Translating Dialects in Search: Mapping between Specialized Languages of Discourse and Documentary Languages

By

Vivien Petras

Magister (Humboldt University Berlin, Germany) 2001

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Information Management and Systems

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor Michael K. Buckland, Chair
Professor Ray R. Larson
Professor Gene I. Rochlin

Spring 2006

The dissertation of Vivien Petras is approved:

Chair_____ Date_____
Michael K. Buckland

_____ Date_____
Ray R. Larson

_____ Date_____
Gene I. Rochlin

University of California, Berkeley
May 2006

Translating Dialects in Search: Mapping between Specialized Languages of Discourse and Documentary Languages

Copyright © 2006
by
Vivien Petras

Abstract

Translating Dialects in Search: Mapping between Specialized Languages of Discourse and Documentary Languages

by

Vivien Petras

Doctor of Philosophy in Information Management and Systems

University of California, Berkeley

Professor Michael K. Buckland, Chair

The biggest problem in searching an information system is to find the appropriate search terms that not only represent the searcher's information need but also match the language used in the information system. This is a translation problem between a specialized dialect of discourse and the documentary language of the information system.

Discourse dialects evolve within specialized communities. They differ from general language and other communities' dialects in terminology (e.g. terms of art, jargon) and grammar patterns. A documentary language is the language used for document representation in an information system. A bibliographic database and its documentary language usually cover more than one domain of discourse.

This dissertation describes a mechanism that will provide a translation aid between specialized languages and the documentary language by suggesting appropriate search terms for a searcher's query in relation to the searcher's domain of discourse. With this kind of vocabulary support in the search process, the different

specialized vocabularies can be disambiguated within the information system. Different perspectives on a topic can be represented to the searcher (based on the different discourses of the topic in the collection), which will help in navigating and exploring this information space more effectively.

The search term recommender system, based on statistical associations between specialized language terms and controlled vocabulary terms, is introduced and its applications for automatic text categorization, query expansion and reformulation, and terminology mapping are described. The search term recommender methodology is tested on three specialties in the Inspec bibliographical database and 33 specialties in the Ohsumed database in a text classification application. It is demonstrated that search term recommender systems are more effective when specialty-based.

Acknowledgements

During the course of my graduate studies and dissertation research, I received help and support from many people for which I am deeply grateful.

First and foremost, I would like to thank my advisor, Michael Buckland, for his unwavering guidance, support and inspiration during my years at Berkeley. As a mentor and teacher, you were always there to motivate and help me and set an example. You were a true Doktorvater in every sense of the word and I count myself lucky to be one of your students. Of course you are right, Professor Buckland!

Not enough thanks can be given to the members of my thesis committee, Ray Larson and Gene Rochlin, for their generous support, invaluable feedback, and responsiveness in writing this dissertation.

The Metadata Research Group at Berkeley has provided a research home for me at every stage of my graduate studies. Besides Michael Buckland and Ray Larson, I would like to thank Kim Carl, Hailing Jiang, Youngin Kim, Lin Muehlinghaus, Natalia Perelman and Jeanette Zerneke for their support and collaboration. In particular, I would like to thank Fred Gey for introducing me to the field of cross-language retrieval and for supporting my research and career in every possible way. I cannot express how much Aitao Chen taught and helped me in information retrieval, programming and computing matters – big and small. Aitao has my deepest gratitude for creating and allowing the use of the Mulir indexing software suite, on which most of the dissertation experiments were run.

I would also like to thank my first advisor, Prof. Walther Umstätter at Humboldt University, Berlin. You always challenged me to think critically as a

student and to question everything. Our discussions sparked my interest in academic research and encouraged me to pursue graduate studies. I cannot thank you enough for your continuing support and interest.

A special thanks to Nancy Van House who introduced me to the field of science and technology studies. You taught me real critical reading and to examine my terminology. Without your influence, my thinking and perspectives as an academic and as a person would be much more limited.

I would like to thank the SIMS faculty, staff and students for creating a stimulating environment to study, research and grow. I deeply appreciate the friendship, camaraderie and collaboration extended to my by the SIMS PhD students. In particular, I would like to thank danah boyd, Rachna Dhamija, Nicolas Duchenaut, Michal Feldman, Megan Finn, Andrew Fiore, Nathan Good, Jens Grossklags, Mahad Ibrahim, Allan Konrad, Paul Laskowski, Dilan Mahendran, Ryan Shaw, Barbara Rosario and Fredrik Wallenberg for their support. A special thanks to Ana Ramirez Chang, Joe Hall, Dan Perkel and Yuri Takhteyev, who, more than anybody else, shared and helped me through the good and the bad days in dissertation writing. Last but not least, a very special thanks to Charis Kaskiris, the other member in our small dissertation support group. Your support, encouragement and compassion was invaluable and is deeply appreciated.

I would like to thank the 1740 Archers - my home away from home. Life in Berkeley would have been so different (and so much less rewarding) without you. Judy and Jeff Kennedy are my second parents. Nobody could have done a better job in making me feel like part of the family - I cannot express how much this means. Go Bears!

There are not enough words to thank Theda Heinks-Maldonado, Kristen Hiestand and Christine Bunzel for their infinite supply of caring, commiseration, and support in all matters big and small. Katrin Schiemenz, my friend and sister-in-arms through high school, college and graduate school, deserves the highest praise for bearing with me for so long. Without you and a healthy dose of triviality, this dissertation might not have been completed.

My sister Bianca told me when I left for Berkeley that she would take over taking care of things now. My reliance on you cannot be overstated. There are not enough ways to thank you and Tobias for your love, care and friendship. Just for the record: you are the best sister in the world.

Above else, I would like to thank my parents, Margitta and Dieter Petras for their unconditional love and support. Mum und Dad, auf Euch kann ich mich immer verlassen. Ihr habt mich gelehrt, niemals aufzugeben und immer nach dem Besten zu streben. Ihr seid mein Vorbild in allen Dingen und ohne Eure Liebe und Unterstützung kann ich mir mein Leben nicht vorstellen.

Without my family, my life and this work would not be possible.

I dedicate this to you.

Table of Contents

List of Tables

List of Figures

Appendix B. Ohsumed Experiments

Chapter 1

Introduction

1.1 Information Retrieval and the Problem of Language

In today's information landscape, finding the right and only the right information for a user's needs has become the proverbial search for the needle in a haystack. For example, for a relatively small information collection (40,000 documents), it was found that only a quarter to half of a percent of the documents could be considered relevant to a given question (Blair, 1996). With collection sizes growing to millions of documents, especially on the internet, the proportion of relevant to non-relevant data in any information system will further decrease. It is the task of the field of information retrieval to find that smaller and smaller "needle" in the haystack.

In information retrieval, not all searching is the same. Finding the “needle” can mean:

- “a known needle in a known haystack;
- a known needle in an unknown haystack;
- an unknown needle in an unknown haystack;
- any needle in a haystack;
- the sharpest needle in a haystack;
- most of the sharpest needles in a haystack;
- all the needles in a haystack;
- affirmation of no needles in the haystack;
- thinks like needles in any haystack;
- let me know whenever a new needle shows up;
- where are the haystacks?; and
- needles, haystacks - whatever.” (Koll, 2000, 16)

Information retrieval systems employ more and more sophisticated algorithms and strategies to target these different needs. However, even though searching in digital information collection has been researched since the 1960s, information retrieval is continuing to face challenges that cannot be overcome with computing power alone. The main difficulty in the search process lies in the intersection between the human searcher and the information system. Search is a communication process between the searcher stating a question and the information system attempting to answer it. In this communication process, one party (the system) possesses a very limited range of capabilities to negotiate understanding and requires precise procedures to operate successfully, whereas the other party (the human) is capable of operating under vastly imprecise conditions and is not used to providing many clues for understanding.

The means of communication in the search process – language - is inherently ambiguous. Language is used to describe the documents in a collection, to describe a

searcher's question, to formulate a search command, to compare the query to documents in the information system, and to present the results set to the user. Each step is an imprecise operation introducing uncertainty: is the language used to describe a topic in a document the same as the language used to describe the topic in a query? Does the language of the document representations really describe what the document is about? Does the language used to describe a topic in a question really describe the concept in the searcher's mind?

For a searcher using an information system, it is difficult to describe something that one is looking for and does not know. The gap between "I'll know it when I see it" and the unambiguous language preferred by the information system constitutes what is called the language problem in information retrieval.

The language problem is a major obstacle to overcome in the search process. This dissertation describes an approach to alleviate language ambiguities in the search process by helping the searcher and the information system negotiate an understanding between the different languages used by the two parties.

Language evolves in communities of discourse. Language within a particular discourse community differs from general language both in terminology (new or technical terms, specialized word usage and different word meanings) and grammar patterns. When searching for topics in a discourse domain, searchers will use the specialized language of the discourse to express their search interest. Information systems also develop specialized languages to represent the documents in their collection. A documentary language is the language used by the information system to describe the

topics covered in the collection's documents. In many bibliographic databases, a documentary language is provided as a controlled vocabulary (e.g. a thesaurus or classification). From this perspective, search is a mapping between the specialized language of the searcher and the documentary language of the information system.

One of the main assumptions of this dissertation is that information systems can usefully provide support for this mapping between individual's specialized languages and the documentary language. Instead of just using the documentary language for document representation and leaving the search term selection to the searcher in the hope that they will use the documentary language, an information system should use the documentary language to initiate an interaction between the user and the system. From the information system's side, the documentary language needs to be presented in such a way that the organizational structure of the collection becomes transparent and vocabulary selection becomes easier.

Most information systems cover more than one domain of discourse. The documentary language of a collection encompasses a broad range of subjects and consequently reflects more than one specialized discourse language. For the searcher, this either constitutes a problem of discovering how the documentary language represents different domains of discourse or of discriminating against search results that do not belong to the discourse area of the question asked. In this thesis, a methodology for vocabulary support – a search term recommender system - is introduced that will aid exploration and discrimination of language in the search process by providing domain-

specific search term suggestions that will map between the searcher language and the documentary language of the information collection.

By supporting a searcher in the formulation of search statements that match the language of the document representations in the collection, an information system can provide insight into the information collection's organization, the information content and the language of document representations. If this process is based on searcher input and is interactive, the communication between a searcher and a system could be substantially improved.

If the information system's vocabulary support can disambiguate different specialized vocabularies within the collection and represents the discourse domains in the document collection adequately, it would aid the searcher in navigating this space effectively. Then the search experience should be truly satisfactory in terms of search speed, ease of use of the system and the conviction that the right portion of the document collection was searched and all relevant documents were found.

1.2 Organization of the Dissertation

In chapter Two, the language problem in information retrieval is described and its consequences for the search process discussed. Three search support strategies for alleviating the language problem are introduced.

In chapter Three, the formations of discourse communities are described with a particular emphasis on scholarly communities. A literature review on three fields

researching specialized languages in the sociology of science and linguistics shows the benefits for certain purposes of focusing on specialized languages within a general language. Domain analysis, another related area introduced in this chapter, is a field in information science that focuses on the domain as the primary analysis. This thesis can be considered an application of the domain-centered approach.

In chapter Four, the role of documentary languages in information collections and search is analyzed. Documentary languages are introduced and their advantages and disadvantages as document representations with full text representations are discussed and compared.

In chapter Five, the different notions of context in information science are presented and the approach of the search term recommender to provide contextual information about an information collection to searchers is described. Then, the search term recommender is introduced and its construction and functions described. The search term recommender belongs to a set of three vocabulary support mechanisms (automatic text categorization, query expansion, and terminology mapping), which are described and contrasted in the final section of the chapter.

In chapter Six, four ways for determining specialties within document collections are specified. Two specialty determination strategies are demonstrated on two scientific bibliographic databases, Inspec, covering research in physics and engineering, and Ohsumed, covering research in the medical fields. An analysis of the specialized languages in the Inspec and Ohsumed collections shows that both the natural language vocabularies and the controlled vocabularies differ between specialties.

In chapter Seven, the effectiveness of the specialty search term recommender system is evaluated in an automatic classification application. It is tested whether search term recommenders with a specialty focus predict better terms from the documentary language than a search term recommender trained on the whole collection. The level of specificity appropriate for optimal performance in term suggestion is determined for six specialties.

In chapter Eight, interactive and automatic methods for presenting specialty search term recommender vocabulary support to users are introduced. The variation in term suggestions for different specialties is analyzed and a preliminary experiment for the automatic selection of specialty search term recommenders in search described.

In chapter Nine, the research is summarized and three directions for further research presented.

Chapter 2

The Language Problem in Information Retrieval

The information retrieval problem in general is “how to obtain the right information for the right user at the right time” (Chu, 2003, 1). Four problem areas can be immediately identified in this statement: the techniques of “obtaining information”, the decision problem of the “right information”, the identification problem of the “right user” and the scheduling problem of the “right time”. These problem areas are not independent of each other but deeply and necessarily interlinked: the “right information” depends on the user and the time; how to obtain information depends on the user, the kind of information sought and the time; the right time depends on the user, the information seeking technique and previously known information; and so on.

This dissertation is primarily concerned with the problem of helping a user to obtain the right information given a certain user context (area of discourse), a particular technique of obtaining information (textual searching and browsing) and a particular information collection (bibliographic databases enhanced with documentary languages). This chapter is concerned with the technique of obtaining information (textual search) and its difficulties. Chapter Three provides an insight into the user context considered here (area of discourse or subdomain) and chapter Four will discuss documentary languages and their features in search.

2.1 The Search Process – Mapping between Searchers and Documents

Any kind of search and retrieval operation can be described as a multi-stage mapping problem: from a searcher's information need to a question, from a question to a search statement (or query), from a search statement to documents, and from documents to an ordered result set. The first two mappings are the tasks of the searcher, the last two mappings the task of the information retrieval system. The field of information retrieval traditionally has been more concerned with the system side, improving the mapping between statements to documents and the ranking of result sets.

From an information retrieval system perspective, a text search is essentially a string (term) comparison between the terms input by the searcher and the terms in a collection's documents. If the search term matches a term in a document, that document

is presented as relevant to the searcher. How the matching operation is performed and how relevance ranking is achieved is based on mostly term characteristics, collection statistics and sometimes user behavior.

Since the 1960s, the theory of information retrieval has come up with a number of mapping and ranking algorithms for text retrieval. For a collection of early papers on the theory of IR, see Sparck Jones & Willet (1997), an overview of current theories can be found in textbooks of the field (see, for example, Baeza-Yates & Ribeiro, 1999; Grossman & Frieder, 2004) and state-of-the-art research is presented annually at the SIGIR (ACM Special Interest Group on Information Retrieval, http://www.acm.org/sigir/) conferences.

The user side of the search process and the two mappings from an information need to a question and then to a search statement are the domain of the field of information seeking, needs and behavior, which studies user's searching behavior and how information is used in particular contexts[1].

The critical issue for the user side of the mapping process is the requirement to state an often rather vague and implicit information need within a sometimes equally indefinite domain area in the distinct and explicit query language of an information retrieval system. How a user comes to identify an information need and formulates a question to be answered by an information system is a relatively under-analyzed question in information science. Belkin (1980; Belkin et al., 1982) speaks of an "Anomalous State

[1] Some good introductions into the field are, Wilson (1999), Case (2002) and many ARIST chapters on this topic, e.g. Paisley (1968), Crane (1971), Crawford (1978), Dervin & Nilan (1986), Hewins (1990), Pettigrew, Fidel & Bruce (2001) and Case (2006).

of Knowledge" - describing the state where a searcher realizes a lack of information concerning an issue and then takes steps to formulate a query. Elsewhere, information need is defined as a "problem", which is "an unknown in a work or situation of a potential user of an information system. A problem signifies that which causes difficulty in finding or working out a solution" (Saracevic & Kantor, 1988).

Commonly, the existence of an "information need" is simply assumed as a prerequisite in the search process without much consideration for the many areas of life where people approach an information system without experiencing a particular need or lack of knowledge, for example, to browse a new subject's literature, to verify a fact, to look up schedule information, to compare prices, etc (see, for example, Erdelez, 1997; Toms, 2000; Solomon, 2002). In the information seeking, needs and behavior field, however, most information retrieval systems studies have been done on research databases with users trying to answer concrete questions, in fact attempting to fulfill an information need in the narrow sense[2].

The information need or lack of knowledge assumed in the search process leads to an interesting dilemma for the user. How do you know what you are searching for and when do you know when you have found it or that it is not in the information collection? This has been called the fundamental paradox of information retrieval where the user is forced to express in concrete words "that which you do not know in order to find it" (Hjerrpe, 1986).

---

[2] Cole & Spink (2004) define information seeking as a subset of information behavior, which describes the "purposive seeking of information in relation to a goal" (Spink & Cole, 2006, 25) and Wilson (2000) shows that this goal-oriented, purposive information seeking approach has been domineering information science research.

From a language point of view, the search process for people looking up information in an information system is really a funneling of a highly ambiguous idea space into very few specific terms for the search interface of the system. This increase in concreteness in term space goes hand in hand with a loss of context in idea space for the searcher. Although the person approaching an information system might only have a vague idea about the concepts he is searching for and the space he is operating in mentally has very fuzzy boundaries, the information system with its computer processing exact standards demands precise statements and sharp boundaries. As a consequence, the actual textual query that is input into the system might be quite different from the question in a searcher's mind, rendering the results put forward by the matching and ranking algorithms less useful for the searcher. The problem of finding the right word to express one's interest is called the language problem of information retrieval.

2.2 The Language Problem in Information Retrieval

Although there has been much research in the last few years to extend information retrieval beyond textual documents (i.e. speech, images, moving images, maps), text retrieval is still often treated as synonymous with (or at least as the primary application of) information retrieval (Sparck Jones & Willett, 1997). Text retrieval is a mapping of an information need expressed in a user query (text) to information content relevant to the information need expressed in a document (text).

This means that information retrieval is, in principal, a language mapping exercise – a system tries to retrieve documents that contain language (terms) equivalent to the language (terms) expressed in the user query (Blair, 2003):

> "We use language in searching for information in two principal ways. We use it to describe what we want and to discriminate what we want from other information that is available to us but that we do not want." (Blair, 2003, 4)

A query expressing a particular information need must fulfill two requirements: (i) it should describe the information need and (ii) it should describe it in a way that discriminates against (excludes) those documents that do not contain relevant information. The first requirement is related to the user-side (input), the second requirement is related to the system side (collection output) of the search process. Both requirements give rise to the language problem in information retrieval: the problem of language mapping (i.e. search) is indeterminate. Indeterminacy in language mapping means that that the probability of a one-to-one mapping between an information need and relevant documents in a collection is low, i.e. the probability of picking terms or phrases that describe an information need perfectly and retrieve the appropriate and relevant (and only those) documents from any given collection at the same time is very small.

The language problem in information retrieval has been recognized from the beginning of automatic information retrieval processing. Bar-Hillel wrote in 1962:

> "Though scientific and technological writers may not make full use of the theoretically unlimited number of ways of expressing their thoughts, put at their disposal by natural languages, they do make use of a large

> enough number to defeat any system based upon simple matching of expressions." (Bar-Hillel, 1962)

One of the early evaluation studies of a large text retrieval system (Blair & Maron, 1985) found large discrepancies between user expectations of how many of the relevant documents they would find and what proportion they actually retrieved. A large part of the problem was the selection of search terms for the retrieval process:

> "Stated succinctly, it is impossibly difficult for users to predict the exact words, word combinations, and phrases that are used by all (or most) relevant documents and only (or primarily) by those documents..." (Blair & Maron, 1985, 295)

More recently, one of the authors of the study claimed that this term selection problem is unlimited:

> "It has been shown empirically (Swanson, 1966), and argued theoretically (Blair, 1990), that the number of different descriptions that can represent the intellectual content of even a relatively short document may have no upper bound." (Blair, 2003, 5)

The indeterminacy of language in information retrieval has also been called "search uncertainty":

> "Search uncertainty is the primary source of problems in information retrieval. Search uncertainty arises because searchers have latitude in choosing terms to express a query and the search strategies they employ in acquiring information." (Chen & Dhar, 1991)

While search uncertainty mainly refers to the searcher side indeterminacy of language in search, the previous quotes show that the problem exists on both sides of the search process: users and system.

2.3 Language Mapping – Information System Perspective

From a system (document collection) standpoint, the central task of information retrieval research is "to understand how documents should be represented for effective retrieval" (Blair, 1990, vii). This is a problem of selecting the right language to describe a document's content effectively so that it can be found by a searcher – a problem of language and meaning.

Preparing documents for representation (and search) in an information retrieval system has run the gamut of not changing a document at all (presenting it in full-text) to abstracting pertinent information from the document and adding content-describing keywords to it. In earlier systems, due to storage and processing constraints, it was necessary to input only short records into a system, resulting in the invention of elaborate subject-describing keyword systems (documentary languages) that would severely restrict the number of possible terms to search on. Nowadays, both full-text collections and collections containing only abstracts or only keywords live side by side. For media other than text, the importance and value of adding keywords to records has become apparent again in the recent success of social tagging systems like Flickr (photos) or del.icio.us (URLs) that let users add keywords to their photos or URLs, thereby increasing the

findability of the item but also providing the opportunity to cluster items around concepts (i.e. the keywords that have been provided for them).

For language mapping, a full-text representation of a document provides the largest possible search space for a user, providing all terms in the document for search. What better to describe the content of a document than the whole content of the document? More text to search increases the chances of making more matches between query terms and terms in documents. Nevertheless, having a large pool of terms to search on does not mean that the probability of search success is automatically higher. The existence of more words might correlate only weakly with the existence of “good” search words. For the process of mapping from query to document and for relevance ranking, enlarging the term pool means introducing many more features that need to be considered for this process, making ranking and retrieval more difficult.

On the other side, having only a few keywords to search on constricts the search space drastically for the user. More effort is necessary to figure out the right search terms that will match documents. However, once an appropriate search term is entered, the retrieval and ranking process is easier because efficient: only documents that are really concerned with the concept described with the search term will be represented by it. If you have only limited terms to choose to represent a document’s content, the selection process and the available term space become even more important. Even for controlled vocabularies, where the term space is severely restricted and the mapping between a concept and a term should be straightforward, the process of mapping is still subject to the ambiguities of language.

2.4 Inter-Indexer & Search Term Selection Inconsistency

Different indexers do not always assign the same term to represent a document. In the library literature this is known as inter-indexer-inconsistency. Even professionally trained indexers will assign different keywords for the same information content. This "vocabulary problem" (Furnas et al., 1987) is far more serious than most users and system creators assume and has been repeatedly demonstrated.

Search term selection inconsistency is the mirror side of inter-indexer inconsistency. As professional indexers will select different keywords to represent a document's content, searchers will select different keywords to represent a question. Search term selection consistency is probably even lower than inter-indexer consistency because searchers do not know the document representation language or the indexing practices in the information collection as well.

For the years from 1961-1971 alone, White (1973) describes 41 reports studying the problem and many other examples can be found up until today (see for example: Fidel, 1985; Furnas et al., 1987; Saracevic & Kantor, 1988; Tonta, 1989; Bates, 1989; Chan, 1989; Humphrey, 1992; Iivonen, 1995; Leininger, 2000). Even though the studies report on different data sets (more or less subject specific, more or less controlled vocabularies) and measurement methodologies (Lancaster, 2003, ch. 5), the measured levels of consistency between searchers or indexers in term selection are astonishingly low.

For general subject collections, Saracevic and Kantor (1988) compared 5 searches

on each of 40 test questions in different DIALOG databases. Researchers used identical search term formulations in only 1.5% of the cases, however on average 27% of the terms assigned overlapped. Furnas et al. (1987) found that the probability that two searchers will use the same term for a concept or book in “five application-related domains” is between 10% and 20%. These numbers have been mirrored by Bates (1989) for library catalog searches, whereas Iivonen (1995) found a slightly higher average (31.2%) for searches in a Finnish database.

For a more subject-specific database with a highly controlled vocabulary, the consistency should be higher because the term and concept space is smaller. For MEDLINE (Medicine), Leonard (1975) analyzed the consistency between ten indexers at the National Library of Medicine working with a complicated controlled vocabulary – the Medical Subject Headings (MESH). He found that consistency between indexers in assigning keywords to documents ranged between 36 and 48% depending on the restrictiveness of measurement and which keyword combinations counted as equal. In a follow-up report, Humphrey (1992) summarized MEDLINE indexing consistency studies, reporting a consistency average for headings assignment of under 49%. Another subject-specific example is a recent analysis of PsycINFO (Leininger, 2000), comparing 60 cases of term assignment with different consistency measurements. Depending on the methodology, the average consistency for term assignment in the PsycINFO records was between 50-60%.

The subject-specific studies were done with highly trained indexers, who knew both the discipline, the concept space and were trained in the use of the controlled

vocabulary. The general area studies described here were done with searchers who were not as fluent in the use of the controlled vocabulary. Assuming a very conservative stance with searchers assumed to have the lowest consistency and trained indexers in a subject-specific database the highest, the probability for any two people of selecting the same keyword for a document or question can be as low as 20% and will not surpass 60%.

2.5 Language Mapping – User Perspective

Other language mapping problems arise for the searcher because of the other mapping that is necessary: between the vague information need in one's head and the vocabulary of the system, and the likelihood of not knowing the system vocabulary or the content (documents) of the system.

From a user's standpoint, the principal difficulty for a searcher is not only to describe his information need in a way that makes sense to him (first mapping) but also to predict "the words or phrases that have been used to represent (index) the document or documents in the system with which he would be satisfied" (Blair, 1990, 9). Only the second mapping is studied in term selection consistency analyses.

The language problem for the searcher derives from three overlapping problem areas:

> (1) the searcher might not know how to state his/her information need,
> (2) the searcher uses language that describes the information need but does not match the language used in the documents relevant to the information need, or
> (3) the searcher uses language that is not discriminating enough and matches irrelevant documents as well as those containing relevant information.

Evidence from the philosophy of language and especially Ludwig Wittgenstein's (1953) later writings suggest that the ambiguity of language is a universal phenomenon. A question is whether the language problem in information retrieval can ever be solved with more sophisticated document representations or more sophisticated and experienced searches.

2.6 Language Ambiguity

In semiotics, C. S. Peirce wrote of the concept of "unlimited semiosis", which describes the problem that "there can never be a necessary and sufficient explanation or description of the meaning of a sign / expression" (Blair, 1990, 137). For our understanding of language, this means that there is never a one-to-one mapping between a sign (term) and concept, nor is there even a limited number of relationships between different concepts and a sign (term).

How do we know which concept is meant by a given expression? Wittgenstein's philosophy of language[3] states that a word's meaning (i.e. the relationship between a concept and an expression / term) arises out of the activity in which it is used. If a word is used in different ways or contexts, even though the term does not change, the meaning of it will. Therefore, the ambiguity of language is a consequence of the ambiguity of language usage or the number of activities, in which language is used.

[3] For an in-depth explanation and application of Wittgenstein's language philosophy in the field of information retrieval, the writings of David Blair (1990, 1992, 2003, 2006) are indispensable.

Zipf (1949) argued that two forces were determining the ambiguity of language: unification occurs because a “general” word can be used in a variety of contexts; and diversification occurs for specific words that appear in few contexts and are more precise in meaning. According to Zipf, the ambiguity of language is based on speaker’s economy, which tries to use as few words as possible in speech. Auditor’s economy is responsible for diversification, i.e. the introduction of more specific words to make understanding easier.

If the meaning of a word is ambiguous and indeterminate in general (we cannot exhaustively enumerate the number of meanings), then the keywords, which represent documents in an information retrieval system, are also ambiguous, even if they are from a controlled vocabulary. It is therefore theoretically impossible to predict with absolute certainty the keyword chosen for a given concept or document, because there is no precise or unique association between terms and concepts. Seen in that light, the argument can be made that by using a controlled vocabulary, the inherent ambiguity of language can be reduced in the context of search, possibly by half (assuming the increase of inter-indexer consistency from ca. 30% for general purpose databases with uncontrolled indexing vocabularies to circa 60% for subject-specific databases with controlled vocabularies is an indication of the reduction in ambiguity.)

How severe is the ambiguity problem? Zipf (1945) analyzed 20,000 words and their “meanings” as they were listed in the Thorndike-Century Senior Dictionary. He found a relationship between the number of different meanings of a word and the square

root of its frequency of occurrence in a text. He argued that the more often a word occurs, the more meanings it will have.

As with inter-indexer consistency, one would expect the ambiguity problem to be less drastic in more subject-specific fields with highly technical languages. Humphrey (2006) reports on a study by Weeber, Mork, & Aronson (2001) where 34 million phrases taken from 409,337 MEDLINE medical records were run against MetaMap (a system that maps phrases to the Unified Medical Language System (UMLS) Metathesaurus). About 3.76 million phrases (11%) were mapped in an exact string match to more than one concept in the thesaurus. Considering that a thesaurus should present concepts in a way that are mutually exclusive and precise, only context (and the surrounding hierarchy in the Metathesaurus) can disambiguate these cases.

Another experiment (Krovetz & Croft, 1992) compared a subject-specific collection (CACM – Communications of the ACM title and abstracts) and a general collection (short TIME Magazine articles) with each other and found that for the subject-specific collection, ambiguity was actually higher (for the CACM collection 4.4 senses per word were found in a dictionary lookup, whereas for the TIME collection only 3.6). For the domain-specific collection, contrary to Zipf's finding of a linear relationship between the number of meanings and word frequency, they found a large number of words with high ambiguity and low frequency. These were words that had less word

senses in the general collection but took on specific meanings in this subject-specific community of discourse[4].

2.7 Language Games & Language Regions

How is meaning disambiguated in every-day language? We learn and use language by means of what Wittgenstein calls “language games”. Language games are regular patterns or structures in which the meaning of a word is explained by its use. According to Wittgenstein, the meaning of a word can only be understood, if the language game, in which it is used, is correctly applied:

> “For Wittgenstein, the criterion for whether you understand the meaning of a word is not whether you have the right idea, but whether you use it correctly in your day-to-day speech and writing.” (Pitkin, 1985, 20)

Contrary to the classical understanding of the meaning of concepts in semiotics, Wittgenstein postulated that meanings are not merely mental entities or ideas because they are dependent on the context of their usage. If meaning is contingent on activity and use, it follows also that meaning “remains ‘nonmonotonic’ or ‘defeasible’ – it is always subject to revision or change” (Pitkin, 1985, 23).

For our focus on specialties and dialects, it is also important to mention Wittgenstein’s notion of “language regions”. Language regions are particular zones in

[4] Examples for the CACM collection contained the words “passing” (as in “message passing”), “parallel,” “closed,” “loop,” and “address”, which take on a special meaning in the computer processing field.

general language with their own grammar and language games, where words are ostensibly used for the same concept but occur in activities not only denoting different meanings but also different prescriptions for usage:

> "To speak of H_2O as the 'chemical for water' is to speak in a confused manner: H_2O is a symbol the rules of whose behaviour are wholly different from those which govern the symbol 'water'. If "water" and "H_2O" were merely labels, they would be used for referring to "the same thing"; but they are not merely labels. They are signals used in radically different language games, performing very different functions." (Pitkin, 1985, 142)

For the representation of documents, the concepts of language games and language regions are important. First of all, they point towards the improbability of a "universal ontology" for everything and everybody – a description of all things occurring in the world, since words and the concepts they denote are context-dependent[5]. Second, since the activity (language region, i.e. jargon) determines language use, subject-specific vocabularies are important for document representation.

For information retrieval and the search process in general, the concept of language games poses some problems. If only context and activity can precisely determine meaning and subject-describing keywords are largely present without context, then the "language of document representation" (Blair, 2002c, 376) is inherently ambiguous. Thus, the goal of any retrieval and document representation system should be

[5] Traditional classification research aimed at creating a single classification system representing the universe of knowledge for every purpose (Miksa, 1998), but Mai (1999) argues that goals are lower now: the "postmodern classification aims at providing a pragmatic tool for specific domains" (Mai, 2004, 39).

to “build as much of the missing activity or institutional context back into the language of representation” (Blair, 1990, 323).

2.8 Consequences in Search

The reality in search looks quite different. Even if special care is taken with document representation from the system side by providing a subject-specific controlled vocabulary to reduce ambiguity of meaning, the searcher could perceive this as a burden or not use the additional help in describing search terms at all.

2.8.1 Term Selection

The searcher is now faced with not only describing his/her information need but also with the additional hurdle of describing it in terms of the controlled vocabulary if he knows about its existence at all. For information retrieval systems that use sparse information for document representation (like library catalogs), the consequences can be dramatic. A 1991 study of academic library catalog use found that 58% of search sessions began with terms that were not included in the system's controlled vocabulary (Peters & Kurth 1991). Markey (1984) categorized 36% of 859 studied subject search attempts in a library catalog as "whatever popped into the searcher's mind", 65% of these random search attempts resulted in no retrievals.

2.8.2 Anchoring Bias

Information retrieval can be seen as decision-making under uncertainty and it is therefore subject to human biases in the decision-making process (Blair, 1980). For the selection of document representations (keywords) and search terms, the anchoring bias can be a particular problem. In a search, the first "early" search terms provide an anchor in finding related terms. Those anchor keywords are rarely changed or taken out of the search - even though they might be unsuitable for a query – because "the searcher overestimates the probability that these first terms will be part of the more successful queries, and, as a consequence, she will keep as many of these first terms as possible in her subsequent queries" (Blair, 2002b, 298). "Anchoring" on unsuitable search terms will prevent even the most sophisticated term expansion strategy or the best subject-specific controlled vocabulary from succeeding in providing better search terms.

Anchoring bias could be a problem for the highly touted ESP Game (http://www.espgame.org, Ahn & Dabbish, 2004), a web application that uses a game to assign keywords to images in a database. For each round, an image is presented to two users and a couple of "taboo words" that cannot be used to index are provided. The two users are playing with each other guessing the terms that the other user will use to index the image. Guessing another person's indexing behavior is prone to the inter-indexer consistency problem and will lead to lowest common denominator indexing (not very specific keywords) but should also help alleviating the problem, because final indexing keywords are only assigned when at least two people agree. The taboo words, however,

could give rise to the anchoring bias because players will naturally try to come up with synonyms to the taboo words first before trying out new concepts overall.

2.8.3 Futility Point Frustration

When the number of retrieved documents reaches a level where the user is not able or willing to evaluate the results he might give up in frustration. The user's needs (how many documents and how specific), the ratio between relevant and irrelevant documents and the number of screens he is willing or motivated to browse through determine the threshold between satisfaction and overload.

Blair (1980) defines the "futility point" where the threshold is reached: If the number of retrieved documents exceeds the user's futility point, then he will give up the search in frustration. If the number of records exceeds the anticipated futility point of the user, he may not even look at any of them. Even if the first few records are satisfactory for the searcher, only the most determined ones will examine more than a certain number (usually not more than one screen).

A search will commonly be stopped when the searcher is satisfied with the documents that he retrieved (Cooper, 1973) and not when all the documents that might be potentially relevant are retrieved. Blair calls this the “satisficing criterion of retrieval” (Blair, 1990, 74-75) following Herb Simon’s work on bounded rationality.

There are only a few cases where satisficing will not stop a retrieval process and exhaustive searching becomes necessary. It is in those areas where knowing everything there is becomes vital for further progress: patent searching, litigation support searching

and (potentially) searching to support academic research. Information system builders have to be aware of the satisficing behavior of their customers and provide search support that will gain successful results in the early stages of search.

2.8.4 Search Failure

The opposite of information overload (exceeding a searcher's futility point) is search failure. Search failure occurs when users do not retrieve a matching document for their query. This either means the information retrieval system retrieves no result at all (an unlikely event in today's large document collections) or - a more complicated case - the query statement retrieves records but none matches the information need of the searcher. In a number of library catalog studies in the 1980s, no-match subject searches were found to range from 35% to a high of 65% of all subject searches (Borgman, 1986; Markey, 1984, 1986). A much smaller (227 searches) but more recent study (Gross & Taylor, 2005) found that 18% of subject searches did not retrieve a valid result (4% retrieved more than 10,000 records and 14% retrieved no records at all).

Nowadays, with the increased number of documents in any information system and the consequent high probability of retrieving at least some documents, the problem of search failure seems more problematic with regard to precision (i.e. retrieving a high number of irrelevant documents with very few relevant documents).

Another search failure can occur when the document representations are too general to become useful in large collections. The problem of not enough specificity or discrimination (to discriminate from other related concepts) in keyword choice is

apparent in fast-growing information systems without a previously agreed-upon vocabulary that would require keyword assignment to be as specific as possible. An example in point is the popular social tagging sytem del.icio.us, where most users started out with very general and broad tags (keywords) only to come back later to realize that every category started to have too many entries (URL assigned with this tag) to provide any real browsing guidance or organizational structure. Re-assignment of tags has become necessary and is considered a chore.

2.8.5 Collection Growth

As document collections grow and documents indexed become larger (i.e. full text search), the indeterminacy of language in the documents and the difficulties of picking the correct search term increases as the number of possible relevant search terms grows as well. This has consequences for the amount of possible different query statements and the amount of documents retrieved. For queries, it has been shown theoretically (Blair, 1990) that the expansion of the search term space does not necessarily improve the retrieval success rate because even a modest increase in the number of terms can lead to a high increase in the number of query statements (if search terms are combined in different ways), each retrieving more result sets. For retrieved sets of documents, another consequence of large document collections and full-text indexing is that any search term or any query - regardless of its suitability for an information need - will retrieve more documents. A larger retrieved set makes it harder to find relevant documents thereby exceeding a searcher’s willingness to evaluate the results.

One more hidden problem of collection growth is that subject-describing keywords may outlive their usefulness because the number of documents added that are assigned with this keyword exceeds a threshold whereby a search would be non-discriminative (Blair, 2003, 6).

Collection growth, the changing nature of the search failure problem, and a probably decreasing satisficing range in the age of Google search (fewer and fewer results are looked at) require more sophisticated language mapping techniques in information retrieval systems.

2.9 Search Support – Alleviating the Language Problem

Over the years a lot of approaches have been suggested to alleviate the language problem in information retrieval. The solutions belong to roughly four areas: expanding or improving document representations, restricting the results set, supporting different search strategies, and expanding or improving the search term space (including feedback techniques). The last two techniques are often combined in the same system.

2.9.1 Exhaustive Indexing

Expanding the document presentations means increasing the term space that is available to describe a document's content. With computer storage and processing not a problem anymore, a lot of information systems switched to using full-text instead of title-abstract-keyword representations of their documents. Even for non-full-text systems,

adding the abstract, additional keywords and all the references has become the norm. Library catalogs are the one exception, although maybe not for long with the inclusion of full scanned-in books in digital formats.

For restricted vocabulary systems (title-abstract-keyword or controlled vocabulary as document representations), Furnas et al. (1987) and Gomez et al. (1990) promote the strategy of "unlimited aliasing", where an unlimited number of controlled vocabulary terms or keywords is assigned to the record. This strategy provides a richer vocabulary for the searcher: the probability of typing a term that is contained in one of the relevant documents is much higher. Gomez et al. (1990) report improved retrieval success.

Numerous experiments and projects have been conducted with adding words from titles, tables of content, and back-of-the-book indexes over the years (e.g. Cain, 1969; Atherton, 1978; Micco & Smith, 1989; Lancaster et al., 1991, Cousins, 1992) and often showed high recall values but low precision. Byrne and Micco (1988) estimated 300% increased retrieval results by enhancing records with an average of 21 additional headings. However, they complained that the large number of records was unmanageable.

