

Dissertation

Activation on the Move: Adaptive Information Retrieval via Spreading Activation

ausgeführt zum Zwecke der Erlangung
des akademischen Grades eines Doktors
der technischen Wissenschaften

unter der Leitung von
ao. Univ. Prof. Dr. Dieter Merkl
E188 Institut für Softwaretechnik
und interaktive Systeme

eingereicht an der Technischen Universität Wien
Fakultät für Technische Naturwissenschaften und Informatik

von

DI Helmut Berger
9302214
Mariengasse 39/2/14
A-1170 Wien

Wien, im Mai 2003

Unterschrift

Contents

1	Introduction	1
2	A Natural Language Information Retrieval System	5
2.1	Introduction	5
2.2	Related Work	6
2.3	The Architecture of the Original System	10
2.3.1	The Knowledge Base	12
2.3.2	Language Identification	14
2.3.3	Error Correction	16
2.3.4	Mapping Natural Language Queries to a Formal Representation	18
2.4	Discussion	19
3	What Users Really Want to Know from Tourism Information Systems	20
3.1	Introduction	20
3.2	Related Work	21
3.3	Field Trial	22
3.3.1	Design Considerations of the Interface	23
3.3.2	Results from the Field Trial	26
3.3.3	Lessons Learned from the Field Trial	37
3.4	Usability Study	38
3.4.1	Test Design and Processing	38
3.4.2	Results from the Usability Study	40

3.4.3	Lessons Learned from the Usability Study	42
3.5	Discussion	42
4	Using Network Structures for Knowledge Representation	44
4.1	Introduction	44
4.2	Related Work	45
4.3	Associative Networks	47
4.4	Spreading Activation	48
4.5	Taming Spreading Activation	51
4.6	Discussion	52
5	An Associative Knowledge Representation Model for Tourism Information	54
5.1	Introduction	54
5.2	Related Work	55
5.3	Associating Domain Knowledge	58
5.3.1	Associative Knowledge Modelling	59
5.3.2	Processing the Network	65
5.3.3	Adapting Associations According to Past User Inter- actions	69
5.3.4	Retrieval Results	74
5.4	Discussion	78
6	Conclusion and Future Work	80

List of Figures

2.1	Screen-shot of the DIETORECS interface	10
2.2	Software Architecture	11
2.3	Layout of the knowledge base	12
2.4	Excerpt of the synonym ontology	13
3.1	Natural language query interface	24
3.2	Standard <i>Tiscover</i> search interface	25
3.3	Result page with matching accommodations and feedback form	26
3.4	<i>Tiscover</i> 's advanced search (part of)	40
4.1	A semantic network example of tourism-related terms	48
4.2	Flowchart of the spreading activation model	49
5.1	Redesigned software architecture	59
5.2	Network structure of abstract concepts	61
5.3	Network structure of concepts at the conceptual layer	62
5.4	Network layer interdependencies	63
5.5	XML representation of concepts in the associative network . .	64
5.6	Knowledge base architecture	66
5.7	Weighted result set determined by constrained spreading ac- tivation	69
5.8	Concept matrix derived from user queries	73
5.9	Result set exemplifying regional dependencies	74
5.10	Result set exemplifying dependencies between cities	75
5.11	Result set exemplifying abstract concepts	76

5.12 Comparing results determined by the original system and the associative network approach	77
5.13 Excerpt of accommodations not associated with a city	78

List of Tables

2.1	Automatic translation of the first verse of <i>The Verve's Bittersweet Symphony</i>	7
2.2	Top ten <i>tri-gram</i> occurrences of German and English text with underscores representing blanks	15
3.1	Origin of queries (derived from the top-level domain of the accessing host)	27
3.2	Manual analysis of language identification accuracy	28
3.3	Word occurrence statistic	30
3.4	Subset of natural language queries obtained during the field trial	30
3.5	Number of concepts per query (counted by manual inspection)	31
3.6	Concepts that have been identified or not identified by the natural language processing module of our interface	33
3.7	Usage of modifiers " <i>and</i> " and " <i>or</i> "	34
3.8	Usage of modifiers " <i>not</i> " and " <i>near</i> "	35
3.9	Combined usage of modifiers	36
3.10	Example scenario from the tourism domain	39
5.1	Concept matrix for a single query	71
5.2	Subset of user queries obtained during the field trial	71
5.3	Concept matrix for multiple queries	72

Abstract

With the increasing amount of information available on the *Internet* one of the most challenging tasks is to provide search interfaces that are easy to use without having to learn a specific syntax. In this thesis a query interface exploiting the intuitiveness of natural language for accessing tourism information is presented.

Furthermore, the results and insights from analyzing the natural language queries collected during a field trial in which the interface was promoted via the homepage of the largest Austrian tourism platform *Tiscover* are described. This analysis shows how users formulate queries when their imagination is not limited by conventional search interfaces with structured forms consisting of check boxes, radio buttons and special-purpose text fields.

In a usability study subjects were asked to identify themselves with pre-defined tourism related scenarios. Subsequently, they had to accomplish the tasks described in these scenarios. On the one hand, subjects were requested to use the standard *Tiscover* interface to fulfill the stipulated tasks, and on the other hand, they were asked to use the natural language interface to retrieve results. The experiences users made during the interaction with the interfaces have been observed and the findings are presented herein.

The major concern of this thesis is the development of a knowledge representation model based on a network structure as motivated by the results of a field trial and a usability study. In particular, an approach based on associative networks, for defining semantic relationships of terms is presented. More precisely, the relatedness of terms is taken into account and it is shown,

how a fuzzy search strategy, performed by a constrained spreading activation algorithm, yields beneficial results and, moreover, suggests closely related matches to users' queries. Thus, spreading activation implicitly implements query expansion.

Kurzfassung

Das *Internet* bietet eine schier unabschätzbare Menge an Information, der es Herr zu werden gilt. Daher wird die Entwicklung von Abfrageschnittstellen, die einfachen und vor allem intuitiven Zugriff auf diese Informationen ermöglichen, mit großem Ehrgeiz verfolgt. In dieser Dissertation wird eine Schnittstelle, die auf eine spezielle Syntax verzichtet um mittels natürlichsprachigen Anfragen Zugriff auf Tourismusinformationen zu bieten, beschrieben.

Die natürlichsprachige Schnittstelle diente als Basis für einen Feldversuch der in Zusammenarbeit mit *Tiscover*, der größten österreichischen Tourismus Plattform, durchgeführt wurde. Die Ergebnisse des Feldversuchs, ermittelt durch die Analyse der natürlichsprachigen Anfragen, werden im Detail beschrieben. Die Auswertungen zeigen in welcher Form Benutzer Anfragen formulieren, wenn diese nicht gezwungen sind ihre Vorstellungen mittels Check-Boxen, Radio-Buttons, etc. auszudrücken.

Zusätzlich wird auf eine *Usability*-Studie eingegangen, in welcher Testpersonen die Aufgabe hatten sich mit Szenarien aus dem Tourismusbereich zu identifizieren. Die Testpersonen mußten die Aufgaben die in diesen Szenarien beschrieben wurden, einerseits mit der Standardsuchschnittstelle von *Tiscover*, und andererseits mit der natürlichsprachigen Suchschnittstelle bewältigen. Die Erfahrungen, die die Testpersonen während der Interaktion mit den Schnittstellen machten, wurden beobachtet und die davon abgeleiteten Ergebnisse werden in dieser Dissertation präsentiert.

Das Hauptaugenmerk liegt auf der Entwicklung eines Netzwerkmodells zur Wissensrepräsentation. Ausgehend von den Resultaten des Feldversuchs

und der *Usability*-Studie, wird ein Ansatz basierend auf assoziativen Netzen beschrieben, welcher die Definition von semantischen Beziehungen zwischen Termen ermöglicht. Diese Beziehungen fungieren als Basis für eine *unscharfe* Suchstrategie die durch einen *Constrained Spreading Activation*-Algorithmus implementiert wird. Da es sich hierbei um keine exakte Suche handelt werden auch verwandte oder ähnliche Resultate in Betracht gezogen und mit der Anfrage des Benutzers in Verbindung gebracht. Die auf dieser Strategie basierende Suche erweitert und verbessert damit die ermittelten Suchergebnisse.

h. to b.

Acknowledgements

To those, bearing my moaning.

To those, inspiring my mind.

To the very *you* of you: You remain special, thus, i am enjoying – still!

Especially “BM”, “MR”, “FR”, “HB”, “LB”, “ER”, “FR”, “KK”, “CG”,
“HM”, “EM”, “HM”, “MD”, “DM”.

Thanx.

Chapter 1

Introduction

Providing easy and intuitive access to information remains still 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. A natural language interface offers easy and intuitive access to information sources, and users are

able to express their information needs in their own words.

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

An approach implementing the functionality in a prototypical manner was used as a basis for a field trial carried out during a ten-day period in March, 2002. One major objective was to investigate the acceptance of interfaces allowing users to pose queries in natural language. Are users actually willing to type natural language sentences to express their information needs? Moreover, it was intended to obtain a broad spectrum of natural language requests to derive suggestions on improving the knowledge stored in the knowledge base of the system.

Furthermore, as another aspect of user-oriented evaluation of our system, we performed a usability study focusing on a comparison of a conventional search interface with the natural language approach. More precisely, subjects were asked to solve tasks described in predefined scenarios with both, the standard *Tiscover* interface and the natural language interface. Experiences users made during the interaction with the interfaces have been observed and, subsequently, analyzed to get a means for comparing both types of interfaces.

The findings obtained from the field trial and the observations made during the usability study motivated the redevelopment of the knowledge base underlying the information retrieval system. In particular, the conceptual model of the knowledge base of the original system was inadequate to model semantic relationships between domain concepts. In order to provide a knowledge representation model allowing to define relations among concepts, 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 known 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. 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.

Furthermore, defining relations between nodes of the associative network is a significant task. Therefore, user queries obtained during the field trial are inspected and potential semantic relationships between information items are derived. In this thesis the idea underlying this approach is described. Moreover, a means for automatic adaptation of associations during system runtime is proposed.

The remainder of this thesis is organized as follows. Chapter 2 describes the architecture of the natural language information retrieval system acting as a basis for the research described in this thesis. In particular, the steps necessary for processing the natural language query are pointed out. In a nutshell, an overview about the knowledge representation model underlying the system is given.

Chapter 3 provides a description of the data collected during a field trial based on the original system as well as the consequences which can be drawn from this data (Berger et al. (2003b); Dittenbach et al. (2002a,b, 2003a)). Moreover, the design goals for the interface are outlined. Next, the findings of a usability study, comparing the standard *Tiscover* interface with the original natural language approach, are presented.

In Chapter 4 an overview about using network structures as a means for knowledge representation is provided. In particular, an extensive description of the structure and application of associative networks is given. Moreover, an algorithm for processing such networks is presented, i.e. spreading activation is discussed in detail.

Then, in Chapter 5 the approach underlying the redeveloped knowledge

base of the information retrieval system is presented. A detailed description of the associative network used to define relations between domain-intrinsic information is provided and exemplified by means of terms of the tourism domain. Moreover, the process of determining exact matches as well as highly associated results via spreading activation through the associative network is detailed (Berger et al. (2003a)). Furthermore, an approach for deriving concept relations from past user interactions is described. Then, results determined via the newly developed approach are presented and discussed.

Finally, Section 6 brings this thesis to a close by giving some conclusions and pointing out some future directions of research.

Chapter 2

A Natural Language Information Retrieval System

2.1 Introduction

Lancaster (1968) pointed out, that “... *an information retrieval system does not inform (i.e. change the knowledge of) the user on the subject of his inquiry. It merely informs on the existence (or non-existence) and whereabouts of documents relating to his request*”.

But how to *inform* the user without knowing her or his real intention? How to determine the real intention of a user, when she or he is forced to choose from a fixed set of options available to her or him? What, if she or he is not able to express her or his request by the means available? Therefore, the development and availability of efficient and appropriate search functions remains still a challenge in the field of database and information systems. Consider, for example, the context of tourism information systems, where intuitive search functionality plays a crucial role for the economic success. Users are often computer illiterate and not familiar with Boolean logic. Addressing this problem, the application of a natural language interface, however, allows these users to express their information needs in an intuitive way.

The remainder of this chapter provides a review of a prototypical natural language information retrieval system, namely Ad.M.In. (abbreviation for Adaptive Multilingual Interfaces). Primarily, the information retrieval system was designed to facilitate multilingual access to tourism information provided by the largest Austrian tourism platform *Tiscover*. Strictly speaking, the objectives of the natural language system have been, first, the development of an information retrieval system that allows query formulation in natural language, second, to allow posing queries in multiple languages, third, to restrict the domain the system operates on and, fourth, to provide access for multiple client devices.

2.2 Related Work

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 Van Rijsbergen (1979); Salton and McGill (1983); Salton (1989); Baeza-Yates and 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 and 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”*.

The Verve's Bittersweet Symphony
'Cause it's a bittersweet symphony, this life. Try to make ends meet. You're a slave to money then you die. I'll take you down the only road I've ever been down. You know the one that takes you to the places where all the veins meet yeah
German Translation
Ursache ist es eine bittersüße Symphonie, dieses Leben. Versuchen Sie, Enden Treffen zu bilden. Sie sind ein Sklave zum Geld dann Sie Würfel. Ich nehme Sie, die einzige Straße niederzuwerfen, die ich überhaupt unten gewesen bin. Sie kennen das, das Sie zu den Plätzen, in denen alle Adern yeah treffen nimmt

Table 2.1: Automatic translation of the first verse of *The Verve's Bittersweet Symphony*

Nevertheless, 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. Multilingual information retrieval systems have to be augmented by mechanisms for query or document translation to support query formulation in multiple languages. 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. Consider, for example, the automatic translation of the first verse of “*The Verve's Bittersweet Symphony*”, performed with *Altavistas' Babelfish*¹, as depicted in Table 2.1.

¹<http://babelfish.altavista.com>

As can be seen, the German translation is a long way off being appropriate, not to mention perfect. Therefore, resolving ambiguities associated with translations remains a crucial task in the field of multilingual information retrieval. Ballesteros and 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 et al. (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.

Natural language information retrieval systems are designed with the goal in mind to relief the user from the burden to formulate queries as lists of keywords with crude Boolean logic to describe the relationship between keywords and restrictions for the documents that shall be retrieved. Instead, the user should be able to ask queries in such a way as she or he would do with an interlocutor. In general, available systems operate on a restricted domain, for instance the tourism domain, to reduce word sense ambiguity and to provide a manageable way for modelling semantic relations between concepts.

An intelligent information agent using semantic methods and natural lan-

guage processing capabilities in order to gather tourism information from the World Wide Web and present it to the user in an intuitive, user-friendly way is described in Staab et al. (1999). The GETESS (German Text Exploitation and Search System) system is used in a restricted domain, like the tourism domain. Information is spread among heterogeneous data sources and formats. The dialog system of GETESS acts like an interlocutor. The user poses her or his question, the system answers and tries to continue the conversation with the user. The natural language component of the system is used in order to, first, linguistically analyze user queries, second, generate the linguistic basis for the extraction of facts from natural language documents and, third, generate natural language responses.

Furthermore, besides the importance of providing appropriate results to user queries, information retrieval systems in the tourism domain should support the user in making decisions and, therefore, suggest several recommendations. The system must take care of user's behavior, i.e. track past user interactions, integrate personal preferences and adapt results to timely phenomena like vacation periods, seasons, etc.

The DIETORECS system (cf. Fesenmaier et al. (2003)) is a recommendation system in the tourism domain that provides several means for querying the system. More precisely, the system supports, on the one hand, the search and selection of specific *travel products* (e.g. a hotel), and on the other hand, building a bundle of products (*travel bag*). Due to the domain diversity, a broad set of attributes is available. Users query the system by choosing, *inter alia*, from a fixed set of attributes represented by option sets or dropdown lists. Unfortunately, this diversity of terms results in a dramatically overloaded search interface (see Figure 2.1). The avoidance of such overloaded interfaces is an argument in favor of query formulation in natural language. Moreover, as O'Brian (2001) pointed out, a sophisticated result determination of, for instance, accommodations goes hand in hand with the ability of expressing intentions in *own* words.

