natural language access to tourism information, *in* W. Winiwarter, S. Bressan & I. K. Ibrahim, eds, 'Proceedings of the 3rd International Conference on Information Integration and Web-based Applications and Services (IIWAS 2001)', Austrian Computer Society, Linz, Austria, pp. 269–276.

Cavnar, W. B. & Trenkle, J. M. (1994), N-gram-based text categorization, *in* 'International Symposium on Document Analysis and Information Retrieval', Las Vegas, NV.

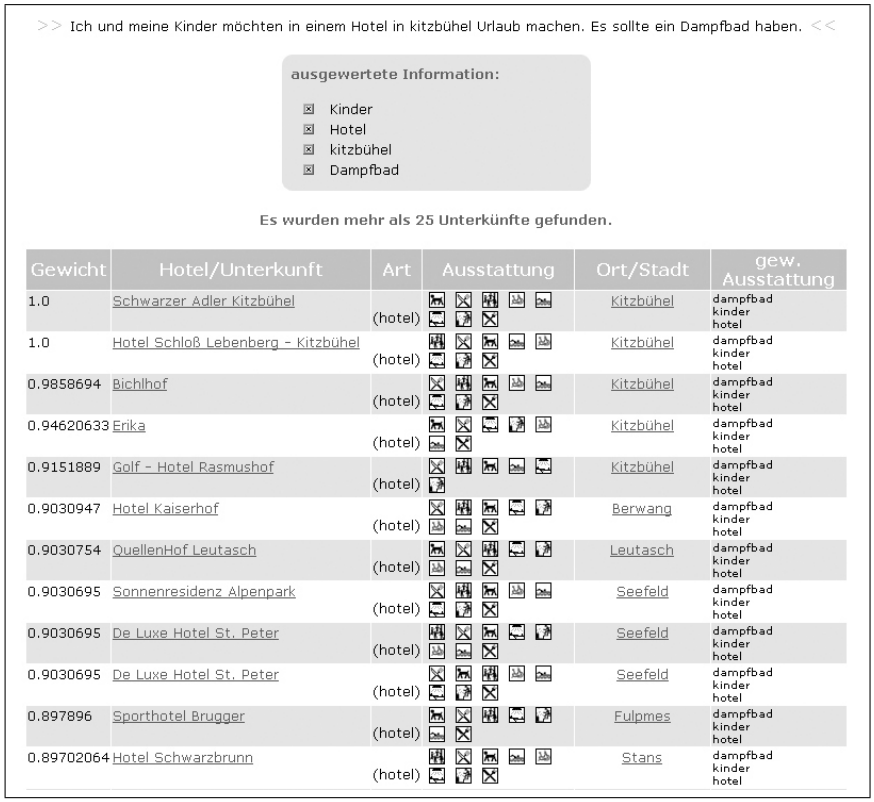

Figure 5: Weighted result set determined by constrained spreading activation

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

Crestani, F. & Lee, P. L. (2000), 'Searching the web by constrained spreading activation', *Information Processing and Management* **36**(4), 585–605.

Croft, W., Lucia, T., Crigean, J. & Willet, P. (1989), 'Retrieving documents by plausible inference: an experimental study', *Information Processing & Management* **25**(6), 599–614.

Croft, W. & Thompson, R. H. (1987), 'I^{3}R: A New Approach to the Design of Document Retrieval Systems', *Journal of the American Society for Information Science* **38**(6), 389–404.

Dittenbach, M., Merkl, D. & Berger, H. (2003), Using a connectionist approach for enhancing domain ontologies: Self-organizing word category maps revisited, *in* 'Proceedings of the 5th International Conference on Data Warehousing and Knowledge Discovery - (DaWaK 2003)'. Accepted for publication.

Harman, D. K. (1995), Overview of the 3rd Text Retrieval Conference (TREC-3), *in* D. K. Harman, ed., 'Proceedings of the 3rd Text Retrieval Conference (TREC-3)', NIST Special Publication 500–225, pp. 1–19.

Hartmann, K. & Strothotte, T. (2002), A spreading activation approach to text illustration, *in* 'Proceedings of the 2nd International Symposium on Smart graphics', ACM Press, pp. 39–46.

Hull, D. A. & Grafenstette, G. (1996), Querying across languages: A dictionary-based approach to multilingual information retrieval, *in* 'Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1996)', pp. 49–57.

Kjeldsen, R. & Cohen, P. (1987), 'The evolution and performance of the GRANT system', *IEEE Expert* pp. 73–79.

Levenshtein, V. I. (1966), 'Binary codes capable of correcting deletions, insertions and reversals', *Soviet Physics Doklady* **10**(8), 707–710.

Philips, L. (1990), 'Hanging on the metaphone', *Computer Language Magazine* **7**(12).

Quillian, M. R. (1968), Semantic memory, *in* M. Minsky, ed., 'Semantic Information Processing', MIT Press, pp. 227–270.

Salton, G. (1989), *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*, Addison-Wesley, Reading, MA.

Salton, G. & McGill, M. J. (1983), *Introduction to Modern Information Retrieval*, McGraw-Hill, New York.

Van Rijsbergen, C. J. (1979), *Information Retrieval*, Department of Computer Science, University of Glasgow.

Xu, F., Netter, K. & Stenzhorn, H. (2000), Mietta - a framework for uniform and multilingual access to structured database and web information, *in* 'Proceedings of the 5th International Workshop on Information Retrieval with Asian languages'.