Brooks (1993) conducted experiments with enhanced records from the LISA, ERIC, and ISA bibliographic databases. His findings also contradict the opinion that more keywords lead to better retrieval results. His main argument is that some descriptors are qualitatively better than others and therefore more useful for retrieval. The un-evaluated adding of possibly low-quality keywords does not necessarily guarantee retrieval success.

Furthermore, “unlimited semiosis” prevails also in the controlled vocabulary /

keyword space, so that the number of aliases, i.e. the words or phrases that can be used to represent the content of a document might have no upper limit.

2.9.2 Restricting the Result Set

To alleviate the information overload problem (too many documents in the result set), users are mostly left to themselves. Large retrieved sets can be reduced by adding additional or more precise query words to the query statement and "ANDing" them in a Boolean search. In most information retrieval systems, large retrieved sets can be reduced by limiting the resulting records with respect to document characteristics like date, language and location. The OASIS system (Buckland et al., 1992) is an example for this technique.

Another approach, which effectively limits the result set, is the partitioning of the search space as proposed by Blair (1994, 2002b). Whereas partitioning the result space occurs only after a search has occurred, splitting the search space divides the search process into two stages. The first stage, which uses “partitioning queries”, divides the document collection into smaller sets that will be searched on. The division can occur by type of publication, time or place restrictions, context of publication (e.g. institutional affiliation etc.) and should describe a “precisely definable region in search space” (Blair, 1994). The second stage, which uses “resolving queries”, operates on the limited document set and should retrieve more relevant documents and help users bringing those documents into focus.

Most efforts today will not go in the direction of limiting result sets because users

are now used to only looking at the first few results in a search anyway. One interesting problem could be the presentation of the result set in a way that shows different sets of documents describing different concepts that were nevertheless retrieved by the same query. Due to collection growth and the satisficing behavior of searchers, document sets that describe a different concept area but are lower-ranked than the prevalent concept area are lost to the searcher. Splitting up the result set in a way to represent several “result spaces” could be very useful.

2.9.3 Supporting Search Strategies and Query Formulation

The best support for formulating a query statement still is a human intermediary who knows both the system’s query statement requirements, the system vocabulary and the subject area of the searcher’s information need. However, human intermediaries are expensive and rare. A number of automatic search intermediaries have been proposed instead.

Automatic search intermediaries are information retrieval system modules that assist the user in formulating and reformulating query statements, in finding the right search strategy, and in identifying the relevant search terms from controlled vocabularies. Automatic search intermediaries are efforts to substitute the human intermediary. Many efforts use very sophisticated user and information system models to support the search process. A few examples shall be listed here.

The I^3R (Intelligent Intermediary for Information Retrieval) system described by Croft and Thompson (1987) uses knowledge-based techniques to assist a user in the

formulation and refinement of query statements. During the search process, a user model incorporating existing domain knowledge and other user-specific characteristics important for the retrieval success is established. Several expert modules help the user in identifying the best retrieval strategies and search terms.

The Metacat system (Chen, 1992) is a knowledge-based retrieval system, which uses a semantic network structure to represent subject knowledge and classification scheme knowledge. Metacat creates a user profile, identifies task requirements, suggests heuristics-based search strategies, performs semantic-based search assistance, and assist online query refinement.

AURA (Associative User Retrieval Aid) incorporates an expert system, which supports the searching process in the MEDLINE and CATLINE databases at the National Library of Medicine (Doszkocs & Sass, 1992). An expert subsystem handles rule-based search strategy modification. Another subsystem was planned to incorporate a neural network that links high-level MeSH classes based on the raw medical subject heading co-occurrence associations.

Khoo and Poo (1994) describe the design of an expert system that focuses on possible search strategies and selection rules for them. The user types a query into the user front-end interface, which responds with the best matching search strategy to use.

Oakes and Taylor (1998) created an automated query system, which will assist pharmacologists in searching the Derwent Drug File (DDF) pharmacological database. The system supports the user with the vocabulary selection (with approximate string matching, morphological analysis, browsing and menu searching), the choice of context

indicators (Boolean operators or context operators), and query reformulation (by using relevance feedback, thesaurus relations between document index terms and term frequency data).

The knowledge-based FIRE system (Brajnik et al., 2002) utilizes a terminological aid module to support query formulation (truncation, controlled vocabulary browsing, relevance feedback) and a rule-based strategic aid module to support different query strategies (area scanning, journal run, author search) offered through hints or advices to the user.

Many of these sophisticated systems report improved retrieval results for their specific application but very few are generalizable over a large number of applications. With Google dominating the search market for general-purpose queries and providing no search guidance whatsoever, these sophisticated systems remain a niche market but should be especially important in information retrieval systems that have complicated query statement requirements or very restricted document representations (e.g. subject-specific databases, patent databases or library catalogs).

It is a question of the application area whether black-boxing the search and ranking process (like Google does) or making the system more transparent (like the application-specific search support systems do) will be more successful in retrieving relevant results and providing a satisfactory search experience for the user. Most probably this depends on the time, effort, specificity of topic and exhaustivity of search requirements of the searcher and the information systems document representation, search statement complicatedness and collection size.

2.10 Conclusion

Wittgenstein's philosophy of language teaches us that meaning arises dynamically in context, in language games. If the surrounding activity is important for understanding language in everyday speech, then context is also important in understanding why query terms fail or succeed given a particular document collection and information retrieval system. To really understand the consequences of a search term for a query and its corresponding result set, we have to know the text and its usage in the document representations (full-text or keywords) or the information retrieval system. Different language games occur in different language regions, which results in varied grammars and language use (concepts are expressed in certain terms associated with a language region). Language regions can determine language games and therefore decide on the usage of a word.

The language problem in information retrieval arises because the surrounding context (or language game in Wittgenstein's terms) is not provided and the language used in documents, document representations and queries becomes ambiguous. From inter-indexer-consistency studies reviewed in this chapter, we know that more subject-specificity in the document collection and / or more controlled vocabulary for the document representations (and also queries) will reduce ambiguity. The more subject-specific an information environment, the more restricted is the language space that we move in and the problem of unlimited semiosis is lessened.

How is meaning explained in everyday language? Demonstrating a word's usage in an activity reduces the ambiguity of the concept by providing necessary context. Understanding is achieved through interaction in learning the appropriate language game. How can information retrieval systems provide this kind of supporting framework to help understand a collection's and information retrieval system's language space? Our focus should be (1) on providing as much context as possible, (2) on making the search process more interactive and (3) on dividing up the language space in search and retrieval to make words appear less ambiguous.

This dissertation approaches these goals from three perspectives. Firstly, by making the structure of document representations more transparent (the controlled vocabulary is used to support search and to show the appropriate language use for the particular information system), more context is provided for the user. Secondly, by providing feedback with vocabulary information, the search process is made more interactive for the searcher. And thirdly, by providing insight into different language spaces or dialects (specialties) within an information system, the search space is divided and words are made less ambiguous. The next chapter provides an overview over specialty communities and dialects in the scientific research environment.

Chapter 3

Specialized Communities & Specialty Languages

This dissertation focuses on mapping specialized languages of searchers to the documentary language of an information system. Specialized languages arise from specialized communities. This chapter describes various approaches to defining specialized communities, chief among them the notion of a scientific research specialty, and the development of specialized languages in them. Finally, the field of domain analysis in information science will be introduced. It regards the "domain" as its object of study and argues for the development of domain-specific and domain-centered approaches to information systems design. This dissertation can be seen as an application of a domain-centered approach in the information science tradition.

The concept of specialized communities encompasses more than just scientific disciplines and research specialties, although these types of specialized communities have been in the forefront in sociological investigations. Because of the stability of scientific structures and the establishment of disciplinary departments in universities, they are easier to study than other more unstable and shifting communities. Most bibliographic databases reflect disciplinary or scientific specialty boundaries. Since the focus of this dissertation is on this set of information retrieval systems, the focus of this chapter will be on disciplines and scientific specialties.

3.1 Disciplines

Although the demarcation and definition of the concept of science has continued to be a topic of research in the philosophy of science (Taylor, 1996), the topic of the discipline (Lemaine, Macleod et al., 1976) seems to have lost some of its attraction to researchers since the 1970s and 1980s (except in Germany, see Stichweh, 1992). However, individual discipline histories have remained popular throughout the years (Oleson & Voss, 1979: Kohler, 1982) and interest in cross- and interdisciplinary work is stronger than ever (Klein, 1990, 1996; Palmer, 1999).

In the introduction to a history of a specific discipline (biochemistry), Kohler defines disciplines as political constructs, created to defend institutions against competing research areas:

> "Disciplines are political institutions that demarcate areas of academic territory, allocate the privileges and responsibilities of expertise, and structure claims on resources. They are the infrastructure of science, embodied in university departments, professional societies, and informal market relationships between the producers and consumers of knowledge". (Kohler, 1982, 1-2)

This definition emphasizes the fact that scientific disciplines encompass various degrees of academic status and are involved in a never-ending competition for scarce resources in the university environment.

Robert Stichweh (2001) focuses more on the demarcating features of disciplines. He describes disciplines as units of internal differentiation within science allowing for the formation of structures (1) in the social system of science, (2) in systems of higher education, (3) as a subject domain for teaching and learning in schools, and (4) as the designation of occupational and professional roles. Specialist communities are systems of communication, meaning that the exchange of ideas and results is one of the more important ingredients for a discipline to maintain existence.

Other than being a political competitor for scarce resources and a differentiating device, a discipline is also defined as a delimited cultural domain, which is:

> "…a socially and culturally defined organizational arrangement that focuses on knowledge production and growth. A discipline can be characterized as an epistemic community whose members have a special frame of reference oriented toward specific abstract objects of investigation." (Lindholm-Romantschuk, 1998, viii)

The character of disciplines varies by the specialization and standardization of tasks and materials, the degree of segmentation, the degree of differentiation into schools,

the hierarchization of sub-units, the impersonality and formality of control procedures, the degree of theoretical coordination, the scope of conflict, and finally the intensity of conflict (Whitley, 1984, 167) [1].

Disciplines, however, are not the place where scientific work really takes place. They represent "historical evolutionary aggregates of shared scholarly interests" (Chubin, Porter, & Rossini, 1986, p. 4) but “disciplinary categories no longer reflect how people think about things” (Geertz, 1983). Chubin (1976) argues that disciplines are a construct to maintain academic departments and train and certificate new scientists, whereas smaller units called specialties are the focal point of research:

> “In short, disciplines form the teaching domain of science, while smaller intellectual units (nestled within and between disciplines) comprise the research domain. Within the sociology of science, these units have been termed "scientific specialties." (Chubin, 1976, 448)

Lenoir (1997) and Knorr-Cetina (1999) also argue that the discipline plays a pedagogical and administrative role while research and problem solving is done in intellectual fields, epistemic communities or specialties and Dogan simply summarizes disciplines as a “cluster of specialties” (Dogan, 2001, 14851) bound together by historical commonalities.

Disciplines today are the overarching constructs used for organization in an academic environment but too general and cumbersome in the research context. Research

[1] Following these characteristics, Whitley defines seven types of scientific fields or disciplines: (1) fragmented adhocracy, (2) polycentric oligarchy, (3) partitioned bureaucracy, (4) professional adhocracy, (5) polycentric profession, (6) technologically integrated bureaucracy, (7) conceptually integrated bureaucracy (Whitley, 1984, 167)

and scientific communication takes place in narrower intellectual fields, which is also the location for the formation of specialized languages and scientific dialects.

3.2 Specialties

A discipline is that portion of approved and taken-for-granted research knowledge found in a textbook – the background knowledge on which specialties within a discipline base their ongoing development on and which serves as a common ground for communication between specialties. Specialties are narrower fields that are not necessarily bounded by the borders of the discipline:

> "Clusters of related research areas constitute specialties whose members are linked by a common interest in a particular type of phenomenon or method. Disciplines, in turn, are composed of clusters of specialties; members of disciplines, however, are affiliated as much by political and social interests as by intellectual concerns. Moreover, members of recognized disciplines are bound to one another by a professional association that organizes scientific meetings and acts as the representative of its members to organizations and institutions outside the discipline." (Crane & Small, 1992, 232)

Specialties are the structures in which a concept is developed to its fruition and at which a regular exchange of information and assessment of results happens. Students are mentored and modeled into researchers at the specialty level and work within its boundaries (Chubin, 1976, 454). The specialty is the prime audience and reference group for results and new scientific developments and it is at this level at which sources for wherewithal and rewards are allocated (Crane & Small, 1992). The specialty is also the

level of organizational structure at which scientific consensus about a question is reached (Mulkay, Gilbert et al., 1975).

Specialties can be small communities with a core of about one or two hundred researchers (Kuhn, 1962 [1996]; Price, 1986) and up to a total membership in the vicinity of a thousand members (Morris, 2005). They can be identified by:

> (1) a firmly established intellectual framework (i.e. seminal papers, research subtopics, experts, centers of excellence, collaboration teams and archival journals (Morris, 2005);
> (2) cognitive and technical standards which confine acceptable innovations within well-defined limits;
> (3) institutionalized mechanisms of recruitment and funding;
> (4) a stable rate of recruitment;
> (5) well-established research teams or groupings of collaborators with acknowledged areas of special competence and interest;
> (6) highly productive and influential scientific leaders who are highly cited in published papers and who occupy a central position in the informal communication system and influence the power relationships in and between specialties (Cappell, Guterbock, 1992);
> (7) opportunities for the provision of findings; and
> (8) opportunities for attaining professional recognition and advancement.
> (Mulkay, Gilbert et al., 1975, 198)

3.3 The Formation of Specialties

If the intellectual landscape undergoes a conceptual change (Wray, 2005) or scientists carve out a new niche and new social role for themselves (Ben-David & Collins, 1966 [1991], 50), a new specialty might be formed. New specialties emerge because existing specialties cannot solve a research problem or - as Kuhn (1962 [1996]) would state it - because the current paradigm is in a crisis. New specialties also come into being when enough researchers decide that a certain research problem or methodology is

worth pursuing and enter into new cooperation and research partnerships (Mulkay, Gilbert et al., 1975). New specialties can also be prompted by funding agencies, which motivate researchers to choose research directions where enough resources are available.

In a 1975 paper, Mulkay, Gilbert et al. present a model of how problem areas or research networks (specialties) go through a three-stage life cycle that consists of an emergence phase, a growth phase and a decline phase. The first exploratory stage is marked by general confusion among the scientists as to what the research area is to which they are contributing. The problems are imprecisely defined and the researchers are insecure as to whether this area is productive enough to warrant entry. At first, communication among researchers is sparse, also because there is a general confusion as to which forum would be the right one to present these results and they tend to be scattered among a variety of disciplinary and general-purpose journals.

The second phase is marked by improved communication and a rapid growth of the research network. Although the number of researchers and papers increases, it does not necessarily mean that the number of innovations and significant findings increases in an equivalent amount. In phase two, research techniques and methodologies get stabilized and researchers commit to the area. The specialty gets saturated. In the third phase, the research network either starts declining and eventually disappears or it shifts its main emphasis on to those tangential problems that are soluble using the methods and theories of the area.

A variant outcome is the fragmentation of the specialty into even more specialized sub-specialties. Fragmentation occurs along substantive, epistemological,

methodological, theoretical, and ideological lines (Dogan, 2001). Another important aspect of specialty development is the recombination of specialties. Recombination occurs when specialties form across disciplinary boundaries around a common research problem. When recombining specialties, concepts, theories, and methods are borrowed and exchanged across specialty and even disciplinary boundaries.

The stages of growth of specialties are manifested in the way new knowledge is produced and handled and the type of published papers distributed. In the beginning stages of a new specialty, a few researchers will publish discovery papers, which are then followed by many empirical knowledge papers. As first, knowledge is loaned from other fields but then it gets extended, applied, consolidated and codified according to the new specialty's intellectual landscape (Tabah, 1999).

3.4 Epistemic Communities, Communities of Practice and Communities of Discourse

There are a number of concepts in the sociology of science literature that are similar to that of a scientific specialty. Of particular interest here are epistemic communities, communities of practice and discourse communities because they all share features with scientific specialties that make them comparable: a set of common beliefs and practices, a set of community members that are attached to the community's paradigm (whatever it may be) and, consequently, a community language, which will distinguish its practices and members from others.

The concept of epistemic communities was introduced by Peter Haas in policy studies. Epistemic communities are “networks of knowledge-based experts” (Haas, 1992, 2) with recognized competence in a particular domain. They can consist of professionals from a variety of disciplines or backgrounds but they share a consensual knowledge base, a set of beliefs (normative and causal) as well as a notion of validity that will cause them to act on certain problems in a similar way. Epistemic communities can be a subset of a discipline (Haas, 1992, 18) but can also span disciplinary boundaries.

Communities of practice (Lave & Wenger, 1991; Davenport & Hall, 2002) are those conglomerates of expert, intermediate and novice practitioners that are constituted by a certain common interest and socio-cultural skills that they all share and which are shared and learned through practicing and participating. Communities of practice are constituted by a common identity, a common set of artifacts and tools used and a common body of knowledge and skills shared by all members of the community (although not necessarily to the same degree). Unlike specialties, communities of practice “in workplaces and schools are mostly ad hoc” (Lave & Wenger, 1991, 78) and do not have formal literatures or institutional structures.

“Discourse community” is a concept from sociolinguistics (Nystrand, 1982; Gumperz, 1982a, 1982b) which draws attention to identity formation through language:

> “The key point of our argument in this book is that social identity and ethnicity are in large part established and maintained through language.” (Gumperz, 1982b)

The main characteristics of discourse communities are the language (terminology, linguistic forms and regulative rules) and the cultural concepts that community members share. A discourse is an instance of general language that is distinguished by features such as genre, register (style), a domain terminology and also document structures ((Morato et al., 2003).

A scientific specialty is also always a discourse community. With the special focus on language use and specialized languages in this dissertation, the primary focus of a specialty in this context will be its discourse or its language use. I will use the terms specialty and discourse community interchangeably throughout this dissertation because of this focus.

3.5 Specialty Languages

Specialist communities are systems of communication, meaning that the exchange of ideas and results is one of the more important ingredients for a discipline or specialty to maintain existence. Communication “is in fact the only general scientific behavior shared by all scientists and scholars” (Lindholm-Romantschuk, 1998, 9).

The means of communication, language, is not only used to express ideas and concepts, but with the means of language communities draw boundaries between “outsiders” and “insiders” - people who can speak the language and people who do not. Language is therefore an important demarcation tool. Learning to move through the

discourse patterns of a community is one of the most important rituals that any newcomer to a community will go through.

The language of a specialty has certain features that go beyond just special terminology, even though specialized vocabulary plays a big part. Three different areas of research have studied specialty languages: (1) sociology of science in its attempt to understand how boundaries of disciplines and specialties are maintained and drawn; (2) computational linguistics and natural language processing in its attempt to represent language structures in a computer processing environment and (3) sociolinguistics and translation studies, which is interested in semantic and stylistic differences in different language areas.

Sociology of science and the two linguistics subfields study language from two opposite perspectives. Whereas the sociology of science studies language as a means of boundary work and interdisciplinary communication that is employed by the members of a specialty community; linguistics studies language as a means to identify and delineate disciplinary or specialty boundaries in communication artifacts.

3.5.1 Languages of Disciplines and Specialties

In his chapter about disciplinary boundaries in “Social Epistemology” (1988), Steve Fuller describes how a discipline (or also a specialty) is bounded by its procedures of adjudicating knowledge claims. Those procedures restrict (1) word usage in the disciplines, (2) borrowings of concepts and terminology permitted from other disciplines, and (3) the appropriate context for justification and discovery.

This is also an interesting explanation of how a specialty can restrict access to its core and how it can defend itself against intruders or other specialties trying to take jurisdiction over this specialty. According to Fuller, three techniques are sufficient for detecting disciplinary boundaries. First, one should look at the disciplines that cover apparently similar knowledge claims. In order to distinguish between these disciplines, one has to examine the different argumentation formats that are used to argue research problems. A second strategy is to look at the metascience implicit in a discipline's argumentation format. This way, one will detect which discipline wields cognitive power over other disciplines (i.e. from which discipline do other disciplines borrow concepts). The third strategy for detecting disciplinary boundaries is to examine the strategies used to synthesize the research of two or more disciplines. A metalanguage is often developed that reduces the claims of the two disciplines to an account that the synthesized audience will understand.

Some of Fuller's arguments also resonate in Abbott's "System of Professions" (1988), where he describes rhetorical strategies that are used to argue claims of one jurisdiction, profession or discipline over another. He calls these cognitive strategies that constitute basic attack mechanisms through which professions (disciplines) try to achieve and maintain power over other jurisdictions and defend their own fields against external and internal forces that try to intrude in theirs.

Reduction is an attacking move made by secure professions to show that some new task is principally reducible to the attacker's already secure discipline. It is often a redefinition of another profession's jurisdiction into the rhetoric structure of one's own.

Another strategy for an attacking discipline is to claim that its treatments (or methods) also apply to problems diagnosed by others. This is a strategy to absorb and incorporate other discipline's problems into one's own and therefore weakening the other's position of necessity of existence. This is comparable to the argumentation format of a discipline in Fuller's book. Metaphors extend one profession's model of inference to others. Although they are not as strong as the reduction strategy, metaphors promote one's own way of imagining problems and therefore impose a certain, advantageous viewing of a claim (similar to Fuller's metalanguage).

In a study on microphysics, Galison (1997) approaches the problem of specialty languages from a different viewpoint and analyzes how different specialties can communicate with each other despite different discourse patterns:

> "My question is […] how, given the extraordinary diversity of the participants in physics - cryogenic engineers, radio chemists, algebraic topologists, prototype tinkerers, computer wizards, quantum field theorists - they speak to each other at all. And the picture (to the extent one simplifies and flattens it) is one of different areas changing over time with complex border zones that sometimes vanish, coalesce, and even burgeon into quasi-autonomous regions in their own right." (Galison, 1997, 63)

This problem was already raised by Kuhn (1962) who questioned whether scientists from different disciplines (or even just different paradigms) could really understand each other and maintain communication (Lindholm-Romantschuk, 1998, 14).

The notion of a trading zone where two or more specialties meet and exchange ideas and concepts is a useful metaphor in this context. In the trading zone, "physicists and engineers are working out a powerful, locally understood language to coordinate

their actions" (Galison, 1997, 833). The language of a trading zone in that sense is an abstract boundary object (Star & Griesemer, 1989) between the specialties[2].

The notion of an "interlanguage" draws once again attention to the variety of "dialects" or languages that need to be straddled and understood in interdisciplinary communication. In chapter Five, I will suggest a way to make different specialty languages visible to the searcher of an information retrieval system and support him in navigating those different language spaces.

3.5.2 Sublanguages

Two areas in the discipline of linguistics deal with topically specialized vocabularies or languages: the area of sublanguages and controlled languages[3] research in computational linguistics (Kittredge & Lehrberger, 1982; Grishman & Kittredge, 1986; Kittredge 2003) and the area of languages for special purposes (LSP), which is further divided into an area devoted to issues of specialized translation and foreign language education and a broader area dedicated to socio-linguistic variation research (Adamzik, 2001).

---

[2] Star & Griesemer's boundary objects are defined as "an analytical concept of those scientific objects which both inhabit several interacting social worlds […] and satisfy the informational requirements of each of them. Boundary objects are objects which are both plastic enough to adapt to local needs and the constraints of the several parties employing them, yet robust enough to maintain a common identity across sites. They are weakly constructed in common use, and become strongly structured in individual-site use." (Star & Griesemer, 1989, 393)

[3] Controlled languages in this area of linguistics (http://www.controlled-language.org) are restricted versions of a natural language engineered for a special purpose, e.g. technical documentation writing (Kittredge, 2003). Controlled languages are utilized for clarifying, disambiguating and standardizing technical jargon and use a well-defined subset of the grammar and the lexicon. Contrary to sublanguages, which more or less naturally evolve, controlled languages are specifically designed and require skill and effort to produce and learn.

Another area in linguistics that deals with specialized vocabulary is the field of dialect studies, which analyzes language variations based on geographic areas. A primary research objective of geographic dialect studies is the production of dialect maps, which show differences in language in different geographic locations. The primary emphasis for dialect studies is the grammatical and syntactic features of the geographical dialects. The focus of this dissertation is to analyze subject-specific dialects and to show variations in language in different topical spaces. The fields of sublanguages and languages for special purposes with their emphasis on terminology are therefore related.

Describing a general natural language in terms of a formal grammar suitable for a natural language processing system is very difficult because of the complexity and variety of an every-day language. The notion of sublanguages is a useful concept in computational linguistics for defining restricted grammars that describe a clearly defined domain (or specialty). Compared to general language grammars, restricted grammars (sublanguage grammars) are more precise and more successful in covering all syntactic structures of their domain, making NLP applications more powerful. In a sublanguage, instances of homography (same spelling, different meaning) are rare, therefore ambiguity is reduced and automatic analysis is substantially easier (Lehrberger, 1986).

Sublanguages were first introduced by Harris (1968, 152) who used the term to distinguish language regions that differed from the general language syntactically or lexically. Hirschman and Sager refined the definition to describe the “particular language used in a body of texts dealing with a circumscribed subject area (often reports or articles of a technical specialty or science subfield), in which the authors of the documents share

a common vocabulary and common habits of word usage" (Hirschman & Sager, 1982, p. 28).

More generally, a sublanguage can be defined "as the language used by a particular community of speakers, say, those concerned with a particular subject matter or those engaged in a specialized occupation" (Bross et al., 1972). These experts communicate about the particular subject matter (restricted domain) in a recurrent situation, i.e. scientific papers in monthly journals (Kittredge, 2003).

It was suggested that texts, which share the same domain of discourse (i.e. about a particular topic) display similar semantic and lexical patterns (Marsh, 1986) whereas texts sharing the same function (e.g. weather forecasts, stock market summaries) display similar syntactic characteristics.

Lehrberger (1982) and Kittredge (2003) list properties of sublanguages that can be exploited in NLP applications: (1) a limited subject matter, (2) a restricted lexicon (including special words not used elsewhere), (3) a small number of distinct lexical classes, (4) a restricted sentence syntax, (5) a deviant sentence syntax, (6) restricted word co-occurrence patterns which reflect domain semantics, (7) a restricted and often deviant text grammar, (8) frequency of occurrence of words and syntax patterns different from the norm for the whole language (high frequency of certain constructions), and (9) the use of special symbols.

Sublanguages also contain general-language structures that are used in the discourse. An intermixing of sublanguage domains (i.e. political issues in a stock-market summary) can increase ambiguity.

Once a sublanguage or domain has been identified (i.e. documents in the domain space collected), detecting sublanguage patterns (e.g. lexical items or syntactic structures) is relatively easily done with statistical methods (Grishman, Nhan et al., 1984). For example, Bonzi (1990) describes a in study of four different scientific sublanguages how the syntactic patterns between soft and hard science disciplines differ.

Utilizing more restricted and therefore more easily processable sublanguages is thought to help in solving problems in machine translation and natural language processing (problems like word sense disambiguation and selection of synonyms) (Luckhardt, 2006). Two very successful applications (machine translation and text extraction) of early sublanguage grammars are the TAUM-MÉTÉO of the University of Montreal, which has been automatically translating weather forecasts from English to French since 1977, and the NYU Linguistic String Project that generated structured database entries of patient information from physicians' treatment and test results summaries (Friedman, 1986).

However, automatic discovery of sublanguages is still a difficult problem (Hirshman, 1986; Grishman, 2001). Sekine (1994) describes a technique of sublanguage detection whereby lexical information is used to find sublanguage clusters in a collection of newspaper articles. More sophisticated work is needed to distinguish linguistic features between sublanguages. Walker & Amsler (1986) use the Longman dictionary's subject codes attached to particular words to cluster NYT articles into particular sublanguage domains. Liddy et al. (1992) use the same technique to classify documents. Losee & Haas (1995) investigate term frequencies in different sublanguages for text classification.

Sublanguages research is the computationally applied version of sociology of science's demarcation work. In both knowledge fields, the delineation from one specialty to the other is the most difficult challenge.

3.5.3 Languages for Special Purposes

Languages for Special Purposes are languages that are used to discuss specialized fields of knowledge (Hoffmann; Kalverkaemper & Wiegand, 1997). They are often defined in contrast to a language for general purposes (LGP), which is the language "we use every day to talk about ordinary things in a variety of common situations" (Bowker & Pearson, 2002). The most distinguishing feature of languages for special purposes are their specialized vocabulary. However, they may also differ from LGPs in having special ways of combining terms or arranging information.

Whereas research in sublanguages seems to focus on syntactic and lexical features (simpler for computational analysis), LSP research directs its attention also to semantic aspects. A recent text (Mayer, 2001) distinguishes cognitive aspects, knowledge representation, terminology, didactics and LSP in academic discourse as major research areas in LSP. Evangelisti Allori (2001) looks at thematic patterning in three different disciplines and finds that different fields of knowledge vary in their application of discursive functions like contrast, manner, or specification.

Another research foray in LSP is the interest in culturally determined structures of LSP texts, for example, the difference in scientific writing style in English compared to other language areas (Baumann, 2001, Gotti, 2001). For example, it has been found that

the English representational style is more “linear” whereas German or other East-European languages favor a more “digressive” style (Kretzenbacher, 2001).

In summary, sociology of science looks at language as a demarcating tool, i.e. specialties use their own language to distinguish themselves from each other, to erect a barrier for outsiders but also to more effectively communicate because ambiguities of language will be decreased through specialized and defined terminology.

Sublanguages in computational linguistics look at language as a characterizing tool, i.e. certain features of language (mainly in written communication) can be used to characterize a field of discourse. Specialized features of sublanguages can then be leveraged to support natural language processing applications like machine translation and word sense disambiguation.

Languages for special purposes look at language as a stylistic tool, i.e. communities can be distinguished along stylistic and discursive lines. What is common to all three approaches is the focus on the languages within general language, which I will call dialects or specialty languages (sublanguages being reserved as a specialized term from a particular discourse community).

3.6 Domain Analysis in Information Science

Domain analysis is a research approach in information science that focuses on the domain (in the sense of a specialty, knowledge domain, discourse community, research subject, or interest area) as its primary unit of analysis (Hjørland, 2002). It argues for

designers of information systems to take the subject domain of a user into account and therefore consider the previous and more specific knowledge of a researcher in the system's responses:

> "The domain-analytic paradigm in information science (IS) states that the best way to understand information in IS is to study the knowledge-domains as thought or discourse communities, which are parts of society's division of labor. Knowledge organization, structure, co-operation patterns, language and communication forms, information systems, and relevance criteria are reflections of the objects of the work of these communities and their role in society. The individual person's psychology, knowledge, information needs, and subjective relevance criteria should be seen in this perspective." (Hjørland & Albrechtsen, 1995, p. 400).

After reviewing the literature in information science and its use of domain, Hjørland and Albrechtsen (1995) cite specialty, discipline, domain or environment as valid equivalents for domains as the units of study and draw the analytic space wide. Tennis (2003) lists a number of other concepts similar to domains studied in information science, among them communities of practice (Davenport & Hall, 2002), Subject Matter (Hjørland, 2001), Discourse Community (Hjørland & Albrechtsen, 1995), Context (Solomon, 2002), Situated Knowledge (Cool, 2001), and Position (Wilson, 1968).

Despite the terminological variety in domain definitions, coming from information science with its traditional focus on textual representations of communication, domains are in large part defined through their communicative activities:

> "The concept of domain cuts across formal definitions and boundaries and focuses on people's activities, collaboration, and common goals when placing documents in discourse communities. [...] The exact boundaries and composition of particular domains is determined through

> an analysis of the domain, focusing on establishing the structures, ontology, and communication patterns present in the domain, that is, the activities that take place, the circumstance under which they can take place, and the constrains imposed by paradigms and current research fronts“ (Mai, 2005, 606).

The main tools of domain analysis as proposed by Hjørland (2002) depend in large part on the communicative output of a community (i.e. domain) and rather less on other activities that domain members are involved in. The eleven approaches to study the different dimensions of a domain in question in information science are (1) literature guides and subject gateways, (2) special classifications, (3) research in indexing and retrieving specialties, (4) empirical user studies, (5) bibliometric studies, (6) historical studies, (7) document and genre studies, (8) epistemological and critical studies, (9) terminological studies, languages for special purposes, discourse studies, (10) studies in structures and institutions in scientific communication, and (11) domain analysis in professional cognition and artificial intelligence. Even though this list represents a subjective and restrictive approach to studying domains, this is the part of community studies that information science is particularly well positioned to analyze.

Even though these eleven approaches help us in describing characteristic features of a domain and domain members' particular practices for information production, information seeking and communication (see, for example Fry, 2006), they will not help to define the boundaries of a domain (which is the subject of the sociology of science) nor do they provide guidance on how to apply a domain-centered approach to the development of information retrieval systems. The new field of domain analysis does

however draw attention to the importance of domains (or communities or specialties) as one interpretive tool to understand people's information practices.

3.7 Conclusion

This dissertation with its focus on supporting domain-specific search behavior by drawing on specialty languages to suggest a better mapping between the dialect of a community and the language of an information retrieval system (see chapter Four) can be seen as an application of the domain-analytic research strand in information science. It goes beyond studying the features of domains to actually proposing an application that will leverage features of a domain (in particular language and subject-specific databases) to improve the search process.

For this purpose, I will define specialty (synonymous to domain and discourse community) as an area of research distinguished by a bounded group (though fuzzy and penetrable) of people (the researchers in the field) interested in a particular problem or concept. The specialty can be characterized by its (1) use of certain theories and methods, (2) particular forms of communication (gathering of findings in particular conferences and literature), and, especially, (3) by its use of a specialized language to describe phenomena in the field.

Price (1986, 65) claims that specialties are not larger than one or two hundred scientists because only a community of this size (or less) can keep up with its own research output. Such a community would have an approximate research output

(literature) between a hundred and five thousand papers during its lifetime (Morris, 2005). Being able to define a specialty in this way provides a starting point for the identification of specialties in document collections, but unfortunately, specialties are not clearly delineated in their initial stages of formation.

New specialties are formed and old specialties are resolved constantly, making it harder for the domain-centered information retrieval researcher to define his or her research field (see also chapter Eight on differentiating specialties in a collection). The three-stage life cycle theory of specialty formation (Mulkay, Gilbert et al. 1975) explains how initial communication between members of a new research community is spread over several channels and publication venues, before separate journals and conferences are founded. With the focus on the specialized language of a research field, it is advantageous to focus on the established specialties – not only because their vocabulary has stabilized and a core terminology has formed but also because we can more easily identify clusters of research output in the literature.

As outlined in chapter Two, the language mapping problem is a problem between the specialized terms used by the searcher and the vocabulary used by the information retrieval system to represent its documents. Whereas this chapter explained why the specialty is the ideal focus for a domain-centered approach (locus of research, communication output and specialized language), the next chapter will provide an overview of the language of information systems.

Chapter 4

The Role of Documentary Languages in Search

This dissertation is concerned with supporting better access to document collections and search by providing more clarity about the language of document representations (i.e. vocabulary used in document description and appropriate for search). In particular, the focus is on access to controlled vocabularies for document representations, because they provide an extra layer of difficulty for the user. Controlled vocabularies are commonly utilized in bibliographic databases to provide a precise and concise subject description of a document's content. Ideally, there is only one term or phrase in the controlled vocabulary to describe any given concept, which gives rise to an exacerbated language-mapping problem especially in non-full-text environments.

If there is only one term used to denote a concept, then the searcher needs to select exactly that term to find relevant documents or not find anything at all if only words from this vocabulary are used in the document representations. On the other hand, if that term is selected in search, then the chance of finding all the relevant documents concerned with the concept and only those is much higher than in a full-text situation where many different terms for the same concept can be used (and needed to be searched on to find all the relevant documents). In a non-full-text situation it is especially important to support the searcher in term selection, particularly when the documentary language is unfamiliar and additional learn effort is required to make that mapping between a question and a search statement.

This chapter is an introduction to documentary languages, of which controlled vocabularies are a subset. The term “documentary language” is used to describe all content-describing document representations but this dissertation will mostly focus on controlled vocabularies. The different types of controlled vocabularies are introduced and the benefits and weaknesses of controlled vocabularies versus free text and uncontrolled keywords will be discussed.

4.1 Bibliographic Representation

Documents that are organized in large collections need to be described in order to provide access to them. A bibliographic language is a special-purpose language to provide such a description. Bibliographic languages can be roughly divided into those

languages that describe the manifestation of a work and those languages that describe the content of the work. A formal bibliographic language is used to describe a work's manifestation (e.g. author, title, source, format, year etc.) and a subject or documentary language is used to describe a work's content (e.g. abstract, summary, keywords or classification codes). Like all languages, also bibliographic languages consist of a vocabulary, semantics, syntax and pragmatics:

> "The vocabulary of a bibliographic language consists of the simple and complex expressions used to name the values of […] entities, attributes, and relationships. Its semantics consists of the relationships among these names […]. Its syntax consists of the ordering relationships among the component elements of complex expressions in the language. Its pragmatics consists of specifications and conditions for the application of the language […]." (Svenonius, 2000, 55)

The vocabulary of a bibliographic language is called metadata - denoting the descriptive function of the bibliographic language as "data about data" - and the semantics, syntax and pragmatics are determined in metadata standards and other more or less sophisticated rule systems. Prominent examples of widely used standards are the International Standard Bibliographic Description, the ISO Standard for Thesaurus Construction (ISO 2788), the Anglo-American Cataloguing Rules, the Library of Congress Subject Cataloging Manual, the Encoded Archival Description (EAD), the Dublin Core or the Resource Description Framework (RDF). For formal bibliographic languages, the vocabulary is often regulated by so-called authority files, determining the correct way to present personal and corporate names, location or resource information.

Documentary languages are too many and to varied to be enumerated here and differ widely between different discourse areas and purposes. Prominent examples from the library world include the Dewey Decimal Classification (DDC), the Library of Congress Classification (LCC), the Library of Congress Subject Headings (LCSH), the Medical Subject Headings (MeSH) and the Getty Arts and Architecture Thesaurus (AAT).

In an information system, formal descriptions are searched in so-called known-item searches, when the author, publisher, year or source is known[1]. Subject metadata searching is more of an imprecise search, where the item is not known and the concept that is looked for needs to be described. Subject searches probably account for more than half of all searches in an information retrieval system but constitute the more difficult half for the searcher[2].

4.2 Documentary Languages

Documentary languages are the content-describing subset of bibliographic languages. They are the vocabulary of the document collection and therefore the source of the language ambiguity problem discussed in chapter Two. Documentary languages

[1] A known-item search does not guarantee search success, because spelling errors or variations, erroneous entries in the information system, or wrong source information can mislead the searcher and cause a mismatch or no match in the information system.

[2] In 1983, a nationwide American library survey found that subject or topical searches accounted for up to 59% of the searching in library catalogs (Matthews, 1983). At the same time, subject searches were reported to be the most problematic search type for the users (43% claimed to have difficulties in formulating a subject search).

vary in vocabulary restrictiveness (from strongly controlled to not controlled terms), size (from a few terms per document representation to full-text) and language type (alphabetic words or classification codes) but also in syntactic, semantic and pragmatic rules (how they are structured in a document representation, how vocabulary terms are related to each other and how they are selected for document representation).