WELCOME TO
Intelligent Recommendation
System for Tourist Decision Making

my travel bags my travel notes existing users new user? partners

travel preferences

Searching for a special travel?
Specify here your travel constraints:

Destination (mark checkbox to validate your destination): ☐

Travel Party:
Family with Children ☐

Duration:
☐ < 3 days
 ☒ 3 - 7 days
 ☐ > 7 days

Budget [€] (accommodation a person a night):
☐ < 15 ☒ 15 - 30
☐ 30 - 60 ☐ > 60

Period of traveling (month/year):
 December 2002

Location:
 Mountain ☐

You want to tighten your travel preferences?
 Make your choice:
 advanced travel wish
 accommodation
 destination
 interests

Recommend the whole travel
 Recommend Travel

[destination preferences]

Destination (mark checkbox to validate your destination): ☐

Map (click on the map to specify your preferred destination):

Country: Italy
 Region: Trentino
 City: Any
 Location: Any

Altitude [m]:
☐ < 300 ☐ 300 - 500
☐ 500 - 1000 ☐ > 1000

What are you planning to do during your holidays?

Sport - Adventure:
☒ Skiing ☐ Boating
☐ Climbing ☐ Trekking
☐ Biking ☐ Diving
☐ Swimming ☐ Hunting
☐ Horseback Riding ☐ Canoeing & Rafting
☐ Tennis ☐ Golfing

Art - Culture:
☒ Historical ☒ Museum ☐ Architectural
☐ Exhibition ☐ Festivals ☒ Markets
☐ Shopping ☐ Nightlife ☐ Ethnic
☐ Theatre ☐ Concert

Leisure - Relax:
☐ Thermal Bath ☐ Therapeutic ☐ Wellness

Save All Recommend Destination

Figure 2.1: Screen-shot of the DIETORECS interface

2.3 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 2.2 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. The spell-checking module transforms words into their sound-

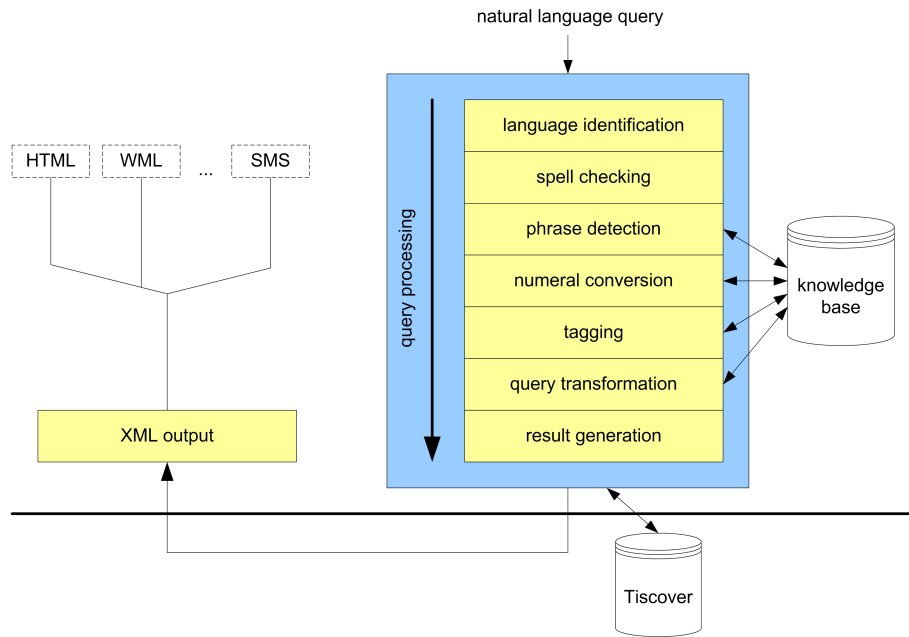


Figure 2.2: Software Architecture

likes and suggests words that are likely to be the correct substitution from the orthographic or phonetic point of view. Subsequently, a phrase recognizer inspects the query and identifies multi-word denominations, as, for instance “*Kirchberg am Wechsel*” or “*holiday flat*”. 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. Finally, the system determines the appropriate result and generates an XML representation for further processing. The XML result set is adapted to fit the needs of the client device.

The remainder of this section describes the components of the natural lan-

guage system. First, the functionality of the knowledge base is illustrated. Next, the language identification process is detailed, followed by a description of the error correction algorithm. Finally, the process of generating the appropriate SQL statement is exemplified.

2.3.1 The Knowledge Base

A major objective of the Ad.M.In. 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.

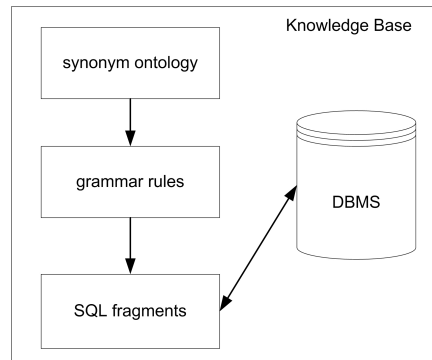


Figure 2.3: Layout of the knowledge base

Figure 2.3 depicts the components of the knowledge base. Basically, the knowledge base consists of separate XML files, whereas the synonym ontology (cf. Figure 2.4) is used to associate terms having the same semantic meaning, i.e. is used to describe 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*”. In the XML represen-

tation, synonyms are expressed by a special tag, namely the `<syn>`-tag.

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*”.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<ontology lang="en">
<classes>
  <class id="0004" word="pension">
    <rewrite>typ Pension</rewrite>
    <syn>guesthouse</syn>
    <syn>guesthouses</syn>
  </class>
  <class id="0005" word="spielplatz">
    <rewrite>einrichtung spielplatz</rewrite>
    <syn>playground</syn>
    <syn>court</syn>
  </class>
</classes>
<modifiers>
  <modifier id="0001" word="ohne">
    <rewrite>ohne</rewrite>
    <syn>without</syn>
    <syn>no</syn>
  </modifier>
  <modifier id="0002" word="und">
    <rewrite>und</rewrite>
    <syn>and</syn>
  </modifier>
</modifiers>
</ontology>
```

Figure 2.4: Excerpt of the synonym ontology

Furthermore, each term is associated with a preferred representation (i.e. a preferred concept). The `<rewrite>`-tag stores the information about the preferred representation. This representation defines the concept more exactly, e.g. in Figure 2.4 the data in the `<rewrite>`-tag of “*playground*” refines

the term to “*einrichtung spielplatz*”. More precisely, each English synonym is mapped to its German equivalent and classified for further processing steps. The knowledge base stores a synonym ontology for each supported language to handle requests in multiple languages, as, for now, English and German. Furthermore, the synonym ontology is used to define a set of modifiers supported by the system. Figure 2.4 exemplifies the definition of the modifiers “*no*” and “*and*”.

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. More precisely, the lightweight grammar is used

- to define permitted positions of operators (for instance, “*and*”, “*or*”, “*not*”, “*near*”) in relation to concepts,
- in analogy to the definition of operators, to restrict the position of quantifiers,
- to describe patterns of permitted strings,
- to associate the informal concept representation with appropriate SQL fragments.

Moreover, a major task is to maintain independency from application logic. Therefore, parameterized SQL fragments are used to build a single SQL statement representing the natural language query. The mapping process is described in detail in Subsection 2.3.4.

2.3.2 Language Identification

To identify the language of a query, an n-gram-based text classification approach (cf. Cavnar and 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*

German		English	
en_	1786	_th	1333
er_	1570	the	1142
de	949	he	928
der	880	_of	592
ie_	779	of_	575
ich	763	_an	439
ein	730	nd_	407
sch	681	_in	389
ch_	642	ion	385
che	599	and	385

Table 2.2: Top ten *tri-gram* occurrences of German and English text with underscores representing blanks

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.

As an example, the top ten *tri-gram* occurrences in the German and English language texts are shown in Table 2.2. In the English text, it can be seen that $\{-th, the, he_, _of \}$ are the most frequent *tri-grams*. Contrarily, in the German text, the most frequent *tri-grams* are endings like $\{en_, er_, _de, der \}$ and *tri-grams* like $\{ich, ein \}$.

To determine the language of a query, the n-gram profile, $n = 1...5$, of the query string is built as described above. The distance between two n-gram

profiles is computed by a rank-order statistic. For each n-gram occurring in the query, the difference between the rank of the n-gram in the query profile and its rank in a language profile is calculated. For example, the *tri-gram* $\{the\}$ might be at rank five in a hypothetical query but is at rank two in the English language profile. Hence, the difference in this example is three. These differences are computed analogously for every available language. The sum of these differences is the distance between the query and the language in question. Such a distance is computed for all languages, and the language with the profile having the smallest distance to the query is selected as the identified language, in other words, the most probable language of the query. If the smallest distance is still above a certain threshold, it can be assumed that the language of the query is not identifiable with a sufficient accuracy. In such a case the user will be asked to specify her or his language manually.

2.3.3 Error Correction

To improve the 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:

- the edit distance between the misspelled word and the suggested word, and
- the edit distance between the misspelled word's *metaphone* and the suggested word's *metaphone*.

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 knowledge 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. Consider, for example, the following query:

I am loking for an acommodation in Kitzbühl featuring a wellness area.

Several misspellings are identified by the error correction algorithm. Both “*loking*” and “*acommodation*” are automatically replaced by their corrected representations. Moreover, even most Austrians don't know how to spell the city of “*Kitzbühl*” correctly. In this case, the phonetic error correction strategy replaces the misspelled with the correct word, i.e. “*Kitzbühel*”.

Another important issue in interpreting the natural language query is to detect terms consisting of multiple words. Proper names like “*Bad Kleinkirchheim*” or substantives 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.

2.3.4 Mapping Natural Language Queries to a Formal Representation

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 relevant 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 substantive 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. To illustrate this processing step, consider the following SQL fragment as the condition for an accommodation being located in or not in a particular city, where @OP is a placeholder for an operator and @PARAM for the city name.

```
SELECT entity."EID" FROM entity WHERE entity."CID" = city."CID" AND
city."Name" @OP @PARAM
```

Depending on modifying terms found in the query as specified in the grammar, the SQL fragment is selected and the parameters are substituted with the appropriate values. The query for an accommodation in Innsbruck produces the following SQL fragment.

```
SELECT entity."EID" FROM entity WHERE entity."CID" = city."CID" AND  
city."Name" = 'Innsbruck'
```

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.

2.4 Discussion

In this Chapter an approach for an information retrieval system that allows query formulation in multiple languages was presented. In order to interact with the system, users are able to express their questions via a single text area by means of a natural language query. First, the natural language query is received by the system and, furthermore, an automatic language identification process based on n-gram text profiling techniques determines the language of the query automatically. Moreover, several processing steps are applied to the query. These processing steps interact closely with the knowledge base of the system. The knowledge base represents the backbone of the system and consists of several ontologies to model linguistic synonymy, define grammar rules and provide a means for generating SQL queries that represent the intentions of the users.

The processing, in turn, is accomplished by a pipeline structure of separate modules. This particular layout facilitates the extension with new or adapted pipeline components and, thus, offers maximum flexibility in changing the underlying domain. The determined output is represented by an XML file, that stores all available domain dependent attributes and can be transformed according to the needs of the client device.

To reduce word sense ambiguity, the approach was set up to operate on restricted domains. In particular, this was exemplified in a prototypical manner for the tourism domain. The system allows natural language access to tourism information provided by the Austrian tourism platform *Discover*.

Chapter 3

What Users Really Want to Know from Tourism Information Systems

3.1 Introduction

In order to test the natural language interface, a field trial was carried out during a ten-day period (March 15 to March 25, 2002). During that time, the interface was accessible via a hyperlink from the *Tiscover* homepage. The time for the trial was chosen deliberately because close to vacation periods, as the Easter week in this case, the traffic at a web-based tourism information system is higher than during other times.

The major objectives for the field trial were, first, to verify whether or not users accept natural language interaction. More precisely, will users actually type natural language sentences to describe their information needs? Second, to gather a broad spectrum of natural language requests for tourism information, now that the users are no longer biased by available tick-boxes, radio buttons or selection lists. Finally, to get an impression of the practical performance of the natural language interface given a real-world setting.

In contrast to the unsupervised field trial, in a usability study subjects

were asked to identify themselves with predefined tourism related scenarios. Subsequently, they had to accomplish the tasks described in these scenarios. On the one hand, subjects were requested to use the standard *Tiscover* interface to fulfill the stipulated tasks, and on the other hand, they were asked to use the natural language interface to retrieve results. The experiences users made during the interaction with the interfaces have been observed and, subsequently, analyzed to get a means for comparing both types of interfaces.

In the remainder of this chapter emphasis is put on the findings from the analysis of the natural language queries collected and processed during the field trial. Moreover, the results of the usability study are presented and discussed in Section 3.4.

3.2 Related Work

Various issues can be addressed by evaluating natural language systems. As Whittaker and Stenton (1989) point out, there is no clear way, of how such evaluations should be carried out and which issues should be regarded as appropriate. Nevertheless, they outline some of them, as for instance, “*coverage*”. “*Coverage*” is concerned with the input the system should be able to handle and how a set of inputs can be determined. Moreover, “*learnability*” deals with the aspect of incompleteness of the capabilities of natural language systems. Each natural language system has its limitations and users will have to learn to communicate within this limitations. Furthermore, “*general software criteria*” like performance, adaptivity, portability are of great importance. The major concerns of the authors were the evaluation of several techniques for assessing natural language interfaces and the effect of using natural language as a means for querying information systems.

Two prototypical studies focusing on learning how language was used for interaction, what language facilities were used and on identifying system requirements in an operational environment are discussed in Ogden and Bernick (1997). The first, carried out by Krause (1980), evaluates the use of

the User Speciality Language (USL) system for answering database queries in a real application. In particular, USL is a natural language interface to access information about grades of students attending a school in Germany. Teachers used the system over a one-year period asking questions, as, for example, if early grades predict later success. 7,300 questions have been asked during this period and Krause focused on 2,121 questions obtained from a single user. It remains unclear, whether or not the user actually received appropriate answers since there is no information provided on this data. A major result of the trial is, that it was easier for the user to recover from semantic than from syntactic errors. More precisely, reformulating questions to circumvent semantic errors turned out to be easier, than if a change of syntactics was necessary (e.g. semantic adaptation: “which student goes to class X” demands a refinement to “which student attends class X”; syntactic adaptation: “show the class X students”).

The second study was carried out by Damerau (1981) and evaluates the Transformational Question Answering (TQA) system. Users were able to ask questions concerning parcels of land in a city over a one-year period. 788 queries were obtained during the trial. Again, no data about result quality and, therefore, the degree of user satisfaction is provided. Results are mainly descriptive but the authors report that users had positive attitudes towards the system and, thus, towards formulating queries in natural language.

3.3 Field Trial

To verify the acceptance and functionality of the natural language system, a field trial, following the motives below, was designed:

- Evaluate, if natural language interfaces are accepted by users; i.e. determine their willingness to formulate queries in natural language.
- Obtain a broad set of queries to determine what users really want to know from tourism information systems.

- Determine, if shallow language processing is sufficient to respond properly to users' requests.
- Moreover, test the stability of the system and its performance in a real operational environment.

The field trial was carried out from March 15 to March 25, 2002. During this time the natural language interface was promoted on and linked from the homepage of the largest Austrian tourism platform, *Tiscover* (cf. Pröll et al. (1998, 2001)). *Tiscover* provides a uniform interface for accessing tourism information of a vast number of independent tourism service providers and regions. In other words, *Tiscover* allows homogenous access via a single interface to tourist information offered by various cities, regions, hotels, etc. Moreover, a generic approach facilitates the integration of individual customer needs into a portal-like environment.

3.3.1 Design Considerations of the Interface

The major design goal at the outset of the project was to provide a simple and easy to use interface. Hence, the interface is dominated by a text-box where the user can enter her or his query and a submit button, the latter one labeled with “ask”. During the field trial short textual descriptions in both German and English in form of sample queries were provided. The sample query “*I am looking for a double room in the center of Salzburg with indoor pool.*” is the only hint on the capabilities of the interface. The intention was to collect a broad range of accommodation requests and, thus, to find out what the users really want. The aim was not to bias the users' imagination when formulating a query. This, admittedly, with the risk of disappointing the user when no or just inappropriate results were found. See Figure 3.1 for a screen-shot of the interface as it looked like during the field trial. For the curious reader we shall note that the interface is still accessible at <http://www.tiscover.at/powersearch>.

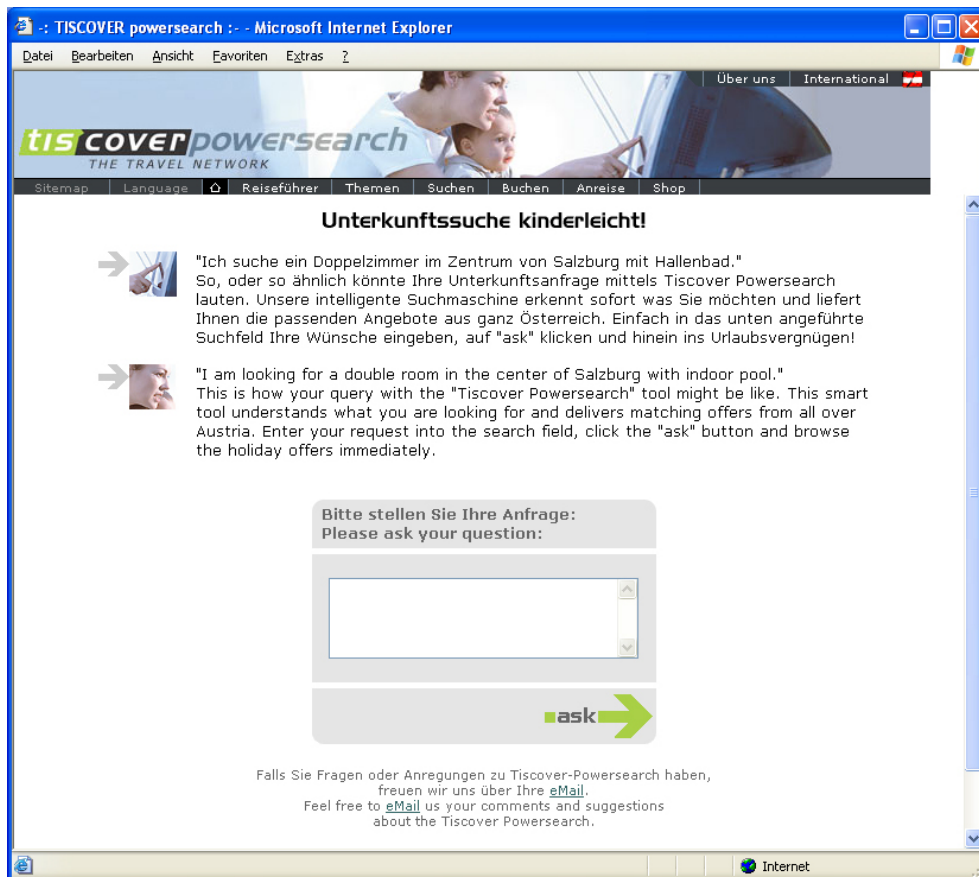


Figure 3.1: Natural language query interface

Figure 3.2 shows the conventional interface of *Tiscover* for searching accommodations. The area (federal state, region, city) can be chosen either by typing the name directly into the text field or via clicking through the hierarchy of names of geographical locations. Further criteria are the name of the accommodation, the chain it belongs to and, perhaps, a particular *theme*, e.g. family hotel, as well as several amenities the accommodation should provide. Note, this list of amenities is rather small compared to the complete information of the *Tiscover* database to keep the interface concise.

We also implemented the look and feel of the *Tiscover* design in order to avoid distraction from the user's task. On the result screen (see Fig-

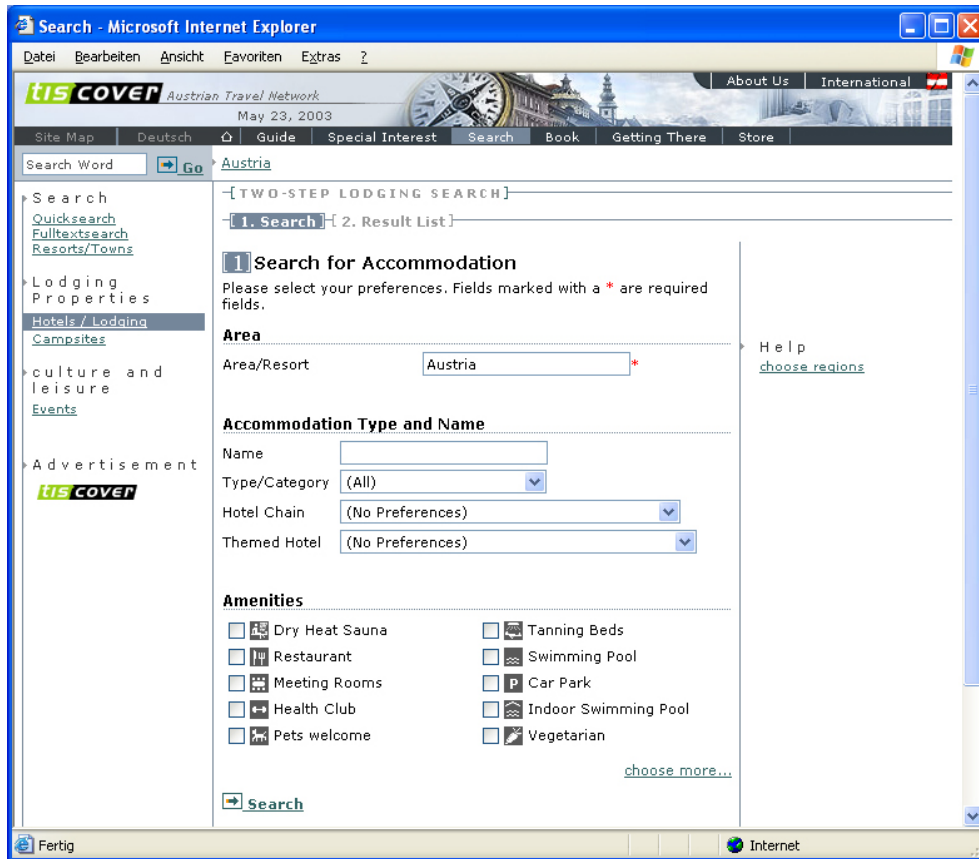


Figure 3.2: Standard *Tiscover* search interface

ure 3.3), the original query as well as the concepts identified by the natural language processing are presented to provide the user with feedback regarding the quality of natural language analysis. Below the list of accommodations matching the criteria, a feedback form where users can enter a comment and rate the quality of the result is provided. After the field trial, it turned out that only 3.37% of the queries have either been annotated or rated where the number of positive and negative comments were almost equal. Due to the unsupervised nature of the trial without any reward for the users, this figure is not surprising because of the additional time it takes to assess the quality of the result and then comment on it. At the bottom of the page, the input field prefilled with the posed query is presented to allow for convenient

query reformulation or refinement. About 10% of the queries were modified by adding or deleting parts of the original query.

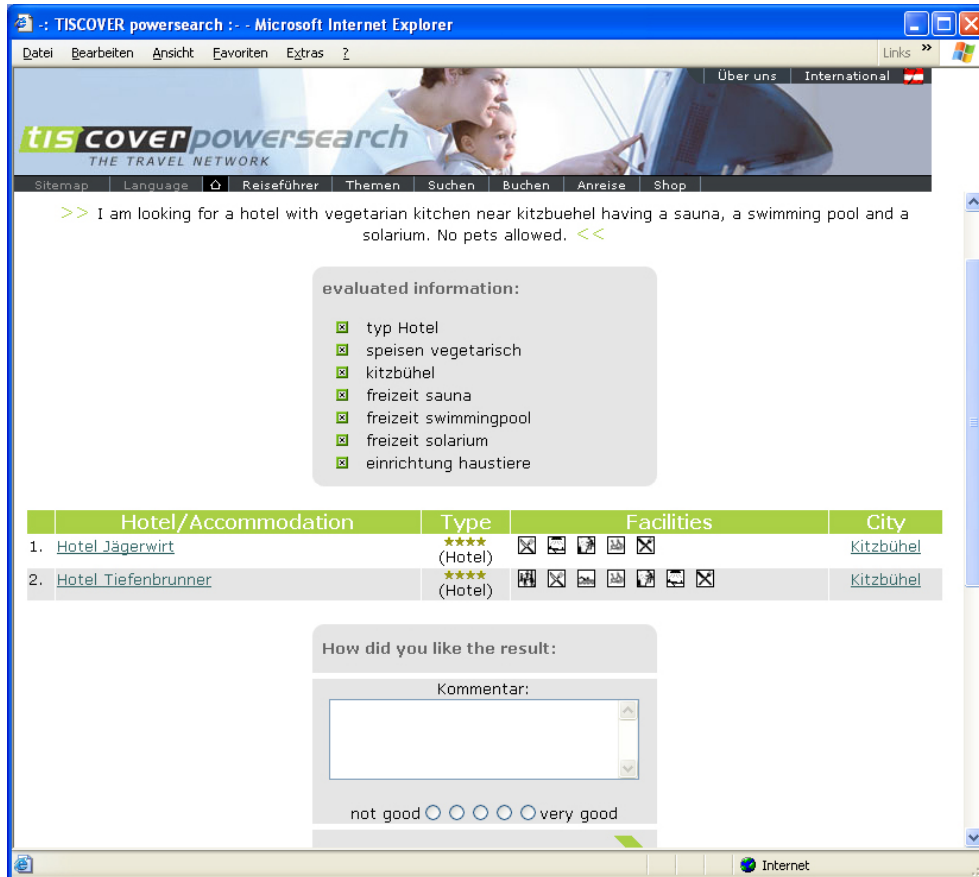


Figure 3.3: Result page with matching accommodations and feedback form

3.3.2 Results from the Field Trial

During the ten-day period 1,425 unique queries were obtained through the 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 the evaluation of the query complexity.

In Table 3.1, a list of countries and the respective numbers of queries is shown. Naturally, most of the queries (39.73%) came from Austrian hosts,

followed by hosts from the *.net* top-level domain, most of which have been identified as German internet service providers by manual inspection. After the 13.13% of queries from the US commercial domain several European countries can be found. A country could not be assigned to 20.42% of the queries because of a non-resolvable domain name.

# of queries (%)	country
566 (39.73%)	Austria
229 (16.07%)	.net (mostly German ISPs)
187 (13.13%)	US commercial
70 (4.91%)	Germany
22 (1.54%)	Switzerland
17 (1.19%)	Italy
14 (0.98%)	Netherlands
8 (0.56%)	UK
6 (0.42%)	Luxembourg
5 (0.35%)	Hungary
4 (0.28%)	Belgium
2 (0.14%)	South Africa
2 (0.14%)	Australia
1 (0.07%)	US military
1 (0.07%)	France
291 (20.42%)	unknown (not resolved)

Table 3.1: Origin of queries (derived from the top-level domain of the accessing host)

Of those 1,425 unique queries, 1,213 (85.12%) were identified as German, 120 (8.42%) were identified as English and 92 (6.46%) were not identifiable, e.g. non-sentence queries like *“hotel salzburg”* that are possible in both languages or just nonsense like *“ghsdfkjg”*. Based on the 1,333 identified queries 85 queries that were not in the scope of the natural language interface were found. Among these were, for example, questions about car rentals and, of course, sex. Obviously, in any kind of publicly available service like this, not all of the people are using it for the intended purpose. However, this number is rather low assuming the rather short description displayed on the start page to give an idea what kind of information can be queried.

To assess the overall quality of the language identification the submitted natural language queries were manually inspected. For each query the system assigns either the most probable language or considers the language of the query as being ambiguous. For example, a German query can either be identified correctly, as English or as ambiguous. In Table 3.2, the actual figures resulting from the manual inspection are provided. Thus, of the 1,213 queries identified as German, 1,210 were correctly identified. However, three in fact English queries have been misclassified as being German. In the third row of the table, it can be seen that of the 92 queries identified as ambiguous, 74 were actually German. This classification error can be explained by the peculiarities of the language identification algorithm based on n-grams. Especially short queries lead to n-gram distributions that do not allow to distinguish between English and German with the required accuracy. In total, of the submitted queries 1,306 were in fact German of which 1,210 were correctly identified. This yields an identification accuracy of 92.6% for the German language. The respective result for the English language is 95.1%.