For a minimal content description, a document's title and authors (sometimes even publisher and publication location) can still give some insight into the topical content of the document. However, early on in bibliographic description research, content description with only title words was contested because title words may not truly express the subject of the work they name and also because works describing the same subject with different title words would not be found together in search (Cutter, 1876). More controlled documentary languages were invented.

One can distinguish several types of documentary languages: hierarchical and faceted classifications, thesauri, subject headings, keywords or tags and natural language full text. In this order, the documentary languages go from very artificial and very controlled to natural and uncontrolled in their vocabulary; and from very restricted in their rules system to no rule system at all (except for the rules of natural language grammar).

4.2.1 Classifications

Classifications try to organize a knowledge area into distinct, mutually exclusive and jointly exhaustive categories. Hierarchical classifications arrange their knowledge

area from broad to most specific, whereas faceted classifications arrange it by orthogonal facets each representing one aspect of a topic. Subjects and subject areas are usually denoted by classification codes consisting of numbers and characters put together in a meaningful way. A subject index is added to a classification to map classification codes to natural language descriptors.

4.2.2 Thesauri & Subject Headings

Thesauri are a form of controlled vocabulary that represents so-called descriptor terms in their hierarchical, equivalent and associational relationships to each other. Non-descriptor terms function as see-references to descriptors. Compared to classifications, thesauri do not attempt to define mutually exclusive categories and have more information on synonymous and related terms. Thesauri commonly use natural language terms for descriptors.

Subject headings are subject describing terms from a controlled list that are assigned to the record to specify the subject of the document. Library catalogs usually have subject headings assigned to their catalog records. Compared to thesauri and classifications, subject heading lists are usually less hierarchically structured, contain a bigger vocabulary and have rules how to pre-coordinate the terms (ordering) in a document representation.

Classifications, thesauri and subject headings are controlled vocabularies. Vocabulary can be controlled in several ways (see also Svenonius, 1986; Bates, 1988): by clustering orthographic variations of a given word under one spelling; by clustering

lexical variants and related word forms under one word form (e.g. singular vs. plural); by regulating word order; by clustering synonymous words under one descriptor; and by disambiguating homographs (through embedding in a hierarchical structure, picking non-homonymous terms for the concept as descriptor, providing scope notes, or adding terms to the homonym descriptor).

Controlled vocabularies bundle documents about the same topic or concept under one concept term, phrase or number. This one term or code is the designated descriptor term or number for a concept. All other terms are non-preferred descriptors and only serve to “lead” to the descriptor. If the lead-in terms are not made available for searching, then only few access points (a known descriptor or code) will lead to relevant documents.

4.2.3 Keywords and Tags

Keywords belong to the group of non-controlled documentary languages. Any natural language words or phrases can be chosen to describe a document’s content. These terms can come from the titles and abstracts of the documents or from a list of terms used before or are chosen at random. The size of the vocabulary for keywords is naturally larger than for classifications or other controlled vocabularies, and so is the number of terms and term variations available to describe the same concept. Keywords provide more freedom and flexibility in word choice than controlled vocabularies and require no initial effort for their creation.

The social tagging systems prevalent in the Web environment like Flickr (photos), Del.icio.us (links) and Last.fm (music) use keywords (called tags) to describe the items

they organize. For these examples, tags are an important subject-describing feature because photos, links or music files are generally not self-explanatory or easily searchable document representations in a text-controlled environment. More interestingly, text-based web services like blogs also started using tags (keywords) to organize documents and describe the content of an item. Tags or blog categories are user-selected terms to represent a blog post. Blog software and blog search engines like Technorati then use tags (instead of the full-text of a blog entry) to cluster and organize similar items. The organization of blog entries with tags or categories is easier than using the full-text of an entry because the most important words and categories have been manually pre-selected and do not need to be painfully extracted from the full-text representation. Even though a controlled vocabulary will probably not be implemented in this environment (Shirky, 2005), there is a trend for individual bloggers to keep the number of their categories small and overseeable.

4.2.4 Full Text

The full text of a document can be regarded as the extreme end of the documentary language spectrum (title and author information being on the other). The full text is a non-controlled, natural-language-based documentary language. Using the full text of a document for its subject representation requires no effort for description; it provides, however, a rather unfocused view of the document's content.

Both controlled vocabularies and non-controlled keywords / tags fulfill a summarizing and categorizing function for document representations that is not easily

reproduced with the full-text of the document and automatic means. As one of the fundamental cognitive mechanisms, categorization reduces the load on memory and simplifies a person's experience with the world (Lakoff, 1987; Jacob, 2004). The facilitation of efficient information storage and organization is one of the reasons for the continued existence of controlled documentary languages side by side with full-text representations. In large collections (and most collections are very large today), the clustering function of controlled vocabularies helps to divide up the search space and provides more certainty in the overall effectiveness of the search. Finding relevant (hopefully all relevant and only relevant) documents is more likely when the searcher has an idea of the space that he is moving in. Organizing that space into subject clusters is an obvious way to provide more structure and relieves the cognitive load of the searcher.

4.3 Purpose of Documentary Languages in Information Retrieval Systems

Documentary languages and especially controlled vocabularies serve several important purposes in information retrieval systems:

(1) they provide a concise topical description of the document (describe what the document is about);
(2) they provide a non-ambiguous term for each concept represented in the collection, that is, they control the subject vocabulary;
(3) they provide an organization scheme for the documents in the collection;
(4) they cluster all relevant documents for a concept under one term;
(5) they provide more searchable text for the user (especially in databases with sparse text in their records like library catalogs); and

(6) they aid retrieval by providing a topical access point that is unambiguous and retrieves a complete and precise document set for a given concept.

Controlled vocabularies as documentary language address several ambiguity problems that occur in natural language retrieval: spelling and lexical variations, homonym and synonym ambiguities and also the complexities and subtleties of meaning variations of related terms (Lansdale & Ormerod, 1994).

Documentary languages not only enable search and retrieval of documents, but they also “facilitate understanding, influence identity, and claim authority” (Morville, 2005, ch. 6). By deciding the document representation’s vocabulary, documentary languages - and controlled vocabularies in particular - determine the way the content of a collection will be viewed (or searched on). They are therefore inherently political, present a subjective view and might disadvantage different perspectives on a knowledge area or discourse (Bowker & Star, 1999; Cornelius, 1996; Hjørland & Albrechtsen, 1999). In a corporate world, product classifications and enterprise vocabularies are used for branding, marketing strategies and support competitive advantages. In a research / academic world, controlled vocabularies can shape and define disciplinary boundaries in document collections and fix subject area vocabularies.

4.4 Human vs. Documentary Categorization

As formal organizational tool, documentary languages represent the vocabulary and knowledge structure of a discourse community or specialty in the collection. For

effective search and retrieval, ideally the organization of the documentary language should reflect the cognitive organization of the knowledge community by its members (or at least the individual's using the document collection). If an individual's cognitive organization of a discourse and the formal organization of the collection do not match, the ultimate language mapping problem occurs. To evaluate a documentary language's effectiveness, these questions need to be asked:

> "(i) Is communication between the information system and the individual influenced by the representation of resources? (ii) Does the organizational structure of the information system cause the individual to adjust her internal cognitive structures? (iii) Does the organization of resources contribute to the creation of meaningful context for information? (iv) Is the meaning of information influenced by the organizational structure of the information system? And (v) What consequences follow from the different organizational structures that can be applied to a collection of information resources?" (Jacobs, 2004, 518)

Human categorization in daily life is flexible and dynamic, with fuzzy and changing boundaries and context-dependent, non-exclusive membership. Documentary categorization (i.e. documentary languages) can be relatively static and permanent, and - depending on the type of documentary language – has clearly delineated albeit arbitrary boundaries and definite criteria for membership. For the types of documentary languages introduced here, free-text keywords or tags are closest to human categorization in their flexibility and non-rigid boundaries, followed by subject headings, thesauri and then classifications.

Being closest to human categorization behavior does not necessarily mean that the documentary language is more effective in the organization of an information system's

documents. The very nature of human categorization – individualistic, ephemeral, context-dependent and dynamic – makes it difficult to establish a more persistent organization that spans both time and space and individual cognitive differences. Because an information system serves many users, its structure is necessarily broader, not as dynamic and tries to reach a wider audience. If the categorization is too flexible, long-term meaningful relationships between concepts cannot be established and the organization of a document collection is not so effective (Jacob, 2004). On the other hand, documentary languages face the problem of pre-establishing knowledge structures and having to help searchers to map from the system's structure to the searcher's cognitive landscape. If this mapping is successful, then more rigid documentary languages are the more effective organization system and should serve as better retrieval tools.

4.5 The Controlled Vocabulary vs. Free Text Debate

The ultimate evaluation measure for different documentary languages must be the effectiveness of the search or browsing process using that language (Dubois, 1987) and so every documentary language's ultimate test bed is the user experience. The debate over the retrieval effectiveness of controlled vocabulary versus free or full text document representations has been going on for some time.

When computer processing was expensive and storage restricted, controlled documentary languages providing only a few terms per document as content description

were thought to solve the constraints by supporting the ranking and matching process effectively and providing only a few but high-quality access points. In the 1960s and 1970s, full-text documents were first stored in bibliographic databases and “free-text” keywords as document representations were promoted. The subject of effective document representations had to be reopened.

The Cranfield II studies (Cleverdon, 1970) were one of the first of a series of retrieval effectiveness evaluations that showed that minimally controlled documentary languages (spelling variations and minimal synonym control) performed as well as fully controlled vocabularies in retrieval. However, they were criticized for their flawed evaluation methodology (Svenonius, 1986). Since then, many studies have been published proving either natural language’s or a controlled vocabulary’s superiority for retrieval performance depending on the document collection, type of queries, evaluation methodology and measurements (for a review, see Rowley, 1994). The only general consensus achieved seems to be that having both natural language (free-text keywords, abstracts) and a controlled vocabulary is the optimal solution providing the benefits of both search and document presentation strategies.

Much has been written about the relative advantages and disadvantages of natural language and controlled vocabularies (for good overviews, see Svenonius, 1986; Dubois, 1987; Rowley, 1994 and Aitchison, Gilchrist & Bawden, 2000). The most important arguments for each document representation strategy are:

Full-text or uncontrolled natural language-based keywords:

(1) create lower costs for producing document representations;

(2) enable simplified searching (no language mapping to a controlled vocabulary necessary);
(3) make either the whole document text or many keywords available for search;
(4) make every word equally retrievable (no preferred terms that have to be looked up);
(5) will not have a misrepresentation of content through human indexing errors because author words are used in search (wrong controlled vocabulary terms assigned);
(6) are more up-to-date: new terms that occur in documents are immediately available in search;
(7) support search for terms that might not occur in a controlled vocabulary but can be very important, e.g. geographic areas, recent topics, specific named objects, value judgments, actions, and individual or psychological characteristics (Markey, K., P. Atherton, et al. 1982); and
(8) allow for easy aggregation of document collections because document representations are in the same documentary language (aggregating collections with two different controlled vocabularies requires a mapping between these vocabularies).

However, they also:

(1) place a greater burden on the searcher to select all relevant terms for search because they do not link synonyms and term variations;
(2) are prone to retrieval failure when concepts that are represented in the documents but are in a different vocabulary than the searcher expects;
(3) hide retrieval failure because in large document collections any search statement will produce results;
(4) make ranking more difficult, because more features in documents need to be considered;
(5) make evaluating result sets more difficult because more documents will be retrieved that are not relevant; and
(6) provide no hierarchical organization of the knowledge area in the database, so browsing can be a problem.

In contrast, controlled vocabularies:

(1) improve the language ambiguity problem by controlling synonyms and near-synonyms, qualifying homographs and providing scope notes to explain concepts;
(2) shift some of the burden of search to the system because the searcher only has to find the correct controlled vocabulary term to find all relevant documents;

(3) provide high-precision search capabilities because only documents concerned with a particular concept will be indexed with it;
(4) are more helpful for exhaustive searching;
(5) provide broader, narrower and related search terms by virtue of their own structure and syntax;
(6) help express concepts in terms that are not used in the documents themselves and can therefore provide more or different vocabulary to search on;
(7) are easier for ranking algorithms because every term is relevant and important;
(8) provide a map of the knowledge areas represented in the document collection; and
(9) support browsing a document collection through their structuring.

However, controlled vocabularies also:

(1) create very high costs for document representations and vocabulary maintenance;
(2) can not keep up-to-date with very new terms in a document collection and very unstable fields that can not be organized in a static way;
(3) put an additional learning effort on the searcher to find the correct controlled vocabulary term to search on;
(4) can introduce indexing errors because the content of a document can be misinterpreted by an indexer;
(5) might present a knowledge area differently from a community of discourse's practice and therefore create a cognitive mismatch in search; and
(6) are also prone to language consistency problems (see chapter Two).

Free-text terms in search improve precision if they are new or very specific so that they would not occur in the controlled vocabulary. A controlled vocabulary, however, will improve both recall - when all the relevant documents for a concept are indexed with the same controlled vocabulary term - and precision, when the correct controlled vocabulary term is picked. However, the trade-off is between the cost of the searcher to find the correct term (and also the cost of maintaining a controlled vocabulary) and the findability of documents without a controlled vocabulary.

The library catalog is one example where this trade-off has been tested. Libraries invest a lot of effort into expanding catalog records with subject headings intended to improve findability of these records. The argument is that catalog records only contain few keywords to search on (i.e. title, author, publisher and eventually notes) so that a controlled vocabulary is necessary to add content-describing terms to the search space. If controlled vocabulary terms are not added to library catalog records, are they still as accessible?

Voorbij (1998) showed in his study of social science and humanities catalog records from the National Library of the Netherlands that 37% of the catalog records were considerably (and an additional 12% slightly) enhanced by the addition of subject headings (that is new words other than title words where added to the search space and they were more specific or less specific than title words). In a second study, title keyword searches and searches with controlled vocabulary terms were compared. The recall for controlled vocabulary searches almost doubled the recall for title searches (86.9% compared to 48.2%), meaning that almost half of the books would not have been found if it had not been for the added terms from the vocabulary.

In a similar study of transaction log data of 186 catalog searches (Gross & Taylor, 2005), it was found that the average proportion of records that would not be retrieved in the absence of subject heading data was 35.9%. For a good third of the sample searches, the recall would have dropped under 50% of what had been retrieved with a controlled vocabulary search. For non-English language materials, where subject headings might be

the only English search terms to search on in the record, this proportion was drastically higher and reached a 100% in some of the searches.

The authors of both studies conclude that subject headings are still essential. However, how controlled the vocabulary really needs to be and whether added keywords are necessary when enough text is available cannot be answered by these studies. The popularity of tagging systems on the internet (a collection with more text than any other) might provide an answer.

4.6 Folksonomies

"Folksonomies" (free-text web-based social tagging systems[3]) are dynamic, non-hierarchical or faceted structures, impose no long-term organization on the items they organize and reflect individual's cognitive structures. Why do they work? Precisely because they do not require an effective organization of the knowledge space that they are working in – this would be impossible since we are talking about the web, all areas of interest and a potentially unlimited numbers and variety of users.

Because users' tags primarily reflect their own cognitive categorization, they work well for the individual. But because there are so many users inputting keywords into the system, the aggregation of items and keywords also creates a powerful resource for the community as a whole. In a sense, the outcome is social but the input process is

[3] „So the user-created bottom-up categorical structure development with an emergent thesaurus would become a Folksonomy?" Posted by Thomas Vander Wal on the members mailing list of the Information Architecture Institute on July 24, 2004. Quoted in Morville (2005, ch. 6).

not really collaborative. The collaborative aspect of these systems (be it collections of references in citeulike.com, collections of urls in del.icio.us, playing lists in last.fm or photos in Flickr) is created by the aggregation and clustering of items and offering indefinite browsing opportunities.

There is no claim to provide a coherent organization to the discourse, not even small specialty discourses are tagged with the same keywords and the language ambiguity problem is wide-spread[4]. New terms are introduced and old terms become obsolete all the time. Because of the ever-changing vocabulary and a searching behavior that follows trends in term selection, older items in the collection will be harder to find even if they are still relevant resources. New users of social tagging systems start out with very broad keywords or tags (Golder & Huberman, 2006), creating categories that are too general and contain too many items to be useful for organization or retrieval. Re-tagging is too painful and requires too much effort from users to be done often and so resources are lost in the tag mix of broad terms.

However, for any given concept, there are still so many resources tagged with the same keyword because of the sheer size of the collection that any search produces results. Contrary to subject-specific bibliographic databases (primary creators of more structured documentary languages), the aim for creators and searchers alike is not to provide all and only (or even just the most) relevant results to the searcher. Serendipitous browsing is required and encouraged by dynamically linking the keywords. Also, URLs and

[4] Mathes (2004) gives a great example for del.icio.us, where the tag „ANT“ will lead to links about Actor-Network-Theory in the field of science and technology studies and Apache Ant, a project building tool in the Java programming language.

references in particular provide the perfect anchoring for aggregation purposes, since they can be used as pivots to identify all keywords associated with an item. For presentation and search, keywords that are linked to the same items are clustered in tag bundles and so minimal vocabulary control can be provided.

Are documentary languages obsolete because of the popularity of the uncontrolled social tagging systems? After all, folksonomies are a kind of documentary language and they might be the best for the purpose they serve: exploratory searches over very large and very varied collections. The merits of documentary languages (and the aggregation of resources with these languages) have been recognized by a large community of users and exist once again side-by-side to full-text search. Over time, the users of social tagging systems have become more sophisticated, introducing tag clusters for minimal vocabulary control, slight hierarchical ordering and using more specific terms for item description so that categories are small and useful for retrieval.

For other collections and other purposes, more controlled and structured documentary language can provide better access for the user, especially if precise and exhaustive result sets are required. Cost-benefit analyses based on user requirements and collection specificity and permanence will decide what type of documentary language should be used for document representation but for many information systems (especially with media other than text), documentary languages are more popular than ever.

4.7 Subject-specific Documentary Languages

Subject-specific documentary languages fulfill a special role in information retrieval systems. As document representations, they encapsulate the vocabulary space for the document collection and provide an organizational structure to it. If they represent the vocabulary of the collection, then they also determine the vocabulary for search and therefore require the user to use that vocabulary if he wants to access the information system. In effect, subject-specific documentary languages are based on the vocabulary of a subject but they also shape that vocabulary in return. For novice members of a research community, subject-specific vocabularies can present a barrier to entry into that field. Documentary languages can help the learning process by introducing not only the vocabulary but also the relationships between individual terms and concepts to the user.

For effective searching, a searcher's cognitive structure of his field must be mapped to the documentary language's structure of the field. If the documentary language is well designed, then it will reflect a knowledge field's conceptual structure and terminology. For experienced users, the mapping process between his concept structure and the documentary language used in the information system should be easier. For a novice, the documentary language can support the alignment process between his cognitive structure and the knowledge structure of the field.

How important is controlling the documentary language for subject-specific information systems? A specific area of discourse will already have a specialized vocabulary (see chapter Three) and the language ambiguity problem could be lessened

for search in specialized databases. The specificity of terminology differs between disciplines and knowledge areas, for example, “the sciences seem not to have as many ways of saying the same thing as do the social sciences and humanities” (Taylor, 1995, 486), requiring possibly a stronger vocabulary control for the social sciences and humanities. Obviously, because of the unlimited semiosis problem seemingly stronger in those areas, creating a good controlled vocabulary that adequately reflects language use in the discourse community becomes so much more difficult.

Bhattacharyya (1974) defines the terminology consistency of a discipline as the stability between a concept in the disciplines and the associated terms referring to it:

> “An operational definition of terminological consistency can be given in terms of the following criteria:
> (i) the degree of control exercised by the organized scientific community in various ways over the use of established terms or in the coinage of new ones.
> (ii) the degree of conformity displayed by individual scientists in relation to any written or customary code of usage of terms.” (Bhattacharyya, 1974)

The more terms refer to a concept (averaged over individual concepts in the discipline), the more unstable and terminological inconsistent is the discipline. It was found that disciplinary vocabularies emphasizing concrete phenomena were more consistent than vocabularies emphasizing abstract phenomena (Bonzi, 1984), confirming the thesis that the sciences will have less synonym control problems as maybe the social sciences or humanities. This result is not surprising if one considers the fact that more abstract phenomena will have fuzzier boundaries as concepts overall and are therefore prone to more descriptive variation.

For all disciplines or specialties, subject-specific information systems already invest a greater effort into organizing their affiliated knowledge areas by delineating the boundaries of the field (which documents are in the collection and which are not). A controlled vocabulary can provide an extra boost for precise and especially exhaustive (high-recall) searching. Depending on the terminological consistency of a field, synonym (and other terminological variations) control can be very important. A controlled vocabulary also provides better browsing capabilities by mapping out the field.

Even subject-specific document collections usually encompass several specialties (or even disciplines). Different specialties can use similar concepts but use different terms in their vocabulary to express them or they can use similar terms for slightly different concepts. The documentary language's biggest function is to disambiguate these terminological confusions and make the relationships between terms and concepts between different specialties apparent. In search, a documentary language needs to be able to distinguish between these specialties and their specific vocabulary dialects to provide focused results.

Employing a controlled vocabulary for document representations can be utilized not only to provide terminological search support but also to provide cognitive mapping support by presenting the imposed structure on the information collection to the user.

4.8 Conclusion

No matter how controlled, documentary languages determine the vocabulary of search. If the vocabulary is controlled and restricted, the mapping between a query statement and relevant documents becomes easier – the burden of vocabulary control is on the system and happens before a search is started. However, the burden of selecting the correct controlled vocabulary term is on the user. If the vocabulary is not controlled, then the searcher's term selection process does not need to involve the learning or understanding of the controlled vocabulary. However, for effective searching, the mapping between a user's question and the search statement and finally the relevant documents becomes an issue of vocabulary control: the searcher now has to find the correct search terms to describe the relevant documents. The burden of vocabulary control is on the searcher and has to happen during the search process. For high-precision, high-recall searching, a controlled vocabulary is therefore recommended:

> "If one just wants to find something regardless of source or quality, keyword will probably do it. However, if one is a researcher looking for the best on a subject or everything on a subject, keyword is quite chancy." (Taylor, 1995, 486)

From a language mapping perspective, controlled vocabularies decrease language ambiguity in a system by determining the vocabulary and the relationship between concepts. In a dynamic language environment, where the meaning of concepts changes and context and viewpoint are important for understanding the organizational landscape of a knowledge field, controlled vocabularies function as boundary objects between

different vocabularies and viewpoints and can provide support in navigating a large information system. On the other hand, this stability in vocabulary control goes hand in hand with inflexibility towards individual users' cognitive organization.

One of the main tenets of this dissertation is that information systems need to provide support for this mapping between individual's vocabularies and the documentary language by going inside out. Instead of just using the documentary language for document representation and leaving term selection to the searcher in the hope that they will use the documentary language; an information system should use the documentary language to initiate an interaction between the user and the system. From the information system's side, the documentary language needs to be presented in such a way that the organizational structure of the collection becomes transparent and vocabulary selection becomes easier.

From a pragmatic standpoint, the first function of a controlled vocabulary in search should be to suggest search terms, show related concepts and provide an overview of the required specificity of terminology for effective search. It is easier for a human to select good search terms that match his interest from a suggested list than to come up with this list in the first place. Anchoring bias and unlimited semiosis in thinking up terms can be alleviated by a controlled vocabulary's power to present search terms in an organized search space and providing its language mapping capability to the user.

If the information system (with the help of the controlled vocabulary) manages to disambiguate between different specialized vocabulary spaces within the information system and represents the knowledge fields in the document collection adequately, it can

aid the searcher in navigating this space effectively. Then the search experience should be truly satisfactory - in terms of search speed, ease of use of the system and the conviction that all relevant documents were found and the right portion of the document collection was searched. The following chapter will introduce a technique where the controlled vocabulary of a document collection is used to provide this type of navigational help to the searcher.

Chapter 5

Translating Specialty Languages into Documentary Languages

The language mapping problem in information retrieval is a translation problem between the searcher's terms and the terms with which documents are represented in the information system. In this chapter, I will introduce a vocabulary support system that will aid the searcher in the translation process by suggesting search terms based on the specialty language of the searcher (gleaned from the searcher input and the specialty language in the information collection) and the documentary language of the system.

Translation of any kind depends a lot on the nature of the interaction – be it between speakers of different languages or between a searcher and an information system. A correct translation requires that terms and phrases be recognized in their particular meaning according to the situation, that is, the meaning of terms needs to be disambiguated with respect to the context in which the utterance occurs. We need to understand the particular language game a user is engaged in (i.e. style of language, meaning of concepts, and availability of terminology) to be able to successfully map between user terms and documentary language terms.

The hypothesis of this dissertation is that the context needed for meaning disambiguation can be approximated by taking the specialty language of the user into account and identifying his area of interest in the document collection. It is grounded on the assumption that document collections can be subdivided into specialty areas based on differences in vocabulary to provide a more accurate mapping between user search terms and the information collection's vocabulary. Dividing up a document collection by specialty or discourse and leveraging the specialty vocabulary to support searching is intended to provide the user with an understanding of the information space he is moving in and enough context to achieve a more successful mapping between his information need and the information system's resources.

This notion of context differs from conventional ideas in the information seeking and retrieval literature, which will be briefly presented in the first section of this chapter. The second section will introduce a search term recommender system that provides vocabulary help (translation) to the searcher on the basis of specialty languages /

documentary languages mappings and specialty spaces in document collections. The third section integrates the search term recommendation system with the larger field of vocabulary support systems and will provide examples of other term suggestion and language mapping applications.

5.1 Context in Information Science

Context is a fashionable notion in information science. Relevance and user satisfaction, the yardsticks of information retrieval system evaluation, are not independently measurable variables but depend on individual user's attitudes and preferences (Schamber, 1994; Mizzaro, 1997). Individual user behavior is guided by the user's context, which ultimately determines search effectiveness and user satisfaction levels. Consequently, effective information systems need to take a users' contexts into account. But what exactly is context and how can an information system take it into account in order to make the search experience more effective and satisfactory?

5.1.1 Definition of Context

Unfortunately, context is an excellent example of language ambiguity: "there is no term that is more often used, less often defined, and when defined defined so variously" (Dervin, 1997, 14). In the literature, context is described in two varieties: context as an independent entity surrounding a person and context as an integrated part of a person's activities, which gives them meaning.

The first perspective on context has been called representational or objectified (Dourish, 2004; Talja et al., 1999) and assumes that context is (1) a form of information, (2) delineable, (3) relatively stable and (4) separate from the activity (Dourish, 2004, 4-5). If context is an attribute or feature of some sort, then the concern of the information system designer has to lie with the encoding and processing of these contextual features in the information system. However, "virtually every possible attribute of person, culture, situation, behavior, organization, or structure has been defined as context" (Dervin, 1997, 14), which requires a careful and deliberate selection of the contextual features that should be considered for the design of an information system. Depending on the application, these could be cognitive or social factors related to a person's task, goals and intentions (Cool & Spink, 2002, 605) or environmental factors characterizing the situation (e.g. location, identity, activity, time) of the person (Dey & Abowd, 2000).

The second perspective on context has been called interactional or interpretive (Dourish, 2004; Talja et al., 1999). It sees context as inextricable from the activity and as the carrier of meaning in human behavior. In this view, context is (1) a relational property between objects and activities, (2) defined dynamically, (3) relevant to particular settings and (4) actively produced from and within the activity (Dourish, 2004, 5). For information systems design, the question is no longer how to program context into the systems responses, but how the system can support the negotiation of context within an activity like search.

Cool distinguishes between 'context' as the framework of meaning and 'situation' as the environment where "interpretive processes unfold, become ratified, change, and

solidify" (Cool, 2001, 11). The objectified view of context describes the situation a person is in, whereas the interpretive view of context describes the mutually constituted relationship between a person and the activity.

5.1.2 The Objectified Notion of Context in Information Science

The objectified notion of context appears in four different levels of research in information science (Cool & Spink, 2002, 606-607):

> (1) the information environment level: where context is seen as the information environment (e.g. institutional, organizational or work task settings), in which information behavior (especially information seeking) takes place;
> (2) the information seeking level: where context is seen as the goal or task a person is trying to achieve and which influences the searching behavior;
> (3) the information retrieval interaction level: where context is seen as user characteristics within search sessions and after which query feedback, relevance judgments and search strategies should be tailored; and
> (4) the query level: where context is seen as the discursive background that determines the true meaning of a query and which needs to be better represented to the system in orders to disambiguate query senses.

At all four levels, context is described as something the user brings to the system and which the system needs to incorporate in order to provide appropriate responses. Given the minimal input an information system normally receives from a user – one or few terms constituting a query – identifying the context of the user on any level is a intractable[1] matter:

> "A particularly difficult problem for IR systems is that of how to understand and to represent the salient aspects of a person's problematic

[1] Wilson (1983) argues that a person might not even understand his own context completely, let alone anyone else's. An information system adds another level of possible interpretive errors.

> situation based on queries that are entered into the IR system." (Cool, 2001, 15)

The objectified notion of context creates some difficult research challenges. How can a user's goal, attitudes and preferences be solicited and incorporated in the search process? From a language mapping view, how can the vocabulary in the search process be adapted to accommodate context?

In a "reference interview", the human intermediary between the documents in the information system and the searcher's information need has an unlimited flexible capability not only taking the searcher's own terms into account and changing them to accommodate different information sources and documentary languages as needed but also to pay attention to other influencing factors that will most likely change the outcome of the search. Knowledge of these other factors, such as the purpose of the search, expected outcomes and previous knowledge of the searcher will help the intermediary tailor and modify the direction of the search and harvest the information system's resources in a more effective way. Better service and results for the searcher arise by virtue of the intermediary knowing the searcher's background and the information system's documentary language, document content, and access structure. The human intermediary becomes context-aware with respect to a particular user.

Considerable research in the areas of user interfaces, human-computer interaction, and information needs and behavior has gone into making the information system more "open", more "context-aware", and more perceptive to multiple variables. A common approach to make an information system context-aware is to build up a searcher's history

and / or a user profile. Search histories and user profiles store a searcher's previous transactions (for an information system, that could be search terms, click-throughs for result sets, number of downloaded documents etc.) with the goal of getting an overview of the user's context: which topics he is interested in, what documents he has already seen, how many results he is willing to look at. More complicated systems try to build up a user profile through pre-transaction questionnaires or rating systems. Based on this profile or history, the information system can show appropriate access point into the content of the database, suggest new or related search terms and show not previously seen documents.

Aside from the fact that the more complicated systems using questionnaires or ratings require considerable time and effort from the searcher to be effective – something that people are unwilling to do when they can type in one or a couple of words into Google and almost always find the amount of information that satisfies them – there are other considerations that make user profiles or histories seem less than ideal when attempting to provide "context".

Even if the information system is capable of building up a history, it will only represent a "true" user profile for a very short time. People's interests and knowledge not only change over time but searchers will often also pursue more than one topic at any given moment in time. Trying to build up a coherent user profile from conflicting search requests representing different search topics is difficult, especially since it is almost impossible for the system to distinguish which search requests belong to which topic.

The matter is even further complicated by the fact that a user will not only query one information system. In a world where searching multiple information systems is the norm, it is unproductive to build a user profile from the queries sent to one information system alone when the "context" and accumulated knowledge of the searcher can change outside of the system.

Even if the system can detect any of these influencing factors and create a context-aware profile of the user, how will it adapt to the user? Most context-aware systems will follow pre-determined rules that prescribe user pathways through a system based on stereotypical user profiles. These attempts, however valuable, can only go so far. A "context-aware" system in this sense is still only a system that has been programmed to expect specific environmental variables, i.e. predetermined solutions that somebody has foreseen. Real context-awareness, as human intermediaries achieve, requires a flexibility and capability to dynamically adapt to unpredicted circumstances that computer systems have not (yet) developed.

5.1.3 The Interpretive Notion of Context in Information Science: User Context and System Context

The interpretive notion of context provides another perspective on the problem of incorporating context in the search process. Information systems researchers have to understand that there is "a marked difference between merely incorporating context and being contextual" (Dervin, 1997, 30). If one cannot fully incorporate a user's attitudes, goals, and preferences, one can try to make the information system more adaptive in a

way that will help the searcher without pre-scripted (and invariably stereotypical) system responses. If the information flow is not one-directional (from the user to the system) but mutual and interactive, then context becomes co-constructed and the search process may be more effective and satisfactory.

A context-sensitive information retrieval system should be flexible enough to allow the user to build up context during the search interaction. Commonly, an information retrieval system is a black box, accepting search statements and retrieving documents without disclosing the inner-workings of the process or even the contents of its information collection. What could be more closed to a potential searcher? Or more non-transparent? The first approach to make an information system more responsive is to open up and reveal its contents and work processes in a way that enables searchers to be adaptive and to adapt the system to their purposes, something that they are much better equipped to do than the system[2].

Before the system can become context-aware in the sense of "understanding" the user's context, it needs to be context-sensitive to its own context or content first. A system's context-awareness is therefore its transparency or the way it manages to reveal its content (i.e. documents, access points etc.) to the searcher:

> "One solution to this problem is to consider how a system can display aspects of its own context – its activity and the resources around which that activity is organized. Looking again at the issue of practice, the

[2] "An intelligent system that tried to outguess the user as to what is needed and how the user might behave would seem to work best with a user who is static rather than capable of learning and adapting. A user with intelligence can be expected to adapt to the system and to try to predict the system's future actions." (Buckland & Florian, 1991, 641)

> goal is to allow the system to reveal a richer picture of activities, and so provide users with a more nuanced interpretation of the meaning of the system's action." (Dourish, 1995, 11)

One way to make the information flow between user and system mutually interactive is to reveal the structure and organization of the information collection to the searcher. In web search, this is rarely possible because the extent and content of the information landscape is unknown. In subject-specific bibliographic information systems, however, the information landscape is bounded and the documentary language that is used for document representations provides a road map to the content of the collection.

The search term recommender system that is introduced in the next section uses the documentary language of the information system to provide context-specific terminology help to the searcher. It is context-specific because it takes the specialty search language of the searcher into account and provides insight into those parts of the information collection that are associated with the concepts represented in a user's search statement. By providing insight into the organization of the collection (through the vocabulary) it is hoped that enough system context is provided to make the searcher more aware of which region of the information landscape he is moving in and consequently achieve a more successful search transaction.

5.2 A Search Term Recommender

The function of a search term recommender system is to suggest terms that will likely increase the chance of a searcher to find relevant documents in the information

collection. Most searchers typically use the first terms that come to their mind (Drabenstott, 2000; Markey, 1984) and do not pay a lot of attention to the vocabulary of the document representations (author language or controlled vocabulary). If the documents in the collection are represented with only a few controlled vocabulary words and / or title and abstract descriptions, the mismatch between searcher terms and document terms results in unsuccessful retrieval:

> "Users of this online catalog search more often by keyword than any other type of search, their keyword searches fail more often than not, and a majority of these users do not understand how the system processes their keyword searches." (Hildreth, 1997, 61)

The goal of the search term recommender is to alleviate this mismatch by suggesting terms that occur in the collection to the searcher and will therefore be more likely to retrieve relevant documents. This kind of vocabulary help also reduces the searcher's need to think of other search terms that might describe his information need. It effectively eases the cognitive load on the searcher since it is much easier for a person to pick appropriate search terms from a list than to come up with search terms by himself[3].

The basic parameters of this search term suggestion system are the controlled vocabulary terms that are used for document representation and the natural language keywords that are input by the searcher. The advantage of suggesting controlled vocabulary terms as search terms is that these terms have been systematically assigned to the documents and so there is a high probability of relevant and precise retrieval results if

[3] "Man's powers of recall are much less robust than his powers of recognition." (Blair 1990, 53). See also Svenonius (1986).

these terms are used instead of whatever natural language keywords the searcher happens to think of (see chapter Four).

The suggested search term recommender system goes a step further by tailoring the suggested search terms to the specific specialty and specialty vocabulary of the searcher. This is achieved by dividing the information collection into subject area specialties and treating the controlled vocabulary terms in these specialty areas as representative of the specialty language in this subject area. Search terms (i.e. controlled vocabulary terms) are recommended according to the subject area that the search is related to. The aim of suggesting subject specialty-focused search terms is not only to predict even more precise and specific search terms but also to introduce a searcher to the specific vocabulary used in this information collection for this subject specialty. From that perspective, the search term recommender system is also a specialty language representation and disambiguation system.

5.2.1 Entry Vocabulary as Search Term Support

The lead-in vocabulary of controlled vocabulary schemes can be seen as a manual, labor-intensive form of a search term recommender system. Relative subject indexes to non-verbal notation systems like classifications and “See”, “See also” or “Use” references for thesauri and subject heading lists have been created to guide the user toward the preferred form of a term or classification code for a concept. Because they provide additional entry points to the controlled vocabulary, there referred to as an entry vocabulary (Lancaster, 1986, 59).

The subject indexes and their vocabularies are based on semantic relationships of synonymy (“Use” or “See” references), hierarchy (broader and narrower terms in thesauri and subject heading lists), and other undefined associational relations (“See also”). They have been criticized for several reasons:

> (1) the descriptive terms of such indexes are limited in number;
> (2) they only include terms that the creators of the index thought of;
> (3) they are labor-intensive and costly both at the creation and maintenance stages;
> (4) they are ordinarily monolingual; and
> (5) because of the high costs of creation and maintenance, they tend to be obsolescent. (Buckland, 2001)

Human intermediaries are another means of providing an entry point to the vocabulary of a system. They can map the searcher's query to the terms of the system's controlled vocabulary because of the associative ability of the human mind to link concepts to terms and vice versa. Human intermediaries have an advantage over manually created indexes to a controlled vocabulary in that they can map any possible query term of a searcher (as opposed to the few terms that the creator of the manual index could think of) to the appropriate system vocabulary term or phrase. Aside from the semantic relationships manually constructed indexes can provide, human intermediaries also recognize associative or functional relationships[4]. Furthermore, they are able to determine the best search terms or phrases out of a list of available terms from the controlled vocabulary - a ranking of the relevance of search terms with respect to the user's information need. But:

[4] For example, in a thesaurus the broad-narrow-term relationship between tree and branch can be determined - but not the functional relationship between tree and squirrel.

(1) most of the time they are not available to the searcher;
(2) they are far more expensive than the manually created index;
(3) they are not always an expert of the subject area a searcher is interested in and will therefore have difficulties associating concepts with system terms; and
(4) they can only have familiarity with a limited number of controlled vocabularies and information retrieval systems and will not be able to make suggestions if they encounter a language they do not understand (i.e., Russian, Chinese etc.).

The automatic search term recommender system suggested here overcomes the limitations of the restricted entry vocabulary of the relative index without the cost of a human intermediary. Its entry vocabulary is based on all natural language terms that occur in the information collection and it can be adapted to any controlled vocabulary system or language available.

5.2.2 Construction of the Search Term Recommender

The search term recommender is created by building a dictionary of associations between two vocabularies: (1) natural language terms and phrases from the subject area's documents in the information collection (e.g. titles, abstracts, authors) and (2) the controlled vocabulary (thesaurus terms, subject headings, classification numbers etc.) used for document representation.

In our implementation, a likelihood ratio statistic is used to measure the association between the natural language terms from the collection and the controlled vocabulary terms to predict which of the controlled vocabulary terms best mirror the topic represented by the searcher's search terms. A searcher's natural language terms will most likely come from the specialty language of his subject area. The natural language

terms from the subject area's documents are thought to represent the specialty language of the subject area and can therefore associate the searcher terms to the controlled vocabulary terms.