	manual analysis			
	german	english	ambiguous	
german	1,210	3	0	1,213
english	22	98	0	120
ambiguous	74	2	16	92
totals	1,306	103		
identification accuracy	92.6%	95.1%		

Table 3.2: Manual analysis of language identification accuracy

To provide some technical information, for the 1,333 processed queries, the mean processing time was 2.63 seconds with a standard deviation of 1.42 seconds. The median of 2.27 seconds shows that there were only a few outliers with longer processing times. Given these figures, it can be safely said that the system is usable regarding its response time. Even with adding a few seconds for data transmission time over the Internet, the response time

still lies below the magic number of ten seconds as suggested by Nielsen (2000). These ten seconds have been measured in usability studies as the approximate maximum attention span of users when waiting for a web page to be loaded before cancelling the request.

Furthermore, the results of two studies analyzing query log files of the large and popular search engines *Altavista*² and *Excite*³ are compared with the results of our analysis, since only few research papers dealing with user behavior in web searches exist. In Jansen et al. (1998) and Silverstein et al. (1998) the authors have shown that the average number of words per query is very small, namely 2.35, interestingly the same in both studies. This indicates that most of the people searching for information on the Internet could improve the quality of the results by specifying more query terms. The field trial revealed the encouraging result of an average query length of 8.90 words for German queries, and of 6.53 for the English queries, see Table 3.3 for details. We regard this as a strong first hint backing our expectations that users accept to type natural language queries.

In more than a half (57.05%) of the 1,425 queries, users formulated complete, grammatically correct sentences whereas only 21.69% used the interface like a keyword-based search engine. The remaining set of queries (21.26%) were partial sentences like “*double room for 2 nights in Vienna*”.

Table 3.4 depicts three German and three English queries obtained during the field trial. No error correction mechanisms have been applied to the queries and, therefore, they represent queries as typed by users. Several of the queries consisted of more than one natural language sentence. This approves the 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 the users is typing complete sentences to express their information needs. Furthermore, the average number of relevant concepts occurring in the German queries is 3.41 with a standard

²<http://www.altavista.com/>

³<http://www.excite.com/>

words/query	# of queries (%)	words/query	# of queries (%)
1	76 (5.33%)	18	12 (0.84%)
2	92 (6.46%)	19	14 (0.98%)
3	117 (8.21%)	20	6 (0.42%)
4	82 (5.74%)	21	9 (0.63%)
5	109 (7.65%)	22	5 (0.35%)
6	147 (10.32%)	23	8 (0.56%)
7	87 (6.11%)	24	2 (0.14%)
8	105 (7.37%)	25	7 (0.49%)
9	98 (6.88%)	26	2 (0.14%)
10	101 (7.08%)	27	3 (0.21%)
11	66 (4.63%)	28	3 (0.21%)
12	75 (5.25%)	29	2 (0.14%)
13	55 (3.86%)	32	1 (0.07%)
14	53 (3.90%)	35	1 (0.07%)
15	30 (2.11%)	37	3 (0.21%)
16	22 (1.54%)	66	1 (0.07%)
17	28 (1.96%)	76	1 (0.07%)

Table 3.3: Word occurrence statistic

doppelzimmer mit zusätzlicher schlafmöglichkeit für 2 kinder in einem schigebiet in salzburg oder tirol. die gesamtkosten für halbpension dürfen pro tag max. 110 euro betragen.
ich möchte in tirol nicht aber im zillertal mountainbiken und paragleiten und suche ein 4 sterne hotel mit hallenbad, sauna und eventuell kinderbetreuung
ich suche ein zimmer fuer 2 erwachsene und 2 kinder im alter von 11 und 14 jahren einschliesslich skipass zum skifahren von 30 maerz bis 6 april
we are looking for ski vacation for 2 adult. neer munich- zalsbourt or insbruck area. ski lesons for beginners and rental equipment. accommodation neer by 11- 16 apr. thanks for your help . mrs michal greenstein
We look for a house at one of the lakes in Austria from July 22 until July 28, 2002. We are a familiy with 2 children of 8 and 11 years and have a dog. We are searching for a house with lake entrance.
please show farms in upper austria that are suited for children an that provide sauna.

Table 3.4: Subset of natural language queries obtained during the field trial

deviation of 1.96, which is still one word per query more than found in the surveys mentioned above. It can be assumed, that, by formulating a query in natural language, users are more specific than compared to keyword-based searches.

To inspect the complexity of the queries, the number of concepts and the usage of modifiers like “*and*”, “*or*”, “*not*”, “*near*” and some combinations of those as quantitative measures are considered. Table 3.5 shows the distribution of the numbers of concepts per query. For example, consider row four of this Table. The entries in this row show the number of queries with three concepts. In particular, we have 310 German and 28 English queries. Note that these figures were derived by manual inspection of the users’ original natural language queries. The majority of German queries contains one to five concepts relevant to the tourism domain with a few outliers of more than 10 concepts. The latter can be explained by people asking for an accommodation in a specific region by enumerating potentially interesting cities and villages.

concepts	query language		
	german	english	totals
0	47	5	52
1	77	28	105
2	272	38	310
3	310	28	338
4	245	12	257
5	137	5	142
6	49	2	51
7	38	1	39
8	18	1	19
9	11	0	11
10	4	0	4
11	1	0	1
17	3	0	3
21	1	0	1
totals	1,213	120	1,333

Table 3.5: Number of concepts per query (counted by manual inspection)

In analogy to Table 3.5, the Tables 3.6 (a) and 3.6 (b) give an indication regarding the quality of the natural language query analysis. In particular, Table 3.6 (a) provides the numbers of identified concepts per query, whereas Table 3.6 (b) that of not identified concepts. Again, the figures given in Table 3.6 (b) were derived by manual inspection. However, most of the concepts not identified, originated from queries falling into the categories of region names, pricing information, room availability and arrival and departure dates. These information were not disclosed to us by *Tiscover* and, thus, were not contained in the part of the database used for the natural language information retrieval system.

Another aspect of the complexity of natural language queries are words connecting concepts logically or modifying their meaning. These modifiers can be compared to operators like “AND”, “OR”, “+” or “−” of web search engines. In Table 3.7 (a) it can be seen, that the distribution of occurrences of the modifier “*and*” corresponds to the number of concepts. In 320 queries the modifier “*and*” was used twice which relates to the occurrence of three concepts per query (cf. Table 3.5). The occurrence statistic includes all implicitly used “*and*” modifiers, i.e. those “*and*”s that are included because of the resulting SQL statement, as well as those explicitly defined, i.e. those “*and*”s that are provided with the natural language query. Just to provide an example, the query

I am looking for a hotel with sauna, solarium and whirlpool in Tyrol

includes one explicitly used “*and*”, and three implicit “*and*” modifiers. Due to the assumption that the underlying semantics of combining concepts is based on the intention to provide facilities somebody wants to have, the “*and*” modifier was defined to be the default logic for combining concepts if no explicitly defined modifier is present. This assumption is made to provide a convenient technique to map the concepts used in a query onto the underlying program logic.

The modifier “*or*” is used far less frequently than “*and*”, as shown in

(a) Concepts identified by the natural language processing

concepts	query language		
	german	english	totals
0	71	14	85
1	104	27	131
2	326	39	365
3	312	24	336
4	201	10	211
5	106	2	108
6	50	2	52
7	19	2	21
8	13	0	13
9	6	0	6
10	1	0	1
16	3	0	3
20	1	0	1
totals	1,213	120	1,333

(b) Concepts not identified by the natural language processing

concepts	query language		
	german	english	totals
0	817	88	905
1	348	29	377
2	45	3	48
3	3	0	3
totals	1,213	120	1,333

Table 3.6: Concepts that have been identified or not identified by the natural language processing module of our interface

Table 3.7 (b). In particular, “*or*” is used in 103 queries only. “*Or*” is mostly used to provide a set of locations or types of accommodations of interest, e.g. “*I am looking for a farm or an apartment in Tyrol or Salzburg*”.

An interesting fact is, that the “*not*”-modifier is used in a very small subset of queries (cf. Table 3.8 (a)). The modifier “*not*” occurs in only 19 German and 3 English queries. This implies, that the vast majority of users formulate their intentions without the need of excluding concepts. In

(a) Usage of modifier “and”

	query language		
and	german	english	totals
1	281	38	319
2	320	29	349
3	246	11	257
4	140	6	146
5	41	1	42
6	33	1	34
7	16	0	16
8	4	0	4
9	2	0	2
10	1	0	1
totals	1,084	86	1,170

(b) Usage of modifier “or”

	query language		
or	german	english	totals
1	67	4	71
2	18	1	19
3	6	1	7
6	1	0	1
8	1	0	1
12	3	0	3
16	1	0	1
totals	97	6	103

Table 3.7: Usage of modifiers “and” and “or”

most of the cases where a “not” is used to exclude a specific property of a region or an accommodation, users wanted to avoid places where pets are allowed as well as accommodations that are *not* particularly well-suited for children, the latter, perhaps, to stress the desire to find a quiet place. Another common use of “not” is to exclude one or more cities from a query where an accommodation in a federal state or region was wanted, e.g. “*I am looking for a hotel in Tyrol, but not in Innsbruck and not in Zillertal.*”

Table 3.8 (b) shows the number of occurrences of the modifier “near” which has been expressed by terms like “around”, “close to” or “near” itself.

Generally, geographical concepts or relations are essential to provide a high-quality tourism information service. Comparing the modifier usage statistics a remarkable detail is noticeable. In 122 out of 1,425 queries (8,6%) the modifier “*near*” is used. This circumstance makes “*near*” to the modifier second-most frequently used, in the queries collected during the field trial. A common way to use “*near*” is to find accommodations in the surroundings of popular sites, cities or facilities, e.g. “*I am looking for a hotel with sauna and pool in St. Anton near the Galzig-Seilbahn*”.

(a) Usage of modifier “*not*”

	query language		
not	german	english	totals
1	12	3	15
2	7	0	7
totals	19	3	22

(b) Usage of modifier “*near*”

	query language		
near	german	english	totals
1	112	9	121
2	0	1	1
totals	112	10	122

Table 3.8: Usage of modifiers “*not*” and “*near*”

Table 3.9 (a) illustrates the combined usage of the modifiers “*and*” and “*or*”. Most commonly used is a combination of one “*or*” and several “*and*” modifiers, e.g. two “*and*” and one “*or*” are used in 17 German queries. As shown in Table 3.9 (b), the usage of “*near*” corresponds with the presence of an “*and*” modifier.

We can say that the sentence complexity, i.e. the frequency of concept combination, is relatively low. In general, queries are formulated on the basis of combining concepts in a simple manner, e.g. “*I am looking for a room with sauna and steam bath in Kirchberg*”. Only a small subset of queries consist of complex sentence constructs that would require a more sophisticated sentence evaluation process. For instance, if the scope or type of the modifier cannot be

		query language		
and	or	german	english	totals
1	1	9	1	10
	2	3	0	3
2	1	17	2	19
	2	3	0	3
3	1	16	0	16
	2	5	0	5
	3	2	1	3
	6	1	0	1
4	1	12	1	13
	2	3	0	3
	12	3	0	3
	16	1	0	1
5	1	8	0	8
	2	1	1	2
	3	2	0	2
6	1	2	0	2
	2	2	0	2
	3	2	0	2
7	1	2	0	2
8	1	1	0	1
	2	1	0	1
totals		96	6	102

(a) Combined usage of modifiers “and” and “or”

		query language		
and	near	german	english	totals
1	1	18	1	19
2	1	32	2	34
3	1	26	1	27
4	1	21	4	25
5	1	7	0	7
	2	0	1	1
6	1	2	1	3
7	1	2	0	2
8	1	1	0	1
totals		109	10	119

(b) Combined usage of modifiers “and” and “near”

Table 3.9: Combined usage of modifiers

determined correctly. As an example, consider the query “*I am looking for an accommodation in Serfaus, Fiss or Ladis*”. For the reader who is not familiar with the geography of Austria, in particular Tyrol in this case, we shall note that “*Serfaus*”, “*Fiss*”, and “*Ladis*” are names of towns, and collectively they refer to an attractive skiing resort. In contrast to the assumption that the default operator of combining concepts is “*and*”, the modifier “*or*” must be used to combine the geographical concepts in this sample query.

The fact that the level of sentence complexity is not very high suggests, that shallow text parsing should be sufficient to analyze the queries emerging in a limited domain like tourism. Nevertheless, regions or local attractions are important information that have to be integrated in such systems. Moreover, users’ queries contained vague or highly subjective criteria like “*romantic*”, “*cheap*” or “*within walking distance to*”. Even “*wellness*”, a label broadly used in tourism nowadays, is far from being exactly defined. These concepts are difficult to model, however, this circumstance is addressed in the approach presented in Chapter 5.

3.3.3 Lessons Learned from the Field Trial