The idea of creating a ranked list of controlled vocabulary terms from natural language search phrases was first developed as the "Classification Clustering" technique by Ray Larson for the Cheshire Information Retrieval System (Larson, 1989, 1991, 1992). Larson used the classification numbers in catalog records to provide further access points for subject searches in an online library catalog. Records with the same classification number are seen as a group (cluster) of similar documents. The subject headings and titles that are assigned to records with the same classification number are used as a lead-in vocabulary for the classification system.

If the user enters a natural language search term, the system will estimate the probability of relevance for this term to a certain classification number. If a relevant classification cluster has been found, the system will then present the user with a number of frequent subject headings that are assigned to documents in this classification cluster. The user can choose important subject headings from this list and can refine the search. With this procedure, the number of potential access points was doubled on average for a record.

The methodology of constructing a general search term recommender without the focus on specialty spaces and vocabularies within a collection has been described in detail by Plaunt and Norgard (1998), and Gey et al. (1999).

5.2.3 Calculating the Association Weight

As the basic technique, a lexical collocation process is used. Lexical collocations are arbitrary and recurrent word combinations (Benson, 1990). If words co-occur with a higher than random frequency, they are associated. By measuring the strength of the association between a natural language term and each of the controlled vocabulary terms in the collection, the order of relevance of the controlled vocabulary terms for a searcher's term can be determined.

For the calculation of the statistical association, a contingency table for natural language terms (A terms) and terms from the controlled vocabulary (B terms) is constructed for all terms in the collection.

AB	A¬B
¬AB	¬A¬B

AB is the combination where both terms occur, A¬B is the event where A occurs without B, in ¬AB B occurs without A, and ¬A¬B denotes a document where neither terms appear. The statistical analysis measures the independence of events A and B from each other. If the hypothesis of independence is rejected (the natural language term and the controlled vocabulary term or phrase co-occur more than randomly), terms A and B are assumed to be associated.

A likelihood ratio proposed by Dunning (1993) is used to calculate the association weight and to predict the order of relevance for certain natural language term / controlled vocabulary term pairs (see Plaunt & Norgard, 1998; Dunning, 1993 for details).

5.2.4 Ranking and Suggesting Controlled Vocabulary Terms

The search term recommender proposed in this dissertation uses the same association calculation between natural language terms and controlled vocabulary terms but inserts another stage where the documents in the collection are subdivided by specialty as will be described in chapter Six. For each specialty in the document collection, separate association weights are calculated for natural language terms and controlled vocabulary terms. Consequently, the same natural language term might be highly associated with a particular controlled vocabulary term in one specialty and less highly associated with it in another specialty. The association weight per specialty effectively reflects the vocabulary use in the specialty and can therefore be used to show profiles of language differences between subject areas.

For example, Table 1 lists the most highly associated controlled vocabulary term from the MESH (Medical Subject Headings) vocabulary suggested by the search term recommender for the search term “heart” for eight subject areas within the medical document collection OHSUMED (described in chapter Six).

Subject Area / Specialty	**Suggested controlled vocabulary term for search term "Heart"**
Allergy and Immunology	Myocardium
Anesthesiology	Heart Rate
Drug Therapy	Heart Failure, Congestive
Geriatrics	Coronary Disease
Gynecology	Heart Rate, Fetal
Medical Oncology	Cholesterol
Pediatrics	Heart Defects, Congenital
Transplantation	Heart Transplantation

Table 1. Vocabulary differences in different specialties. Most highly associated MESH (Medical Subject Headings) term for the search term "Heart" in 8 specialties.

If the search statement contains more than one term, then controlled vocabulary terms are suggested based on the absolute association weight per specialty for all of the terms in the search statement. The absolute association weight of a controlled vocabulary term with respect to a search phrase is the sum of the individual association weights for the controlled vocabulary term and individual terms in the search statement.

For example, Table 2 shows how the association rank for the two most highly associated controlled vocabulary terms was calculated for a three-word query in the specialty Allergy and Immunology for the medical database OHSUMED.

Search statement: "IgE-mediated food allergy"		
Search term	Association weight for controlled vocabulary term	
	IgE	Food Hypersensitivity
IgE-mediated	*1230.39*	*10.34*
food	*9.44*	*730.67*
allergy	*24.06*	*345.16*
Absolute weight	*1263.89*	*1086.17*
Association rank	**1**	**2**

Table 2. Calculation of association rank for a three-word search statement.

This absolute weight calculation occurs over all controlled vocabulary terms associated with a search term. The controlled vocabulary terms with the highest absolute weight for the specialty and search statement are then suggested.

5.2.5 Applications for the Search Term Recommender

The search term recommender can take any input search term of the user and associate it with relevant controlled vocabulary terms regardless of metadata vocabulary type (thesaurus, subject heading list, or classification), language, or size.

The purpose of any search term recommender will be to map one vocabulary to another. Mapping a search term to the information system's controlled vocabulary is the primary function of the search term recommender. However, other mappings are also possible, in particular between different controlled vocabulary schemes (terminology mapping) and between documents and the controlled vocabulary (automatic text categorization). Examples for these mappings will be described in the next section.

Another important application is the mapping of a foreign natural language to a controlled vocabulary or natural language to a foreign controlled vocabulary. Because the mappings are based on statistical associations, if there is a training corpus available that contains natural language terms and controlled vocabulary terms in two different natural languages, then (with appropriate language processing) mappings can be created.

Advantages of search term recommender systems of this type include:

(1) they can be generated easily and quickly;
(2) they are not limited to one or only a few languages (like a human intermediary) but can be generated for any language for which a text corpus and a respective controlled vocabulary exist (and they do not even this have to be in the same language); and
(3) because they can be generated quickly, they are relatively easy to update (see also Buckland, 2001).

However, because the search term recommender is solely based on statistical association, it cannot be assumed that it has the qualitative and multitudinous associative capabilities of a human expert search intermediary. Of course, no automatic method could claim that.

On the other hand, because search term recommenders are based on statistical associations between search words and system vocabulary terms, the relationship established between them can be not only semantic but also functional. Search term recommenders enrich the conventional semantic relationships that can be found in controlled vocabularies (synonymy, hierarchy, equivalence). Functional relationships are those associations between concepts that are not conventional synonymy, part-whole or hierarchical correlations but rather concepts that belong to the same context or language

game. For example, semantically related concepts for “soccer field” include markings (part), sporting venue (broader) or pitch (synonym). Functionally related concepts could include ball, coach and world championship – all terms that would be good suggestions for more specific searches depending on the information need of a searcher. Functional relationships may describe a certain subject area or problem better than mere semantic relationships and may therefore be of more use to the searcher (Buckland, 2001).

A search term recommender could also be used as an additional aid for human intermediaries because they suggest a range of relevant vocabulary terms that could be used in the search and relieve the searcher of the burden of coming up with them himself.

Search term recommenders that take the specialties represented in an information system into account can improve retrieval performance in two ways. Firstly, by concentrating on mapping a searcher’s specific vocabulary to the database’s documents vocabulary – especially if they are different – will increase the matching ratio between query terms and document terms and therefore improve the retrieval performance. Additionally, targeting the search towards a specific domain within the collection’s documents associated with the query’s specialty vocabulary will make the search more precise because (a) primarily documents relevant to a searcher’s area of interest will be searched, and (b) polysemous meanings across different subject areas will be disambiguated. Specialty search term recommenders can be helpful in solving the homograph problem because they determine different controlled vocabulary terms for a given concept depending on meaning and context in the specialized subject area.

Focusing on specialty vocabularies and utilizing specialty search term recommenders can simultaneously act to restrict as well to expand a search strategy: it can focus the search by reducing the search space to documents relevant to a particular specialty or providing the searcher with more specific search terms related to his or her query; and it can expand the search by suggesting alternative or more general search terms to the searcher. Every specialty in the search term recommender represents a different, more specific information space in the collection. Making them available to the user for search should help to focus and specify an information need.

It is obvious that a method that will associate user search terms with appropriate system vocabulary terms will result in more successful searches in terms of number of documents and exhaustiveness of search. It is also foreseeable that most users, who are often unaware of the existence of such helpful system vocabularies, will welcome help in the form of vocabulary support like this.

5.3 Vocabulary Support Systems

The search term recommender system introduced here is related to a whole range of vocabulary support systems for information retrieval. In particular, this section will describe systems that use controlled vocabularies for query expansion, automatic text categorization and terminology mapping.

5.3.1 Automatic Text Categorization

Automatic text categorization or classification is the process of automatically sorting documents into predetermined categories (Sebastiani, 2004). An example of a text classification task is to sort email into spam and non-spam folders. This binary classification task is relatively simple compared to the multi-class categorization tasks that require the classifier to choose between multiple categories to assign a document to. The task is even harder if a document can be assigned several categories. This is the case for most document representation systems in bibliographic information collections (a document will have several subject headings or thesaurus terms assigned).

Automatic classification systems that assign controlled vocabulary terms to documents will have to be compared to the work of human indexers, who are more costly but also more adaptive and precise in their category selection. A big drawback for the evaluation of automatic classification systems is that human indexers – whose category assignments serve as the gold standard for comparison – do not choose the same categories for the same documents consistently (this is the inter-indexer consistency problem described in chapter Two). This means that any evaluation of the performance of an automatic classifier is only reliable to a certain degree, depending on the consistency of the original human indexers.

The areas of machine learning and information retrieval have both contributed different algorithms determining how an automatic classifier will be trained to predict categories based on document features. For recent surveys of the field, see Yang (1999), Kjersti et al. (1999) and Sebastiani (2004).

The search term recommender can be used as automatic classifier because documents that have to be categorized can be treated as search statements. Given a title or abstract of a document, the search term recommender then suggests the most highly associated controlled vocabulary terms for the document. The search term recommender can be applied to the most difficult categorization task: assigning several controlled vocabulary terms from a large category system (in terms of numbers of controlled vocabulary terms that are in theory available for assignment). The specialty search term recommender faces a two-stage categorization problem: first the appropriate specialty has to be selected and then the appropriate controlled vocabulary terms can be suggested.

This is similar to work done by Frank and Paynter (2004) who attempted to classify 20,000 internet resources (web pages) with Library of Congress Subject Headings (LCSH) into the hierarchical Library of Congress Classification (LCC) system. For each document, they first predicted the top-level category from the LCC hierarchy (one of 21) and then chose the appropriate classification code from within that category. When tested on 50,000 catalog records, this classifier (a support vector machine algorithm trained on 800,000 records) achieved an accuracy of 55% in assigning the correct LCC code to the record.

The work of Larson (1992) on classification clustering, which was the inspiration for the search term recommender, also falls into this category. Larson classified 283 catalog records of books within one LCC top-level category. The classifier was based on a training set of 30,000 catalog records that were clustered into virtual documents (containing title phrases and LCSH subject headings) each representing a LCC category.

Larson tested several information retrieval algorithms that would find the most similar LCC cluster (category) for a given test catalog record. The best approach achieved an accuracy of 46.6% in assigning the correct LCC category to a catalog record and provided plausible assignments for most of the rest.

The OCLC project Scorpion (Thompson et al., 1997; Godby & Stuler, 2003) is similar to Larson's classification clustering approach. The task is to assign classification codes from the LCC to web resources and other full-text documents. However, whereas the training documents for Larson were created from titles and subject headings in catalog records, Scorpion uses the text in the LCC classification schedules itself to create a textual representation for each category to train on. Scorpion uses the same log-likelihood statistic for calculating associations as the search term recommender does.

Pharos (Dolin, 1998) also assigns LCC categories to different information resources based on training documents using text from the classification schedules and subject headings of catalog records. Whereas Larson used information retrieval algorithms and Scorpion used Dunning's (1993) log-likelihood statistics, Dolin used latent semantic indexing for calculating associations.

Humphrey (1998, 1999) uses journal subject descriptors to assign keywords to a number of information resources. The journal subject descriptors (keywords describing the content of a journal) are extracted from a serials authority database and matched with the terms in the journal articles using a co-occurrence formula. Based on the training set, new documents can be indexed using the statistical calculations between keywords in articles and subject descriptors. Humphrey claims that the advantage of this approach is

that the training set does not need previous indexing or a controlled vocabulary to match the search terms.

This journal descriptor approach was used in one set of evaluations of the search term recommender system described in the next chapter. In order to select subject specialties (the first stage for the search term recommender process), journal descriptors were used to subdivide the collection on the basis of documents belonging to a certain journal set and therefore subject area.

Other classification applications include Leung and Kan (1997) who describe a statistical learning approach for the automatic indexing of controlled vocabulary terms by using a positive (includes records that have a particular controlled vocabulary term assigned) and a negative (includes records that have not the particular controlled vocabulary term assigned) training set for the co-occurrence calculation. Also Chen et al. (1995, 1997, 1998) who introduce the concept space approach for domain-specific automatic thesaurus generation and information retrieval. Co-occurrence data and other term filtering methods (i.e. dictionary lookup) are used to identify candidate controlled vocabulary terms from particular concept spaces. Concept spaces are similar to the specialty approach of the search term recommender in that they try to cluster controlled vocabulary terms around subject areas and predict more specific terms.

A version of the search term recommender system (without the additional zooming in on specialties) was used by Plaunt & Norgard (1998) to classify over 450 documents consisting of titles and abstracts from the INSPEC database. They found that the results were comparable to manual indexing by humans. The next chapter will

describe automatic classification experiments evaluating the performance of the specialty search term recommender introduced earlier in the chapter.

5.3.2 Query Expansion

Query expansion is the process by which the original search statement is augmented with additional terms that are thought to improve the retrieval performance. The result of a query expansion process can also be a query reformulation where original query terms are dropped and new terms are added to the search statement. Automatic (adding query terms without user input), interactive (the information retrieval system suggests new search terms and the user selectively adds them to the query), and manual (the user thinks of new search terms that would complement the query) query expansion techniques have been researched for decades (for a good overview, see Efthimiadis, 1996).

Automatic query expansion has been mostly discussed in the context of blind feedback, which adds new terms from the top ranked documents retrieved with the original search statement to a query (Robertson & Sparck Jones, 1976; Salton et al., 1985) or highly evolved expert systems, which add new search terms from a variety of other resources and filtering techniques (e.g. Gauch & Smith,1993; Doszkocs & Sass, 1992).

Controlled vocabularies are mainly used for manual or interactive query expansion (for an overview see Shiri et al., 2002a, 2002b), where controlled vocabulary terms are suggested to the user to supplement or substitute the original search terms.

The search term recommender can be used for automatic and interactive query expansion. By suggesting controlled vocabulary terms from a variety of specialties, it is thought that interactive query expansion where the searcher reviews the subject areas where relevant documents can be found and adds appropriate search terms from the area he is interested in will lead to the most effective retrieval results.

Whereas blind feedback from high-ranked-document terms has generally proved to improve retrieval results (e.g. Chen & Gey, 2004), the literature reports mixed results on query expansion with controlled vocabularies.

Jones et al. (1995) report on experiments expanding queries with descriptors from the INSPEC thesaurus but found no improved performance for the expanded queries. For the medical literature, Srinivasan (1996) achieved improved retrieval results when expanding a query with MESH subject headings, whereas Hersh et al. (2000) did not increase their performance when expanding queries with controlled vocabulary terms from the UMLS (Unified Medical Language System) Metathesaurus. More recently, Sihvonen & Vakkari (2004) experimented with query expansion from the ERIC thesaurus and report improved results but only if the subject area is sufficiently known to the searcher. Joho et al. (2004) presented query terms in concept hierarchies and report an increased retrieval performance and especially a speedier search process for the user. Suomela & Kekäläinen (2005) also presented concept terms in a hierarchy but found that their users preferred to search without the added controlled vocabulary support.

A version of the search term recommender system (without the focus on specialties) was used for automatic query expansion in the European Cross-Language

Evaluation Forum's domain-specific retrieval track (Petras, 2004; Petras, 2005). I found that automatically adding terms from the collection's thesaurus suggested by the search term recommender generally increased the retrieval performance for both German and English retrieval.

5.3.3 Terminology Mapping

Mapping controlled vocabularies to each other presents a big interoperability issue between information systems. Federated search across information systems with different document representations (i.e. controlled vocabularies) can only be successful if a searcher can move between those systems without the same language problem (matching the search terms with the document representations) occurring over and over again. In the past years, terminology mapping, that is the mapping between two or more controlled vocabularies by identifying the terms, concepts and hierarchical relationships that are equivalent to each other, has become a growing research field.

In order to map between different controlled vocabularies, various factors have to be considered:(1) the extent of overlap and coverage in the subject matter between the vocabularies; (2) the level of specificity of terms; (3) the degree of pre/post-coordination in term assignment; (4) how the vocabulary represents equivalence, hierarchical, and other relationships; (5) different word uses due to different natural languages or the chosen language level (e.g. common versus scientific names); and (6) the differences in meaning because of different conceptualizations or target audiences (see Lancaster & Smith, 1983; Doerr, 2001; Vizine-Goetz, 2004).

The end product of a terminology mapping project comes in three different flavors: a metathesaurus or integrated vocabulary, which subsumes several controlled vocabularies under one umbrella (e.g. the UMLS Metathesaurus); several micro-thesauri or satellite vocabularies connected to a more abstract superstructure (e.g. the various extensions to the Library of Congress Subject Heading like the Thesaurus for Graphic Materials or the Legislative Indexing Vocabulary); or a macro-vocabulary, which sits in-between vocabularies and function as abstract switching mechanisms between them. More terminology mapping projects are listed in Zeng & Chang (2004).

Zeng & Chang (2004) also distinguish between eight methods for terminology mapping:

> (1) Derivation / Modeling: a specialized or simpler vocabulary is developed with an existing, more comprehensive vocabulary as an initial model;
> (2) Translation / Adaptation: some controlled vocabularies consist of terms translated from a controlled vocabulary in a different language;
> (3) Satellite and Leaf Node Linking: specialized thesauri are treated as satellites of a superstructure;
> (4) Direct mapping: establishing equivalence between terms in different controlled vocabularies or between verbal terms and classification numbers;
> (5) Co-occurrence mapping: works at the application level, when documents are indexed with two controlled vocabularies, they can be linked through the common terms of the documents;
> (6) Switching: in translating equivalent terms in different vocabularies, a switching language or scheme may be used as an intermediary;
> (7) Linking through a temporary union list: terms that are not conceptual equivalents but are closely related linguistically may be linked to enhance retrieval; and
> (8) Linking through a thesaurus server protocol: thesaurus services accept queries in the format specified by the protocol, run the queries against their local thesaurus, and return results to the application that initiated the query in a standard format or in a customized extended format (Zeng & Chang, 2004, 381-385).

The search term recommender system can be used as a terminology mapping service. According to the list above, it would be an application of the co-occurrence mapping methodology. Co-occurrence mapping alleviates some of the critical factors (level of specificity, word uses, differences in meaning) mentioned earlier because the controlled vocabularies and their terms are already linked through the documents.

An important prerequisite for co-occurrence mapping between two controlled vocabularies is the existence of documents that contain controlled vocabulary terms from both of them. Most library catalog records in the United States have both Library of Congress Subject Headings and Library of Congress Classification codes assigned to them, making catalog records an ideal test bed for mapping. Major libraries in different countries will own the same books but assign controlled vocabulary terms from their own indexing schemes. Co-occurrence mapping has been used in the MACS (Multilingual Access to Subject) project, where three subject-heading vocabularies (Schlagwortnormdatei in German, RAMEAU in French, and the LCSH in English) were mapped to each other by virtue of having catalog records in these different systems (Landry, 2001, 2004). Because the search term recommender's association algorithm is language-independent, it could be used for these kinds of mappings.

A previous version of the search term recommender system was used to map between the patent classification codes of the very different U.S. and International Patent Classifications.

5.4 Conclusion

The search term recommender can provide the searcher with an insight into the context within which the system operates (opening the black box) by providing a representation of its internal knowledge structure (controlled vocabulary). By effectively asking “Did you mean:” and then providing what the system assumes are related terms to the query, the search term recommender opens up a conversation with the searcher, which will hopefully result in a more precise and effective search.

By leveraging the specialty language used in different knowledge areas, the search space can be more clearly delineated and choices between different senses can be made apparent to the searcher.

The search term recommender can be used as a query expansion mechanism, an automatic categorization application or as a terminology mapping device. Its effectiveness in any of these applications is dependent on the quality of the documents used for training the mappings between natural language terms and controlled vocabulary terms and the performance of the statistical association algorithm.

The general effectiveness of statistical association techniques and especially co-occurrence statistics for query expansion has been doubted because of poor, non-discriminative term selection (Croft & Thompson, 1987; Peat & Willet, 1991). Being based on statistical co-occurrence, the search term recommender is prone to fall under this critique, however, the focus on specialties should improve the term selection process.

Furthermore, in an interactive search situation, searchers might be annoyed by another intermediate step in the search process before they arrive at a result – especially if the search term recommender is not suggesting the appropriate search terms. (The searcher will have a preconceived notion of what the “right” search terms are). An evaluation of prediction rates and successful searches is therefore necessary.

The next chapters will analyze a series of experiments evaluating the effectiveness of the search term recommender system for subject-specific retrieval. In particular, it will answer the following questions: How different are specialties within a discipline when it comes to their vocabulary? How can a specialty be determined in a document collection? How well can the search term recommender predict controlled vocabulary terms? And how specific should a specialty search area be for effective retrieval?

Chapter 6

Determining Specialties in an Information Collection

The specialty search term recommender methodology of associating natural language terms in queries or documents with controlled vocabulary terms is based on the assumption that different subject specialties can be identified within an information collection. It is hypothesized that the different specialty vocabularies can be leveraged to provide a more precise insight into the subject areas covered in the collection and support a more effective search.

Subject area specialties within information collections can be identified by more than just different vocabularies. Bibliographic information collections, which are the

focus of this work, contain documents describing published works by researchers in different subject areas. Those documents can be distinguished by the different topics that they describe, by the authors (researchers working in different fields), by publication source (e.g. journals covering different fields), by different writing and formatting styles or the terminology used. Most bibliographic information collections cover a restricted set of disciplines or subject areas, consequently the first subject-distinguishing criterion for every document is whether it is included in the collection or not.

This chapter will first describe ways to distinguish specialties within an information collection (with special emphasis on bibliographic databases). Then it will describe the methods for determining specialties in the two bibliographic databases that were used for experimentation in this dissertation: Inspec, covering mostly the Physics and engineering fields and Ohsumed, a subset of Medline covering the medical field. The third part of the chapter will explore the question of whether the vocabulary is really different in different specialties – an assumption on which the functioning of the search term recommender is based. If the terminology does not vary sufficiently between different specialties, then the search term recommenders trained on particular specialties will not be more predictive for that specialty than a search term recommender trained on the general collection. Chapter Seven analyzes whether specialty search term recommenders trained on the more specific vocabulary of a specialty are more effective in recommending terms than general ones.

6.1 Determination of Specialties in a Document Collection

The identification and selection of documents belonging to a particular specialty in order to build a search term recommender is no trivial problem. Even if the searcher is able to describe his or her subject area of interest, the search term recommender system still has to determine which documents fit the description and target the search term recommender construction towards it.

An optimal determination technique would translate a searcher's explanation of his or her specialty into a selection algorithm for the STR (search term recommender) creation process. If the searcher's description is focused on identifying documents in an information collection, the description might contain descriptive keywords for the relevant problems, journal titles, important conferences, academic departments or researchers' names. Considering the language problem, this is once again a difficult exercise for any user-system interaction.

In an ideal case, a customized search term recommender could be dynamically created for every search space (i.e. specialty) a user of the STR system is interested in. However, even if a system could dynamically select documents according to some parameter, this would still require a lot of pre-search effort from the searcher in defining this space, something he might not be willing to do. It is imaginable that a STR system recognizes a searcher's preferred specialty by the search terms entered (requiring no further effort from the searcher), but only if they are specific enough to discriminate

against other specialties (e.g. the term “black hole” might indicate an interest in the Astrophysics specialty within a Physics information collection).

It is more realistic to assume that for any document collection, particular specialty areas could be pre-selected and search term recommenders will be created for them. The searcher can then opt to use one of more of these specialty STRs for vocabulary support, depending on whether the pre-selected specialties correspond with his internal cognitive map of the fields. Four ways to pre-select specialties from an information collection are proposed in this dissertation: domain terminology; clustering by publication sources; clustering by subject-specific classification; and bibliometric or social network analysis.

6.1.1 Domain Terminology

The domain terminology approach tries to assemble documents for a specialty STR by identifying characteristic specialty keywords or phrases. For this approach it is necessary to assemble a domain vocabulary (words characteristic of the specialty) specific enough to be distinguished from a general vocabulary and other subject area’s domain vocabularies.

A simple way to define a specialty vocabulary for a STR is to manually compile a list of domain-specific words and then find documents that include them. In a preliminary experiment with specialty STRs, Kim (1998) used a title word search to find documents in two bibliographic databases that contained a particular domain-determining word (in that example the word “water”). This is a very expensive approach in terms of time, labor and effort. Several researchers have tried to define domain vocabularies automatically.

Haas and colleagues (Haas & He, 1993; Losee & Haas, 1995; Haas, 1997) tried to define sublanguage vocabularies by determining "domain seed words" from a text with a dictionary approach (missing terms that did not occur in the dictionary, subject codes terms, and domains names like "computer screen" in the domain of computers in general) and to expand this list with candidate terms that were situated near the seed words in a text. A document's affiliation with a particular specialty can be infered by looking for sublanguage usage within the document.

Damerau (1990, 1993) describes a method of creating a domain vocabulary list by taking a set of texts from a predefined and already categorized source, and extracting content-bearing words by applying a word frequency count and some filtering methods.

Others used dictionaries with subject codes attached to particular words to compare words and phrases from documents with the coded terms in the dictionary (Liddy et al., 1992; Walker & Amsler, 1986). Words and phrases from the documents that matched the coded terms in the dictionary were clustered depending on the subject area and a frequency measure was used to determine the subject of the document.

Because of the ambiguous nature of language, the domain terminology approach is very imprecise. Even technical terms occurring only in certain disciplines will still occur in more than one specialty – especially in related specialties – which makes the categorization of documents into a particular specialty difficult.

6.1.2 Publication Source

Whereas domain words are a very fine-grained way of defining a specialty in a document collection, the publication source approach is broader. In scholarly communication, research is commonly published in subject-specific outlets. Most bibliographic databases contain documents assembled from many different publication sources (journals, conference proceedings, books, technical reports etc.). A specialty can be identified by selecting the publication sources that a document from a particular subject area will typically appear in. This can be done manually or in a semi-automatic way from predefined lists.

Predefined lists of publication sources for particular subject areas in academic disciplines are not hard to find. Every researcher will be able to name the most prominent journals and conferences in his or her area. Several bibliographic database providers also publish lists of journals grouped by subject area. For example, the Institute for Scientific Information publishes annual Journal Citation Reports[1] that group journals by predefined subject areas in the sciences and social sciences. The Institution of Engineering and Technology (the producer of Inspec) offers a list of journals in five subject areas for its current awareness service[2] in Physics and engineering. The National Library of Medicine publishes a list of the journals indexed for the Medline bibliographic database[3] that have been assigned subject-specific keywords that can be used to group the journals by

[1] http://portal.isiknowledge.com/portal.cgi?DestApp=JCR&Func=Frame
[2] http://www.iee.org/Publish/inspec/ProdCat/Aware/Keyabs.cfm
[3] http://www.nlm.nih.gov/tsd/serials/terms_cond.html

specialty. These lists can be used to automatically extract documents from a collection belonging to certain specialties for STR creation.

In a preliminary experiment with specialty STRs, Kim (1999) used the Journal Citation Report from the Institute for Scientific Information to assemble a list of journals for a certain specialty.

Humphrey (1998, 1999) used the National Library of Medicine journal subject descriptors to group documents by specialty (for an automatic classification application). In this dissertation, the Medline journal descriptors were used to group documents into 33 specialties in the Ohsumed medical collection.

Fully automatic techniques for clustering journals for STR creation are also imaginable. Journals could be grouped by using domain-specific title words or citation patterns. This would allow for a more flexible determination of specialty areas because it does not rely on predetermined lists from which one had to choose but it is prone to the dangers of the domain terminology approach.

Because the articles in research journals reflect patterns of scholarly work and communication (see also chapter Three), this approach is probably more representative for specialty identification than the domain terminology approach. It also lends itself to more interactive STR creation, if desired, because it is easier for a user to select journals from a list of all journals in the collection than to come up with domain words for the determination of specialties.

6.1.3 Bibliometric or Social Network Analysis

A specialty can be described by the community of researchers that work in it. Social network analysis (Wasserman & Faust, 1994) is one tool to identify a community of people who have some characteristic features in common, but for the identification of specialties in document collection a more specifically document-centric approach would be preferred.

For manual creation of a specialty STR, a user could point to a group of scholars and authors that make up a specialty's core membership, which will then be used to find documents authored by these people.

A specialty community can also be identified from a starting publication by following the references to or from other authors. With the help of citation indexes, this tracing of citations is possible both backwards and forwards in time. One can assume that citations or references are somehow related to the specialty that the document is associated with. Bibliometrics (or scientometrics) is a discipline concerned with analyzing these patterns in published scholarly communication (White & McCain, 1989). By analyzing citing patterns, one can detect subject area structures. Examples can be found in White & McCain (1998) and Small (1999). Recently, more efforts have been made in connecting the document-centric analyses of bibliometrics with the more qualitative methods of social network analysis (White et al., 2004) and discourse analysis in the area of applied linguistics (White, 2004).

Social network or bibliometric analysis would define specialties by the members of the community. This would overcome the problems associated with the domain

terminology (ambiguities of language) and publication source approaches (documents can be published in general-purpose journals or in out-of-domain sources), however it also requires a lot of analytic effort to identify these structures. Another problem with the bibliometric approach is that identifying a community of people that represents a specialty does not necessarily provide the STR creator with a name, description or even content explanation for that specialty, which in turn makes it hard to represent different specialties to the user of a system for selection. On the other hand, bibliometric analysis will be able to identify new and interdisciplinary specialties, which is difficult to do with the other approaches.

6.1.4 Subject-specific Classification

If a document collection has a subject-specific classification associated with it, specialty areas can be determined alongside classification categories. The top-level categories of a classification usually circumscribe specialties within an information collection. Since every document will be assigned with one or more classification categories, the identification of specialties for STR creation can be relatively straightforward. In this dissertation, the top-level categories of the Inspec classification scheme were used to define three specialties in the Inspec document collection.

This approach is probably the easiest for automatic processing purposes, because the documents are already grouped by classification categories. Nevertheless, it will not be effective in identifying specialties if the classification of the information collection is not hierarchical (faceted classifications represent aspects of every specialty in each of

their facets) or if the categories do not conform to specialty boundaries and are either too broad or too narrow.

Each of these approaches to determine specialties in an information collection will result in at least slightly different specialty definitions (as identified by the documents that will be associated with them). For the practical purposes of specialty STR construction, an analysis of the affordances of the information collection can point to the most suitable approach. The next section will describe two experimental document collections and specialty identification approaches used in this dissertation.

6.2 Inspec and Ohsumed – Two Subject-specific Bibliographic Databases

6.2.1 Inspec

Inspec is a multidisciplinary bibliographic database that covers over 3,800 scientific journals, 2,200 conference proceedings, and other books, reports, and dissertations in the disciplines Physics, electrical and electronic engineering, computers and control, information technology and (since 2004 a more comprehensive coverage of) production, manufacturing and mechanical engineering. As of 2006, the Inspec database contains over 8 million bibliographic documents and is growing at the rate of over 450,000 records each year[4].

[4] http://www.iee.org/publish/inspec/about/

A typical Inspec document contains bibliographic information about an article or another publication such as author, title, source, publication year, etc. Additionally, it contains an abstract, descriptors from the Inspec thesaurus, and classification codes from the Inspec classification. The Inspec thesaurus[5] now contains circa 9,000 descriptors (after an extensive revision in 2004) and many more entry vocabulary and cross-reference terms. The Inspec classification[6] was revised in 2004 to add one more top-level category (Manufacturing & Production Engineering) to its four main sections of Physics, Electrical & Electronic Engineering, Computers & Control, and Information Technology for Business.

For experiments in this dissertation, over 400,000 Inspec documents were downloaded via the Melvyl system of the University of California. Table 3 gives an overview for the Inspec test collection used for search term recommender creation. The number of unique terms represent the title terms from the bibliographic documents and does not include 631 stopwords (most common words like prepositions, articles etc.) and is calculated after stemming (removing word endings like plural –s to arrive at a common word stem).

Number of documents	427,340
Number of unique terms	60,601
Number of unique Inspec descriptors	8,447

Table 3. Inspec collection numbers.

[5] http://www.iee.org/publish/support/inspec/document/thes/
[6] http://www.iee.org/Publish/Support/Inspec/Document/Class/index.cfm

Because Inspec includes the Inspec classification, we used the subject-specific classification approach to define specialties in the collection. Three specialty search term recommenders were created:

> (1) a specialty STR for Physics trained on documents from the Inspec collection that would have a classification code assigned starting with the letter A (top-level category);
> (2) a specialty STR for Electrical & Electronic Engineering with documents containing classification codes starting with B (top-level category); and
> (3) a specialty STR Computers & Control with documents containing classification codes starting with C (top-level category).

An Inspec document can be assigned more than one category code. If a document was assigned classification codes from more than one specialty (e.g. a classification code starting with A for Physics and classification code starting with B for Electrical Engineering for the same document), then the document was grouped into both specialties. Because research specialties overlap, this is a more realistic representation of the specialty spaces in an information collection than forming mutually exclusive categories.

Because the top-level category D (Information Technology) contains a lot fewer documents and top-level category E (Manufacturing & Production Engineering) was newly added, they were not included for experiments. For an overview of the collection numbers used for the Inspec specialty search term recommenders, see Appendix A, table A1.

On average, each Inspec document contains circa 7 Inspec descriptors. However, in a sample of over 120,000 Inspec test documents, several hundred documents were found with over 20 Inspec descriptors and one document with 44 descriptors. Table A2

shows the distribution of Inspec descriptors per document for a sample of the Inspec database. For the STR construction process, this means that every title word is associated with circa seven controlled vocabulary words (descriptors).

The specialty search term recommenders were constructed by using the title words from the Inspec documents as representatives of the vocabulary in the discourse in that specialty and the Inspec descriptors as representatives of the specialty controlled vocabulary. The most common and non-subject-specific words (i.e. 631 stopwords) were removed from the titles because they will not be discriminative for the search term recommendation process. All title words were also stemmed so that different spelling variations and word endings of the same term will result in the same term representation. The Inspec descriptors were recorded as is.

6.2.2 Ohsumed

The Ohsumed[7] collection (Hersh et al., 1994) is a subset of the Medline (Medical Literature Analysis and Retrieval System Online) bibliographic database published by the National Library of Medicine. Medline[8] contains approximately 13 million bibliographic documents covering 4,800 international journals in the biomedicine and health disciplines.

Ohsumed contains circa 350,000 bibliographic Medline documents from 270 mostly clinically oriented journals over a five-year period (1987-1991). A typical

[7] Ohsumed stands for Oregon Health Sciences University Medline collection.
[8] http://www.nlm.nih.gov/pubs/factsheets/medline.html

Ohsumed document contains author, title, source and publication type information as well as abstracts and Mesh indexing terms.

Mesh is the National Library of Medicine's controlled vocabulary for Medline bibliographic documents[9]. As of 2006, it contains 22,997 subject headings and more than 130,000 entry term and see reference terms. Every Medline document will be assigned 5-20 Mesh headings from the vocabulary[10]. Those Mesh headings can be subdivided into major Mesh headings, which indicate the main focus of the paper, and minor headings, which indicate secondary topics. A Mesh heading consists of a descriptor, subheadings (qualifiers, which indicate specific aspect of a concept represented by the descriptor), check tags (parameters of subject content like gender), publication types and age group headings.

For example, below are some of the Mesh headings assigned to an article titled "Gaucher disease with pulmonary involvement in a 6-year-old girl: report of resolution of radiographic abnormalities on increasing dose of imiglucerase" in the Medline database:

Major headings with subheadings:	*Gaucher Disease/complications *Lung Diseases/etiology
Minor headings with subheadings:	Dose Response Relationship Lung Diseases/radiography
Check tags:	Case Report; Female; Human
Age group headings:	Child
Publication type:	Journal Article

For this collection, the publication source method for determining specialties in the collection was used. The National Library of Medicine's List of Journals indexed for

[9] http://www.nlm.nih.gov/pubs/factsheets/mesh.html
[10] http://www.hsl.creighton.edu/HSL/searching/MeSH-Intro.html

Medline[11] contains a subject descriptor for each journal listed. For each journal in the Ohsumed collection, the subject descriptor was looked up[12]. Then, the documents were grouped by journals belonging to the same subject area as described by the journal descriptor. For conformity reasons, only those specialty collections containing more than 2 journals and containing between 4,000 and 10,000 documents were used for specialty STR construction. This resulted in 33 different specialties identified in the Ohsumed collection forming a subcollection of circa 170,000 documents as a whole.

For the specialty STR construction, title words from the bibliographic documents were used to represent the specialty natural language and only the major descriptors (without subheadings, tags, etc.) from the Mesh headings were used to represent the special controlled vocabulary in the collection. Table 4 gives an overview for the Ohsumed test collection used for specialty search term recommender creation. The number of terms represents the title terms from the bibliographic documents and does not include 631 stopwords and is calculated after stemming.

Number of documents	168,463
Number of unique terms	39,762
Number of unique Mesh Headings	12,140

Table 4. Ohsumed collection numbers.

On average, the Ohsumed test collection's documents were assigned three major Mesh headings. The distribution of Mesh headings was much more uniform than the

[11] http://www.nlm.nih.gov/tsd/serials/terms_cond.html
[12] In some cases, more than one journal descriptor was assigned to a journal. In those cases, the first and primary journal descriptor was selected.

descriptor assignment for the Inspec collection: all documents were assigned between one and eleven Mesh headings (2 out of 18,700 sample documents contained 11 Mesh headings). The distribution of Mesh headings per document (analyzed over a sample of 18,700 Ohsumed documents) can be found in Appendix B, table B2.

A list of the 33 specialties identified and their respective collection numbers can be found in table B1.

6.3 Differences in Vocabulary

Most communities of discourse (specialties) have their own domain language for describing concepts. Language within a domain includes general usage words, words adapted through metaphor and given a specialized local meaning, and new words coined within the specialty. For a specialty search term recommender to work, it is assumed that documents in an information collection will reflect the different terminologies of the specialties and that searchers will use their specialty vocabulary to express their search interest. Specialty STRs utilize the different terminologies and word usages to associate the words with controlled vocabulary terms from the information collection.

This part analyzes the terminology in three specialties from the Inspec and Ohsumed collections. Both the natural language vocabulary (in form of title terms from the collection documents) and the controlled vocabulary within the specialties are looked at. If the specialty vocabularies were different from each other, one would expect that the overlap between specialty terminologies is small and that a different set of controlled

vocabulary terms (representing different concepts in the information space of the collection) is assigned with the documents. Some terms (especially disciplinary terms) will be shared between specialties but will occur with different frequencies reflecting the focus of the specialty. General language terms will occur in all specialties, which is the reason why there will always be an overlap between any kinds of discourse space within a language.

For the analysis, the most common general language terms (i.e. 631 stopwords) were removed from the titles in all specialty collection documents. All title terms were extracted from the specialty collection and stemmed to remove spelling and word ending variations. The controlled vocabulary terms were extracted from each collection as is and their frequency in the collection was counted.

For the natural language terms, the differences between specialties were analyzed by calculating the overlap in terms between the specialty collections and the number of unique terms for each collection out of the sum of all terms in all three collections. For the controlled vocabulary terms, the same calculation was performed for all three collections. Additionally, the frequency of occurrence for controlled vocabulary terms in the specialty collections was analyzed. Because the controlled vocabulary is a smaller vocabulary than the natural vocabulary in a collection, it is expected that controlled vocabulary terms overlap more in the specialties than natural language terms. For each specialty collection, the thirty most often occurring controlled vocabulary terms were extracted and their frequency of occurrence in the other specialty collections checked. It is assumed that controlled vocabulary terms that occur very frequently reflect a topical

focus of the collection. These terms should occur in their specialty much more frequently than in the other specialties, even if the general overlap between controlled vocabulary terms is high.

6.3.1 Inspec

For the Inspec collection, the three specialty collections in the fields of Physics, Electrical & Electronic Engineering, and Computers & Control were analyzed. Table 5 shows the number of documents, unique terms and number controlled vocabulary terms in these three collections.

Specialty Collection	Number of documents	Number of unique terms	Number of unique Inspec descriptors
Physics	193,596	42,334	7,506
Electrical & Electronic Engineering	128,068	29,911	7,443
Computers & Control	105,676	28,193	6,245

Table 5. Inspec specialty collection numbers.

Even though the Electrical Engineering specialty collection has 21% more documents than the Computers & Control collection, the number of unique terms occurring in each one is about equal. The total number of Inspec descriptors in the Inspec general collection is 8,400 and most of them occur in each specialty collection, showing the expected overlap between controlled vocabulary terms.

The total number of unique terms occurring in the three Inspec specialty collections is 60,601, so every specialty collection contains about half of the total number of terms in the collection (Physics even contains 70%).

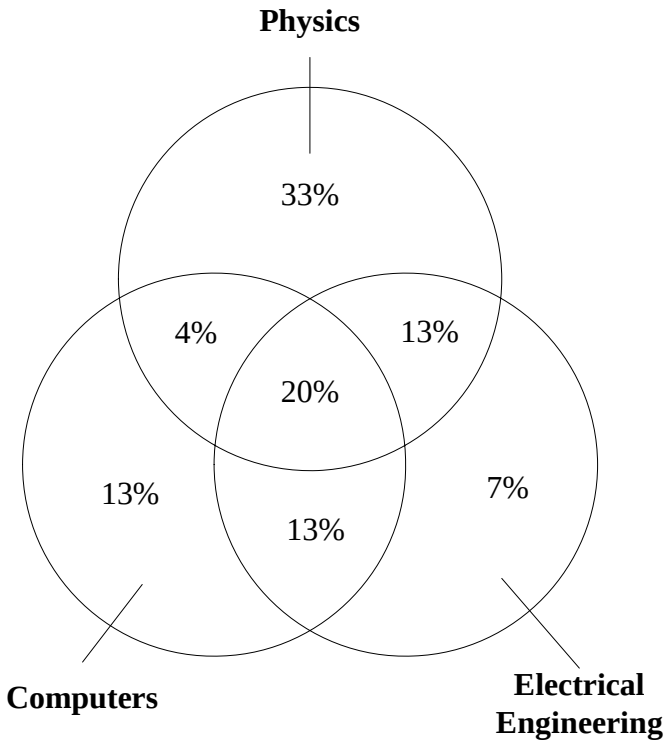

Figure 1. Overlap between natural language terms in three specialty collections in Inspec (total terms analyzed: 60,601).

Figure 1 shows that the three collections overlap completely in only 20% of the terms analyzed. All general language terms not already extracted will fall into this category. Another 30% of the terms are shared between two collections but not three. Physics contains the most unique terms not occurring in other collections. This is not surprising because it is also the largest collection in terms of documents and number of terms. Physics and Computers & Control seem to be more distinct from each other because they have fewer terms in common (out of the total number of terms in the two

collections, only 25% are shared). Electrical Engineering and Computers & Control seem to be more related because about 43% of their terms are shared between them. Electrical Engineering seems to be a bridging specialty between Physics and Computers & Control because it also shares a lot of terms with Physics (around 38%) and has very few unique terms itself.

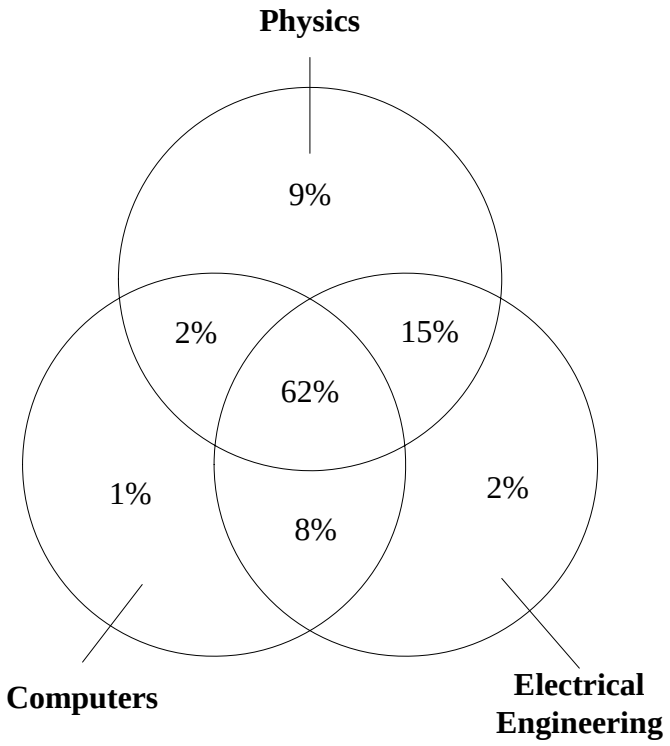

Figure 2. Overlap between controlled vocabulary terms in three specialty collections in Inspec (total terms analyzed: 8,447).

For the controlled vocabulary term overlap, the picture looks different. In total, 8,447 controlled vocabulary terms occur in all three collections and the majority of them occur in every one (89% in Physics, 88% in Electrical Engineering, and 74% in Computers & Control). The overlap between controlled vocabulary terms is therefore

substantial (shown in figure 2): 62% of the controlled vocabulary terms occur in all three collections, and another 25% occur in two collections.

Once again, Physics and Computers & Control seems to be more distinct: 66% of the controlled vocabulary terms occurring in both collections are shared, whereas Electrical Engineering shares about 79% of the controlled vocabulary terms with Physics (out of all terms occurring in both collections). It is once again a bridging specialty, also sharing 79% of its controlled vocabulary terms with the Computers & Control specialty.

Even though the majority of controlled vocabulary terms overlap in the three specialties, they occur in different frequencies in each collection. Looking at the 30 most often occurring Inspec descriptors in each of the collections (table A3 lists the terms) clearly shows that the specialties focus on different topics in their documents. Table 6 shows how the frequencies of occurrence of controlled vocabulary terms differ for each collection. The most frequent Inspec descriptors in the Physics specialty collection occur four times as often in Physics documents than they occur in Electrical Engineering and 222 times as often than in the Computers & Control specialty collection.

In evaluating these numbers, one obviously has to take the collection sizes into account. The Physics collection is almost twice as large as the Computers & Control collection, however, even dividing the factor by half shows a large difference in frequency of occurrence. The distinctiveness of each collection is also displayed in the inverse relationship. The 30 most often occurring controlled vocabulary terms in the Computers & Control collection occur 98 times more often in this collection than in the Physics collection.

Physics		
	Electrical & Electronic Engineering	4
	Computers & Control	222
Electrical & Electronic Engineering		
	Physics	33
	Computers & Control	80
Computers & Control		
	Physics	98
	Electrical & Electronic Engineering	4

Table 6. Factor by which the 30 most frequent assigned Inspec descriptors in the Physics, Electrical Engineering or Computer specialty collection occur more often than in the other 2 specialty collections.

Electrical Engineering seems to be in the middle as was also shown by the natural language term distribution. Both Physics and computer & control vocabulary terms occur only 4 times more often in those collections as they occur in Electrical Engineering. Electrical Engineering terms occur more frequently in the other collections than vice versa.

On average (over all three collections), the 30 most frequent Inspec descriptors in one specialty collection occur 73 times more often in this collection than in the other two.

6.3.2 Ohsumed

For the Ohsumed collection, three out of the 33 specialty collections were selected for vocabulary analysis: Communicable Diseases, Gynecology and Orthopedics. Table 7 shows the collection numbers for these three collections. The specialty collections in the Ohsumed collection are much smaller than in the Inspec collection, and

each contains a much smaller percentage of the total number of controlled vocabulary terms in the collection. The total number of controlled vocabulary terms in the three Ohsumed collections is 5,376. Gynecology contains the most Mesh headings out of the total number with 62%. The Orthopedics specialty collection contains only 30% and the Communicable Diseases collection only 45%. Considering these numbers, it can be assumed that the overlap between these collections with respect to the controlled vocabulary will not be as high.

Specialty Collection	Number of documents	Number of unique terms	Number of unique Mesh Headings
Communicable Diseases	4,973	5,151	2,412
Gynecology	8,268	6,728	3,347
Orthopedics	5,812	4,913	1,631

Table 7. Ohsumed specialty collection numbers.

The Gynecology collection is much larger than the other two collections, both in terms of documents but also with respect to the number of unique terms and Mesh headings occurring in the collection. The other two collections are of similar size, but Orthopedics seems to be a more homogeneous collection than the Communicable Diseases specialty because fewer Mesh headings occur in it.

The total number of unique terms in these three collections is 11,663. Like the Inspec specialty collections, each collection contains about half of the total number of terms. Figure 3 shows the overlap for the three collections in the Ohsumed collection. The overlap between specialties is smaller than in the Inspec collections: 30% are shared between all three or two of the specialties. Comparing two specialty collections at a time,

the overlap seems to be around the same with slightly less term overlap between the Communicable Diseases and Orthopedics collections (21% versus 26% between Communicable Diseases and Gynecology and 25% between Gynecology and Orthopedics), so these specialties seem to be less related to each other.

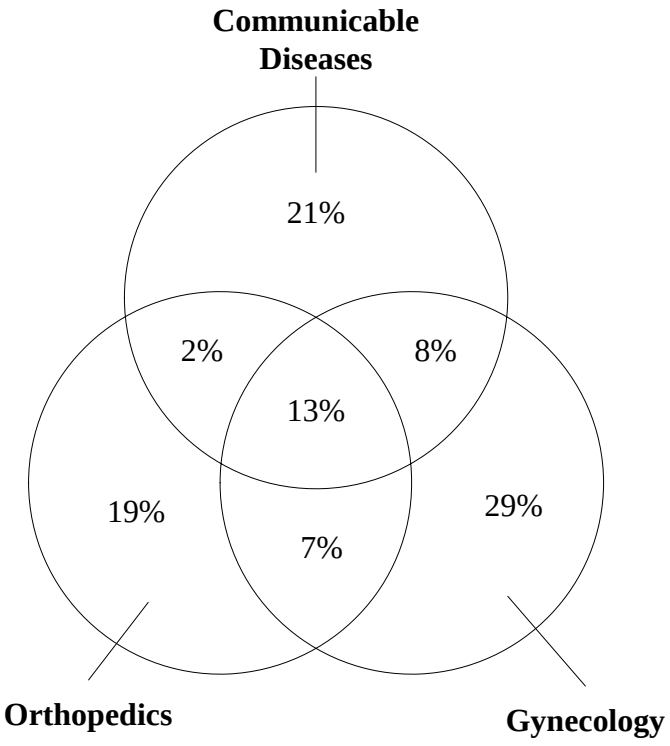

Figure 3. Overlap between natural language terms in three specialty collections in Ohsumed (total terms analyzed: 11,663).

Figure 4 shows the overlap between the Mesh headings in the three specialty collections. As expected, the overlap between these specialties is much lower than in the Inspec collection. Only 8% of the controlled vocabulary terms are shared between all three specialties. One can see in the controlled vocabulary distribution that Communicable Diseases and Orthopedics (overlap between Mesh headings in these two collections is 14%) are slightly more distinct from each other than either Communicable

Diseases and Gynecology (overlap is 21%) or Gynecology and Orthopedics (overlap is also 21%).

The controlled vocabulary term overlap between the three collections is only slightly higher (32%) than the natural language term overlap (30%) indicating that the controlled vocabularies are as distinct as the natural language vocabularies between those collections. The controlled vocabulary of the Ohsumed collection is also more varied than the Inspec controlled vocabulary. For every Inspec descriptor, there are about five natural language terms, whereas for every Mesh heading, there are only 2 natural language terms in the Ohsumed collection showing that the distribution of Mesh headings is larger than the distribution of Inspec descriptors.

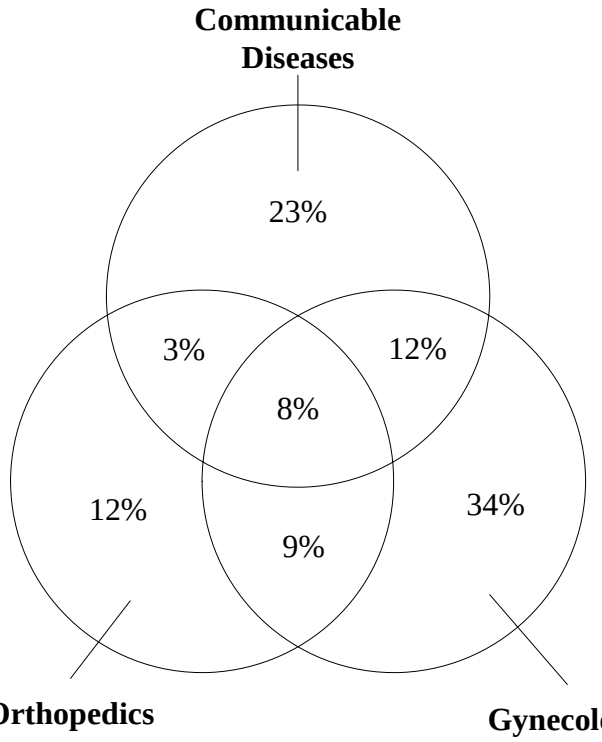

Figure 4. Overlap between controlled vocabulary terms in three specialty collections in Ohsumed (total terms analyzed: 5,376).

Although the analysis of the overlap between the controlled vocabulary terms has already shown the variations in vocabulary between the three Ohsumed collections, the frequency distributions were also tested. For each specialty collection, the 30 most often occurring Mesh headings were extracted (see table B3). Table 8 shows the factor by which the most frequently assigned Mesh headings in the Communicable Diseases, Gynecology or Orthopedics specialty collections occur more often than in the other 2 specialties.

Communicable Diseases		
	Gynecology	41
	Orthopedics	80
Gynecology		
	Communicable Diseases	137
	Orthopedics	154
Orthopedics		
	Communicable Diseases	155
	Gynecology	135

Table 8. Factor by which the 30 most frequent assigned Mesh headings in the Communicable Diseases, Gynecology or Orthopedics specialty collection occur more often than in the other 2 specialty collections.

Averaged over all three specialty collections, the 30 most frequent Mesh headings in one specialty collection are about 117 times more often assigned in this collection than in the other two collections. The average number is so high for these collections because the controlled vocabularies do not overlap as much as they do in the Inspec collections. In fact, about half of the most frequent Mesh headings in one collection were not assigned in the other collections.

6.4 Conclusion

Specialties can be identified in various ways within information collections. This chapter has proposed four ways for determining specialties within document collections and has described experiments on two different bibliographic databases. For the Inspec collection, specialties were determined by the grouping of documents within a subject-specific hierarchical classification. For the Ohsumed / Medline collection, specialties were selected according to journal descriptors placing documents in particular subject areas.

For the Inspec collection, three specialty fields were identified, for the Ohsumed collection, thirty-three smaller and more distinctive specialties were determined. An analysis of the terminology within the specialty collection shows that both the natural language vocabularies and the controlled vocabularies differ between specialties. Differences in vocabulary depend on the relatedness of the specialties to each other and the variety of terms to choose from.

Differences in vocabulary are a basic assumption of the specialty search term recommender methodology introduced in this dissertation. If the terminology is different in different specialties, then different controlled vocabulary terms will be associated with natural language terms put forward in a search statement and a more focused search vocabulary support is possible. The next chapter will describe experiments designed to evaluate the effectiveness of specialty search term recommenders.

Chapter 7

The Search Term Recommender as Automatic Classification System

Search term recommenders can be used in several ways: for query expansion and reformulation (recommending controlled vocabulary terms as search terms), for automatic classification (recommending controlled vocabulary terms for subject description of a document), and terminology mapping (linking two controlled vocabularies with each other). This chapter describes experiments designed to evaluate the effectiveness of the search term recommender methodology when used for automatic classification.

There are other techniques for helping searchers with the STR such as query expansion and terminology mapping tools, but automatic classification has a substantial advantage with respect to evaluation. Compared to query expansion experiments, automatic classification analyses do not require additional effort to find test cases. For query expansion experiments, test search statements have to be created, which requires additional human effort and knowledge of the search and document space of the collection. Test search statements could also be gathered from query log files from an information system but these are rarely available and also require additional processing. For automatic classification experiments, test cases can be gathered from the information collection itself, by sampling a number of documents from the database.

Furthermore, compared to query expansion or terminology mapping experiments, the evaluation itself is cost-effective and does not require additional human effort. For query expansion experiments, the original search statements have to be retrieved against the collection and then the effectiveness of the search evaluated. Later, the expanded search statements have to be retrieved against the collection and the effectiveness of this search has to be evaluated. Evaluation of the retrieval results with respect to their relevance for a searcher has to be done by humans because of the necessity of reading every document to estimate its relevance for a query. For terminology mapping, every mapping needs to be analyzed by a human indexer familiar with both controlled vocabularies to estimate the correctness or suitability of the suggested relationship between terms.

This human effort in evaluation can be avoided in automatic classification experiments because the test cases come with a measure of relevance attached: every test document contains the original, human-assigned controlled vocabulary terms, which will be used in the evaluation phase to compare with the STR-suggested controlled vocabulary terms.

Two questions regarding the effectiveness of specialty search term recommenders as automatic classification application will be addressed in this chapter:

(1) will a search term recommender with a specialty focus predict better controlled vocabulary terms than a search term recommender trained on the whole collection (i.e. does it make sense to emphasize specialty vocabularies within a larger information space?); and

(2) how focused can a specialty STR really be and what are the practical limitations of specialization in a collection (i.e. when is specific too specific?)?

For the evaluation, the search term recommender systems for the Inspec and Ohsumed collections introduced in chapter Six were used.

7.1 Evaluation Measures

In information retrieval research, the effectiveness of retrieval systems is analyzed in terms of effective and correct retrieval of relevant documents. Relevant documents are those documents that “match” a search question (e.g. fulfill the information need stated in the question).

In automatic classification research, the effectiveness of classification applications is analyzed in terms of effective and correct classification of documents into pre-determined categories. The effectiveness of an automatic classification application can be analyzed with respect to how effectively the relevant (i.e. correct) categories are determined for a set of test documents. The relevant categories are those categories that "match" the document's content. This can be operationalized by examining the application's ability to choose the same classification categories for a document that a human indexer would. For evaluation, those categories that were previously assigned to the document are presumed to be relevant. Seen from this perspective, automatic classification is a specific case of an information retrieval problem and can be evaluated with information retrieval measures.

In information retrieval, retrieval effectiveness is traditionally evaluated with two measures: recall and precision. Recall describes the number of retrieved relevant documents out of all relevant documents in the collection with respect to a particular query, a measure of completeness of retrieval. Precision describes the number of relevant retrieved documents out of all retrieved documents, a measure of the quality of retrieval.

$$\text{Recall} = \text{Relevant} \cap \text{Retrieved} / \text{Relevant}$$
$$\text{Precision} = \text{Relevant} \cap \text{Retrieved} / \text{Retrieved}$$

In effect, recall describes the hit-rate of the retrieval system (did it find all the relevant documents) and precision describes the accuracy of the system (did it find only relevant documents).

For the evaluation of the search term recommender methodology as an automatic classification application, the traditional information retrieval measures of recall and precision are used – however, not measured in retrieved documents but in suggested controlled vocabulary terms. Recall is defined as the proportion of the STR-suggested controlled vocabulary terms that are originally assigned terms out of all originally assigned controlled vocabulary terms. Precision is defined as the proportion of all STR-suggested controlled vocabulary terms that were originally assigned to a document (see also Plaunt & Norgard, 1998).

$$\text{Recall} = \text{Originally assigned} \cap \text{Suggested} / \text{Number of Originally assigned}$$
$$\text{Precision} = \text{Originally assigned} \cap \text{Suggested} / \text{Number of Suggested}$$

Recall and precision can be measured at different cut-off levels of suggested controlled vocabulary terms. The cut-off level determines how many controlled vocabulary terms the search term recommender suggests. At a cut-off level of 1, only the single controlled vocabulary term with the highest association weight will be suggested. Different cut-off levels influence the recall and precision measures. For example, if the original document had three controlled vocabulary terms assigned and the cut-off level for suggested terms is 1, then the recall can never be more than 0.33 because only a third of the originally assigned controlled vocabulary terms could have been found. Recall is increasing with higher cut-off levels (the more controlled vocabulary terms are suggested the more relevant terms will be found), but precision can fluctuate depending on how well the STR predicts the controlled vocabulary terms. Table 9 shows how recall and

precision levels change for an example where the original document had four controlled vocabulary terms assigned.

Cut-off level	Originally controlled vocabulary terms suggested	Recall	Precision
1	1	0.25	1
2	1	0.25	0.5
3	1	0.25	0.33
4	2	0.5	0.5
5	3	0.75	0.6
6	4	1	0.66
7	4	1	0.57

Table 9. Example for recall and precision calculation. One relevant controlled vocabulary term is suggested at cut-off level 1 (highest ranked term), another one at cut-off level 4, another one at 5 and the last one at cut-off level 6.

In the Ohsumed collection, documents have been assigned an average of three controlled vocabulary terms. In the Inspec collection, documents have been assigned on average about seven controlled vocabulary terms. Comparatively, suggesting controlled vocabulary terms for the Inspec collection will be more difficult, because more correct controlled vocabulary terms have to be found. For the recall/precision calculations, cut-off levels between 1 and 15 terms for the Inspec collection and cut-off levels between 1 and 10 for the Ohsumed collection were calculated. All precision and recall measures were averaged over the total number of test documents with which each search term recommender was evaluated.

Additionally, a combined “oracle” recall/precision measure was calculated. For each document, the cut-off level was set at the number of originally assigned controlled vocabulary terms. At this cut-off level, the number of suggested terms is, by definition, the same as the number of assigned terms. Therefore, the proportion of correctly

suggested terms constitutes not only the recall performance but also the score for precision.

Oracle Recall/Precision = Originally assigned ∩ Suggested / Number of Originally assigned (= Number suggested)

The oracle recall measure shows how well the STR would classify a document if the required number of terms was known beforehand. It is also a convenient combination of both recall and precision measures. For the example in table 9, the oracle recall/precision measure is 0.5: out of the four originally assigned controlled vocabulary terms, the STR suggests two correctly when it can predict four, so recall is 2 out of 4 = 0.5 and precision is also 2 out of 4 = 0.5

As a matter of interest, the proportion of “perfect” classifications was also calculated. Perfect classifications are those cases where all the suggested terms are correct at the oracle cut-off level: recall and precision are both 1.0 as would have been the case in the previous example if all 4 (instead of just 2) of the suggested terms had been correct. In other words, the percentage of perfect classifications represents those cases where the STR assigns exactly the same controlled vocabulary terms as a human indexer has done when the number of term suggestions is set to the number of originally assigned terms.

The recall and precision measures used for evaluation of the search term recommenders are conservative measures. Because they are automatically calculated, these measures cannot take into account value judgments that a human evaluator would make. For example, if the STR suggests a more specific or a broader controlled

vocabulary term than the originally assigned, this controlled vocabulary term could be counted as a relevant or correct classification for a test document. However, the automatic evaluation will discard the suggested term as non-matching and therefore irrelevant.

Inter-indexer consistency describes the overlap of controlled vocabulary term assignments between two or more human indexers. Studies have shown that inter-indexer consistency lies anywhere between 30% and 60% between two indexers (see chapter Two). If human indexers suggest only a third to a half of the term assignments that another indexer has made, it is sensible to assume that a search term recommender system will not exceed these numbers. Even when the search term recommender predicts the exact same controlled vocabulary terms as the human indexer has assigned and therefore achieves a high performance evaluation, another human indexer would probably suggest slightly different controlled vocabulary terms and evaluate the search term recommender's suggestions differently. With these caveats in mind, the precision and recall numbers calculated in this chapter should serve as a comparison tool (between two different automatic classification applications) but do not necessarily represent a realistic picture of the predictive power of the search term recommender methodology if evaluated by several users.

7.2 Specialty vs. General Search Term Recommenders

Most bibliographic databases cover a whole range of academic fields. A search term recommender system focusing on specialties will divide the document space of the database into (potentially overlapping) specialty areas and draw on the specific vocabulary of the specialty collections to make controlled vocabulary search term suggestions. The focus on specialty vocabulary spaces is based on the assumption that term suggestions by specialty search term recommenders will be more precise and specific than that of a search term recommender based on the general collection. This assumption was tested on two search term recommender systems for the Inspec and Ohsumed collections.

For the Inspec and Ohsumed test databases, the classification performance of specialty search term recommenders was compared with the performance of a general search term recommender trained on the general collection (documents from all specialties). The individual performance of every specialty STR versus the general STR was analyzed and the results were then averaged over all specialties. The specialty collections, on which the specialty search term recommenders were based, were determined as described in chapter Six (3 for the Inspec and 33 for the Ohsumed collection).

For every specialty identified in the two bibliographic test databases, a number of test documents were extracted. The test documents were not included in the search term recommender training specialty collection (used for STR creation). The number of test

documents was roughly 10% of the number of documents in the specialty collection (see table A1 for the Inspec specialties and table B1 for the Ohsumed specialties). Due to the different sizes of the specialty collections in the Inspec and Ohsumed collections, the number of test documents used for evaluation was very different. The average number of test documents per specialty in the Inspec collection was 14,245, whereas the average number of test documents in the Ohsumed collection was 568. In total, 42,735 test documents were extracted for the 3 Inspec specialties and a total of 18,733 test documents were extracted for the 33 Ohsumed specialties.

Recall and precision values for several cut-off levels were calculated. For the Ohsumed collection, values for cut-off levels between 1 and 10 suggested Mesh headings were analyzed. The Inspec collection contains documents with a higher average number of controlled vocabulary terms (7 Inspec descriptors versus 3 Mesh headings per document), so the cut-off levels for the Inspec collections were set from 1 to 15 suggested Inspec descriptors.

7.2.1 Inspec

Table 10 shows the recall and precision values at 15 cut-off levels for the specialty and general search term recommenders for the automatic classification exercise in the Inspec collection. The values for the specialty search term recommenders are averaged over all three specialties (for individual specialty STR comparisons, see tables A4 and A5).

On average, the specialty search term recommenders improve over the general search term recommender with respect to suggesting correct controlled vocabulary terms. Recall (hitrate of prediction) improves by 11.1% and precision (accuracy of prediction) also improves by 9.4%.

Cut-off level	Recall SSTR	Recall GSTR	Precision SSTR	Precision GSTR
1	0.0812	0.0709	0.4288	0.3820
2	0.1368	0.1200	0.3672	0.3302
3	0.1777	0.1582	0.3230	0.2932
4	0.2111	0.1893	0.2907	0.2653
5	0.2393	0.2154	0.2652	0.2428
6	0.2635	0.2375	0.2446	0.2240
7	0.2853	0.2566	0.2280	0.2084
8	0.3045	0.2745	0.2137	0.1955
9	0.3221	0.2912	0.2016	0.1848
10	0.3378	0.3061	0.1909	0.1752
11	0.3520	0.3195	0.1813	0.1666
12	0.3651	0.3320	0.1729	0.1590
13	0.3772	0.3439	0.1652	0.1523
14	0.3886	0.3549	0.1584	0.1463
15	0.3989	0.3650	0.1521	0.1408

Table 10. Recall and precision at each cut-off level between 1 and 15 Inspec descriptors for specialty (SSTR) and general (GSTR) search term recommender in the Inspec collection (averaged over 3 specialties).

As figure 5 shows, the specialty search term recommenders will have found as many correct controlled vocabulary terms at a cut-off level of 12 as the general STR at a cut-off level of 15. The improvement is especially marked within the first three predicted controlled vocabulary terms: recall improves over the general STR by 13.6% and precision by 11.2%. This shows that a specialty search term recommender will not only predict more correct controlled vocabulary terms but it will also be more precise in the

highly-ranked predictions than a general STR. In an interactive search environment where the searcher will choose from a list of suggested terms, it is especially important to be precise in the upper regions of a result list. The specialty STRs show potential to do so.

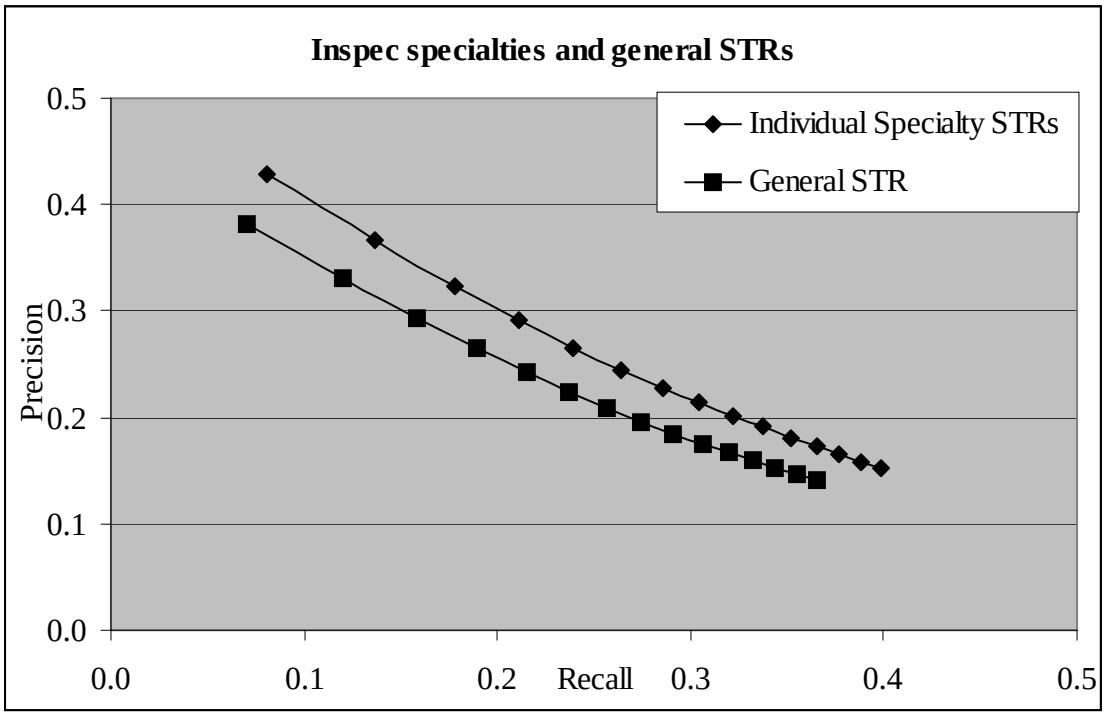

Figure 5. Recall and precision at 15 cut-off levels for specialty (SSTR) and general (GSTR) search term recommenders (averaged over 3 specialties) in the Inspec collection.

The higher precision of the specialty STR in controlled vocabulary term suggestion can also be shown for the combined recall/precision measure at the oracle level (improvement of 10.6%) and especially in the number of perfect classifications as shown in table 11 (for individual specialty STRs, compare table A6). The specialty search term recommenders suggest all the correct controlled vocabulary terms 40% more often than the general STR.

	Oracle Recall / Precision	% Perfect Classifications
Specialty STRs	0.2477	0.60%
General STRs	0.2240	0.43%

Table 11. Oracle recall/precision and percentage of perfect classifications for the specialty (SSTR) and general (GSTR) search term recommenders in the Inspec collection. At the oracle cut-off level, the cut-off level was set at the number of original Inspec descriptors predicted for each document (averaged over 3 specialties).

The performance of all three specialty STRs is very similar, however the Computers & Control specialty STR predicts slightly more correct controlled vocabulary terms (higher recall) than the other two STRs. As the vocabulary analysis in chapter Six has shown, the specialty collection for Computers & Control not only contains fewer documents than the other two specialties but also fewer controlled vocabulary terms (74% of the total terms in this specialty versus 89% and 88% in the other two). The fewer controlled vocabulary terms are assigned in the collection, the fewer controlled vocabulary terms are available to choose from in the prediction stage, which will make association decisions easier. However, the number of documents in the specialty collection should also affect the STR effectiveness. More documents will provide more training data, which should make the prediction better. This relationship will probably counterbalance the effect of the fewer controlled vocabulary terms for prediction effectiveness.

7.2.2 Ohsumed

Table 12 shows the recall and precision values at 10 cut-off levels for the specialty and general search term recommenders for the automatic classification exercise

in the Ohsumed collection. The values for the specialty search term recommenders are averaged over all 33 specialties (see table B4 for individual specialty STR performances at the cut-off level of 10).

Cut-off level	Recall SSTR	Recall GSTR	Precision SSTR	Precision GSTR
1	0.2307	0.1789	0.6031	0.4717
2	0.3519	0.2786	0.4780	0.3791
3	0.4281	0.3485	0.3969	0.3218
4	0.4820	0.4026	0.3398	0.2816
5	0.5198	0.4439	0.2963	0.2508
6	0.5491	0.4783	0.2629	0.2267
7	0.5724	0.5075	0.2364	0.2070
8	0.5919	0.5314	0.2149	0.1905
9	0.6087	0.5529	0.1973	0.1768
10	0.6229	0.5709	0.1823	0.1648

Table 12. Recall and precision at each cut-off level between 1 and 10 Mesh headings for specialty (SSTR) and general (GSTR) search term recommenders in the Ohsumed collection (averaged over 33 specialties).

Because the average number of assigned controlled vocabulary terms is lower than in the Inspec collection (3 versus 7), especially the recall but also precision values at the various cut-off levels can be expected to be higher. The results show that the improvement of the specialty search term recommenders over the general search term recommender is much higher in this collection. The specialty STRs achieve a 17.1% better recall rate and a 18.1% better precision rate. As with the Inspec collection, the specialty STRs show a more marked improvement in the highest ranked predictions: for the three first suggested controlled vocabulary terms, the improvement over the general STR is 26% for recall and 25.6% for precision. Figure 6 shows that the specialty STRs

will have found as many correct controlled vocabulary terms at a cut-off level of 7 as the general STR at a cut-off level of 10.

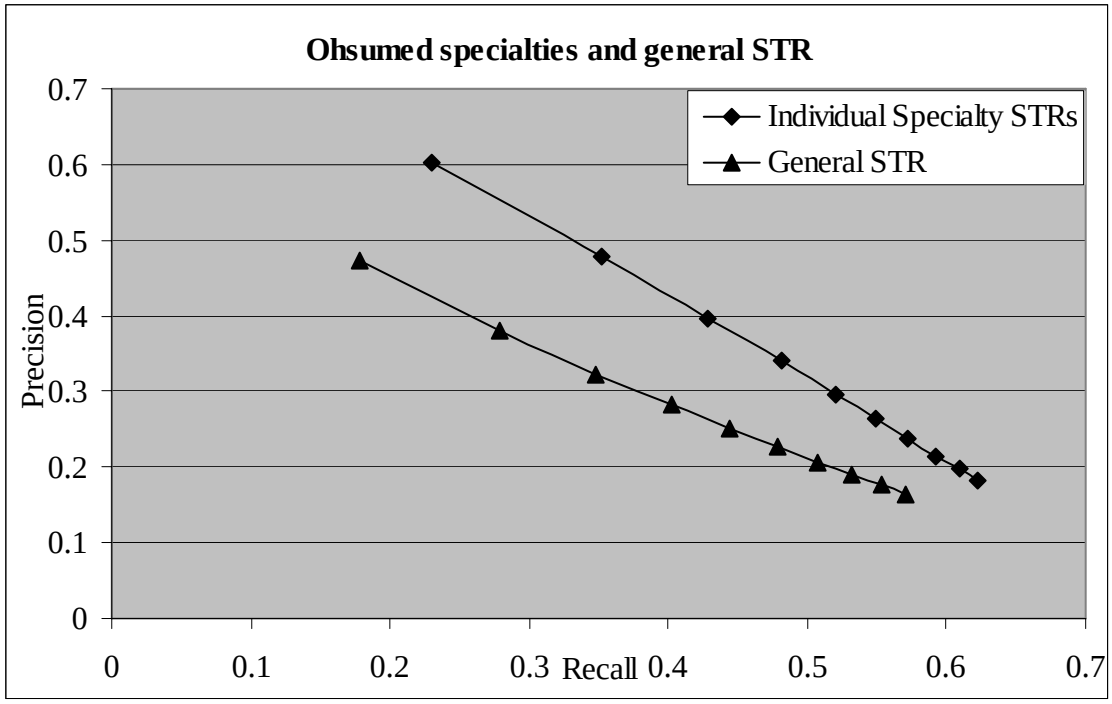

Figure 6. Recall and precision at 10 cut-off levels for specialty (SSTR) and general (GSTR) search term recommenders (averaged over 33 specialties) in the Ohsumed collection.

The oracle combined recall/precision measure and the percentage of perfect classifications presented in table 13 also show the higher improvement of the specialty STRs compared to the general STR in the Ohsumed collection for individual specialties, see tables B5 and B6). The oracle recall/precision rate improves by 24.4% for the specialty STRs and the specialty STRs will achieve 61.1% more perfect controlled vocabulary term predictions than the general STR.

	Oracle Recall / Precision	% Perfect Classifications
Specialty STRs	0.4151	10.26%
General STRs	0.3336	6.37%

Table 13. Oracle recall/precision and percentage of perfect classifications for the specialty (SSTR) and general (GSTR) search term recommenders in the Ohsumed collection. At the oracle cut-off level, the cut-off level was set at the number of original Mesh headings predicted for each document (averaged over 33 specialties).

Unlike the three specialty STRs in the Inspec collection, the 33 Ohsumed specialty STRs achieve different prediction performances. Whereas the average is 0.4151, the lowest oracle recall/precision rate is achieved by the Nursing specialty (0.2687), whereas the highest is achieved by the Psychiatry specialty STR (0.5308). Psychiatry also has the highest number of perfect classifications with 17.8% of its predictions, whereas Endocrinology has the lowest with 4.5% (see tables B5, B6).

With more specialty STR performances to compare, one can look at the impact of the number of documents, terms and controlled vocabulary terms in the specialty collection on the STR performance. It is hypothesized that the number of documents will positively influence the performance (the more training documents, the better the performance), whereas the number of Mesh headings will influence it negatively (more Mesh headings available in a specialty will make the prediction more difficult).