Basically, the information retrieval system fulfilled its intended purpose but several flaws that implied a redevelopment of parts of the system were recognized. Strictly speaking, the underlying knowledge base reached its limitations as the integration of semantic relations between information items is difficult in the non-hierarchic structure of the original prototype. Regarding the queries obtained during the field trial, the flat structure limits the power of the system and, therefore, restricts the ability of providing appropriate recommendations. Furthermore, the integration of subjective criteria, region and location dependent information implies the use of an alternative approach that combines the features of the original system with the demands derived from the results of the field trial.

3.4 Usability Study

In addition to the field trial a usability study, comparing the standard *Tiscover* interface with the natural language approach, was carried out. In particular, by carrying out the usability study, users implicitly judged both interfaces according to the following criteria:

- **Learnability** indicates how *easy* users learn to interact with an interface. As a means for determining the degree of *learnability*, the time a user takes to accomplish certain tasks is measured. Therefore, the time to deal with the whole set of features of the interface as well as to interact properly is taken into account (cf. Nielsen (1994)).
- **Memorability** defines the ability to remember features of an interface, i.e. how easy it is to accomplish a certain task if the user is already familiar with the interface.
- An **error** defines an action that hinders the accomplishment of a pre-defined task. The number of errors made is counted and expressed by the error rate.
- **Satisfaction** determines the degree of contentment during interacting with the interface. To measure this subjective criterion, users are simply asked for their opinion. Finally, the statements are compared with each other to derive a set of more general opinions.

3.4.1 Test Design and Processing

Basically, the test itself consists of first, a pilot test, used to check the integrity and design of the test, and second, the usability test. The aim of the test design is, to get a means for comparing the standard *Tiscover* interface with the natural language approach. Moreover, the main objectives of the usability study are subsumed as follows:

- Evaluate the acceptance of natural language interfaces.

- Determine, in which way users interact with natural language interfaces.
- Gather information about the degree of satisfaction of the results determined; both with the standard interface and the natural language interface.
- Furthermore, derive hints for further improvements.

After evaluating the results of the pilot test, 17 subjects took part in the usability test. The guideline through the test was a questionnaire containing information about the two types of interfaces, questions to gather some background data about the subjects themselves and a detailed description of the five scenarios the subjects had to accomplish.

An important task of the usability study was the development of five scenarios common to the tourism domain. Each of these scenarios acts as starting point for the following search process. More precisely, the users were asked to accomplish the search process with both the standard *Tiscover* interface and the natural language interface.

<p>Imagine, you and your partner would like to enjoy your holidays in a pension and you take your two dogs with you. Moreover, you and your partner are fond of extensive bicycle tours but you do not want to take your own bicycles with you.</p>

Table 3.10: Example scenario from the tourism domain

Table 3.10 depicts one of the five scenarios the subjects accomplished during the usability study. Moreover, the subjects used the interfaces in alternating order, i.e. for the first scenario subjects started their search using the natural language interface followed by the standard *Tiscover* interface. For the second scenario subjects started with the standard interface followed by the natural language interface. A comprehensive report about the test

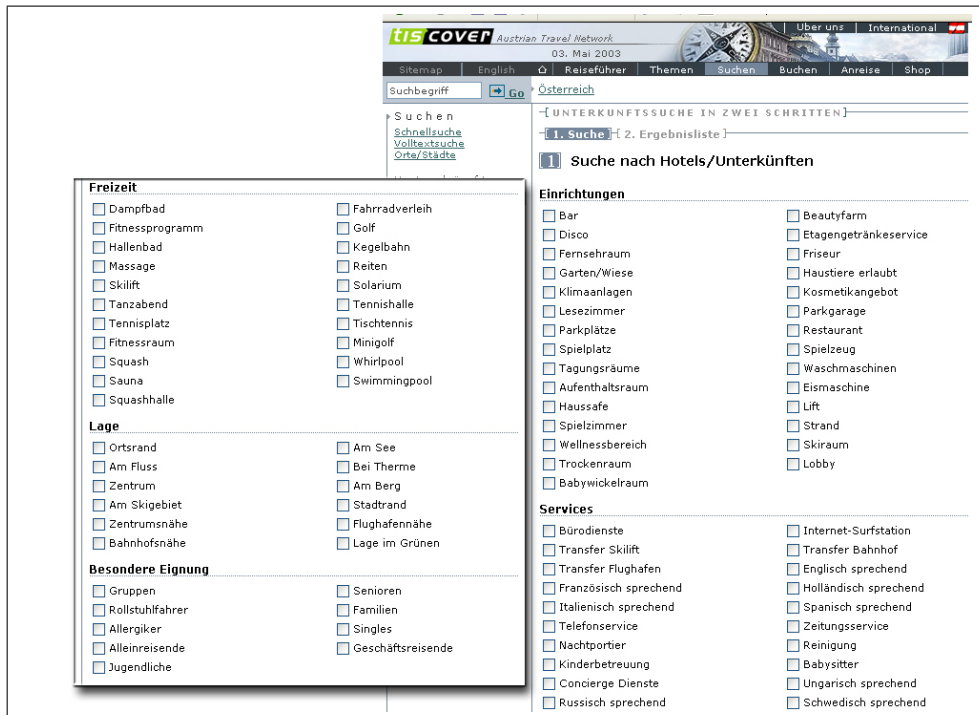


Figure 3.4: *Tiscover*'s advanced search (part of)

procedure, the questionnaire and the scenarios can be found in Pribernig (2003).

3.4.2 Results from the Usability Study

The usability study revealed that users found it complicated to formulate queries with the standard *Tiscover* interface. More precisely, it was difficult to map the scenarios onto formal search queries as stipulated by this kind of interface. This might be due to the fact, that at the time the usability study was carried out, only a very limited number of features was accessible via the standard interface. At the time of writing, *Tiscover* adapted its strategy and integrated a broad set of additional attributes available on a separate advanced search page.

The newly integrated advanced search page of *Tiscover* (cf. Figure 3.4,

two overlaid screenshots of the interface still don't show all available options), on the one hand, might solve several problems caused during the usability study, yet, on the other hand, new problems arise. Without regarding the language inconsistency of the page, it's questionable, if users might be overwhelmed by such overloaded interfaces. A single page containing nearly 100 checkboxes doesn't encourage clarity (cf. Raskin (2000)). The natural language approach eliminates overloaded interfaces by using a single text field offering the same functionality with the additional advantage for the user of expressing intentions in her or his own words. Therefore, the vast majority of users considered the natural language interface as a *more comfortable* means for expressing their intentions and to phrase queries based on the scenarios.

Furthermore, subjects showed broad acceptance for formulating queries in natural language and stated the preference for using this kind of interfaces. Moreover, the subjective sensation of usability of the natural language interface surpasses the one of the standard interface. Although, in contrast to the assumed intuitiveness of natural language interfaces, subjects requested the integration of a detailed explanation of the functionality and features of the interface. More precisely, some subjects were uncertain about how to phrase a query in natural language.

An interesting phenomenon was the observation of a change in formulating natural language queries during the usability test. At the beginning, most subjects posed complete, grammatically correct, sentences. After a while they formulated partial sentences and finally, the natural language interface was adapted to a form-based keyword search interface.

In order to evaluate the quality of the results, subjects were asked to judge their satisfaction about the matches obtained. The results determined via the standard *Tiscover* interface were to the greatest possible extent unsatisfying. This circumstance goes hand in hand with the restrictedness of the interface and, therefore, the limited possibility to map scenarios to queries. In contrast, subjects are more satisfied with the results obtained via the natural language interface. The simplicity to map scenarios to natural language

queries leads to an improvement in the subjective impression of result quality.

3.4.3 Lessons Learned from the Usability Study

The results of the usability study indicate again that users are willing to accept natural language interfaces. In particular, the simple way of mapping the user's intention to a representative query favors the natural language interface. Moreover, the results determined via the natural language approach showed a higher degree of satisfaction than those retrieved by the standard *Tiscover* interface.

Nevertheless, the natural language approach showed several flaws. First, the limited quality of the phrase recognizer inhibits the power of the system, i.e. improving the recognition rate of multi-word denominations enhances the result quality. Second, as already mentioned in the previous section, the integration of sophisticated geographic data was stipulated by the subjects. Consider, for example, that a user travels from Vienna to Salzburg, and asks the natural language system to provide results about accommodations along the highway she or he is driving on. In order to fulfill this request, geographic coordinates of the course of the highway, and of cities along or in the vicinity of the highway must be available. Finally, subjects criticized the lack of a comprehensible feedback about how the results have been determined as well as which features have been evaluated. Moreover, a relaxation strategy applied to query terms providing additional recommendations was motivated.

3.5 Discussion

The findings of a field trial and the observations made during a usability study have been discussed in this chapter.

During the field trial, carried out for a ten-day period in March, 2002, about 1,400 queries, most of which in German language, have been gathered. Most importantly, the users are willing to type natural language queries to express their information needs. This observation is approved by a compari-

son with web-search engines, where the average number of words per query is substantially smaller than with the original information retrieval system. Moreover, the complexity of these queries is higher than with standard web-search engines. Furthermore, the distribution of various modifiers as well as the combinations of modifiers extracted from the queries has been shown.

The expectation that shallow language processing is sufficient given a limited application domain is backed by the fact that most of the query concepts which had their counterpart in the knowledge base where successfully extracted from the natural language query.

Moreover, by way of this field trial allowing natural language descriptions of information needs as opposed to the strictly limited variability of tabular-based information entry, an impression of what users actually look for was received.

Nevertheless, the field trial revealed that regions or local attractions are important information that have to be integrated in such systems. Moreover, users' queries contained vague or highly subjective criteria like "*romantic*", "*wellness*", "*cheap*" or "*within walking distance to*". These concepts are difficult to model in the knowledge base of information systems and this problem, among others, is addressed in the remainder of this thesis.

In addition to the field trial, a usability study was carried out. The major objective of the study was to compare the standard *Tiscover* interface with the natural language approach. Therefore, subjects have been asked to identify with predefined scenarios from the tourism domain and then to accomplish the stipulated tasks described in the scenarios. These tasks were carried out with both, the standard interface and the natural language interface. Users stated that it is much easier to map their intentions via the natural language interface than with the standard interface. Moreover, the results determined via the natural language approach showed a higher degree of satisfaction than those retrieved by the standard *Tiscover* interface. Finally, a relaxation strategy applied to query terms providing additional matches was motivated.

Chapter 4

Using Network Structures for Knowledge Representation

4.1 Introduction

Representing domain knowledge played ever since a crucial role in the field of artificial intelligence. Crestani and Van Rijsbergen (1997) point out, that the significance of an information item in information retrieval can only be fully captured by considering its semantic relationship with other items.

Based on the supposed structure and functionality of human memory, several approaches for knowledge representation have been developed. Associative networks represent, for instance, a model for defining relations between information items. However, besides the need to represent knowledge in a convenient way, techniques to determine appropriate results have to be applied on these network structures. Spreading of activation which is initially placed on some of the information items can accomplish this stipulated task.

The remainder of this chapter describes the basic structure of semantic and associative networks. Moreover, the pure spreading activation algorithm as a means for information processing in such networks is presented. In order to get control about the spreading process several constraints can be applied and are comprehensively detailed in Section 4.5.

4.2 Related Work

In knowledge representation a decision has to be made, whether a symbolic or a sub-symbolic knowledge representation model will be chosen. In particular, a symbolic domain knowledge representation uses symbols to represent information items. Each symbol corresponds to an information item and what kind of information the symbol is supposed to represent. Nevertheless, symbolic representation models show several drawbacks. First, the difficulty to determine an appropriate domain-structure and level, second, handling the dynamic aspect of domain knowledge, and, third, a symbolic model represents the viewpoint of an expert on the domain and not that of the user.