For the purposes of analysis, the 33 specialty collections were divided into two groups (determined by the median): the group of specialties that contained relatively many training documents and a group with relatively fewer numbers of documents. The averaged oracle recall/precision rates over all specialties in each group were compared. As predicted, the collections with more training documents achieve a 5.4% higher

recall/precision rate than the specialty STRs with relatively fewer training documents. This relationship can also be shown for the number of unique terms occurring in the collections: the group with relatively more terms to train on for STR construction achieves a 4.6% higher recall/precision oracle rate than the collections with relatively fewer terms. However, the relationship between the number of Mesh headings in the collection and the STR performance is not so clear. The group of specialty collections with relatively fewer Mesh headings to choose from achieves a higher oracle recall/precision rate as predicted, but the improvement is marginal (1.3%).

It can be assumed that collections with more documents and terms will also contain more Mesh headings, so the effects of these factors might counterbalance each other. A third comparison including both the number of documents or terms and the number of Mesh headings was therefore devised. The ratio between the number of training documents or terms and Mesh headings was calculated for each collection, showing the variation of Mesh headings in the collection. If there are more documents or terms per Mesh heading in the collection, the Mesh heading selection or assignment is not as varied. The collections were again divided into two groups, one with more Mesh heading variation per document or terms and one with relatively less variation. The comparison of the oracle recall/precision values shows that less Mesh heading variation improves the performance of the STR as predicted: the group with less Mesh heading variation in the specialty collection achieves more than a 9% better oracle rate than the group with more Mesh heading variation.

7.2.3 Summary

The evaluation of the search term recommenders for the Inspec and Ohsumed collection has shown that focusing on smaller fields within information collections can achieve a better performance. For the automatic classification exercise, the general performance (measured in recall and precision) could be improved by about 10% in the Inspec collection and about 25% in the Ohsumed collection.

The Inspec classification exercise was more difficult because more original controlled vocabulary terms were assigned to a document and had to be predicted correctly. Additionally, the specialties selected in the collection were broader in scope and overlapping, which made the association weight calculation less decisive. This is probably the reason why the Ohsumed search term recommender system performed better overall.

It was also suggested that the number of controlled vocabulary terms and the number of documents and terms available for training influence the performance of the search term recommender. More training documents and terms generally improve the performance of the search term recommender, whereas more controlled vocabulary terms (i.e. a bigger variation of controlled vocabulary terms in the collection) make the prediction and selection stage harder.

7.3 Specificity of Specialty Search Term Recommenders

If focusing on the specialty vocabulary will improve the suggestions of the search term recommender system, the question arises how focused or specialized a search term recommender should be to achieve optimal performance. How specific can a specialty be described for specialty selection for the STR construction process? And when is specific too specific?

The effects of serendipity in search cannot be overlooked. Overly specialized search spaces (specialty collections) might restrict the searcher's perspective too much to offer a reasonable overlook of a subject area. Similarly, too specialized search term recommenders might not offer optimal search term suggestions or confuse the searcher by offering too many different access routes into a collection.

The four specialty selection methods introduced in chapter Six can be used to define specialties of various sizes and specificity, but at some point they will cease to delimit the boundaries of a specialty and become overly selective. For example, the grouping of specialties by publication source can only go as deep as an individual journal and the grouping by classification can only go as specific as the classification hierarchy allows.

The specificity of the specialty itself is ultimately a function of the discourse domain. Specialties cannot be identified arbitrarily on a continuous line of more and more specific discourse areas. Rather, they are socially-formed clusters of various specificity levels (i.e. the discipline, the research area, the sub-specialty, the work group etc.). If

those levels are not reflected in the specialty determination mechanism (e.g. the different levels of the hierarchy do not represent the different levels of specificity), then using those methods to define more specific specialties will be unrealistic and, probably, misleading..

7.3.1 Training Collection Size

The more specialized a specialty is, the fewer documents from a collection can be categorized into this area and consequently the training set for the specialty STR construction becomes smaller.

Small training sets can influence the statistical reliability of the association weight calculation. If there are not enough data available to train the STR correctly, the statistically calculated term associations might be skewed. Small training sets could also give a skewed distribution of controlled vocabulary terms available and might not represent a realistic frequency distribution of the terms in the collection. Research has shown that the accuracy for automatic classification applications commonly increases with an increasing number of training data, however with diminishing returns after a threshold (e.g. Ginsparg et al., 2004; Frank & Paynter, 2004). This was also confirmed by the Ohsumed experiments described in the previous section.

The fewer documents in a specialty collection, the fewer terms from the specialty vocabulary are represented. If a user searches for a term not represented in the collection, the search term recommender cannot recommend controlled vocabulary terms because the search term was not available for the association process. This problem should

virtually disappear if sufficiently large and current document collections can be used for specialty STR constructions. Until then, the requirement of a sufficient number of training documents available for STR construction serves as a lower boundary for how narrow specialized search term recommenders can be.

7.3.2 Specificity of Specialties

In this section, a number of experiments are presented that test the performance of specialty search term recommenders at different levels of specificity. In the previous section is has already been shown that specialty search term recommenders achieve a better performance than a general search term recommender in an automatic classification exercise. The focus of this analysis is on the specificity of the specialty STR.

For this purpose, more specialties were determined for STR construction and evaluation. The new specialty search term recommenders represent specialties at a more specific or focused level than the original specialty STRs. The selection followed the same strategies as the original determination methodology for the two experimental collections. For the Inspec collection, more specific specialty collections were determined according to more specific classification categories (following the hierarchical structure of the classification). For the Ohsumed collection, one publication source (i.e. one journal) from the specialty collection was selected for more specific specialty STR construction.

Every new specialty collection is a subset of a broader one with the general collection as the upper bound. This allows for a comparison of specificity along two criteria (classification hierarchy and publication source). For evaluation of the different specialty STRs, new test documents were extracted from the most specific specialty collection (they were not used for constructing the STRs). The prediction power of each specialty STR was tested with this test set, measuring the performance of the STR at each level of specificity.

7.3.2.1 Inspec

For the Inspec collection, the performance of the search term recommender system was tested at four levels of specificity in three different research fields. In addition to the general Inspec STR and the three specialty STRs for Physics, Electrical Engineering and Computers & Control, six more specialized STRs were created following the categorization system of the Inspec classification (two more for each field). The following specialty search term recommenders were created:

> (1) a Nuclear Physics STR trained on documents with assigned classification codes in the A2 category (one level more specific than the Physics specialty);
> (2) a Nuclear Structure STR trained on documents with assigned classification codes in the A21 category (two levels more specific than the Physics specialty);
> (3) a Components, Electron Devices and Materials STR trained on documents with assigned classification codes in the B2 category (one level more specific than the Electrical Engineering specialty);
> (4) a Passive Circuit Components STR trained on documents with assigned classification codes in the B21 category (two levels more specific than the Electrical Engineering specialty);
> (5) a Computer Hardware STR trained on documents with assigned classification codes in the C5 category (one level more specific than the Computers & Control specialty); and

(6) a Circuits & Devices STR trained on documents with assigned classification codes in the C51 category (two levels more specific than the Computers & Control specialty).

Table A1 provides an overview over the collection and test document numbers for each of these search term recommenders. The test documents were extracted from the three most specific specialties: Nuclear Structure, Passive Circuit Components and Circuits & Devices. The specificity level's effect on performance was tested on a total of 2,425 documents. Recall and precision values at 15 cut-off levels for four different levels of specificity were calculated: the general STR, the specialty STRs used for performance evaluation in the previous section, the "sub-specialty" STR (denoting the STR that is one level more specific than the specialty STR) and the "sub-sub-specialty" STR (denoting the STR that is two levels more specific than the specialty STR).

Table 14 shows the recall and precision levels at 4 cut-off levels for the different search term recommenders (averaged over all three Inspec fields). For individual performances, see tables A7-A10.

Cut-off level	Sub-sub-specialty STR	Sub-specialty STR	Specialty STR	General STR
Recall				
1	0.1279	0.0949	0.0512	0.0417
7	0.3997	0.3465	0.2243	0.1865
10	0.4650	0.4196	0.2764	0.2330
15	0.5337	0.4985	0.3454	0.2910
Precision				
1	0.6547	0.4995	0.2853	0.2408
7	0.3216	0.2764	0.1845	0.1564
10	0.2656	0.2372	0.1596	0.1361
15	0.2072	0.1915	0.1334	0.1133
Oracle Recall / Precision				
	0.3626	0.2987	0.1900	0.1573
% Perfect classifications				
	2.19%	0.94%	0.26%	0.25%

Table 14. Average recall and precision at 4 cut-off levels, recall/precision at the oracle cut-off level and percentage of perfect classifications for four levels of specificity in the Inspec collection (averaged over 3 specialties).

In figure 7, the performance at several levels of specificity can be seen more clearly. The improvement between the general STR and the specialty STR is a little higher in this experiment than in the previous experiments (which evaluated the performance of these two STRs on more test cases): the specialty STR improves over the general STR with circa 20% in recall and 17.6% in precision. However, the search term recommenders trained on more specific specialty collections improve even on the specialty STR.

The biggest performance jump was achieved when training on a specialty collection that was one level more specific than the Physics, Electrical Engineering or Computers & Control collections. Here, recall and precision improved on the specialty STR by over half (55.8% in recall and 51.3% in precision).

Again, the performance improvement is especially marked in the first three ranked controlled vocabulary terms: the sub-specialty STRs improve on the specialty STRs by 72.6% in recall and 63.3% in precision. Compared to a general STR trained on the whole collection, the prediction power of these STRs doubles (a 111% improvement in recall and a 92.3% improvement in precision).

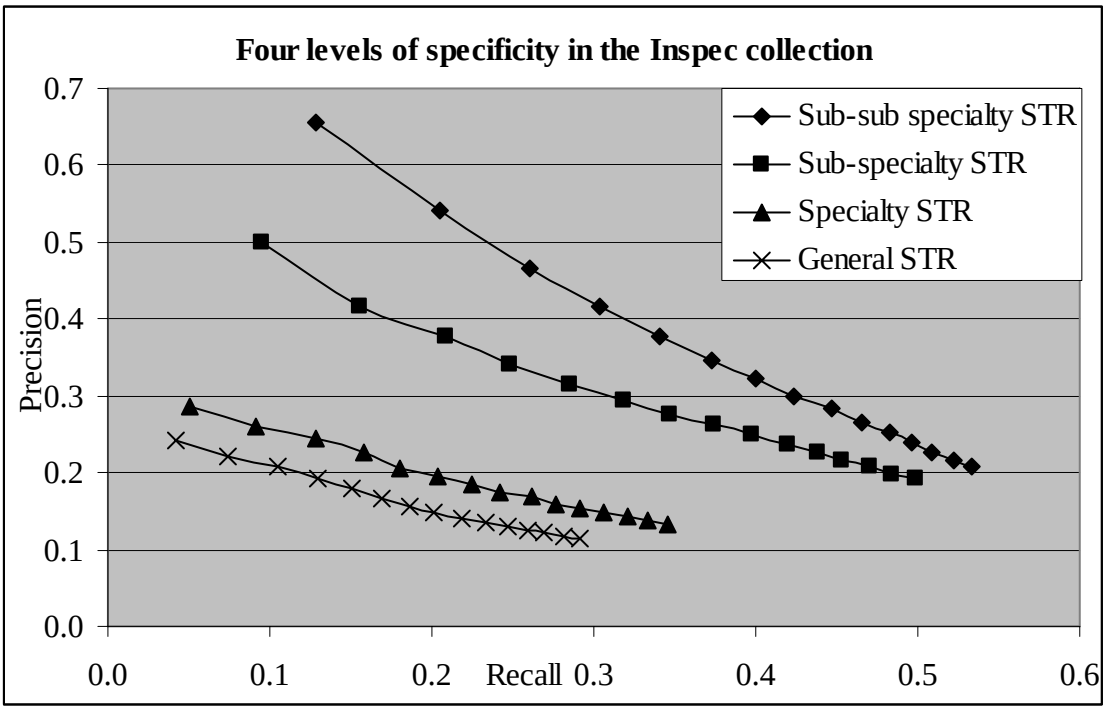

Figure 7. Recall and precision at 15 cut-off levels for four levels of specificity in the Inspec collection (averaged over 3 specialties).

Increasing the specificity even further improves the performance of the STR but not by as much as can be seen in figure 7 (improvement over the sub-specialty STR is circa 16% in precision and recall). This result indicates that more specific search term recommenders will achieve better performance in their specific sub-fields but with diminishing returns.

The individual performances for the fields of Physics, Electrical Engineering and Computers & Control (figures A1-A3) show more clearly how the effectiveness of the search term recommenders at different levels of specificity varies according to collection-specific characteristic. Whereas increasing the specificity in the Physics collection does not seem to make a difference at the second and third level, for Electrical Engineering, going as specific as the third level (sub-sub specialty) clearly increases the performance of the STR over all others.

7.3.2.2 Ohsumed

For the Ohsumed collection, three levels of specificity for search term recommender construction were evaluated. For performance evaluation, three specialties (which were also used for vocabulary analysis in chapter Six) were selected: Communicable Diseases, Gynecology, and Orthopedics. For each specialty, the performance of the general STR, the specialty STR and a sub-specialty STR was compared. The more specific search term recommenders were created based on documents that were published in one journal within the broader specialties of Communicable Diseases, Gynecology and Orthopedics. The following sub-specialty search term recommenders were created:

> (1) an Infectious Diseases STR based on documents referencing articles published in the “The Journal of Infectious Diseases” (one level more specific than the Communicable Diseases specialty);
> (2) an Obstretrics STR based on documents referencing articles published in the “Obstetrics & Gynecology” journal (one level more specific than the Gynecology specialty); and

(3) a Clinical Orthopedics STR based on documents referencing articles published in the "Clinical Orthopaedics & Related Research" journal (one level more specific than the Orthopedics specialty).

Test documents were extracted from the most specific specialty collections (745 test documents total) and the automatic classification performance of each STR was evaluated. For an overview of the collection and test document numbers for the Ohsumed specificity experiments, see table B7. Table 15 shows the recall and precision levels at 3 cut-off levels for search term recommenders at three levels of specificity in the Ohsumed collection (for individual performances, see tables B8-B11).

Cut-off level	Journal STR	Specialty STR	General STR
Recall			
1	0.2236	0.2192	0.1599
3	0.4436	0.4373	0.3233
10	0.6411	0.6622	0.5745
Precision			
1	0.6068	0.5934	0.4396
3	0.4230	0.4171	0.3092
10	0.1923	0.1994	0.1718
Oracle Recall / Precision			
	0.4317	0.4274	0.3228
% Perfect classifications			
	9.07%	8.97%	4.31%

Table 15. Average recall and precision at 3 cut-off levels, recall/precision at the oracle cut-off level and percentage of perfect classifications for three levels of specificity in the Ohsumed collection (averaged over 3 specialties).

The graphical representation of the recall and precision levels in figure 8 (and figures B1-B3 for the individual fields) shows more clearly that increasing the specificity for the Ohsumed search term recommender system does not achieve a better performance in the automatic classification exercise. The improvement rate for the specialty STR over

the general STR with this new set of test documents is similar to the one reported in the previous experiment with about 25% for recall and precision.

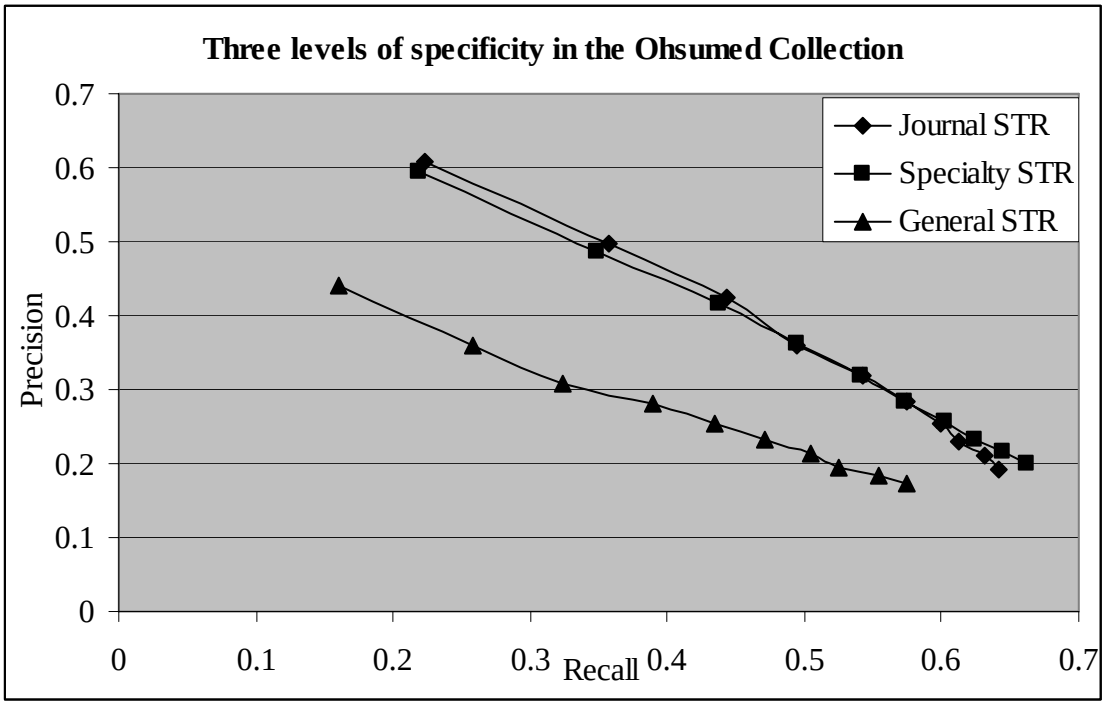

Figure 8. Recall and precision at 10 cut-off levels for three levels of specificity in the Ohsumed collection (averaged over 3 specialties).

The sub-specialty STR trained on documents from one journal did not improve over the original specialty STR for the three fields analyzed. Although the individual performances show a very slight improvement for the journal STR, averaged over all three fields, no improvement can be detected except in the first suggestions. For the first three suggested controlled vocabulary terms, the precision and recall rates of the Journal STR improve over the broader specialty STR by a negligible 2%.

7.3.3 Summary

At which level of specificity a search term recommender should be constructed will be determined by collection-specific heuristics and practical training set size limitations. For the Inspec collection, it was shown that more specific search term recommenders than the three relatively broadly constructed STRs for Physics, Electrical Engineering and Computers & Control will improve performance in an automatic classification exercise, however higher levels of specificity achieve diminishing returns.

For the Ohsumed collection, it was shown that increasing the level of specificity past the 33 specialty STRs already constructed will not improve performance. This result is not surprising because the 33 specialties determined at the first level of specificity already represent a much more focused search space than the three specialties in the Inspec collection.

Furthermore, the individual journals selected to represent sub-specialties might be more specific when it comes to their publication source but not necessarily more specific in their coverage of the discourse within the specialty. In that case, using journals for STR construction might not increase the specificity in the vocabulary. This would explain its lack of improvement in prediction power. A better performance for the search term recommender is achieved when the more specific specialty collections contains natural term–controlled vocabulary term associations that were not apparent in the broader specialty selection. Using the documents from one publication source (e.g. the Obstetrics & Gynecology journal) might not manifest any more specific term associations than were already determined in the specialty (e.g. Gynecology).

Even though the methodologies for selecting specialties within a collection (described in chapter Six) are targeted towards finding specialties with different vocabularies, when taken to an extreme, they might not produce specialty collections with a more focused vocabulary – the most important feature for the search term recommenders to work. In a practical application of the search term recommender methodology, an analysis of the vocabulary at different levels of specificity is necessary to determine the appropriate level for STR construction.

7.4 Conclusion

The experiments described in this chapter have shown that focusing on the specialty vocabulary for search term suggestion will increase the preciseness of the suggestion. An automatic classification exercise has demonstrated that substantial improvements in hitrate (recall) and accuracy (precision) can be achieved by specialty search term recommenders compared to general search term recommenders in two test collections.

An analysis of the specificity level appropriate for search term recommender construction has suggested that the method of determining the specialty, the number of documents available to train the search term recommender and vocabulary differences in the different levels of specificity determine which level of specificity is appropriate for optimal performance in term suggestion.

The specificity level also determines the number of different specialty search term recommenders constructed for an information collection. In the search process, a large number of specialty search term recommenders to choose from for vocabulary support might confuse the searcher. If the appropriate specialty search term recommenders are selected automatically, a large selection to choose from increases the difficulty of the selection process. The selection of search term recommenders in the search process is the focus of the next chapter.

Chapter 8

Selection of Specialty Search Term Recommenders in the Search Process

The use of specialty search term recommenders for vocabulary support adds more steps to the search process. After entering the original search statement, a searcher utilizing the search term recommender system can review suggested terms and add or substitute them in the query before retrieving documents. When several specialty search term recommenders are provided, there is also the additional prior step of deciding which STR, if any, to use for a given query or for a search session. The selection of appropriate search term recommenders is dependent on the purpose of the search (exploratory or

specific), the specialty focus of the searcher and the manner of selection (manual, interactive or automatic).

If a search is targeted towards exploring an information space, a specialty search term recommender system can provide insight into the different domains of discourse (i.e. specialties) covered in the collection. When a searcher enters a query, the various specialty STRs can display the different controlled vocabulary terms associated with the topic (represented by the search statement). Every specialty STR in effect represents a different context or discussion perspective for a topic, which is expressed in the different controlled vocabulary terms that are associated with it in each specialty. As an exploratory tool, the search term recommender system can be used to find discourse areas that are related (e.g. discussing a topic similarly as reflected by the associated controlled vocabulary terms) or to show the number of different discourses around a topic (as represented by the specialties that suggest distinct controlled vocabulary terms). In this case, the specialty search term recommenders best representing different or related discourses need to be selected.

If a search is more focused and the searcher only wants to spend a very limited amount of time and effort on query formulation, the selection of the specialty STR that will predict the “best” vocabulary terms for the search becomes very important. In this case, the search term recommender system needs to focus on providing targeted help in selecting one or few specialty search term recommenders most appropriate for a searcher’s query.

This chapter will discuss the selection process for specialty search term recommenders in search and describe some preliminary experiments with an automatic selection system.

8.1 Manual and Interactive Specialty Search Term Recommender Selection

Information retrieval with search term recommenders as vocabulary support technique introduces another step in the search process, which might also introduce new errors. The search term recommender system could suggest unsuitable search terms, lead the searcher in the wrong direction (away from his original focus) or cause frustration because of the additional time and effort required to reformulate the search. The most difficult problem occurs when the specialties determined in the construction process do not match the searchers concept space. In this case, the vocabulary support in the search process might create confusion for the searcher (because of the cognitive mismatch) instead of alleviating it.

To avoid adding another search process step for the searcher, the selection of specialty search term recommenders and the query reformulation could be automated according to some parameters. However, automating this selection and reformulation process (and thereby black-boxing the association process with the controlled vocabulary) undermines the original purpose of the search term recommender technique, which is to provide more transparency and insight into the information system.

Automatically selecting a specialty may also introduce more probability for errors. The association between the terms in the query and the controlled vocabulary of the collection is a statistically based process with an error rate for suggestion that is collection-dependent (see experiments in chapter Seven). If the search term recommenders used for vocabulary support are automatically selected, this process adds another layer of difficulty and increases the margin of error.

This additional error margin can be avoided if the selection process is left to the user of the system. Because the specialties are predetermined in the construction process, the search term recommender system for a particular information collection can provide the searcher with a list of specialties to choose from if vocabulary support is wanted. This manual selection process will not frustrate more focused searchers who would want to avoid the vocabulary support phase altogether (e.g. because they know the vocabulary of the system, have a simple and specific question or not enough time).

If many specialty search term recommenders exist for a given information collection (in the Ohsumed collection, 33 specialties were determined; the specificity experiments for the Inspec collection showed that specialties more specific than the division into three fields were preferable), then the manual selection can become cumbersome for the searcher and hard to present for the information system interface.

A solution for this problem is to introduce an interactive search term recommender selection process. An interactive algorithm would suggest a ranked list of specialty search term recommenders based on the original query (retrieved through an automated process) and then present it to the user for a final choice. This would reduce

the search space but still provide enough context for the user to make an informed decision about the choice of vocabulary support used. For any practical application of the search term recommender technique, this is the recommended solution.

8.2 Automatic Specialty Search Term Recommender Selection

The ideal application of a specialty search term recommender system is the automatic, dynamic and on-the-fly creation of a specialty search term recommender according to the domain of discourse represented by the searcher's question and the use of this STR for vocabulary support for this particular subject area. This solution, if achievable, would avoid the necessity of selection an appropriate specialty STR altogether.

If the searcher wants focused and quick vocabulary support without adding the selection step to the search process, the search term recommender system could select the most-fitting specialty search term recommender automatically for the searcher. Given the lack of information about a searcher's context in the search process (only a few search terms are input), this approach seems not only difficult to implement but also unadvisable (see chapter Five).

Nevertheless, automatically selecting the specialty STR for vocabulary support eases the burden of choice for the user and does not require extra effort from the searcher in the search process. The rest of the chapter will describe some ideas and preliminary experiments on how to achieve this.

The problem of automatically choosing the correct specialty search term recommender (from query terms) can be interpreted as a routing problem in text categorization (given a query, predict the correct specialty instead of controlled vocabulary terms) or as a special case of a distributed collection selection problem.

French et al. (2002) already studied collection selection algorithms together with specialty STRs for the Ohsumed test collection. Several more collection selection algorithms could be suitable for this kind of task (see French et al. (2002) for a listing).

The other area of interest in this context is automatic text classification. The problem of classifying a query into a specialty is not so different from predicting controlled vocabulary terms for a query. However, given the smaller numbers of categories (fewer specialties than controlled vocabulary terms) and the larger amount of training data per category (the whole text collection), different automatic classification algorithms should be tested for this problem.

Because the search term recommender technique can be used for automatic classification, one could imagine using the same statistical methods used in our search term recommender technology for the categorization problem in selecting specialty STRs. Three different methods for using the search term recommender technique in the selection process can be envisioned.

8.3 Variation in Controlled Vocabulary Term Suggestions

The first method for the STR selection process is using the variation in controlled vocabulary term suggestions expected from the different specialty search term recommenders. It was hypothesized earlier that differences in the vocabulary in the specialties lead to different controlled vocabulary term suggestions, therefore making the search term recommendations more precise for a given specialty.

These differences in term suggestions can be used for the specialty STR selection process by presenting those specialties to the user, which represent the most distinct vocabulary spaces for a given topic. This should be advantageous for an exploratory search application where the search term recommender system provides the searcher with an overview over the different contexts (i.e. specialties) in the collection.

Practically, the search statement will be sent to each specialty STR and the controlled vocabulary term suggestions will be analyzed. Those specialty STRs that suggest controlled vocabulary terms that haven’t been suggested by others should be presented to the searcher. The selection process should emphasize finding those specialties that are most distinct from each other.

A brief analysis of the Inspec and Ohsumed search term recommender systems shows that differences in controlled vocabulary suggestions can be detected.

8.3.1 Inspec

For the Inspec collection, 100 random test documents were extracted from each of the specialties in Physics, Electrical Engineering and Computers & Control, which were used in the automatic classification exercise. The 300 documents (titles) were searched against each specialty search term recommender and the overlap between suggested controlled vocabulary terms was analyzed. For every specialty search term recommender, the first seven suggested controlled vocabulary terms (average number of assigned Inspec descriptors for a document) were looked at.

Table 16 shows the number of unique controlled vocabulary terms suggested out of a pool of 21 suggestions from the three specialty search term recommenders.

Number of unique suggested terms	**Queries**	
10	4	1.33%
11	6	2.00%
12	16	5.33%
13	24	8.00%
14	44	14.67%
15	56	18.67%
16	47	15.67%
17	43	14.33%
18	31	10.33%
19	18	6.00%
20	11	3.67%
Average number of unique terms: 15.6		

Table 16. Variations in controlled vocabulary term suggestion (7 terms are suggested) in-between three Inspec specialty STRs (averaged over 300 queries).

Because each STR suggests seven controlled vocabulary terms, seven unique terms in the pool would indicate complete overlap between the specialty STR

suggestions, that is each specialty STR suggests exactly the same controlled vocabulary terms. Accordingly, 21 unique suggestions would indicate that every STR suggests completely different terms. Table 16 shows that there is neither complete overlap nor complete distinctiveness. For this test set in the Inspec collection, about two thirds of the suggested controlled vocabulary terms are unique, providing enough distinctiveness to show different contexts in the specialties.

Because the Inspec collection had only three specialties determined for different search term recommenders, the overlap between the specialty STR's controlled vocabulary term suggestions can be represented graphically. Figure 9 shows the overlap between the three specialties in controlled vocabulary term suggestion. The figure shows the similarities in controlled vocabulary term suggestion patterns and vocabulary differences in the specialty collections (as described in chapter Six). The Physics and Computers & Control specialties will not suggest as many similar controlled vocabulary terms as the Physics and Electrical Engineering or the Computers & Control and Electrical Engineering specialties. This supports the hypothesis that differences in vocabulary will lead to different term suggestions (chapter Six showed that Physics and Computers & Control also had a more distinct vocabulary, whereas Electrical Engineering functioned as a bridging specialty).

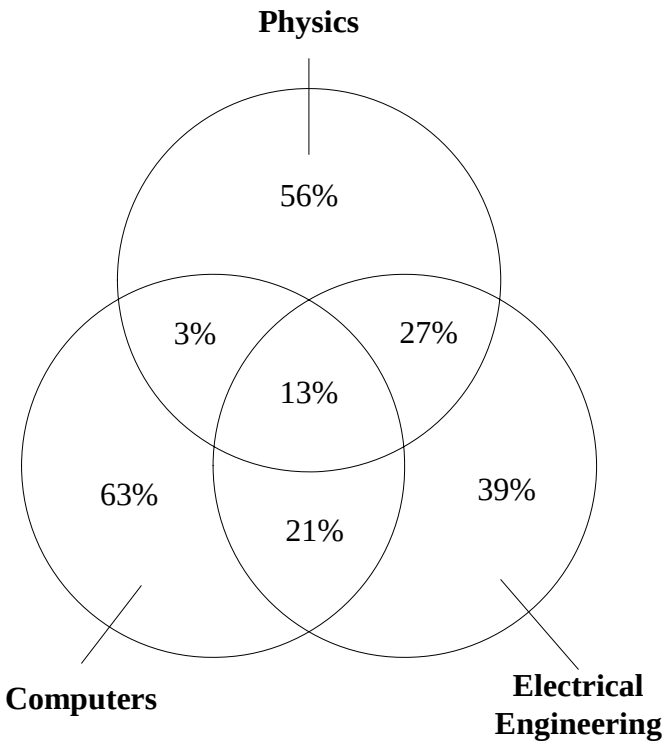

Figure 9. Overlap in controlled vocabulary term suggestion between three Inspec specialty STRs (averaged over 300 queries). Overlap is calculated out of 7 suggested controlled vocabulary terms for each specialty STR.

8.3.2 Ohsumed

For the Ohsumed collection, 10 random test documents were extracted from each of the specialty collections used in the automatic classification exercise. The titles of the 330 test documents were sent against each of the 33 specialty search term recommenders and the controlled vocabulary suggestions analyzed. The first three suggested Mesh headings (average number of Mesh headings per Ohsumed document) were examined.

For 47 test documents, not every specialty recommender suggested three controlled vocabulary terms (e.g. because the title terms could not be found in the vocabulary used for specialty STR training or no associated controlled vocabulary terms

could be found). Table 17 shows the overlap in controlled vocabulary term suggestions for the 287 remaining test documents.

Number of unique suggested terms	**Queries**	
30-39	10	3.53%
40-49	45	15.90%
50-59	65	22.97%
60-69	88	31.10%
70-79	51	18.02%
80-89	16	5.65%
90-98	8	2.83%
Average number of unique terms: 61.9		

Table 17. Variations in controlled vocabulary term suggestion (3 terms are suggested) in-between 33 Ohsumed specialty STRs (averaged over 283 queries).

If every specialty search term recommender suggests three Mesh headings, then a total of three unique suggested controlled vocabulary terms would indicate complete overlap between all 33 specialty STRs. In contrast, a total of 99 unique suggested terms would indicate completely distinct specialties. As in the Inspec analysis, there was neither complete overlap nor complete isolation between the different specialty search term recommenders. Again two thirds of the suggested vocabulary terms were unique. A few extreme cases could be found, where over 90 unique terms were suggested. This resonates with the analysis of the vocabulary differences in the three Ohsumed specialty collections described in chapter Six. Their natural and controlled vocabularies were more distinct than in the Inspec specialties.

These analyses show that distinct specialties (with respect to their controlled vocabulary term suggestions for the same query) can be ascertained. This suggests that

the different term suggestions can be used to represent different subject areas for the STR selection process. However, if multiple distinct specialty STRs can be identified, the problem of ranking and presenting the most appropriate STRs to the user might still remain (especially if there is no overlap in the suggestions).

8.4 Predicting the Specialty Search Term Recommender

The search term recommender technique itself can be used to predict the appropriate specialty search term recommenders given a particular query in two ways. Firstly, a search term recommender can be constructed that will suggest a specialty STR instead of controlled vocabulary terms given a search statement as input. For both the Inspec and Ohsumed collections, a “Prediction Search Term Recommender” was built based on the specialties determined for the automatic classification exercise (predicting a specialty STR out of 3 specialties for Inspec and 33 for Ohsumed).

Secondly, the association weight measure can be used to predict the specialty. The association weight (described in chapter Five) calculates the degree of association between a natural language term and a controlled vocabulary term for a particular specialty. The total association weight between a query (consisting of several natural language terms) and a given controlled vocabulary term is calculated by adding the individual association weights for each query term and the controlled vocabulary term together. Because of vocabulary differences in the specialties, it can be assumed that

different controlled vocabulary terms with different association weights will be suggested for any given query.

If the association weight measures the degree of association (i.e. how closely the natural language terms and controlled vocabulary terms are related), then the degree of association between a query and a specialty search term recommender can also be determined. The degree of association between a query and a specialty search term recommender is represented by the total association weight of all controlled vocabulary terms suggested for the query (i.e. adding all association weights). The higher the total association weight, the more related the query is to a particular specialty because the query terms and controlled vocabulary terms are more strongly associated. Specialty search term recommenders can be selected by ranking the STRs according to the total association weight associated with a query. Similar to the term suggestion method for STR selection, a query would have to be searched against each specialty STR in order to determine the association weights for each specialty.

For the Inspec and Ohsumed collections, the 300 and 330 respective test documents for each collection that were used for the term suggestion variation analysis were used for experiments testing the two STR selection methods. First, the effectiveness of the two methods of specialty selection was evaluated. For every test document, the specialty suggestion from the Prediction STR and the specialty suggestion determined from the total association weight were analyzed. The selections were compared to the original specialty the test documents came from.

Using the original specialty as a measurement yardstick can only be as precise as the original specialty determination process described in chapter Six. Documents could belong to another specialty than the one to which they were originally assigned. This is another reason why the selection of the specialty STR should be an interactive process between the searcher and the search term recommender system. The hit-rate for each selection process is represented at several cut-off levels, showing at which point in a ranked list the searcher would find the original specialty.

As a preliminary experiment, we used the two selection methods to determine the specialty STR used for term suggestion for the automatic classification exercise described in chapter Seven. The results of the automatic classifications achieved with the selected specialty STRs were compared with the automatic classification results for the “perfect” specialty recommender selection (determined by the original specialty the test document came from) and the general search term recommender. The general search term recommender based on the general collection represents the choice of vocabulary support system if the best appropriate specialty search term recommender could not be determined.

8.4.1 Inspec

For Inspec, the Prediction STR predicted the specialty search term recommender for a particular query out of the three specialties Physics, Electrical Engineering, and Computers & Control. Table 18 shows the hit-rate (recall) of the Prediction STR and the total association weight methodology for suggesting the specialty based on 300 test

documents. There are only two cut-off levels because a maximum number of three specialty STRs could be predicted (at the cut-off level 3, the recall must be 1).

Cut-off level	Recall Prediction STR	Recall Association Weight
1	0.2833	0.3333
2	0.7000	0.6667

Table 18. Recall of Specialty Prediction STR and association weight calculation at 2 cut-off levels. Percentage of records where the classification-selected specialty was predicted within the first-ranked predictions.

The association weight methodology achieves better results in predicting the correct specialty STR but both methods do not achieve an effective selection performance (picking the correct specialty out of three in one third of the cases is not better than random selection). One of the reasons might be that the three specialties in the Inspec collection are overlapping and share more vocabulary compared to the Ohsumed collection making them harder to distinguish for any given query. It is possible that most of the documents could have occurred in another specialty than the one they were originally assigned.

Consequently, the performance in the automatic classification exercise is expected to be low because the automatic selection process of the specialty STR already introduced a high error. This effect might be balanced by the specialty STRs suggesting more precise controlled vocabulary terms than the general STR but aggravated by the potential mismatching of controlled vocabulary terms by another specialty STR. Table 19 shows the recall/precision value at the oracle level and the percentage of perfect classifications for the “perfect” specialty STR (determined by the original specialty the

test document occurred in), the general STR (no specialty STR was chosen), the specialty STR selected by degree of association and the specialty STR determined by the Prediction STR (for recall and precision values at 15 cut-off levels, see table A11).

	Specialty STR	**General STR**	**Association Weight**	**Prediction STR**
Oracle Recall / Precision				
	0.2638	0.2474	0.2411	0.1455
% Perfect classifications				
	0.67%	0.33%	0.67%	0.33%

Table 19. Automatic determination of the specialty in the Inspec collection for 300 queries. Oracle recall/precision and percentage of perfect classifications for Specialty STR: predicting the specialty by classification (perfect prediction); General STR: no specialty prediction; Association Weight: predicting the specialty by highest association weight (over 3 specialties); and Prediction STR: predicting the specialty with a STR for the 3 Inspec specialties.

Figure 10 shows that the Prediction STR method performs much worse in the automatic classification exercise than the other three methods. The association weight method achieves a 66% better performance than the Prediction STR method even though it only predicted the “correct” specialty STR in 17% more cases than the Prediction STR. This suggests that the association weight might indicate more about a specialty than the original specialty determination by classification code. However, even the association weight method performs only as well as the general search term recommender, demonstrating that for the Inspec collection, a specialty prediction hitrate of 33% is not precise enough.

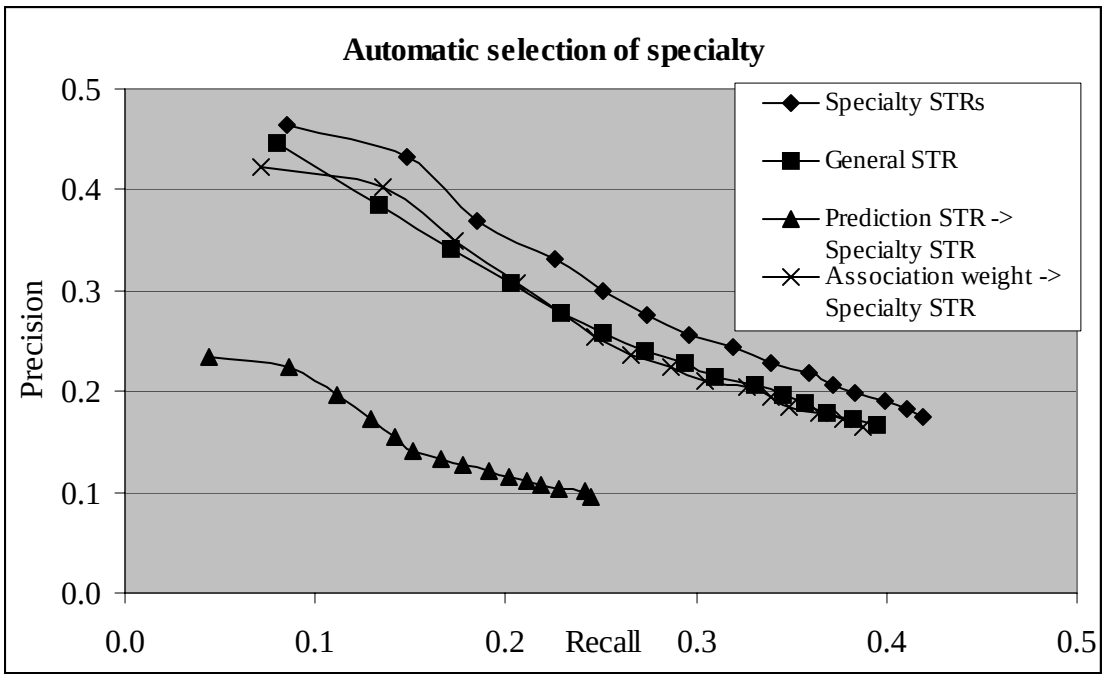

Figure 10. Automatic determination of the specialty in the Inspec collection for 300 queries. Recall and precision at 15 cut-off levels for Specialty STR: predicting the specialty by journal descriptor (perfect prediction); General STR: no specialty prediction; Association Weight: predicting the specialty by highest association weight (over all 3 specialties); and Prediction STR: predicting the specialty with a STR for the 3 Inspec specialties.

The general search term recommender performs a little better than average for the test set of 300 documents - the specialty search term recommender (perfect selection) achieves only a 6% better performance rate (compared to 10% overall in the original classification exercise). This might be a function of the randomly selected test documents not being representative of the general collection.

The specificity experiments in chapter Seven have shown that specialty search term recommenders more specific than the broad division into three Inspec specialties achieve higher performance rates than the original Inspec specialty STRs. In a second

round of experiments, the two specialty STR selection methods were therefore tested on more specific specialties.

The Prediction STR was constructed based on the sub-specialty collections introduced in chapter Seven (Nuclear Structure, Components, Electron Devices and Materials, and Computer Hardware instead of the original Physics, Electrical Engineering and Computers & Control) and the total association weight for a query was also calculated based on these sub-specialty STRs. Even though the test documents were taken from the more general specialties, this strategy improved the hitrate for the correct specialty STR selection (44% more correct specialty STR predictions for the Prediction STR and 31% more correct predictions for the association weight method) and brought the prediction rates of the two methods closer together (the association weight method is now only 6.5% better at predicting the correct specialty).

Table 20 also shows the considerable improvement of the Prediction STR method of STR selection for the automatic classification exercise.

	Specialty STR	**General STR**	**Association Weight**	**Prediction STR**
Oracle Recall / Precision				
	0.2638	0.2474	0.2441	0.2264
Perfect classifications				
	0.67%	0.33%	1.2%	1%

Table 20. Automatic determination of the specialty in the Inspec collection using sub-specialties for 300 queries. Oracle recall / precision and percentage of perfect classifications for Specialty STR: predicting the specialty by classification (perfect prediction); General STR: no specialty prediction; Association Weight: predicting the specialty by highest association weight (over 3 sub-specialties); and Prediction STR: predicting the specialty with a STR for 3 sub-specialties.

However, the improved performance of the association weight method for STR selection did not transfer to the automatic classification results (only 1.2% improvement over the previous result). Of particular interest are the percentages of perfect classifications achieved for the four methods tested in the automatic classification exercise. Although low, both the association weight and the Prediction STR method achieved a better performance for the test set of documents than the “perfect” specialty STR and the general STR. Some test documents were categorized into another specialty than the original specialty and the appropriate specialty STR suggested better controlled vocabulary terms than the original specialty STR would have done.

The results indicate that the automatic selection methods introduced in this chapter might propose specialties to a user that could not have been recognized by the specialty determination methods utilized for specialty STR construction. In that case, an interactive specialty STR selection process could provide more insight into the different specialty vocabulary spaces for the searcher.

8.4.2 Ohsumed

For the Ohsumed collection, the Prediction STR and association weight method predicted the most appropriate specialty STR out of the 33 specialty STRs created for the automatic classification exercise described in chapter Seven. This is a more difficult selection problem than selecting one specialty out of three as was the case in the Inspec collection. For the association method, the first three suggested controlled vocabulary

terms (average number of assigned terms) were used to calculate the total association weight. Table 21 shows the hitrate for the two selection methods at 10 cut-off levels.

Cut-off level	Recall Prediction STR	Recall Association Weight
1	0.4207	0.3606
2	0.5925	0.5212
3	0.6705	0.6061
4	0.7151	0.6667
5	0.7445	0.7273
6	0.7690	0.7455
7	0.7899	0.7788
8	0.8057	0.8152
9	0.8197	0.8273
10	0.8338	0.8545

Table 21. Recall of Specialty Prediction STR and association weight calculation at 10 cut-off levels in the Ohsumed collection. Percentage of records where the journal-descriptor-selected specialty was predicted within the first-ranked predictions.

It is difficult to compare these results to the Inspec results because of the different task complexities in the selection process. Nevertheless, considering that the “correct” specialty is one out of 33 possible, predicting the correct specialty within the first three suggestions in about two thirds of the cases tested should be regarded as a good result.

Table 22 shows the oracle recall/precision values and the percentage of perfect classifications for the four different methods of specialty STR selection (for recall and precision values at 15 cut-off levels, see table B12).

	Specialty STR	General STR	Association Weight	Prediction STR
Oracle Recall / Precision				
	0.3897	0.3115	0.2816	0.2563
Perfect classifications				
	10.30%	5.45%	6.67%	5.15%

Table 22. Automatic determination of the specialty in the Ohsumed collection for 330 queries. Oracle recall / precision and percentage of perfect classifications for Specialty STR: predicting the specialty by journal descriptor (perfect prediction); General STR: no specialty prediction; Association Weight: predicting the specialty by highest association weight (over all 33 specialties); and Prediction STR: predicting the specialty with a STR for the 33 Ohsumed specialties.

Similar to the Inspec experiments, the association weight method of specialty STR selection achieves a 22% performance improvement compared to the general STR for perfect classifications indicating that it can be more precise than a general vocabulary support system but figure 11 shows that the perfect specialty STR selection and the general STR outperform the other two methodologies in the automatic classification exercise.

The association weight method and the general STR achieve similar results at the first cut-off level but the association weight method causes performance to drop at other cut-off levels.

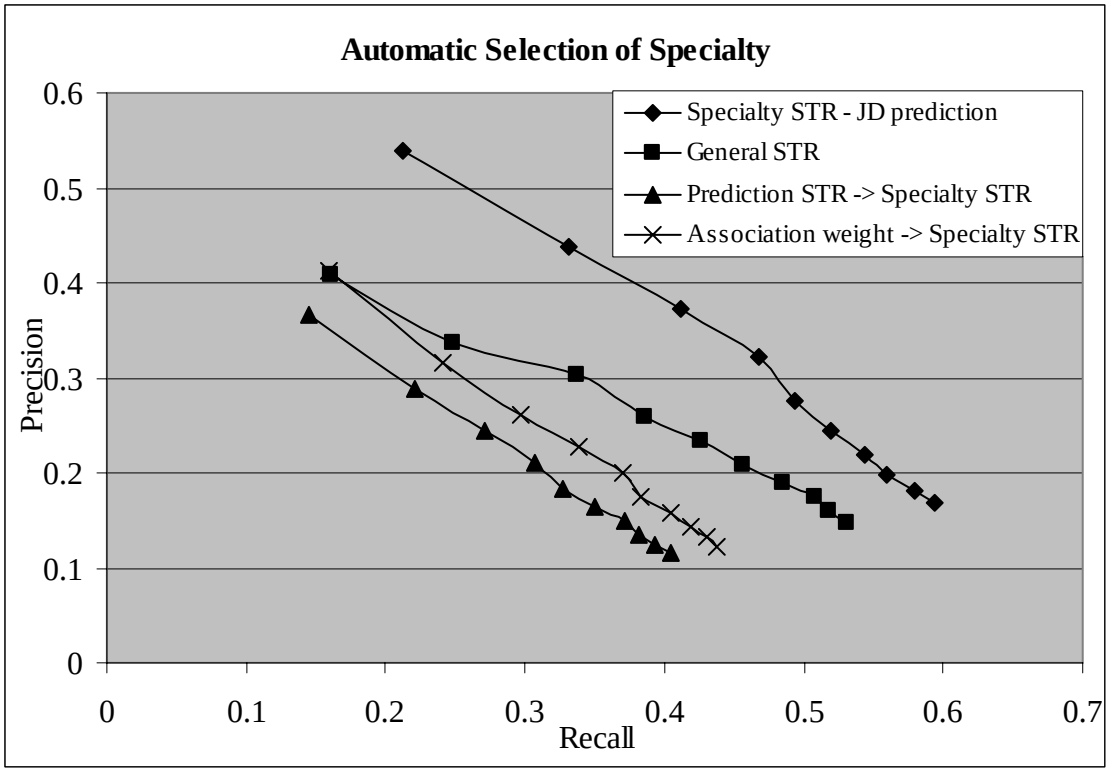

Figure 11. Automatic determination of the specialty in the Ohsumed collection for 330 queries. Recall and precision at 10 cut-off levels for Specialty STR: predicting the specialty by journal descriptor (perfect prediction); General STR: no specialty prediction; Association Weight: predicting the specialty by highest association weight (over all 33 specialties); and Prediction STR: predicting the specialty with a STR for the 33 Ohsumed specialties.

8.5 Conclusion

This chapter discussed how specialty search term recommenders could be presented to a user in the search process. Because the automatic selection of appropriate specialty search term recommenders obscures the information system's procedures for vocabulary support and does not provide insight into the different contexts (specialties)

represented in an information collection, the manual or interactive selection methods that will involve the user's input are generally preferred.

The methods for automatic specialty STR selection introduced and evaluated in this chapter demonstrate that automatic specialty STR selection might increase the error margin in vocabulary suggestion to a point where it is advisable to utilize a general search term recommender instead of a potentially misleading specialty search term recommender. This is another argument for using manual and interactive selection in the search process.

Other collection selection methods might achieve better results in specialty selection. Furthermore, an interactive selection process utilizing automatic specialty selection methods might provide new insights into the specialty vocabularies of the collection. Some of the results of the automatic selection evaluation indicate that the suggestion of specialties with the search term recommender technology can show relationships in the vocabulary that might not be detected otherwise.

It was also hypothesized that the automatic selection methodology used should be related to the optimal level of specificity for specialties in the STR construction process. The effectiveness of a selection methodology for specialty STRs will depend on the variety of specialties to choose from, their level of specificity, and the level of user involvement.

How and when a specialty search term recommender is chosen should be left to the entity who is most adaptive in the search process and who can best determine its most effective use: the searcher.

Chapter 9

Conclusion

9.1 A Specialty Focus on the Language Mapping Problem

The fundamental research question of information retrieval is how to obtain the right information for the right user in a digital environment. Text retrieval, the primary application area in information retrieval, is concerned with the searching of text documents in information collections. Searching in an information collection can be considered a mapping or matching process. The question of a searcher is matched with documents in the collection that are thought to be relevant to the question. On a closer look, this matching is a multi-stage process with subsequent information loss at each step. From the original formation of a question in a searcher's mind to the formal search

statement input into the information system's search interface, the concept space of the question in the searcher's mind is reduced to a few search terms received by the system. Search terms are then mapped to documents contained in the database, which themselves are term representations derived from an author's concept space. If the content of the documents in the information collection is not represented with the author language (full-text of the document) but with a controlled vocabulary, the available term space for concept representation is even further reduced.

Text retrieval from an information retrieval system's perspective is a term comparison (matching) between the terms input by the searcher and the terms in documents. If a search term matches a document term, that document is presented as relevant to the searcher. The real difficulty in searching, however, is not this matching process; it is the mapping between the concept space of the searcher's question and the concept space of the document's content, which is complicated by the limitations of the number of search terms representing the question and the number of terms used for document content representation. This has been called the language problem in information retrieval.

The language or language mapping problem is indeterminate: there is no upper bound to the number of possible descriptions for a question or a document's intellectual content. In conversations between humans, an interactive dialogue, in which the meaning of words and the appropriate terminology is negotiated, can resolve language ambiguities. Language games demonstrate a word's correct usage and meaning in the activity they occur. By taking the context of the conversation into account, the language

problem is alleviated. In the search process, the interaction between a searcher and an information system is rarely a conversation, that is, a mutual negotiation of meaning and context. Because language ambiguities cannot be resolved, the language problem is exacerbated.

How can information retrieval systems help searchers in the mapping between their questions and the information collection's answers (i.e. documents)? By providing insight into the information collection's organization, content and document representations, an information retrieval system helps a user to negotiate the formulation of search statements. If this process is based on searcher input and responsive, the interaction could approach a conversational style and the language problem should be mitigated.

In order to alleviate the language problem, information retrieval systems should focus on representing documents in their contexts in the collection, making the search process more interactive, and, especially, dividing up the language space in search to make words appear less ambiguous.

This dissertation describes an approach to resolve language ambiguities in search that divides the language space of the collection, provides insight into the language of document representations of the information retrieval system and, in its recommended application, interacts with the searcher to achieve a more effective retrieval experience. The search term recommender system introduced in this dissertation is based on three assumptions.

First, it is assumed that restricting the language space by focusing on subject-specific areas within an information collection will provide more precise and focused access (i.e. vocabulary support) into the collection. It is argued that the more subject-specific an information environment (or conversation), the more restricted the language space available (the appropriate language games restrict word usage and disambiguate word meanings). If the language space is restricted, the language mapping problem is alleviated.

Second, it is assumed that providing vocabulary support in the form of term suggestions from the document representations of an information system (i.e. the controlled vocabulary) will initiate the kind of interactive negotiation of the language space necessary to generate an effective search. Utilizing the controlled vocabulary terms in the search process will ease the burden of language mapping between a searcher and the information collection and provide a more effective search vocabulary.

Third, it is assumed that providing subject-specific vocabulary support will also provide insight into the information collection's organization and discourse areas covered. For a searcher, an overview of different discourse areas related to a topic (represented by a search statement) will help to reach a deeper understanding of the discussion around a topic and the information collection in general.

Drawing on research in the sociology of science and in linguistic studies on sublanguages and languages for special purposes, the research specialty was identified as a starting point for the search term recommender system's discourse-specific approach to search. A specialty (synonymous to domain and discourse community) is defined as an

area of research characterized by certain theories and methods, particular forms of communication between community members and particularly by its use of a specialized language to describe phenomena in the field.

9.2 Contributions

With the focus on the specialized languages of research fields, a first application area for the search term recommender system can be determined: search support in information collections targeted towards scholarly communities, i.e. library catalogs and bibliographic subject databases. The search term recommender methodology was evaluated on two scientific bibliographic databases, Inspec, covering research in physics and engineering, and Ohsumed, covering research in the medical fields.

The core function of the search term recommender system (described in chapter Five) is to calculate the degree of association between natural language terms and controlled vocabulary terms in different specialty areas of a collection. The degree of association specifies functional relationships between natural language terms and controlled vocabulary terms creating a mapping between terms that will be used for vocabulary support in search. Search term recommenders can be applied in three areas: for automatic text classification, for query expansion and search support, and for terminology mapping between controlled vocabularies.

Focusing on discourse areas within an information collection in order to provide a more accurate mapping between user search terms and the information collection’s

vocabulary is grounded on the premise that it is possible to subdivide document collections into specialty areas. A second premise for the search term recommender to function effectively is that specialty areas in information collections contain different specialty vocabularies. The mapping process between specialty vocabularies in the collection and search terms should reveal different discourses associated with a topic (represented by the query).

This dissertation addresses the following questions:

(1) How can specialties be determined in an information collection?
(2) Do different specialties have different specialty vocabularies?
(3) Is focusing on specialty collections more effective for vocabulary support?
(4) What is the appropriate degree of specificity for a specialty in order to provide optimal vocabulary support?
(5) How could the appropriate specialties for a question be identified in the search process?

Chapter Six specifies four ways for determining specialties within document collections: through domain terminology, publication sources, bibliometric or social network analysis, and subject-specific classifications. For the Inspec collection, specialties were determined by the grouping of documents within a subject-specific hierarchical classification. For the Ohsumed collection, specialties were selected according to journal descriptors placing documents in particular subject areas.

An analysis of the specialty vocabularies in the Inspec and Ohsumed collections shows that both the natural language vocabularies and the controlled vocabularies differ between specialties. Differences in vocabulary depend on the relatedness of the specialties to each other and the size and variety of the specialty vocabularies.

In chapter Seven, the effectiveness of the specialty search term recommenders is evaluated in an automatic classification application. It was tested whether search term recommenders with a specialty focus predict better controlled vocabulary terms than a search term recommender trained on the whole collection. The analysis shows that substantial improvements in hit-rate (recall) and accuracy (precision) can be achieved by specialty search term recommenders compared to general search term recommenders in the Inspec and Ohsumed collections.

The level of specificity appropriate for optimal performance in term suggestion is determined by identifying the specialty in the collection, the number of documents available to train the search term recommender, and vocabulary differences in the different levels of specificity. This is shown in an analysis of six specialty search term recommenders at different levels of specificity.

When specialty search term recommenders are utilized for search support, a search term recommender system should help the searcher identify important discourse areas associated with his question. By identifying the specialties covering a particular topic in a query, the specialty search term recommenders associated with a specialty can display related controlled vocabulary terms, effectively showing the discourse on the topic with the help of the document representations.

In the case of a focused search, or if the search term recommender application does not allow for user interaction, the specialty recommender(s) most appropriate for a particular query need to be selected. Chapter Eight discusses interactive and automatic methods for presenting specialty search term recommender vocabulary support to users.

The methods for automatic specialty selection might increase the error margin in vocabulary suggestion to a point where it is advisable to utilize a general search term recommender instead of a potentially misleading specialty search term recommender.

9.3 Future Work

There are at least three directions where research on search term recommender methodology could continue. From an algorithm perspective, the calculation of the association weight is just one method to statistically resolve co-occurrence patterns in data, here co-occurrence of natural language and controlled vocabulary terms to resolve the degree of association. The areas of text classification and collection selection in distributed search offer a wide array of algorithms that could be applied to the search term recommender problem. The optimal performance of the search term recommender algorithm is collection and application dependent. Different association algorithms could consequently achieve performance improvements; however, the determination of the optimal algorithm in information retrieval or text classification traditionally seems to be a heuristic rather than a theoretical one. One direction to continue research on search term recommenders is therefore to explore various association algorithms for different collections and to determine a list of impact factors for the performance effectiveness of the algorithms depending on collection and application characteristics.

From a user interface perspective, the search term recommender methodology posits some challenges for the usability and structure of a search interface. In this

dissertation, the effectiveness of the search term recommender has been tested for automatic text classification, but for vocabulary support in an interactive search environment, the appropriate interface construction is important. Another direction to continue research on search term recommenders is therefore to explore how to integrate the search term recommender methodology effectively but unobtrusively into search interfaces for bibliographic databases and to determine ways to visualize the discourse areas suggested by different specialty search term recommenders in the results interface. Only an effective user interface will guarantee the continued and successful use of the search term recommender system for query formulation.

From a collection and application area perspective, the current implementations of the search term recommender methodology rely on the relatively well-structured organization of bibliographic databases and controlled vocabularies as document representations to perform the mapping between searcher terms and information system terms. In theory, the core function of the search term recommender of calculating a degree of association can be applied to any kind of co-occurrence patterns in textual documents. The third direction to continue research on search term recommender is therefore to explore applications that extend the reach beyond the mapping to a controlled vocabulary.

For bibliographic databases, it could be equally interesting for a searcher to map his topics (expressed in a query) to authors, other documents, or publication sources that target discourses related to the topic. Furthermore, search term recommenders could be researched for other types of information collections. A challenge for research lies in

determining specialty areas in collections that are not bibliographic in nature. The domain terminology approach described in chapter Six is one strategy that could be applied to any kind of textual collection, whereas the classification or bibliometric approaches are targeted towards more structured bibliographic information systems. For hyperlinked information collections like the Internet, a network analysis of link structures could help in determining areas of discourse, therefore extending the application area of the search term recommender almost indefinitely.

Like the number of potential descriptions of a concept, the number of potential applications and research opportunities for the search term recommender may have no upper bound.

References

Abbott, A. D. (1988). *The system of professions: An essay on the division of expert labor*. Chicago: University of Chicago Press.

Adamzik, K. (2001). Ist die Linguistik eine "anglophon geprägte" Disziplin? Eine Analyse am Beispiel der Fachsprachenforschung. In *Language for special purposes: perspectives for the new millennium*. F. Mayer (Ed.), 3-35. Tübingen: G. Narr.

Ahn, L. v. and L. Dabbish (2004). Labeling images with a computer game. In *Proceedings of the SIGCHI conference on human factors in computing systems, Vienna, Austria*. New York: ACM Press, 319-326.

Aitchison, J., A. Gilchrist, et al. (2000). *Thesaurus construction and use: A practical manual*. Chicago: Fitzroy Dearborn Publishers.

Atherton, P. (1978). *Books are for use: Final report of the Subject Access Project to the Council on Library Resources.* Syracuse, New York: Syracuse University School of Information Studies.

Baeza-Yates, R. and B. d. A. Ribeiro-Neto (1999). *Modern information retrieval*. New York; Harlow, England: ACM Press; Addison-Wesley.

Bar-Hillel, Y. (1962). Theoretical aspects of the mechanization of literature searching. In *Digital information processors*. W. Hoffmann (Ed.), 406-443. New York: Wiley Interscience Publ.

Bates, M. J. (1988). How to use controlled vocabularies more effectively in online searching. *Online* 12(6): 45-56.

Bates, M. J. (1989). Rethinking subject cataloging in the online environment. *Library Resources & Technical Services* 33(4): 400-412.

Baumann, K.-D. (2001). Cognitive turn in LSP research. In *Language for special purposes: Perspectives for the new millennium*. F. Mayer (Ed.), 87-102. Tübingen, G. Narr.

Belkin, N. J. (1980). Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science* 5: 133-143.

Belkin, N. J., R. N. Oddy, et al. (1982). ASK for information retrieval. Part 1: Background and theory. *Journal of Documentation* 38(2): 61-71.

Ben-David, J. and R. Collins (1966 [1991]). Social factors in the origins of a new science: The case of psychology. In *Scientific growth: essays on the social organization and ethos of science*. J. Ben-David and G. Freudenthal (Eds.). Berkeley: University of California Press.

Benson, M. (1990). Collocations and general-purpose dictionaries. *International Journal of Lexicography* 3(1): 23–35.

Bhattacharyya, K. (1974). The effectiveness of natural language in science indexing and retrieval. *Journal of Documentation* 30(3): 235-54.

Blair, D. C. (1980). Searching biases in large interactive document-retrieval systems. *Journal of the American Society for Information Science* 31(4): 271-277.

Blair, D. C. and M. E. Maron (1985). An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the Association for Computing Machinery* 28(3): 289-299.

Blair, D. C. (1990). *Language and representation in information retrieval*. Amsterdam; New York: Elsevier Science Publishers.

Blair, D. C. (1992). Information retrieval and the philosophy of language. *Computer Journal* 35(3): 200-207.

Blair, D. C. (1994). Will it scale up? Thoughts about intellectual access in the electronic networks. In *Gateways, gatekeepers, and roles in the information omniverse: proceedings of the third symposium: November 13-15, 1993, the Washington Vista Hotel, Washington, DC*. A. Okerson, D. Mogge, Association of Research Libraries.et al. (Eds.). Washington, DC: Association of Research Libraries Office of Scientific and Academic Publishing.

Blair, D. C. (1996). STAIRS redux: Thoughts on the STAIRS evaluation, ten years after. *Journal of the American Society for Information Science* 47(1): 4-22.

Blair, D. C. (2002). The challenge of commercial document retrieval. Part II: a strategy for document searching based on identifiable document partitions. *Information Processing & Management* 38(2): 293-304.

Blair, D. C. and S. O. Kimbrough (2002). Exemplary documents: a foundation for information retrieval design. *Information Processing & Management* 38(3): 363-379.

Blair, D. C. (2003). Information retrieval and the philosophy of language. *Annual Review of Information Science and Technology* 37: 3-50.

Blair, D. C. (2006). *Wittgenstein, language and information: "Back to the rough ground!"* Springer.

Bonzi, S. (1984). Terminological consistency in abstract and concrete disciplines. *Journal of Documentation* 40(4): 247-263.

Bonzi, S. (1990). Syntactic patterns in scientific sublanguages: A study of four disciplines. *Journal of the American Society for Information Science* 41(2): 121-131.

Borgman, C. L. (1986). Why are online catalogs hard to use - Lessons learned from information-retrieval studies. *Journal of the American Society for Information Science* 37(6): 387-400.

Bowker, G. C. and S. L. Star (1999). *Sorting things out: Classification and its consequences.* Cambridge, Mass.: MIT Press.

Bowker, L. and J. Pearson (2002). *Working with specialized language: A practical guide to using corpora.* London; New York: Routledge.

Brajnik, G., S. Mizzaro, et al. (2002). Strategic help in user interfaces for information retrieval. *Journal of the American Society for Information Science and Technology* 53(5): 343-358.

Brooks, T. A. (1993). All the right descriptors - a test of the strategy of unlimited aliasing. *Journal of the American Society for Information Science* 44(3): 137-147.

Bross, I. D. J., P. A. Shapiro, et al. (1972). How information is carried in scientific sub-languages. *Science* 176: 1303-1307.

Buckland, M. and D. Florian (1991). Expertise, task complexity, and artificial intelligence: A conceptual framework. *Journal of the American Society for Information Science* 42(9): 635-643.

Buckland, M., M. H. Butler, et al. (1992). OASIS: A front-end for prototyping catalog enhancements. *Library Hi Tech* 40: 7-22.

Buckland, M. (2001). Vocabulary, Intermediaries, and Retrieval Performance. In *Information in a networked world: Harnessing the flow. 64th Meeting of the American Society for Information Science and Technology, Nov 3-8, 2001, Washington, DC,* 112-117. Medford, NJ: Information Today.

Byrne, A. and M. Micco (1988). Improving OPAC subject success: the ADFA experiment. *College and Research Libraries* 49: 432-441.

Cain, A. M. (1969). Thesaural problems in an on-line system. *Bulletin of the Medical Library Association* 57(3): 250-259.

Cappell, C. L. and T. M. Guterbock (1992). Visible colleges: The social and conceptual structure of sociology specialties. *American Sociological Review* 57(2): 266-273.

Case, D. O. (2002). *Looking for information: A survey of research on information seeking, needs, and behavior*. San Diego, Calif.: Academic Press.

Case, D. (2006). Information behavior. *Annual Review of Information Science and Technology* 40: 293-327.

Chan, L. M. (1989). Inter-indexer consistency in subject cataloging. *Information Technology and Libraries* 8(4): 349-358.

Chen, A. and F. Gey (2004). Multilingual information retrieval using machine translation, relevance feedback and decompounding. *Information Retrieval* 7(1-2): 149-182.

Chen, H. C. and V. Dhar (1991). Cognitive process as a basis for intelligent retrieval-systems design. *Information Processing & Management* 27(5): 405-432.

Chen, H. C. (1992). Knowledge-based document-retrieval - Framework and design. *Journal of Information Science* 18(4): 293-314.

Chen, H. C., T. Yim, et al. (1995). Automatic thesaurus generation for an electronic community system. *Journal of the American Society for Information Science* 46(3): 175-193.

Chen, H. C., T. D. Ng, et al. (1997). A concept space approach to addressing the vocabulary problem in scientific information retrieval: An experiment on the worm community system. *Journal of the American Society for Information Science* 48(1): 17-31.

Chen, H. C., J. Martinez, et al. (1998). Alleviating search uncertainty through concept associations: Automatic indexing, co-occurrence analysis, and parallel computing. *Journal of the American Society for Information Science* 49(3): 206-216.

Chu, H. (2003). *Information representation and retrieval in the digital age*. Medford, N.J.: Published for the American Society for Information Science and Technology by Information Today.

Chubin, D. E. (1976). Conceptualization of scientific specialties. *Sociological Quarterly* 17(4): 448-476.

Chubin, D. E., A. L. Porter, et al. (1986). Interdisciplinary research: The why and the how. In *Interdisciplinary analysis and research*. D. E. Chubin, A. L. Porter, F. A. Rossini and T. Connolly (Eds.), 3–10. Mt Airy: Lomond.

Cleverdon, C. W. (1970). *The effect of variations in relevance assessments in comparative experimental tests of index languages*. Cranfield, UK: Cranfield Institute of Technology. (Cranfield Library Report No. 3)

Cole, C. and A. Spink (2004). A human information behavior approach to a philosophy of information. *Library Trends* 52(3): 617-628.

Cool, C. (2001). The concept of situation in information science. *Annual Review of Information Science and Technology* 35: 5-42.

Cool, C. and A. Spink (2002). Issues of context in information retrieval (IR): An introduction to the special issue. *Information Processing & Management* 38(5): 605 - 611.

Cooper, W. S. (1973). On selecting a measure of retrieval effectiveness. *Journal of the American Society for Information Science* 24(March-April): 87-100.

Cornelius, I. (1996). Information and Interpretation. In *CoLIS 2. Second international conference on conceptions of library and information science: Integration in perspective, October 13-16, 1996, Copenhagen, The Royal School of Librarianship.*

Cousins, S. A. (1992). Enhancing subject access to OPACs: Controlled vocabulary vs. natural language. *Journal of Documentation* 48: 291-309.

Crane, D. (1971). Information needs and uses. *Annual Review of Information Science and Technology* 6: 3-39.

Crane, D. and H. Small (1992). American sociology since the seventies: The emerging identity crisis in the discipline. In *Sociology and its publics: The forms and fates of disciplinary organization*. T. C. Halliday and M. Janowitz (Eds), 197-234. Chicago: University of Chicago Press.

Crawford, S. (1978). Information needs and uses. *Annual Review of Information Science and Technology* 13: 61-81.

Croft, W. B. and R. H. Thompson (1987). I^3r - a new approach to the design of document-retrieval systems. *Journal of the American Society for Information Science* 38(6): 389-404.

Cutter, C. A. (1876). Library catalogues. In *Public libraries in the United States of America: Their history, condition and management.* Washington, D.C.: U.S. Government Printing Office. (Special Report, U.S. Bureau of Education)

Damerau, F. J. (1990). Evaluating computer-generated domain-oriented vocabularies. *Information Processing & Management* 26(6): 791-801.

Damerau, F. J. (1993). Generating and evaluating domain-oriented multi-word terms from texts. *Information Processing & Management* 29(4): 433-447.

Davenport., E. and H. Hall (2002). Organizational knowledge and communities of practice. *Annual Review of Information Science and Technology* 36: 171-227.

Dervin, B. and M. Nilan (1986). Information needs and uses. *Annual Review of Information Science and Technology* 21: 3-33.

Dervin, B. (1997). Given a context by any other name: methodological tools for taming the unruly beast. In *Proceedings of an international conference on information seeking in context, Tampere, Finland*, 13 - 38.

Dey, A. K. and G. D. Abowd (2000). Towards a better understanding of context and context-awareness. In *Workshop on the what, who, where, when, and how of context-awareness. The 2000 conference on human factors in computing systems (CHI 2000), The Hague, The Netherlands, April 3, 2000.*

Doerr, M. (2001). Semantic problems of thesaurus mapping. *Journal of Digital Information* 1(8).

Dogan, M. (2001). Specialization and recombination of specialties in the social sciences. In *International encyclopedia of the social & behavioral sciences.* N. J. Smelser and P. B. Baltes (Eds), 14851-14855. Amsterdam; New York: Elsevier.

Dolin, R. A. (1998). *Pharos: A scalable distributed architecture for locating heterogeneous information sources.* Ph.D. thesis. University of California, Santa Barbara.

Doszkocs, T. E. and R. K. Sass (1992). An associative semantic network for machine-aided indexing, classification and searching. In *Advances in classification research, Vol. 3. Proceedings of the 3rd ASIS SIG/CR Classification Research Workshop*, 15-35. Medford, NJ: Learned Information.

Dourish, P. (1995). Accounting for system behavior: Representation, reflection and resourceful action. In *Third decennial conference on computers in context CIC' 95, Aarhus, Denmark.*

Dourish, P. (2004). What we talk about when we talk about context. *Personal and Ubiquitous Computing* 8(1): 19-30.

Drabenstott, K. M. (2000). Web search strategies. In *Saving the user's time through subject access innovation*. W. J. Wheeler (Ed.). Champaign, IL: University of Illinois, Graduate School of Library and Information Science.

Dubois, C. P. R. (1987). Free text versus controlled vocabulary: A reassessment. *Online Review* 11(10): 243-253.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19(1): 61-74.

Efthimiadis, E. N. (1996). Query expansion. *Annual Review of Information Systems and Technology* 31: 121-187.

Erdelez, S. (1997). Information encountering: A conceptual framework for accidental information discovery. In *Information seeking in context: Proceedings of an international conference on research in information needs, seeking and use in different contexts, Tampere, Finland, 1996,* 412-421. London: Taylor Graham.

Evangelisti Allori, P. (2001). Conceptual and genre-specific constraints: How different disciplines select their discoursal features. In *Language for special purposes: perspectives for the new millennium*. F. Mayer (Ed.), 70-79. Tübingen, G. Narr.

Fidel, R. (1985). Individual variability in online searching behavior. In *48th annual meeting of the American Society for Information Science*. White Plains, NY: Knowledge Industry Publications.

Frank, E. and G. W. Paynter (2004). Predicting Library of Congress classifications from Library of Congress subject headings. *Journal of the American Society for Information Science and Technology* 55(3): 214-227.

French, J., A. Powell, et al. (2002). Exploiting manual indexing to improve collection selection and retrieval effectiveness. *Information Retrieval* 5(4): 323-351.

Friedman, C. (1986). Automatic structuring of sublanguage information. In *Analyzing language in restricted domains: Sublanguage description and processing*. R. Grishman and R. Kittredge (Eds), 85-102. Hillsdale, N.J.: L. Erlbaum Associates.

Fry, J. (2006). Scholarly research and information practices: A domain analytic approach. *Information Processing & Management* 42: 299-316.

Fuller, S. (1988). *Social epistemology*. Bloomington [Ind.]: Indiana University Press.

Furnas, G. W., T. K. Landauer, et al. (1987). The vocabulary problem in human system communication. *Communications of the ACM* 30(11): 964-971.

Galison, P. L. (1997). *Image and logic: A material culture of microphysics*. Chicago; London: University of Chicago Press.

Gauch, S. and J. B. Smith (1993). An expert system for automatic query reformation. *Journal of the American Society for Information Science* 44(3): 124-36.

Geertz, C. (1983). *Local knowledge: Further essays in interpretive anthropology*. New York: Basic Books.

Gey, F., H. Chen, et al. (1999). Advanced search technology for unfamiliar metadata. In *Proceedings of the Third IEEE metadata conference, April 1999, Bethesda, Maryland.*

Ginsparg, P., P. Houle, et al. (2004). Mapping subsets of scholarly information. *Proceedings of the National Academy of Sciences* 101(Suppl. 1): 5236-5240.

Godby, C. J. and J. Stuler (2003). The Library of Congress classification as a knowledge base for automatic subject categorization. In *Subject retrieval in a network environment*. I. C. McIlwane (Ed.), 163-169. Munich, Germany: K.G. Saur.

Golder, S. and B. A. Huberman (2006). The structure of collaborative tagging systems. *Journal of Information Science* 32(2): 198-208.

Gomez, L. M., C. C. Lochbaum, et al. (1990). All the right words - Finding what you want as a function of richness of indexing vocabulary. *Journal of the American Society for Information Science* 41(8): 547-559.

Gotti, M. (2001). The rise of the experimental essay. *Language for special purposes: perspectives for the new millennium*. F. Mayer (Ed.), 459-465. Tübingen, G. Narr.

Grishman, R., T. N. Nhan, et al. (1984). Automated determination of sublanguage syntactic usage. In *COLING 84 (Tenth international conference on computational linguistics), Stanford, CA*, 96-100.

Grishman, R. and R. Kittredge, Eds. (1986). *Analyzing language in restricted domains: sublanguage description and processing.* Hillsdale, N.J.: L. Erlbaum Associates.

Grishman, R. (2001). Adaptive information extraction and sublanguage analysis. In *Workshop on adaptive text extraction and mining. Seventeenth international joint conference on artificial intelligence (IJCAI-2001), Seattle, WA; August 5, 2001*.

Gross, T. and A. G. Taylor (2005). What have we got to lose? The effect of controlled vocabulary on keyword searching results. *College & Research Libraries* 66(3): 212-30.

Grossman, D. A. and O. Frieder (2004). *Information retrieval: Algorithms and heuristics*. Dordrecht; Norwell, MA: Springer.

Gumperz, J. J. (1982a). *Discourse strategies*. Cambridge [UK]; New York: Cambridge University Press.

Gumperz, J. J. (1982b). *Language and social identity*. Cambridge [UK]; New York: Cambridge University Press.

Haas, S. W. and S. He (1993). Toward the automatic identification of sublanguage vocabulary. *Information Processing & Management* 29(6): 721-732.

Haas, S. W. (1997). Disciplinary variation in automatic sublanguage term identification. *Journal of the American Society for Information Science* 48(1): 67-79.

Haas, P. M. (1992). Introduction: epistemic communities and international policy coordination. *International Organization* 46(1): 1-35.

Harris, Z. S. (1968). *Mathematical structures of language*. New York: Interscience Publishers.

Hersh, W., C. Buckley, et al. (1994). OHSUMED: An interactive retrieval evaluation and new large test collection for research. In *Proceedings of the ACM SIGIR conference*, 192-201.

Hersh, W., S. Price, et al. (2000). Assessing thesaurus-based query expansion using the UMLS metathesaurus. In *Proceedings of the 2000 American Medical Informatices Association (AMIA) Symposium*, 344-348.

Hewins, E. T. (1990). Information need and use studies. *Annual Review of Information Science and Technology* 25: 145-174.

Hildreth, C. R. (1997). The use and understanding of keyword searching in a university online catalog. *Information Technology and Libraries* 16: 52-62.

Hirschman, L. and N. Sager (1982). Automatic information formatting of a medical sublanguage. In *Sublanguage: Studies of language in restricted semantic domains.* R. Kittredge and J. Lehrberger (Eds.). Berlin; New York: W. de Gruyter.

Hirschman, L. (1986). Discovering sublanguage structures. In *Analyzing language in restricted domains: Sublanguage description and processing.* R. Grishman and R. Kittredge (Eds.), 211-234. Hillsdale, N.J.: L. Erlbaum Associates.

Hjerrpe, R. (1986). Project HYPERCatalog: Visions and preliminary conceptions of an extended and enhanced catalog. In *Intelligent Information Systems for the Information Society*. B. C. Brookes (Ed.). New York: North-Holland.

Hjørland, B. and H. Albrechtsen (1995). Toward a new horizon in information science – Domain analysis. *Journal of the American Society for Information Science* 46(6): 400-425.

Hjørland, B. and H. Albrechtsen (1999). An analysis of some trends in classification research. *Knowledge Organization* 26(3): 131-139.

Hjørland, B. (2001). Subject access points in electronic retrieval. *Annual Review of Information Science and Technology* 35: 249-298.