Contrarily, in the sub-symbolic representation model the correspondence between information items and symbols is not present. Several different atomic elements build the knowledge representation structure. Moreover, information is stored without explicitly associating information items with parts of the knowledge representation structure (cf. Rumelhart and Norman (1988)). Due to the fact, that sub-symbolic representation structures are often adaptive, this model provides a means to overcome some of the problems related to symbolic representation strategies in information retrieval.

One approach of representing knowledge in information retrieval is to use a network model. A network expresses the semantics of information items in terms of associations with other information items of the domain knowledge. Organizing knowledge in a network structure has its roots in the field of psychology. More precisely, to a great extent the structure of a knowledge network is similar to the supposed functionality of human memory. Repovs (2002) summarizes several approaches of defining a specific type of memory, namely semantic memory. In contrast to the research performed in the field of cognitive psychology (focus was laid on the question of how concepts are represented in memory), the first model of the structure of semantic memory and its processing techniques were introduced by Quillian (1968). Information items (concepts) in semantic memory are represented as nodes in a complex network, structured in a hierarchical manner. Each

concept is directly connected to its sub-concepts and to its super-concepts. Moreover, sub-concepts inherit all attributes associated with super-concepts.

Anderson (1983b) describes a similar network model of semantic memory based on concepts represented by nodes. These nodes are connected via links possessing different weights. Weights between nodes are adapted according to the frequency of their use, i.e. the less frequent a link is processed the lower is its associated weight. This approach was introduced as part of the ACT (Adaptive Control of Thought) theory (cf. Anderson (1983a))

In order to process network structures as mentioned above a technique called *spreading activation* can be applied. Crestani (1997) describes the application of spreading activation techniques in the field of information retrieval. A comprehensive overview about spreading activation is given in Preece (1981). Basically, a set of nodes act as activation sources and this activation spreads along the links in a wave-like manner through the network and stimulate adjacent nodes until a termination condition is met. Moreover, Salton and Buckley (1988) mention the application of spreading activation over network structures as a means for a technique called *associative retrieval*. The *associative linear retrieval model* dates back to Salton (1968) and is based on the idea of expanding the original user's query by exploiting statistically determined term to term, term to document, and document to document associations. More precisely, a network is used to refine and expand users' requests based on semantic associations defined in the network.

Another attempt exploiting semantic relationships between information items was introduced by Campbell and Van Rijsbergen (1996). Ostensive information retrieval is an approach to search information without the use of a query and follows the assumption below: "*A user may not be good at describing his information need but is, by definition, able to identify something that is relevant to him - in fact, he is the only agent capable of doing so*". In particular, a user points at relevant documents and other relevant ones, which are estimated to be similar in semantic content, are suggested. In an interactive process other relevant documents are identified by a user

and by resolving semantic relations between documents additional ones are retrieved.

Finally, it should be mentioned, that these network structures are representatives of *connectionist network models*. In particular, connectionism defines a *type of modeling* based upon networks. Connectionist models are sometimes referred to as *Parallel Distributed Processing* (abbreviated as PDP) models or networks. Moreover, connectionist systems are known as *neural networks*. For a detailed explanation on connectionist models see Zell (1994); Kröse and Van der Smagt (1996). The curious reader will find in Oard (1994) a good starting point for references on applications of neural networks in information retrieval.

4.3 Associative Networks

As mentioned in the previous section, 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,

- nodes that express knowledge in terms of concepts,
- concept properties, and
- the hierarchical sub-super class relationship between these concepts.

Each concept in a semantic network represents a semantic entity; the 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

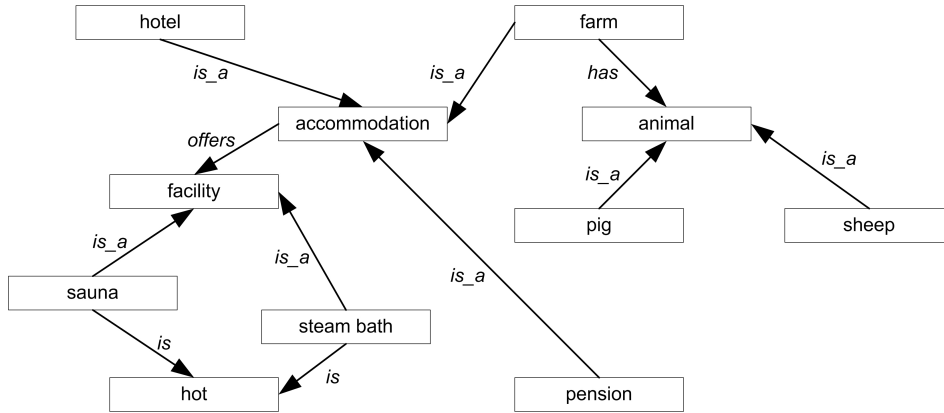


Figure 4.1: A semantic network example of tourism-related terms

to all subsequent nodes. An example of a semantic network is depicted in Figure 4.1.

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.

4.4 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)

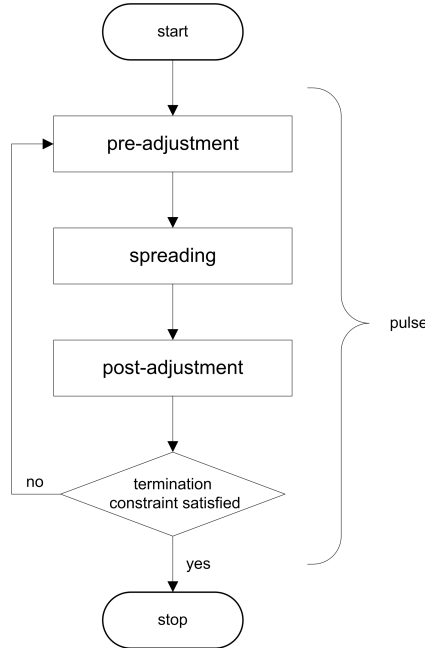


Figure 4.2: Flowchart of the spreading activation model

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* (cf. Figure 4.2). 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:

$$I_j(p) = \sum_i^k (O_i(p-1) \cdot w_{ij}) \quad (4.1)$$

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:

$$O_i = f(I_i) \quad (4.2)$$

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.

4.5 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 is spread based 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.

4.6 Discussion

Knowledge representation is a crucial task in the field of artificial intelligence. In this chapter a description about a knowledge representation model based on a network structure is given. More precisely, the structure of semantic networks as well as a more general approach, namely associative networks was presented. It was pointed out that network structures are used to associate information items according to their semantic relationships. Links between information items in an associative network are weighted according to the degree of the relatedness of these information items.

Moreover, a technique to process such network structures, namely spreading activation, was detailed. To recall, activation which is initially placed on some of the information items is propagated via weighted links to adjacent nodes of the network. Several constraints to control the process of activation propagation have been described.

Chapter 5

An Associative Knowledge Representation Model for Tourism Information

5.1 Introduction

The importance of understanding what users really want to know from information retrieval systems remains a crucial task in the field of information retrieval. An approach leaving the means of expression in users' hands, narrows the gap between users' needs and interfaces used to express these needs. Therefore, a natural language interface allows for easy and intuitive access to information sources.

Still, the core element remains the underlying knowledge representation model. On the one hand, the conceptual model has to offer adequate performance during the search process and, on the other hand, the knowledge representation model must allow for easy integration of additional domain relevant information. To achieve this, an approach based on associative networks reflecting the relationship of information items in the tourism domain is used. Thus, the approach described in this chapter, incorporates a means for knowledge representation allowing for the definition of semantic relation-

ships of domain-intrinsic information. Due to the network structure of the knowledge representation model, implicit query expansion enriches the result set with additional matches.

Nevertheless, setting up such associative networks, i.e. to define appropriately weighted associations between information items, remains a non-trivial task. Hence, the application of natural language interfaces assists in gathering information about users' real interests and in expanding and refining the vocabulary. Moreover, it is shown how past user interactions are used to derive semantic relations between information items.

The remainder of this chapter is organized as follows. First, an overview about the tradition of spreading activation in information retrieval is given. In Section 5.3 the approach based on associative networks is described. Moreover, the processing technique and results obtained are discussed. Finally, an approach for deriving associations between concepts is described in Section 5.3.3.

5.2 Related Work

One of the first information retrieval systems using constrained spreading activation was GRANT. Kjeldsen and 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:

$$\text{IF } x \text{ AND } R(x, y) \rightarrow y \quad (5.1)$$

$R(x, y)$ 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 et al. (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 incidentally. The I³R system acts as a search intermediary (cf. Croft and 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 reached. 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 the ostensive approach with the as-

sociative retrieval approach is described in Crestani and 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 the 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 and 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.

Moreover, 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 et al. (2003b) 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 knowledge stored in the associative network.

5.3 Associating Domain Knowledge

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

- interactions users made during past sessions,
- personal preferences of users,
- semantic relations of domain intrinsic information, and,
- 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. Figure 5.1 depicts the layout of the new pipeline structure. In analogy to the original pipeline, the first three processing steps are accomplished. Next, the *initialization*-element associates concepts extracted from the query with

nodes of the associative network. These nodes act as activation sources. Subsequently, the *spreading*-element implements the process of activation propagation. Finally, the *evaluation*-element analyzes the activation level of the associative network determined during the spreading phase and produces a ranking according to this activation level.

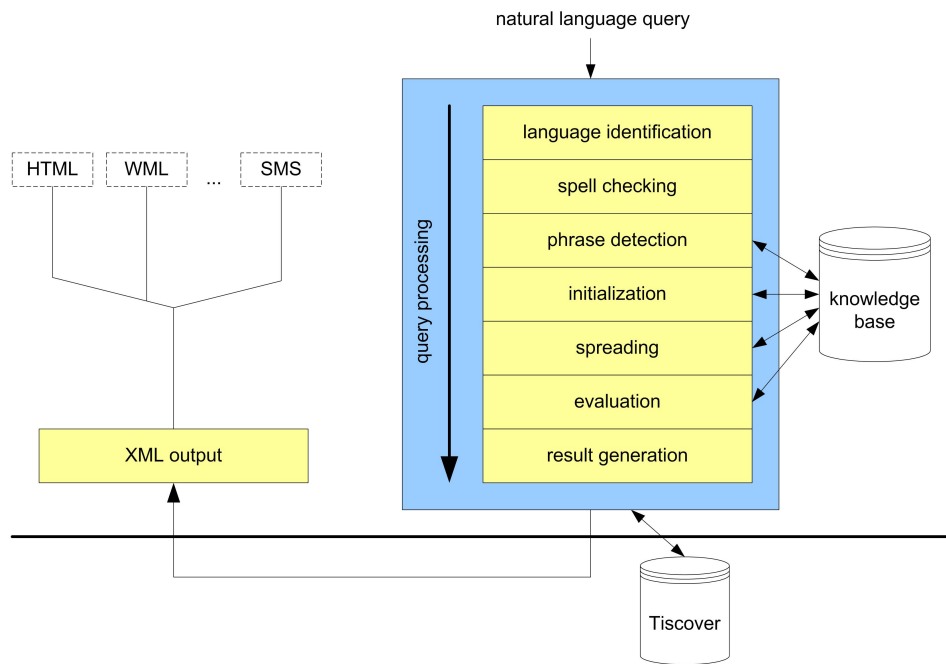


Figure 5.1: Redesigned software architecture

The following sections provide a detailed report about the theoretical background of the approach.

5.3.1 Associative Knowledge Modelling

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 looks as follows:

Entity: "Hotel Wellnesshof"

Attributes: <category, 4>; <facility, sauna>; <facility, solarium>; ...

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. Figure 5.2 depicts the abstract concept "*wellness*". The bold ellipse symbolizes the membership in the group of abstract concepts. Moreover, a second abstract concept is illustrated, namely "*swim facility*". Links between abstract concepts and concepts of other layers are indicated by a solid line. Each weighted link indicates the strength of semantic relationship. The weights assigned in Figure 5.2 are for demonstration purpose only.
- **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. In Figure 5.3 a set of concepts and their weighted associations illustrate the structure of the network. Again, connections and weights are only exemplary.
- **Entity layer:** Finally, the entity layer associates entities with information items (entity attributes) of the conceptual layer, e.g. an entity

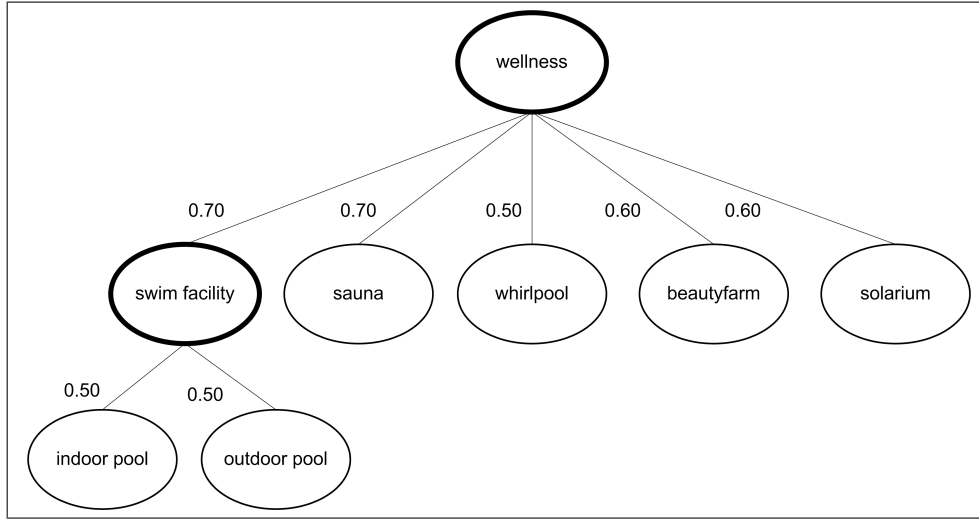


Figure 5.2: Network structure of abstract concepts

possessing the attribute “*sauna*” is associated with the *sauna*-node of the conceptual layer.

To get a better picture of the interdependencies associated with the layers introduced above see Figure 5.4. Each layer holds a specific set of concepts. Abstract concepts associate concepts at the same or at the conceptual layer. Information items 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. Information items at the entity layer are not associated with items at the same layer. Consider, for example, the abstract concept “*indoor sports*” and the concrete 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

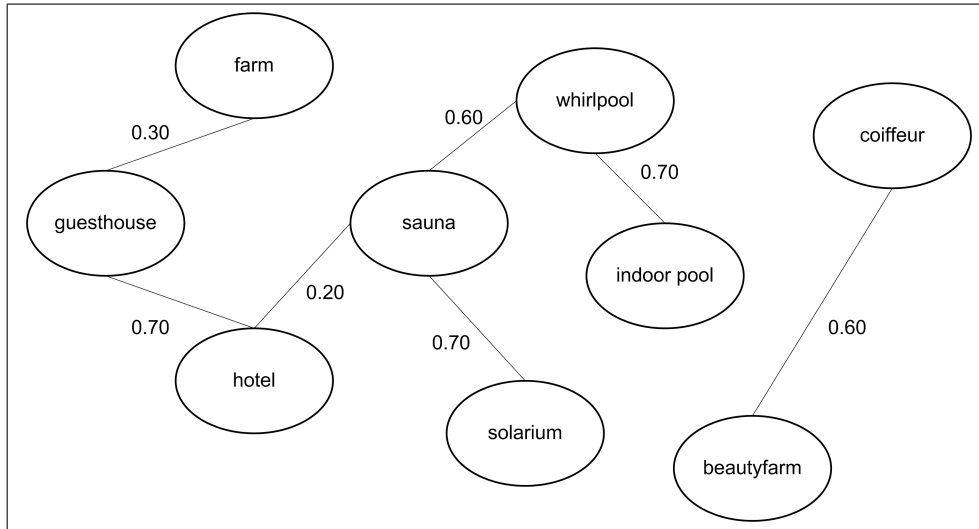


Figure 5.3: Network structure of concepts at the conceptual layer

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.

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. Figure 5.5 depicts an excerpt of the XML representation of a set of sample concepts. Consider, for example, the concept “*dampfbad*”. Each concept is described by a unique id (e.g. id=“*dampfbad*”) and an initial activation level (e.g. initial=“1.0”). Furthermore, each concept has a tag named *role*. Three different role types can be distinguished:

- **Concrete** concepts are used to represent information items at the conceptual layer. More precisely, concrete concepts refer to entity attributes. The concept “*sauna*” in Figure 5.5 illustrates an example of a concrete concept.
- Concepts with an **abstract** role refer to terms at the abstraction layer.

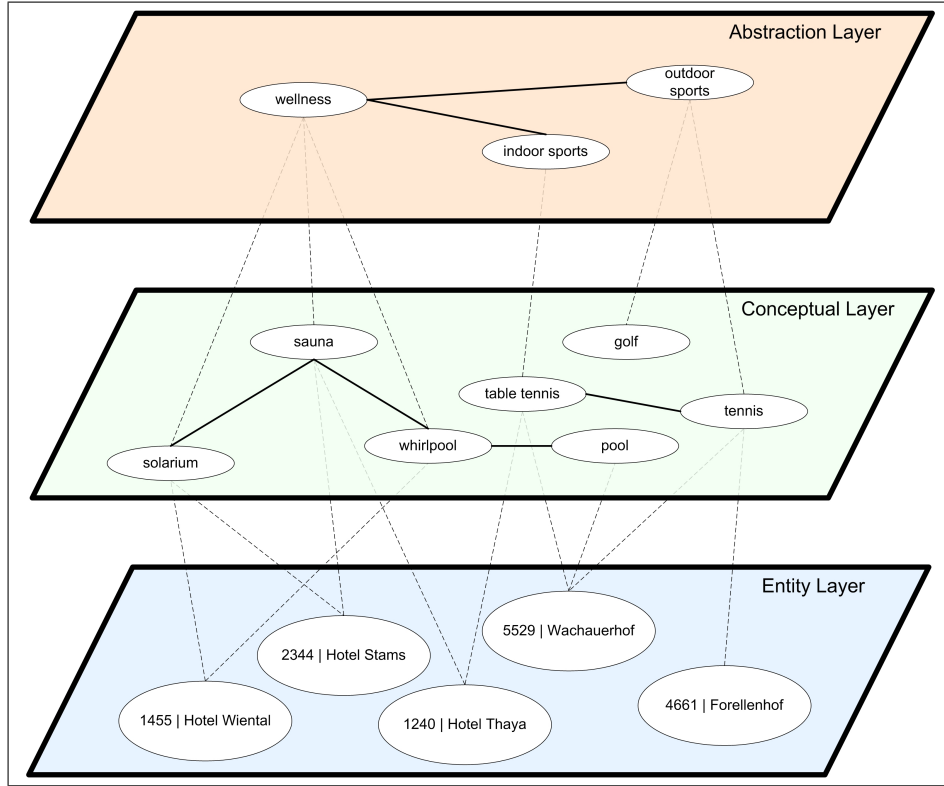


Figure 5.4: Network layer interdependencies

The concept “*wellness*” in Figure 5.5 illustrates an example of a abstract concept.

- 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 *initial* tag’s 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 “*dampfbad*”. 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

```

<concept id="dampfbad" role="concrete" initial="1.0">
  <connectedTo id="sauna" weight="w1"/>
  <lang id="de">
    <syn>dampfbad</syn>
  </lang>
  <lang id="en">
    <syn>steam bath</syn>
  </lang>
</concept>
<concept id="sauna" role="concrete">
  <connectedTo id="solarium" weight="w2"/>
  <lang id="*">
    <syn>sauna</syn>
  </lang>
</concept>
<concept id="wellness" role="abstract">
  <parentOf id="sauna" />
  <parentOf id="solarium" />
  <parentOf id="dampfbad" />
  <parentOf id="beautyfarm" />
  <lang id="*">
    <syn>wellness</syn>
  </lang>
</concept>
<concept id="nicht" role="modifier" initial="-1.01">
  <lang id="de">
    <syn>nicht</syn>
    <syn>ohne</syn>
    <syn>kein</syn>
  </lang>
  <lang id="en">
    <syn>not</syn>
    <syn>without</syn>
  </lang>
</concept>

```

Figure 5.5: XML representation of concepts in the associative network

concepts. These relations will be detailed later in this section.

Determining appropriate weights to apply them to links is a crucial and non trivial task. Weights associated with links are, on the one hand, determined during a manual tuning process and, on the other hand, influenced by subjective preferences of users. Moreover, the evaluation of past user interactions provides a starting point for defining the weights of associations. As will be described in Subsection 5.3.3 concepts extracted from user queries are used to derive a network of relations.

Moreover, a set of concepts representing a particular domain is described in a single XML file and act 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.

5.3.2 Processing the 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 5.6 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 5.6. This hotel offers a “*sauna*”, a “*steam bath*” and a “*solarium*” and is, therefore, linked to the corresponding concepts at the conceptual layer.

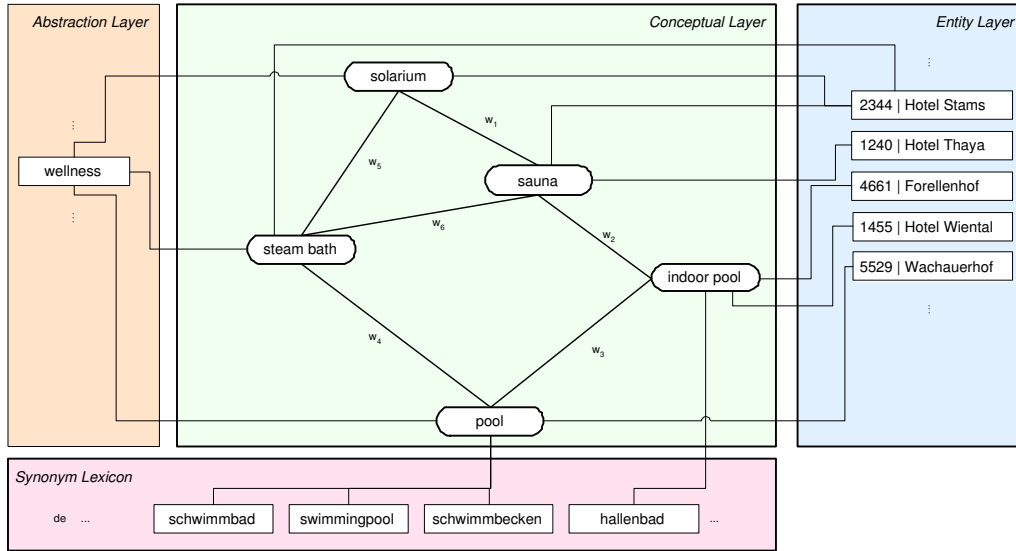


Figure 5.6: Knowledge base architecture

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 abstract or conceptual layer. These concepts act as activation sources and, subsequently, the activation process is initiated and activation spreads according to the algorithm as outlined below.

At the beginning, the role of each concept is evaluated. Depending on its role, different initialization strategies can be 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 the value defined in the *initial* tag of the concept. 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 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 the following formula:

$$O_i(p) = \begin{cases} 0 & \text{if } I_i(p) < \tau, \\ \frac{F_i}{p+1} \cdot I_i(p) & \text{otherwise, with } F_i = (1 - \frac{C_i}{C_T}) \end{cases} \quad (5.2)$$

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 4.5), 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 4.5). 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 5.6). 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 5.7 depicts the results determined for the query

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

In this particular case, the entities “*Schwarzer Adler Kitzbühel*” and “*Hotel Schloss Lehenberg – 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*”, “*dampfbad*” and “*kinder*” 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, on the one hand,

>> Ich und meine Kinder möchten in einem Hotel in kitzbübel Urlaub machen. Es sollte ein Dampfbad haben. <<

ausgewertete Information:

- ☒ Kinder
- ☒ Hotel
- ☒ kitzbübel
- ☒ Dampfbad

Es wurden mehr als 25 Unterkünfte gefunden.

Gewicht	Hotel/Unterkunft	Art	Ausstattung	Ort/Stadt	gew. Ausstattung
1.0	Schwarzer Adler Kitzbübel	(hotel)		Kitzbübel	dampfbad kinder hotel
1.0	Hotel Schloß Lebenberg - Kitzbübel	(hotel)		Kitzbübel	dampfbad kinder hotel
0.9858694	Bichlhof	(hotel)		Kitzbübel	dampfbad kinder hotel
0.94620633	Erika	(hotel)		Kitzbübel	dampfbad kinder hotel
0.9151889	Golf - Hotel Rasmushof	(hotel)		Kitzbübel	dampfbad kinder hotel
0.9030947	Hotel Kaiserhof	(hotel)		Berwang	dampfbad kinder hotel
0.9030754	QuellenHof Leutasch	(hotel)		Leutasch	dampfbad kinder hotel
0.9030695	Sonnenresidenz Alpenpark	(hotel)		Seefeld	dampfbad kinder hotel
0.9030695	De Luxe Hotel St. Peter	(hotel)		Seefeld	dampfbad kinder hotel
0.9030695	De Luxe Hotel St. Peter	(hotel)		Seefeld	dampfbad kinder hotel
0.897896	Sporthotel Brugger	(hotel)		Fulpmes	dampfbad kinder hotel
0.89702064	Hotel Schwarzbrunn	(hotel)		Stans	dampfbad kinder hotel

Figure 5.7: Weighted result set determined by constrained spreading activation

provides a means for exact information retrieval and, on the other hand, incorporates a fuzzy search strategy that determines closely related matches to the user's query.

5.3.3 Adapting Associations According to Past User Interactions

In addition to a predefined network structure based on assumptions made by domain engineers, user interactions should have some influence on the

relatedness of domain concepts. Based on the idea, that user queries contain concepts associated with each other on a semantic and intentional basis, one major objective of the redevelopment was to derive information from past user interactions. More precisely, user queries are automatically inspected to determine relations between concepts. In the following section the basic idea of this adaptation strategy is described.

Concept Matrix

User queries collected during the field trial have been used as a basis for analyzing associations between concepts. First, words that are not contained in the domain ontology are eliminated from the query. Then, each remaining concept of the query is associated with each other. The weight of the association is determined by the number of co-occurrences of concepts. Consider, for example, the following concepts:

{hotel, pension, wellness, sauna, solarium, playground}

By means of these concepts the generation of a concept matrix will be illustrated. First, the following query consisting of a subset of the concepts mentioned above is inspected:

I am looking for a hotel with a wellness area, especially with a sauna and a solarium as well as a playground in Vienna.

Then, all terms that are not concepts as well as location identifiers, i.e. references to particular towns or geographical regions, are removed from the query resulting in the following refined representation.

{hotel, wellness, sauna, solarium, playground}

In this case, concept associations as shown in Table 5.1 are identified. Each row as well as each column contains a concept. If an association between two concepts is determined, the association counter is increased, i.e. the associa-

tion counter represents the number of associations between two concepts.

	hotel	pension	wellness	sauna	solarium	playground
hotel	–	–	1	1	1	1
pension	–	–	–	–	–	–
wellness	–	–	–	1	1	1
sauna	–	–	–	–	1	1
solarium	–	–	–	–	–	1
playground	–	–	–	–	–	–

Table 5.1: Concept matrix for a single query

The example query, for instance, indicates *relatedness* between the concepts “*wellness*” and “*sauna*”. Table 5.2 contains a subset of queries obtained during the field trial. Two rows of the Table describe a query and the identified concepts, respectively.

Query	I am looking for a pension with sauna and solarium in Imst.
Concepts	{pension, sauna, solarium}
Query	Show me all hotels with a wellness area and a playground.
Concepts	{hotel, wellness, playground}
Query	Show me all hotels or pensions in Innsbruck.
Concepts	{hotel, pension}
Query	I am searching for a hotel with a wellness area.
Concepts	{hotel, wellness}

Table 5.2: Subset of user queries obtained during the field trial

If these queries are decomposed and evaluated as described above, a concept matrix as depicted in Table 5.3 is obtained. We shall note, that the concept matrix illustrated in Table 5.1 acts as a basis for the results presented in Table 5.3. The query subset determines a concept matrix that defines a set of associations reflecting users’ demands, as, for instance, facilities commonly used in combination (cf. “*sauna*” and “*solarium*” as well as “*hotel*” and “*wellness*”).

In order to generate a comprehensive concept matrix, 1,213 German queries obtained during the field trial have been decomposed and analyzed.

	hotel	pension	wellness	sauna	solarium	playground
hotel	–	1	3	1	1	2
pension	–	–	–	1	1	–
wellness	–	–	–	1	1	2
sauna	–	–	–	–	2	1
solarium	–	–	–	–	–	1
playground	–	–	–	–	–	–

Table 5.3: Concept matrix for multiple queries

We shall note, that three in fact English queries have been misclassified as being German and are, thus, included in the set of 1,213 German queries. A subset of concepts associated more than ten times is depicted in Figure 5.8. Taking a look at the relations derived, reveals, for instance, that the concept “*sauna*” is highly associated with the concept “*hotel*”. Moreover, a *recreational aspect* of the concept “*sauna*” can be derived. It is related to “*hallenbad*”, “*mountain biking*”, “*solarium*”, “*swimming pool*”, etc. Another example is the concept “*gleitschirmfliegen*”. In this special case, a semantic relationship between action sports can be derived. More precisely, “*gleitschirmfliegen*” is highly associated with “*mountain biking*” and “*rafting*”.

The associations derived, assist in setting up and refining the associative network according to semantic relations defined by users themselves. Moreover, automatic and implicit adaptation according to user interactions during system runtime allows

- the integration of new semantic relations between concepts, and,
- the adjustment of weights of existing associations.

This automatic domain adaptation process is still at an experimental stage. Nevertheless, promising intermediate result have been observed. As a first step, the automatic determination and integration of links between unassociated concepts has been incorporated. Basically, users’ queries are decomposed and associations between concepts are derived by the method

	alpin-ski	am see	bauernhof	betten	dz	ez	ez,dz,mz,suiten	ferienwohnung	gleitschirmfliegen	golf	halbpension	hallenbad	haustiere	hotel	internetzugang	kategorie	kinder	mountainbiken	orts-/stadtzentrum	parkplätze	pension	rafting	sauna	see	solarium	swimmingpool	tagungsraeume	whirlpool
alpin-ski	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	10	-	-	-	-	-	-	-	-	-	-	-
am see	-	-	-	-	-	-	-	-	-	-	-	-	-	14	-	-	-	-	-	-	-	-	-	-	-	-	-	-
bauernhof	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	18	-	-	-	-	-	-	-	-	-	-	-
betten	-	-	-	-	11	-	22	42	-	-	-	-	-	13	-	11	-	-	-	-	-	-	13	-	-	-	-	-
dz	-	-	-	-	-	-	-	-	-	-	11	65	-	47	-	21	20	-	66	-	13	-	54	-	10	-	-	-
ez	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	14	-	-	-	-	13	-	-	-	-	-	-	-
ez,dz,mz,suiten	-	-	-	-	-	-	-	-	-	-	-	-	-	13	-	-	11	-	-	-	-	-	17	-	-	-	-	-
ferienwohnung	-	-	-	-	-	-	-	-	-	-	-	-	15	-	-	-	16	-	-	-	-	-	11	-	-	-	-	-
gleitschirmfliegen	-	-	-	-	-	-	-	-	-	-	-	19	-	20	-	18	-	18	-	-	-	10	18	-	-	-	-	-
golf	-	-	-	-	-	-	-	-	-	-	-	-	-	20	-	-	-	-	-	-	-	-	-	-	-	-	-	-
halbpension	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
hallenbad	-	-	-	-	-	-	-	-	-	-	-	-	-	63	-	31	11	22	41	-	-	12	42	-	-	-	-	13
haustiere	-	-	-	-	-	-	-	-	-	-	-	-	-	17	-	-	22	-	-	-	13	-	-	-	-	-	-	-
hotel	-	-	-	-	-	-	-	-	-	-	-	-	-	11	19	120	40	22	-	29	-	12	116	-	16	22	20	17
internetzugang	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	14	-	-	-	-	-	-	18	-
kategorie	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	22	-	-	-	-	-	-	-	-	-	-
kinder	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	12	37	-	-	-	-	-
mountainbiken	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	12	20	-	-	-	-	-
orts-/stadtzentrum	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
parkplätze	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	13	-	-	-	14	-
pension	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
rafting	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	12	-	-	-	-	-
sauna	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	20	14	-	15
see	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
solarium	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
swimmingpool	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
tagungsraeume	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
whirlpool	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Figure 5.8: Concept matrix derived from user queries

described above. More precisely, if two unassociated concepts are contained in a query they are automatically associated and the link is weighted with a default value.

Furthermore, the automatic adaptation of weights associated with links turned out to be a quite difficult task. However, in analogy to the method described above, the adaptation process has been implemented. Each query is analyzed, links between concepts are determined and the associated weight is strengthened or weakened accordingly, i.e. the weight is simply increased or decreased by one. Hence, the generated network reflects the co-occurrence of concepts derived from user queries. Thus, the adaptive network might be integrated besides the predefined network and influences the final ranking position of the results.

5.3.4 Retrieval Results

The redevelopment of the information retrieval system enhanced the search results in several ways. First of all, the integration of regional dependencies and the definition of relations between cities and towns provides a means for sophisticated retrieval of location dependent information. In particular, federal states of Austria are associated according to their relative geographical positions, i.e. the association between Styria and Burgenland is stronger than the association between Styria and Vorarlberg. Hence, the results retrieved and depicted in Figure 5.9 for the following query reflect this regional dependencies.

I am looking for a wellness hotel in Burgenland with a playground for my kids.

Gewicht	Hotel/Unterkunft	Art	Ausstattung	Ort/Stadt	gew. Ausstattung
1.0	Steigenberger Avance Hotel	(hotel)		Bad Tatzmannsdorf	spielplatz kinder
0.96468437	Hotel Rogner Birdie Therme	(hotel)		Stegersbach	spielplatz kinder
0.947035	Vila Vita Hotel und Feriendorf Pannonia	(hotel)		Pamhagen	spielplatz
0.94700944	Hotel "Krutzler"	(hotel)		Heiligenbrunn	spielplatz kinder
0.9294456	Steigenberger Golf & Thermalhotel	(hotel)		Bad Tatzmannsdorf	spielplatz kinder
0.91171086	Villa Kunterbunt	(hotel)		Jennersdorf	spielplatz kinder
0.82349026	"Das Schmidt"	(hotel)		Mörbisch am See	spielplatz kinder
0.81174964	Austria Trend Seehotel Rust	(hotel)		Rust	spielplatz kinder
0.7992684	Seehotel Jägerwirt	(hotel)		Turracher Höhe	spielplatz kinder
0.7992684	Rogner-Bad Blumau	(hotel)		Blumau	spielplatz kinder

Figure 5.9: Result set exemplifying regional dependencies

Top ranked is the accommodation “*Steigenberger Avance Hotel*” in the city of “*Bad Tatzmannsdorf*” which is located in the federal state Burgenland. The 9th match (“*Seehotel Jägerwirt*”) is located in the federal state of Styria.

Moreover, associations between cities and towns enhance the granularity of search results. More precisely, cities located close to each other (as mentioned in the previous section, within a circumference of 15km) are connected by a first order relation. Furthermore, the weights associated with the links reflect the distance between cities within this circumference. The results depicted in Figure 5.10 are based on the following query and illustrates this effect.

Show me all hotels in Igls with a beauty farm.

Gewicht	Hotel/Unterkunft	Art	Ausstattung	Ort/Stadt	gew. Ausstattung
1.0	Best Western Sporthotel Igls	(hotel)		Igls	beautyfarm hotel
0.9295131	F.X. Mayr-Zentrum Parkhotel Igls	(hotel)		Igls	beautyfarm hotel
0.882632	Lanserhof Gesundheitszentrum	(hotel)		Lans	beautyfarm hotel
0.858958	Sporthotel Igls	(hotel)		Igls	hotel
0.78847116	Hotel Astoria	(hotel)		Igls	hotel
0.78847116	Hotel Gruberhof	(hotel)		Igls	hotel
0.78847116	Hotel-Pension Sonnenhof	(hotel)		Igls	hotel
0.78847116	Hotel Batzenhäusl	(hotel)		Igls	hotel
0.78847116	Hotel Römerhof	(hotel)		Igls	hotel

Figure 5.10: Result set exemplifying dependencies between cities

Obviously, the first three matches offer a “*beauty farm*”, whereas the third match is situated in the city of “*Lans*“ which is located in the vicinity of “*Igls*”.

As motivated by the findings of the field trial, the integration of subjective and abstract terms augment the capabilities of the system. Consider, for example, the following query:

I am searching for a wellness hotel in Fulpmes with a fitness room. Pets should be allowed.
































Gewicht	Hotel/Unterkunft	Art	Ausstattung	Ort/Stadt	gew. Ausstattung
1.0	Best Western Sporthotel Igls	(hotel)	     	Igls	haustiere fitnessraum hotel
0.9617425	Aktivhotel DONNERHOF	(hotel)	    	Fulpmes	haustiere hotel
0.9617425	Alte Post	(hotel)	    	Fulpmes	haustiere fitnessraum hotel
0.95368326	Sporthotel Igls	(hotel)	    	Igls	haustiere fitnessraum hotel
0.92536205	Hotel Klosterbräu	(hotel)	    	Seefeld	haustiere fitnessraum hotel
0.9022148	Sonnenresidenz Alpenpark	(hotel)	    	Seefeld	haustiere fitnessraum hotel

Figure 5.11: Result set exemplifying abstract concepts

The abstract term “*wellness*” is implicitly expanded and refined by the system according to the weighted links in the associative network. Subsequently, the results shown in Figure 5.11 are determined by spreading activation through the network and, finally, presented to the user. A remarkable effect is, that in this special case, an accommodation in the city of “*Igls*” is top ranked and not the city of “*Fulpmes*”, as requested in the query. Due to the fact, that the city of “*Igls*” is located within a circumference of some kilometers and the amenities of the accommodation exceed those available in comparable accommodations in “*Fulpmes*”, the system judged the “*Best Western Sporthotel Igls*” to be the best matching.

Nevertheless, the adapted information retrieval system provides the functionality known from the original system. As expected, results determined by the original system are part of the result set of the associative network approach. Figure 5.12 illustrates a comparison of matches retrieved by both, the original system (depicted as the overlaid result set) and the new approach. The result set of the associative network approach is enriched with related matches.

Moreover, an interesting aspect is the possibility to put special emphasize on favored concepts. More precisely, by posing queries containing multiple occurrences of the same preferred attribute, its significance increases according to the number of occurrences. Certainly, this circumstance represents an

Gewicht	Hotel/Unterkunft	Art	Ausstattung	Ort/Stadt	gew. Ausstattung
1.0	Hotel Linserhof	(hotel)		Imst	spielplatz hotel
0.9771895	Alpenblick	(hotel)		Imst	spielplatz hotel
0.93157935	Romantik Hotel Post	(hotel)		Imst	spielplatz hotel
0.90830564	Kinderhotel Lärchenwald	(hotel)		Arzl im Pitztal	spielplatz hotel
0.8854951	Ferienhotel Bergland	(hotel)		Arzl im Pitztal	spielplatz hotel
0.86289406	Panoramahotel Gurgltaler Hof	(hotel)		Tarrenz	spielplatz hotel
0.8402929	Hotel Winkler	(hotel)		Imst	spielplatz hotel
0.79438424	Hotel Gasthof zum Hirschen	(hotel)		Imst	hotel

	Hotel/Unterkunft	Art	Ausstattung	Ort/Stadt
1.	Hotel Winkler	*** (Hotel)		Imst
2.	Hotel Linserhof	**** (Hotel)		Imst
3.	Romantik Hotel Post	**** (Hotel)		Imst
4.	Alpenblick	(Hotel)		Imst

Figure 5.12: Comparing results determined by the original system and the associative network approach

inappropriate means for expressing user preferences. But a further enhancement might be the integration of the possibility for users to favor several concepts by explicitly increasing a significance value assigned to the concept.

The implementation of Boolean operators in the original system caused serious problems and had to be accomplished with great effort. Due to the nature of the associative network and the refinement of the search strategy to fit the needs of a fuzzy approach, this problem was implicitly resolved. More precisely, results are ranked according their relevance to the query based on semantic associations between concepts contained in the query. In the case that a user's query consists of several concepts combined by Boolean "and"s, results providing all stipulated concepts are ranked first. Furthermore, if concepts are combined by Boolean "or"s, results are determined in the same way. This is due to the fact, that the "or"-operator does not demand exclusiveness.

During the evaluation of the approach, inconsistency in the data stored in the database was revealed by accident. The data stored in the database

represents the status of the *Tiscover* database as available in October, 2001. In order to illustrate the effect of the “not”-operator, the following query produced a strange result set.

I am searching for an accommodation not in Austria.














Gewicht	Hotel/Unterkunft	Art	Ausstattung	Ort/Stadt	gew. Ausstattung
0.0	Hotel Post	(hotel)	   		
0.0	Mitter	(bauernhof)			
0.0	Reif	(bauernhof)			
0.0	Kully-Hube	(jugendherberge)			
0.0	Leitmesner	(bauernhof)			
0.0	Richter	(bauernhof)			
0.0	Mukonig	(bauernhof)			
0.0	Schirlbauer	(bauernhof)	 		
0.0	Clavara	(bauernhof)			
0.0	1. Erdäpfelpension Österreichs Landgasthof Ranklleiten	(pension)	    		

Figure 5.13: Excerpt of accommodations not associated with a city

Assuming that all accommodations are associated with at least one city of Austria, a result set containing accommodations ranked according to their amenities was expected. Surprisingly, the system retrieved a result set as depicted in Figure 5.13. Several accommodations throughout Austria have not been correctly associated with a city and, therefore, failed to be retrieved with the original system.

5.4 Discussion

Due to the necessity, urged by the field trial, to exchange the knowledge base of the original natural language system, an alternative approach based on associative networks was integrated. In particular the approach addressed

the following issues: first, to implement a knowledge representation model that takes semantic relatedness of terms into account, second, to implement spreading activation as a means that evaluates the relatedness of these terms and, therefore, provides implicit query expansion, and, finally, to offer a flexible method of defining relationships between terms to unleash the ability to retrieve 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.

Furthermore, the spreading activation algorithm used to process the network structure has been presented. Moreover, the constraints used to control spreading of activation through the network were discussed.

However, the determination of weights associated with links between information items remains a non-trivial task. An automatic adaptation process of weights based on past users queries was proposed. Moreover, the queries collected during the field trial provided a good starting point for, first, determining important information items, and, second, to derive the strength of associations between these information items.

Finally, several result sets have been presented to illustrate the functionality of the associative network approach.

Chapter 6

Conclusion and Future Work

A natural language system based on an approach described in Berger (2001); Berger et al. (2001) has been reviewed in this thesis and, furthermore, provided the basis for the research presented herein. Basically, the reviewed system offers multilingual access to information on a restricted 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. In this particular case the system operates on the tourism domain. In order to get a picture of the software architecture underlying the original information retrieval system, a detailed report was given in Chapter 2.

The findings obtained during a field trial and the results derived from a usability study motivated the redevelopment of the knowledge base underlying the original system. Strictly speaking, the knowledge representation model used in the original system was inadequate to model semantic relationships between domain-intrinsic information. Therefore, a knowledge representation model based on associative networks was chosen to replace the flat, unassociated knowledge base underlying the original system.

An approach for modelling semantic relationships by means of a network structure as well as to set up such networks appropriately were the major concerns in this thesis. Therefore, Chapter 4 provided a detailed review of knowledge representation based on network models. Moreover, a processing

technique for such networks, namely spreading activation, and constraints to get control over the spreading process have been discussed.

In this thesis main focus was laid on the development of 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.

Nevertheless, determining appropriate associations between concepts of the associative network has great impact on the quality of results. In order to obtain a starting point for associating domain concepts, an approach based on analyzing past user interactions was proposed in this thesis. Therefore, the concept co-occurrence in a query is used as a measure of the degree of relatedness. Furthermore, an automatic adaptation process of weights associated with links was introduced. More precisely, based on the co-occurrence of concepts, weights between nodes of the associative network are strengthened or weakened.

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 concepts. Users should be able to assign a degree of significance to concepts. 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 emphasize concepts, 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 concept. 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 derived 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.

Bibliography

- J. R. Anderson. *The Architecture of Cognition*. Harvard University Press, Cambridge, MA, 1983a.
- J. R. Anderson. A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, 22:261–295, 1983b.
- R. A. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, Reading, MA, 1999.
- L. Ballesteros and W. B. Croft. Resolving ambiguity for cross-language retrieval. In *Research and Development in Information Retrieval*, pages 64–71, 1998.
- R. K. Belew. Adaptive information retrieval: Using a connectionist representation to retrieve and learn about documents. In N. J. Nicholas J. Belkin and C. J. Van Rijsbergen, editors, *Proceedings of the 12th International Conference on Research and Development in Information Retrieval (SIGIR’89)*, pages 11–20. ACM, 1989. ISBN 0-89791-321-3.
- H. Berger. Adaptive multilingual interfaces. Master’s thesis, Vienna University of Technology, 2001.
- H. Berger, M. Dittenbach, and D. Merkl. Activation on the move. In *Proceedings of the 14th International Conference and Workshop on Database and Expert Systems Applications (DEXA 2003)*, 2003a. Accepted for publication.

- H. Berger, M. Dittenbach, and D. Merkl. Querying tourism information systems in natural language. In *Proceedings of the 2nd International Conference on Information System Technology and its Applications*, 2003b. Accepted for publication.
- H. Berger, M. Dittenbach, D. Merkl, and W. Winiwarter. Providing multilingual natural language access to tourism information. In W. Winiwarter, St. Bressan, and I. K. Ibrahim, editors, *Proceedings of the 3rd International Conference on Information Integration and Web-based Applications and Services (IIWAS 2001)*, pages 269–276, Linz, Austria, September 10–12 2001. Austrian Computer Society.
- I. Campbell and C. J. Van Rijsbergen. The ostensive model of developing information needs. In *Proceedings of the 2nd International Conference on Conceptions of Library Science (COLIS-96)*, pages 251–268, Kobenhavn, DK, 1996.
- W. B. Cavnar and J. M. Trenkle. N-gram-based text categorization. In *International Symposium on Document Analysis and Information Retrieval*, Las Vegas, NV, 1994.
- F. Crestani. Application of spreading activation techniques in information retrieval. *Artificial Intelligence Review*, 11(6):453–582, 1997.
- F. Crestani and P. L. Lee. Searching the web by constrained spreading activation. *Information Processing and Management*, 36(4):585–605, 2000.
- F. Crestani and C. J. Van Rijsbergen. A model for adaptive information retrieval. *Journal of Intelligent Information Systems (JIIS)*, 8(1):29–56, 1997.
- W. Croft, T. Lucia, J. Crigean, and P. Willet. Retrieving documents by plausible inference: an experimental study. *Information Processing & Management*, 25(6):599–614, 1989.

- W. Croft and R. H. Thompson. I³R: A New Approach to the Design of Document Retrieval Systems. *Journal of the American Society for Information Science*, 38(6):389–404, 1987.
- F. J. Damerau. Operating statistics for the transformational question answering system. *American Journal of Computational Linguistics*, 7:30–42, 1981.
- M. Dittenbach, D. Merkl, and H. Berger. Free speech for tourists. In A. Wenn, M. McGrath, and F. Burstein, editors, *Proceedings of the 13th Australasian Conference on Information Systems (ACIS 2002)*, volume 3, pages 1145–1154, Melbourne, Australia, December 4–6 2002a.
- M. Dittenbach, D. Merkl, and H. Berger. What customers really want to know from tourism information systems but never dared to ask. In *Proceedings of the 5th International Conference on E-Commerce Research (ICECR-5)*, Montréal, Canada, October 23–27 2002b.
- M. Dittenbach, D. Merkl, and H. Berger. A natural language query interface for tourism information. In A. J. Frew, M. Hitz, and P. O’Connor, editors, *Proceedings of the 10th International Conference on Information Technologies in Tourism (ENTER 2003)*, pages 152–162, Helsinki, Finland, January 29–31 2003a. Springer-Verlag.
- M. Dittenbach, D. Merkl, and H. Berger. 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)*, 2003b. Accepted for publication.
- D. R. Fesenmaier, F. Ricci, E. Schaumlechner, K. Wöber, and C. Zanella. DI-ETORECS: Travel Advisory for Multiple Decision Styles. In A. J. Frew, M. Hitz, and P. O’Connor, editors, *Proceedings of the 10th International Conference on Information Technologies in Tourism*, pages 232–241, Helsinki, Finland, January 29–31 2003. Springer-Verlag.

- D. K. Harman. Overview of the 3rd Text Retrieval Conference (TREC-3). In D. K. Harman, editor, *Proceedings of the 3rd Text Retrieval Conference (TREC-3)*, pages 1–19. NIST Special Publication 500–225, 1995.
- K. Hartmann and Th. Strothotte. A spreading activation approach to text illustration. In *Proceedings of the 2nd International Symposium on Smart graphics*, pages 39–46. ACM Press, 2002. ISBN 1-58113-555-6.
- D. A. Hull and G. Grafenstette. 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)*, pages 49–57, 1996.
- B. J. Jansen, A. Spink, J. Bateman, and T. Saracevic. Real life information retrieval: A study of user queries on the web. *SIGIR Forum*, 32(1):5–17, 1998.
- R. Kjeldsen and P. Cohen. The evolution and performance of the GRANT system. *IEEE Expert*, pages 73–79, 1987.
- J. Krause. Natural Language Access to Information Systems. An Evaluation Study of its Acceptance by End Users. *Information Systems*, 5:297–319, 1980.
- B. Kröse and P. Van der Smagt. *An Introduction to Neural Networks*. 8th edition, 1996.
- F. W. Lancaster. *Information Retrieval Systems: Characteristics, Testing and Evaluation*. John Wiley, New York, 1968.
- V. I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.
- J. Nielsen. *Usability Engineering*. Kaufmann, M., New York, 1994.
- J. Nielsen. *Designing Web Usability: The Practice of Simplicity*. New Riders Publishing, 2000.

- D. W. Oard. Neural networks in information filtering and retrieval. Technical report, University of Maryland, 1994.
- P. O'Brian. Dynamic travel itinerary management: The ubiquitous travel agent. In *Proceedings of the 12th International Australasian Conference on Information Systems*, Coffs Harbour, Australia, 2001.
- W. C. Ogden and P. Bernick. *Handbook of Human-Computer Interaction*, chapter Using Natural Language Interfaces, pages 137–161. Elsevier Science, 1997.
- L. Philips. Hanging on the metaphone. *Computer Language Magazine*, 7 (12), 1990.
- S. Preece. *A spreading activation model for Information Retrieval*. PhD thesis, University of Illinois, Urbana-Champaign, USA, 1981.
- M. Pribernik. Usability-Studie für ein Suchmaschineninterface. Master's thesis, Vienna University of Economics and Business Administration, 2003.
- B. Pröll, W. Retschitzegger, and R. Wagner. TIScover – Eine generische Plattform für webbasierte Tourismusinformation. *Informatik Forschung und Entwicklung*, 16:1–13, 2001.
- B. Pröll, W. Retschitzegger, R. Wagner, and A. Ebner. Beyond traditional tourism information systems: TIScover. *Information Technology and Tourism*, 1, 1998.
- M. R. Quillian. Semantic memory. In M. Minsky, editor, *Semantic Information Processing*, pages 227–270. MIT Press, 1968.
- J. Raskin. *The Humane Interface: New Directions for Designing Interactive Systems*. Addison-Wesley, 2000.
- G. Repovš. Mechanisms and structure of semantic memory. In *Proceedings of the 4th International Conference of Information Society*, pages 50–53, 2002.

- D. E. Rumelhart and D. A. Norman. *Steven's handbook of experimental psychology*, chapter Representation in memory, pages 511–587. John Wileys & Sons Ltd., New York, 2. edition, 1988.
- G. Salton. *Automatic information organization and retrieval*. Mc Graw Hill, New York, 1968.
- G. Salton. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, Reading, MA, 1989.
- G. Salton and C. Buckley. 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*, pages 147–160, Grenoble, France, June 1988.
- G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 1983.
- C. Silverstein, M. Henzinger, H. Marais, and M. Moricz. Analysis of a very large AltaVista query log. Technical Report 1998-014, digital Systems Research Center, 1998.
- S. Staab, C. Braun, I. Bruder, A. Dsterhft, A. Heuer, M. Klettke, G. Neumann, B. Prager, J. Pretzel, H. Schnurr, R. Studer, H. Uszkoreit, and B. Wrenger. GETESS - searching the web exploiting german texts. In *Cooperative Information Agents*, pages 113–124, 1999.
- C. J. Van Rijsbergen. *Information Retrieval*. Department of Computer Science, University of Glasgow, 1979.
- S. P. Whittaker and S. P. Stenton. User studies and the design of natural language systems. In *Proceedings of EACL '89*, pages 116–123, Kobenhavn, DK, 1989.

- F. Xu, K. Netter, and H. Stenzhorn. 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*, 2000.
- A. Zell. *Simulation Neuronaler Netze*. Addison-Wesley(Deutschland), 1994.

Curriculum Vitae

Helmut Berger

Mariengasse 39/14

1170 Wien – Austria

helmut.berger@ec3.at

Date of birth: July 29th, 1975

Place of birth: Klagenfurt, Austria

Marital status: single

Education

Master of Science, Computer Science, Vienna University of Technology, Wien–Austria, October 2001. Thesis: “Adaptive Multilingual Interfaces”. Advisor: Dr. Dieter Merkl.

Study of Computer Science, Vienna University of Technology, Wien–Austria, 1993–2001.

General qualification for university entrance, Klagenfurt–Austria, 1993.

Comprehensive secondary school, Klagenfurt–Austria, 1985–1993.

Elementary school, Klagenfurt–Austria, 1981–1985.

Research Experience

Masters Research, Vienna University of Technology, November 2000–October 2001. Developed a prototypical multilingual natural language information retrieval system for the tourism domain. Dr. Dieter Merkl, Institut für Softwaretechnik und interaktive Systeme.

Researcher, Electronic Commerce Competence Center – EC3, October 2001–present. Multilingual natural language interfaces, speech technology, Semantic Networks, Ontology development.

Publications

H. Berger, M. Dittenbach, and D. Merkl. Activation on the Move. In *Proceedings of the 14th International Conference and Workshop on Database and Expert Systems Applications (DEXA 2003)*, 2003. Accepted for publication.

M. Dittenbach, D. Merkl, and H. Berger. 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)*, 2003. Accepted for publication.

H. Berger, M. Dittenbach, and D. Merkl. Querying Tourism Information Systems in Natural Language. In *Proceedings of the 2nd International Conference on Information System Technology and its Applications*, 2003. Accepted for publication.

M. Dittenbach, D. Merkl, and H. Berger. A Natural Language Query Interface for Tourism Information. In A. J. Frew, M. Hitz, and P. O'Connor, editors, *Proceedings of the 10th International Conference on Information Technologies in Tourism (ENTER 2003)*, pages 152-162, Helsinki, Finland, January 29-31 2003. Springer-Verlag.

M. Dittenbach, D. Merkl, and H. Berger. Free Speech for Tourists. In A. Wenn, M. McGrath, and F. Burstein, editors, *Proceedings of the 13th Australasian Conference on Information Systems (ACIS 2002)*, volume 3, pages 1145-1154, Melbourne, Australia, December 4-6 2002.

M. Dittenbach, D. Merkl, and H. Berger. What Customers Really Want to Know from Tourism Information Systems but Never Dared to Ask. In *Proceedings of the 5th International Conference on E-Commerce Research (ICECR-5)*, Montreal, Canada, October 23-27 2002.

H. Berger, M. Dittenbach, D. Merkl, and W. Winiwarter. Providing Multilingual Natural Language Access to Tourism Data. In W. Winiwarter, St. Bressan, and I. K. Ibrahim, editors, *Proceedings of the 3rd International Conference on Information Integration and Web-based Applications and Services (IIWAS 2001)*, pages 269-276, Linz, Austria, September 10-12 2001. Austrian Computer Society.

Presentations

M. Dittenbach, D. Merkl, and H. Berger. A Natural Language Query Interface for Tourism Information. In A. J. Frew, M. Hitz, and P. O'Connor, editors, *Proceedings of the 10th International Conference on Information Technologies in Tourism (ENTER 2003)*, pages 152-162, Helsinki, Finland, January 29-31 2003. Springer-Verlag.

Type: contributed paper

Conference name: 10th International Conference on Information Technologies in Tourism (ENTER 2003)

Berger, H., Dittenbach, M., Merkl, D., and Winiwarter, W. Providing Multilingual Natural Language Access to Tourism Information. In Winiwarter, W., Bressan, S., and Ibrahim, I. K., editors, *Proceedings of the 3rd International Conference on Information Integration and Web-based Applications*

and Services (IIWAS 2001), Linz, Austria, September 10-12 2001. OCG.

Type: contributed paper

Conference name: 3rd International Conference on Information Integration and Web-based Applications and Services

Organizer: Software Competence Center Hagenberg, Johannes Kepler University of Linz and National University of Singapore