Hjørland, B. (2002). Domain analysis in information science. Eleven approaches - Traditional as well as innovative. *Journal of Documentation* 58(4): 422-462.

Hoffmann, L., H. Kalverkaemper, et al. (1997). *Fachsprachen / Languages for special purposes: Ein Internationales Handbuch zur Fachsprachenforschung und Terminologiewissenschaft / An international handbook of special languages & terminology research*. Berlin: Walter De Gruyter

Humphrey, S. M. (1992). Indexing biomedical documents: From thesaural to knowledge-based retrieval systems. *Artificial Intelligence in Medicine* 4(5): 343-371.

Humphrey, S. M. (1998). A new approach to automatic indexing using journal descriptors. In *Proceedings of the ASIS Annual Meeting 35*, 496-500.

Humphrey, S. M. (1999). Automatic indexing of documents from journal descriptors: A preliminary investigation. *Journal of the American Society for Information Science* 50(8): 661-674.

Humphrey, S. M., W. J. Rogers, et al. (2006). Word sense disambiguation by selecting the best semantic type based on journal descriptor indexing: Preliminary experiment. *Journal of the American Society for Information Science and Technology* 57(1): 96-113.

Iivonen, M. (1995). Consistency in the selection of search concepts and search terms. *Information Processing & Management* 31(2): 173-190.

Jacob, E. K. (2004). Classification and categorization: A difference that makes a difference. *Library Trends* 52(3): 515-540.

Joho, H., M. Sanderson, et al. (2004). A study of user interaction with a concept-based interactive query expansion support tool. In *ECIR 2004*. S. McDonald and J. Tait (Eds.), 42-56. Berlin Heidelberg: Springer. (LNCS 2997)

Jones, S., M. Gatford, et al. (1995). Interactive thesaurus navigation: Intelligence rules OK? *Journal of the American Society for Information Science* 46(1): 52-9.

Khoo, C. S. G. and D. C. C. Poo (1994). An expert-system approach to online catalog subject searching. *Information Processing & Management* 30(2): 223-238.

Kim, Y. (1998). *Sensitivity of entry vocabulary modules to subdomains*. Metadata Research Program: Technical Report.

Kim, Y. (1999). *Analysis of the subdomain sensitivity of the EVM*. Metadata Research Program: Technical Report.

Kittredge, R. and J. Lehrberger (1982). *Sublanguage: Studies of language in restricted semantic domains*. Berlin; New York: W. de Gruyter.

Kittredge, R. I. (2003). Sublanguages and controlled languages. In *The Oxford handbook of computational linguistics*. R. Mitkov (Ed.), 430-447. Oxford; New York: Oxford University Press.

Kjersti, A. and L. Eikvil (1999). *Text categorisation: A survey*. Norwegian Computing Center: Technical Report.

Klein, J. T. (1990). *Interdisciplinarity: history, theory, and practice*. Detroit: Wayne State University Press.

Klein, J. T. (1996). *Crossing boundaries: Knowledge, disciplinarities, and interdisciplinarities*. Charlottesville, Va.: University Press of Virginia.

Koll, M. (2000). Track 3: Information retrieval. *Bulletin of the American Society for Information Science* 26(2): 16-18.

Knorr-Cetina, K. (1999). *Epistemic cultures: How the sciences make knowledge*. Cambridge, Mass.: Harvard University Press.

Kohler, R. E. (1982). *From medical chemistry to biochemistry: The making of a biomedical discipline*. Cambridge [UK]; New York: Cambridge University Press.

Kretzenbacher, H. L. (2001). Looking backward - Looking forward - Still looking Good? On style in academic communication. In *Language for special purposes: Perspectives for the new millennium*. F. Mayer (Ed.), 443-458. Tübingen, G. Narr.

Krovetz, R. and W. B. Croft (1992). Lexical ambiguity and information retrieval. *ACM Transactions on Information Systems* 10(2): 115-141.

Kuhn, T. S. (1962[1996]). *The structure of scientific revolutions.* Chicago, IL: University of Chicago Press.

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind.* Chicago: University of Chicago Press.

Lancaster, F. W. and L. Smith (1983). *Compatibility issues affecting information systems and services.* Paris: UNESCO, General Information Programme and UNISIST. (PGI-83/WS/23)

Lancaster, F. W. (1986). *Vocabulary control for information retrieval*. Arlington, Va.: Information Resources Press.

Lancaster, F. W., T. H. Connell, et al. (1991). Identifying barriers to effective subject access in library catalogs. *Library Resources and Technical Services* 35(4): 377-391.

Lancaster, F. W. (2003). *Indexing and abstracting in theory and practice*. London: Facet.

Landry, P. (2001). The MACS project: Multilingual access to subjects (LCSH, RAMEAU, SWD). *International Cataloguing and Bibliographic Control* 30(3): 46-9.

Landry, P. (2004). Multilingual subject access: The linking approach of MACS. *Cataloging and Classification Quarterly* 37(3/4): 177-191.

Lansdale, M. W. and T. C. Ormerod (1994). *Understanding interfaces: A handbook of human-computer dialogue*. London; San Diego: Academic Press.

Larson, R. R. (1989). An automatic method of enhancing topical searching for online catalogs based on classification clustering. In *The Annual Review of OCLC Research, July 1988- June 1989*. Dublin, Ohio: OCLC Online Computer Library Center, Inc.

Larson, R. R. (1991). Classification clustering, probabilistic information retrieval, and the online catalog. *Library Quarterly* 61(2): 133-173.

Larson, R. R. (1992). Experiments in automatic Library of Congress classification. *Journal of the American Society for Information Science* 43(2): 130-148.

Lave, J. and E. Wenger (1991). *Situated learning: Legitimate peripheral participation*. Cambridge [UK]; New York: Cambridge University Press.

Lehrberger, J. (1982). Automatic translation and the concept of sublanguage. In *Sublanguage: Studies of language in restricted semantic domains*. R. Kittredge and J. Lehrberger (Eds.). Berlin; New York: W. de Gruyter.

Lehrberger, J. (1986). Sublanguage analysis. *Analyzing language in restricted domains: Sublanguage description and processing*. R. Grishman and R. Kittredge (Eds.), 19-38. Hillsdale, N.J.: L. Erlbaum Associates.

Leininger, K. (2000). Interindexer consistency in PsycINFO. *Journal of Librarianship and Information Science* 32(1): 4-8.

Lemaine, G., R. Macleod, et al., Eds. (1976). *Perspectives on the emergence of scientific disciplines*. The Hague, Chicago: Mouton; Aldine.

Lenoir, T. (1997). *Instituting science: The cultural production of scientific disciplines*. Stanford, Calif.: Stanford University Press.

Leonard, L. E. (1975). *Inter-indexer consistency and retrieval effectiveness: measurement of relationships*. Ph.D. Thesis. Champaign, IL: University of Illinois, Graduate School of Library Science.

Leung, C. H. and W. K. Kan (1997). A statistical learning approach to automatic indexing of controlled index terms. *Journal of the American Society for Information Science* 48(1): 55-66.

Liddy, E. D., W. Paik, et al. (1992). Use of subject field codes from a machine-readable dictionary for automatic classification of documents. In *Advances in classification research, Vol. 3. Proceedings of the 3rd ASIS SIG/CR Classification Research Workshop*, 83-100. Medford, NJ: Learned Information.

Lindholm-Romantschuk, Y. (1998). *Scholarly book reviewing in the social sciences and humanities: The flow of ideas within and among disciplines*. Westport, Conn.: Greenwood Press.

Losee, R. M. and S. W. Haas (1995). Sublanguage terms - Dictionaries, usage, and automatic classification. *Journal of the American Society for Information Science* 46(7): 519-529.

Luckhardt, H.-D. (2006). *The sublanguage approach: How can different special domains be dealt with*. http://is.uni-sb.de/studium/handbuch/infoling/ambi/sublanguage. Accessed 3/25/2006.

Mai, J.-E. (1999). A postmodern theory of knowledge organization. In *Knowledge: Creation, organization and use. Proceedings of the ASIS annual meeting*. Medford, NJ: Information Today.

Mai, J.-E. (2004). Classification in context: Relativity, reality, and representation. *Knowledge Organization* 31(1): 39-48.

Mai, J.-E. (2005). Analysis in indexing: Document and domain centered approaches. *Information Processing & Management* 41(3): 599-611.

Markey, K., P. Atherton, et al. (1982). An analysis of controlled vocabulary and free text search statements in online searches. *Online Review* 4: 225-236.

Markey, K. (1984). *Subject searching in library catalogs: Before and after the introduction of online catalogs*. Dublin, Ohio: OCLC Online Computer Library Center.

Markey, K. (1986). Users and the online catalog: Subject access problems. In *The Impact of online catalogs*. J. R. Matthews (Ed.), 35-69. New York: Neal-Schuman Publishers.

Marsh, E. (1986). General semantic patterns in different sublanguages. In *Analyzing language in restricted domains: Sublanguage description and processing*. R. Grishman and R. Kittredge (Eds.), 103-127. Hillsdale, N.J.: L. Erlbaum Associates.

Mathes, A. *Folksonomies - Cooperative classification and communication through shared metadata*. http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html (2004). Accessed 3/20/2006.

Matthews, J. R., G. S. Lawrence, et al. (1983). *Using online catalogs. A nationwide survey: A report of a study sponsored by the Council on Library Resources*. New York, NY: Neal-Schuman.

Mayer, F., Ed. (2001). *Language for special purposes: Perspectives for the new millennium*. Tübingen, G. Narr.

Micco, M. and I. Smith (1989). Designing a workstation for information seekers. *Reference Librarian: Expert Systems in Reference Services* 23: 135-152.

Miksa, F. L. (1998). *The DDC, the universe of knowledge, and the post-modern library*. Albany, N.Y.: Forest Press.

Mizzaro, S. (1997). Relevance: The whole history. *Journal of the American Society for Information Science* 48(9): 801-832.

Morato, J., J. Llorens, et al. (2003). Experiments in discourse analysis impact on information classification and retrieval algorithms. *Information Processing & Management* 39(6): 825-851.

Morris, S. A. (2005). Manifestation of emerging specialties in journal literature: A growth model of papers, references, exemplars, bibliographic coupling, cocitation, and clustering coefficient distribution. *Journal of the American Society for Information Science and Technology* 56(12): 1250-1273.

Morville, P. (2005). *Ambient findability*. Sebastopol, Calif.: Farnham, O'Reilly.

Mulkay, M. J., G. N. Gilbert, et al. (1975). Problem areas and research networks in science. *Sociology-the Journal of the British Sociological Association* 9(2): 187-203.

Nystrand, M. (1982). *What writers know: The language, process, and structure of written discourse*. New York: Academic Press.

Oakes, M. P. and M. J. Taylor (1998). Automated assistance in the formulation of search statements for bibliographic databases. *Information Processing & Management* 34(6): 645-668.

Oleson, A. and J. Voss (1979). *The Organization of knowledge in modern America, 1860-1920*. Baltimore: Johns Hopkins University Press.

Paisley, W. J. (1968). Information needs and uses. *Annual Review of Information Science and Technology* 3: 1-30.

Palmer, C. L. (1999). Structures and strategies of interdisciplinary science. *Journal of the American Society for Information Science* 50(3): 242-253.

Peat, H. J. and P. Willett (1991). The limitations of term cooccurrence Data for query expansion in document-retrieval systems. *Journal of the American Society for Information Science* 42(5): 378-383.

Peters, T. A. and M. Kurth (1991). Controlled and uncontrolled vocabulary subject searching in an academic library online catalog. *Information Technology and Libraries* 10(3): 201-211.

Petras, V. (2004). GIRT and the use of subject metadata for retrieval. In *Multilingual information access for text, speech and images: 5th workshop of the Cross-Language Evaluation Forum, CLEF 2004, Bath, England, 15-17 September 2004*, 298-309.

Petras, V. (2005). How one word can make all the difference - Using subject metadata for automatic query expansion and reformulation. In *Working notes for the CLEF 2005 Workshop, 21-23 September, Vienna, Austria*.

Pettigrew, K. E., R. Fidel, et al. (2001). Conceptual frameworks in information behavior. *Annual Review of Information Science and Technology* 35: 43-78.

Pitkin, H. F. (1985). *Wittgenstein and justice: on the significance of Ludwig Wittgenstein for social and political thought*. Berkeley: University of California Press.

Plaunt, C. and B. A. Norgard (1998). An association-based method for automatic indexing with a controlled vocabulary. *Journal of the American Society for Information Science* 49(10): 888-902.

Price, D. J. d. S. (1986). *Little science, big science-- and beyond*. New York: Columbia University Press.

Robertson, S. E. and K. Sparck Jones (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science* 27(3): 129-146.

Rowley, J. E. (1994). The controlled versus natural indexing languages debate revisited: A perspective on information retrieval practice and research. *Journal of Information Science* 20(2): 108-19.

Salton, G., E. A. Fox, et al. (1985). Advanced feedback methods in information retrieval. *Journal of the American Society for Information Science* 36(3): 200-210.

Saracevic, T., P. Kantor, et al. (1988). A study of information seeking and retrieving. 1. Background and methodology. *Journal of the American Society for Information Science* 39(3): 161-176.

Saracevic, T. and P. Kantor (1988). A study of information seeking and retrieving. 2. Users, questions, and effectiveness. *Journal of the American Society for Information Science* 39(3): 177-196.

Saracevic, T. and P. Kantor (1988). A study of information seeking and retrieving. 3. Searchers, searches, and overlap. *Journal of the American Society for Information Science* 39(3): 197-216.

Schamber, L. (1994). Relevance and information behavior. *Annual Review of Information Science and Technology* 29: 3-48.

Sebastiani, F. (2004). Text categorization. In *Text mining and its applications*. A. Zanasi (Ed.). Southampton, UK: WIT Press.

Sekine, S. (1994). A new direction for sublanguage N.L.P. In *International conference on new methods in language processing, Manchester - England, 1994*.

Shirky, C. (2005). *Ontology is overrated: Categories, links, and tags*. http://www.shirky.com/writings/ontology_overrated.html. Accessed 3/20/2006

Shiri, A. A., C. Revie, et al. (2002a). Thesaurus-enhanced search interfaces. *Journal of Information Science* 28(2): 111-22.

Shiri, A. A., C. Revie, et al. (2002b). Thesaurus assisted search term selection and query expansion: A review of user centered studies. *Knowledge Organization* 29(1): 1-19.

Sihvonen, A. and P. Vakkari (2004). Subject knowledge improves interactive query expansion assisted by a thesaurus. *Journal of Documentation* 60(6): 673-690.

Small, H. (1999). A passage through science: Crossing disciplinary boundaries. *Library Trends* 48(1): 72-108.

Solomon, P. (2002). Discovering information in context. *Annual Review of Information Science and Technology* 36: 229-264.

Sparck Jones, K. and P. Willett (1997). *Readings in information retrieval*. San Francisco, Calif.: Morgan Kaufman.

Spink, A. and C. Cole (2006). Human information behavior: Integrating diverse approaches and information use. *Journal of the American Society for Information Science and Technology* 57(1): 25-35.

Srinivasan, P. (1996). Retrieval feedback in MEDLINE. *Journal of the American Medical Informatics Association* 3(2): 157-167.

Star, S. L. and J. R. Griesemer (1989). Institutional ecology, 'translations' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-1939. *Social Studies of Science* 19(3): 387-420.

Stichweh, R. (1992). The sociology of scientific disciplines. *Science in Context* 5: 3-15.

Stichweh, R. (2001). History of scientific disciplines. In *International encyclopedia of the social & behavioral sciences*. N. J. Smelser and P. B. Baltes (Eds.), 13727-13731. Amsterdam; New York: Elsevier.

Suomela, S. and J. Kekäläinen (2005). Ontology as a search-tool: A study of real users' query formulation with and without conceptual support. In *ECIR 2005*. D. E. Losada and J. M. Fernández-Luna (Eds.), 315-329. Berlin, Heidelberg: Springer. (LNCS 3408)

Svenonius, E. (1986). Unanswered questions in the design of controlled vocabularies. *Journal of the American Society for Information Science* 37(5): 331-340.

Svenonius, E. (2000). *The intellectual foundation of information organization*. Cambridge, Mass.: MIT Press.

Swanson, D. R. (1966). *Studies of indexing depth and retrieval effectiveness*. Unpublished report, National Science Foundation Grant GN 380.

Tabah, A. N. (1999). Literature dynamics: Studies on growth, diffusion, and epidemics. *Annual Review of Information Science and Technology* 34: 249-286.

Talja, S., H. Keso, et al. (1999). The production of context in information seeking research: A metatheoretical view. *Information Processing and Management* 35: 751–763.

Taylor, A. G. (1995). On the subject of subjects. *The Journal of Academic Librarianship* 21: 484-91.

Taylor, C. A. (1996). *Defining science: A rhetoric of demarcation*. Madison, Wis.: University of Wisconsin Press.

Tennis, J. T. (2003). Two axes of domains for domain analysis. *Knowledge Organization* 30(3/4): 191-195.

Thompson, R., K. Shafer, et al. (1997). Evaluating Dewey concepts as a knowledge base for automatic subject assignment. In *Proceedings of the second ACM international conference on digital libraries*, 37-46.

Toms, E. G. (2000). Serendipitous information retrieval. In *First DELOS workshop on information seeking, searching and querying in digital libraries, Zurich, Switzerland, December 11-12, 2000*, 17-20.

Tonta, Y. (1991). A study of indexing consistency between Library of Congress and British Library catalogers. *Library Resources & Technical Services* 35(2): 177-185.

Vizine-Goetz, D., C. Hickey, et al. (2004). Vocabulary mapping for terminology services. *Journal of Digital Information* 4(4): Article No. 272, 2004-03-11.

Voorbij, H. J. (1998). Title keywords and subject descriptors: A comparison of subject search entries of books in the humanities and social sciences. *Journal of Documentation* 54(4): 466-476.

Walker, D. E. and R. A. Amsler (1986). The use of machine-readable dictionaries in sublanguage analysis. In *Analyzing language in restricted domains: sublanguage description and processing.* R. Grishman and R. Kittredge (Eds.), 69-83. Hillsdale, N.J.: L. Erlbaum Associates.

Wasserman, S. and K. Faust (1994). *Social network analysis: Methods and applications*. Cambridge; New York: Cambridge University Press

Weeber, M., J. G. Mork, et al. (2001). Developing a test collection for biomedical word sense disambiguation. In *Proceedings of the American Medical Informatics Association annual symposium (AMIA 2001)*. Philadelphia: Hanley & Belfus.

White, H. D. and K. W. McCain (1989). Bibliometrics. *Annual Review of Information Science and Technology* 32: 99-168.

White, H. D. and K. W. McCain (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. *Journal of the American Society for Information Science* 49(4): 327-355.

White, H. D. (2004). Citation analysis and discourse analysis revisited. *Applied Linguistics* 25(1): 89-116.

White, H. D., B. Wellman, et al. (2004). Does citation reflect social structure? Longitudinal evidence from the "Globenet" interdisciplinary research group. *Journal of the American Society for Information Science and Technology* 55(2): 111-126.

White, L. (1973). *Minimizing variabilities in indexing*. Tucson, AZ: Arizona University. (Research Report No. NTZS PB-237-989)

Whitley, R. (1984). *The intellectual and social organization of the sciences*. Oxford, [UK]; New York: Clarendon Press.

Wilson, P. (1968). *Two Kinds of power: An essay on bibliographic control*. Berkeley, CA: University of California Press.

Wilson, P. (1983). *Second-hand knowledge: An inquiry into cognitive authority*. Westport, Conn.: Greenwood Press.

Wilson, T. D. (1999). Models in information behaviour research. *Journal of Documentation* 55(3): 249-270.

Wilson, T. D. (2000). Human information behavior. Informing Science 32(2): 49-56.

Wittgenstein, L. (1953). *Philosophical investigations*. New York: Macmillan.

Wray, K. B. (2005). Rethinking scientific specialization. *Social Studies of Science* 35(1): 151-164.

Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Information Retrieval* 1(1/2): 69-90.

Zeng, M. L. and L. M. Chan (2004). Trends and issues in establishing interoperability among knowledge organization systems. *Journal of the American Society for Information Science and Technology* 55(5): 377-395.

Zipf, G. K. (1945). The meaning-frequency relationship of words. *Journal of General Psychology* 33: 251-256.

Zipf, G. K. (1949). *Human behavior and the principle of least effort; an introduction to human ecology*. Cambridge, Mass.: Addison-Wesley Press.

Appendix A. Inspec Experiments

Specialty Collection	Number of documents	Number of unique terms in titles	Number of unique Inspec Descriptors	Number of test documents
General	427,340	60,601	8,447	
Physics	193,596	42,334	7,506	19,360
Nuclear Physics	16,236	8,951	4,162	
Nuclear Structure	2,781	898	2,004	783
Electrical & Electronic engineering	128,068	29,911	7,443	12,807
Components, Electron Devices and Materials	35,497	11,946	5,515	
Passive circuit components	2,012	2,793	2,149	572
Computers & Control	105,676	28,193	6,245	10,568
Computer Hardware	34,115	13,984	4,732	
Circuits and Devices	3,790	3,530	2,116	1,070

Table A1. All Inspec collection numbers.

Descriptors / Document		Descriptors / Document	
2	1.59%	12	2.77%
3	6.29%	13	1.80%
4	12.12%	14	1.18%
5	14.91%	15	0.77%
6	15.09%	16	0.50%
7	13.22%	17	0.30%
8	10.87%	18	0.16%
9	8.09%	19	0.10%
10	5.87%	20	0.06%
11	4.20%	21	0.04%
Average number of Inspec descriptors per document: 6.99			

Table A2. Descriptor distribution in 121,248 Inspec test documents.

Physics	Electrical & Electronic Engineering	Computers & Control
III-V semiconductors	III-V semiconductors	Computational complexity
X-ray diffraction	Silicon	Digital simulation
Organic compounds	Elemental semiconductors	Neural nets
Silicon	Gallium arsenide	Internet
Elemental semiconductors	Semiconductor growth	Real-time systems
Gallium arsenide	Indium compounds	Optimisation
High-temperature superconductors	Aluminium compounds	Learning (artificial intelligence)
Annealing	Semiconductor epitaxial layers	Control system synthesis
Silicon compounds	Optimisation	Probability
Transmission electron microscopy	Asynchronous transfer mode	Object-oriented programming
Infrared spectra	Annealing	Statistical analysis
Barium compounds	Gallium compounds	Feedback
Scanning electron microscopy	Probability	Computational geometry
Crystal structure	Digital simulation	User interfaces
Photoluminescence	Photoluminescence	Medical image processing
Semiconductor growth	Silicon compounds	Parallel algorithms
Aluminium alloys	Semiconductor thin films	Human factors
Indium compounds	Semiconductor device models	Genetic algorithms
Ceramics	Medical image processing	Robust control
Aluminium compounds	Filtering theory	Feature extraction
Yttrium compounds	Finite element analysis	Performance evaluation
Crystal microstructure	Semiconductor lasers	Production control
Semiconductor epitaxial layers	Cellular radio	Matrix algebra
Monte Carlo methods	X-ray diffraction	Parallel programming
Iron alloys	Integrated circuit testing	Image segmentation
Nanostructured materials	Telecommunication traffic	Formal specification
Strontium compounds	VLSI	Filtering theory
Diffusion	Computational complexity	Computer vision
Surface topography	Image reconstruction	Parameter estimation
Nickel alloys	Statistical analysis	Adaptive control

Table A3. 30 most frequent controlled vocabulary terms in three Inspec specialties.

Cut-off level	Physics	Electrical & Electronic Engineering	Computers & Control
1	0.0789	0.0763	0.0883
2	0.1349	0.1274	0.1481
3	0.1763	0.1670	0.1900
4	0.2088	0.2006	0.2239
5	0.2354	0.2277	0.2547
6	0.2588	0.2517	0.2800
7	0.2788	0.2727	0.3043
8	0.2970	0.2912	0.3252
9	0.3140	0.3085	0.3440
10	0.3297	0.3233	0.3605
11	0.3436	0.3369	0.3754
12	0.3562	0.3503	0.3886
13	0.3678	0.3621	0.4016
14	0.3792	0.3730	0.4135
15	0.3899	0.3827	0.4242
	General		
1	0.0666	0.0669	0.0791
2	0.1137	0.1139	0.1324
3	0.1509	0.1510	0.1728
4	0.1821	0.1828	0.2030
5	0.2086	0.2091	0.2285
6	0.2299	0.2311	0.2516
7	0.2485	0.2509	0.2705
8	0.2659	0.2681	0.2896
9	0.2817	0.2850	0.3068
10	0.2958	0.3001	0.3225
11	0.3094	0.3132	0.3358
12	0.3222	0.3249	0.3490
13	0.3336	0.3366	0.3614
14	0.3443	0.3475	0.3727
15	0.3543	0.3580	0.3827

Table A4. Recall at 15 cut-off levels for the individual specialty search term recommenders (SSTR) vs. the general search term recommender (GSTR) for each specialty in the Inspec collection.

Cut-off level	Physics	Electrical & Electronic Engineering	Computers & Control
1	0.4332	0.4436	0.4095
2	0.3765	0.3775	0.3477
3	0.3345	0.3327	0.3017
4	0.3008	0.3016	0.2696
5	0.2731	0.2757	0.2468
6	0.2514	0.2548	0.2276
7	0.2333	0.2377	0.2130
8	0.2186	0.2227	0.1998
9	0.2061	0.2103	0.1885
10	0.1955	0.1989	0.1784
11	0.1859	0.1889	0.1693
12	0.1772	0.1804	0.1612
13	0.1694	0.1724	0.1538
14	0.1626	0.1653	0.1473
15	0.1563	0.1586	0.1412
	General		
1	0.3793	0.3960	0.3707
2	0.3293	0.3440	0.3171
3	0.2943	0.3059	0.2794
4	0.2681	0.2788	0.2491
5	0.2472	0.2551	0.2260
6	0.2281	0.2356	0.2082
7	0.2123	0.2199	0.1929
8	0.1994	0.2060	0.1811
9	0.1884	0.1951	0.1710
10	0.1785	0.1850	0.1621
11	0.1701	0.1758	0.1537
12	0.1628	0.1676	0.1468
13	0.1560	0.1606	0.1405
14	0.1499	0.1543	0.1349
15	0.1443	0.1486	0.1294

Table A5. Precision at 15 cut-off levels for the individual specialty search term recommenders (SSTR) vs. the general search term recommender (GSTR) for each specialty in the Inspec collection.

Oracle Recall / Precision		
Physics	**Electrical & Electronic Engineering**	**Computers & Control**
0.2478	0.2464	0.2489
General		
0.2205	0.2251	0.2265
% Perfect Classifications		
Physics	**Electrical & Electronic Engineering**	**Computers & Control**
0.70%	0.38%	0.72%
General		
0.46%	0.22%	0.61%

Table A6. Recall/precision at the oracle cut-off level and percentage of perfect classifications for the 3 specialty (SSTR) and general (GSTR) search term recommenders in the Inspec collection.

Cut-off level	Sub-sub-specialty STR	Sub-specialty STR	Specialty STR	General STR
Recall				
1	0.1279	0.0949	0.0512	0.0417
2	0.2046	0.1553	0.0915	0.0746
3	0.2602	0.2092	0.1285	0.1052
4	0.3042	0.2484	0.1574	0.1296
5	0.3402	0.2852	0.1803	0.1512
6	0.3727	0.3183	0.2040	0.1694
7	0.3997	0.3465	0.2243	0.1865
8	0.4234	0.3743	0.2424	0.2016
9	0.4470	0.3974	0.2613	0.2182
10	0.4650	0.4196	0.2764	0.2330
11	0.4821	0.4382	0.2910	0.2469
12	0.4960	0.4534	0.3058	0.2589
13	0.5090	0.4703	0.3213	0.2696
14	0.5218	0.4844	0.3332	0.2809
15	0.5337	0.4985	0.3454	0.2910
Precision				
1	0.6547	0.4995	0.2853	0.2408
2	0.5411	0.4168	0.2599	0.2200
3	0.4660	0.3769	0.2440	0.2073
4	0.4157	0.3408	0.2254	0.1922
5	0.3771	0.3157	0.2068	0.1793
6	0.3468	0.2949	0.1952	0.1661
7	0.3216	0.2764	0.1845	0.1564
8	0.2997	0.2616	0.1749	0.1476
9	0.2826	0.2489	0.1680	0.1418
10	0.2656	0.2372	0.1596	0.1361
11	0.2517	0.2267	0.1530	0.1309
12	0.2386	0.2157	0.1472	0.1257
13	0.2267	0.2071	0.1428	0.1210
14	0.2164	0.1988	0.1377	0.1170
15	0.2072	0.1915	0.1334	0.1133

Table A7. Average recall and precision at 15 cut-off levels for four levels of specificity in the Inspec collection (averaged over 3 specialties).

Cut-off level	Nuclear Structure STR	Nuclear Physics STR	Physics STR	General STR
Recall				
1	0.1350	0.1164	0.0399	0.0185
2	0.2089	0.1777	0.0659	0.0346
3	0.2633	0.2357	0.0884	0.0559
4	0.3076	0.2819	0.1071	0.0738
5	0.3445	0.3199	0.1230	0.0881
6	0.3760	0.3583	0.1393	0.1013
7	0.4064	0.3876	0.1524	0.1143
8	0.4280	0.4225	0.1676	0.1260
9	0.4532	0.4497	0.1835	0.1419
10	0.4739	0.4764	0.1989	0.1533
11	0.4938	0.4978	0.2108	0.1675
12	0.5073	0.5152	0.2274	0.1782
13	0.5209	0.5377	0.2406	0.1875
14	0.5360	0.5497	0.2518	0.1989
15	0.5511	0.5657	0.2659	0.2080
Precision				
1	0.6462	0.5722	0.1992	0.0983
2	0.5243	0.4489	0.1705	0.0907
3	0.4466	0.4010	0.1520	0.0979
4	0.3978	0.3630	0.1395	0.0958
5	0.3617	0.3338	0.1292	0.0927
6	0.3301	0.3123	0.1226	0.0888
7	0.3080	0.2906	0.1151	0.0854
8	0.2848	0.2773	0.1108	0.0835
9	0.2682	0.2654	0.1078	0.0829
10	0.2529	0.2548	0.1047	0.0810
11	0.2400	0.2431	0.1015	0.0802
12	0.2270	0.2317	0.0999	0.0781
13	0.2156	0.2245	0.0974	0.0763
14	0.2063	0.2141	0.0942	0.0750
15	0.1987	0.2060	0.0932	0.0734
Oracle Recall / Precision				
	0.3556	0.3264	0.1267	0.0854
% Perfect classifications				
	2.17%	1.15%	0.13%	0.00%

Table A8. Average recall and precision at 15 cut-off levels, recall/precision at the oracle cut-off level, and percentage of perfect classifications for four levels of specificity in the Physics specialty.

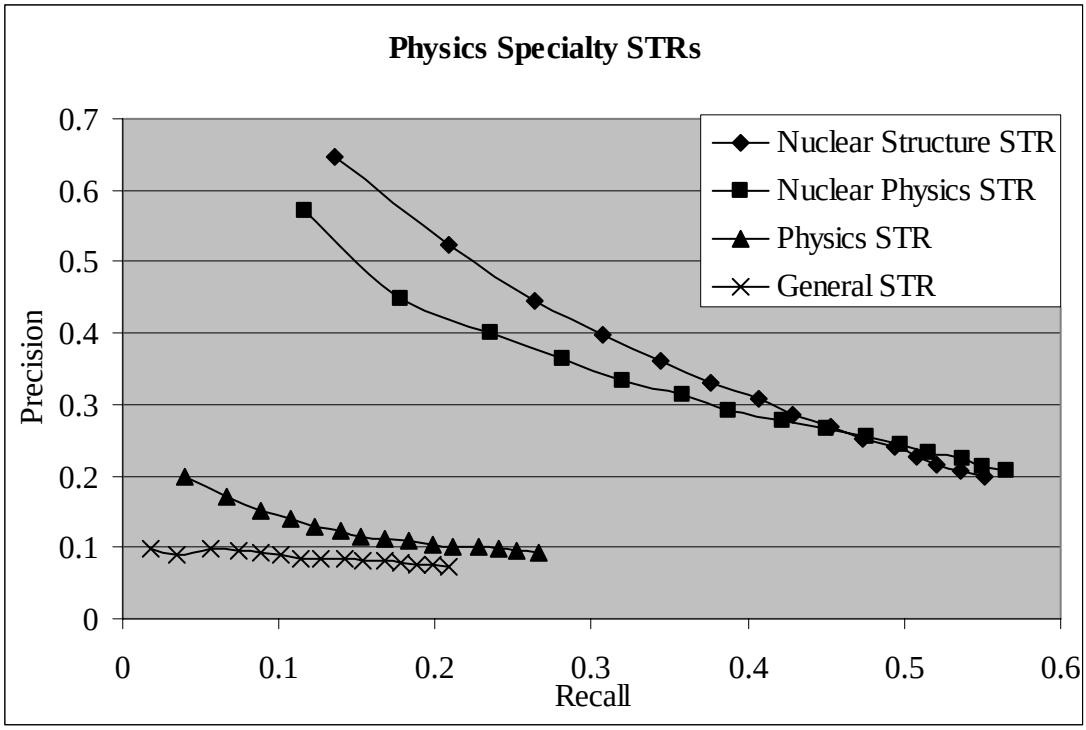

Figure A1. Recall and precision at 15 cut-off levels for four levels of specificity in the Physics specialty.

Cut-off level	Passive Circuits Components STR	Components, Electron Devices & Materials STR	Electrical & Electronic Engineering	General STR
Recall				
1	0.112318	0.0653	0.0451	0.0393
2	0.175809	0.1083	0.0889	0.0721
3	0.231194	0.1508	0.1293	0.0989
4	0.279051	0.1824	0.1604	0.1234
5	0.312452	0.2170	0.1845	0.1465
6	0.344561	0.2459	0.2116	0.1634
7	0.371669	0.2731	0.2330	0.1796
8	0.397815	0.2948	0.2529	0.1923
9	0.423522	0.3146	0.2702	0.2064
10	0.439839	0.3330	0.2835	0.2210
11	0.456027	0.3493	0.2979	0.2309
12	0.467855	0.3617	0.3113	0.2444
13	0.481017	0.3739	0.3269	0.2531
14	0.490863	0.3889	0.3385	0.2625
15	0.500672	0.4022	0.3471	0.2712
Precision				
1	0.6486	0.4161	0.3042	0.2727
2	0.5271	0.3514	0.2998	0.2544
3	0.4709	0.3246	0.2885	0.2354
4	0.4305	0.3016	0.2679	0.2190
5	0.3934	0.2860	0.2458	0.2042
6	0.3648	0.2689	0.2346	0.1879
7	0.3397	0.2562	0.2230	0.1761
8	0.3197	0.2415	0.2115	0.1646
9	0.3038	0.2302	0.2016	0.1571
10	0.2857	0.2191	0.1900	0.1505
11	0.2711	0.2109	0.1813	0.1429
12	0.2570	0.2006	0.1732	0.1377
13	0.2448	0.1915	0.1680	0.1315
14	0.2326	0.1848	0.1620	0.1266
15	0.2223	0.1789	0.1550	0.1225
Oracle Recall / Precision				
	0.3573	0.2464	0.2120	0.1647
% Perfect classifications				
	2.45%	0.17%	0.00%	0.00%

Table A9. Average recall and precision at 15 cut-off levels, recall/precision at the oracle cut-off level, and percentage of perfect classifications for four levels of specificity in the Electrical & Electronic Engineering specialty.

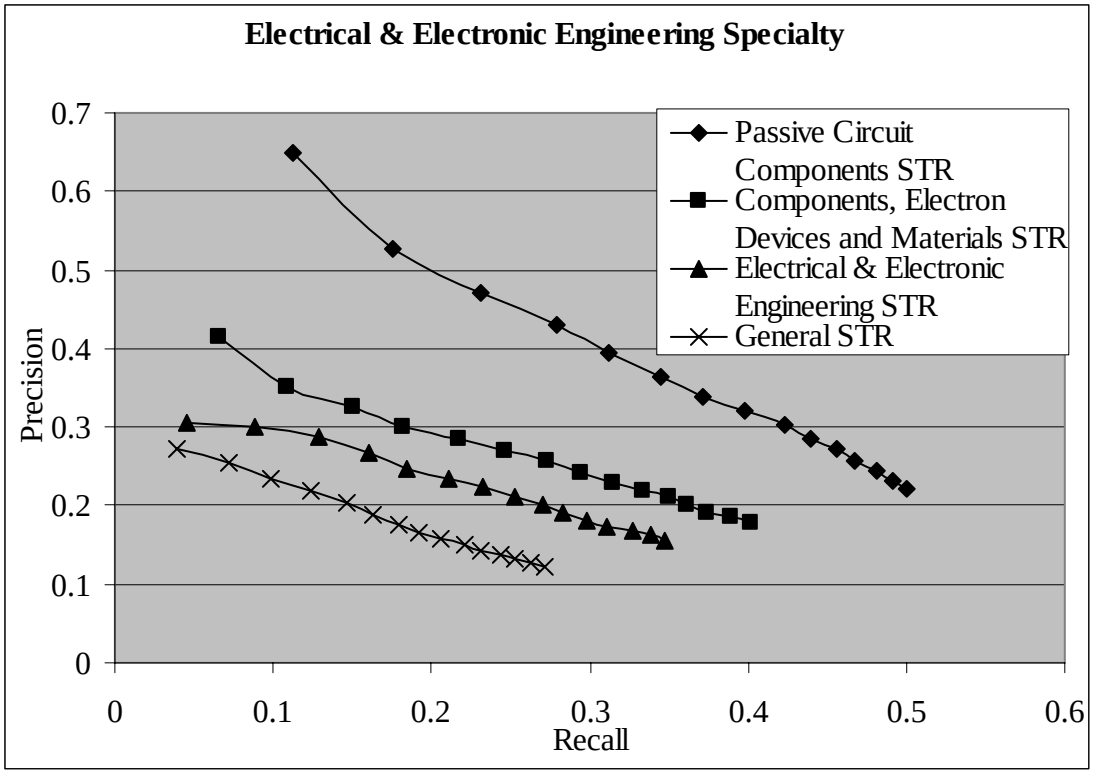

Figure A2. Recall and precision at 15 cut-off levels for four levels of specificity in the Electrical & Electronic Engineering specialty.

Cut-off level	Circuits & Devices STR	Computer Hardware STR	Computers & Control STR	General STR
Recall				
1	0.1365	0.1031	0.0686	0.0672
2	0.2290	0.1800	0.1196	0.1170
3	0.2861	0.2412	0.1679	0.1608
4	0.3260	0.2809	0.2047	0.1916
5	0.3638	0.3186	0.2335	0.2191
6	0.3977	0.3507	0.2612	0.2435
7	0.4210	0.3787	0.2873	0.2657
8	0.4443	0.4057	0.3067	0.2864
9	0.4642	0.4280	0.3302	0.3064
10	0.4811	0.4494	0.3468	0.3246
11	0.4965	0.4675	0.3642	0.3423
12	0.5130	0.4832	0.3786	0.3541
13	0.5252	0.4993	0.3962	0.3682
14	0.5386	0.5145	0.4093	0.3813
15	0.5493	0.5276	0.4233	0.3938
Precision				
1	0.6692	0.5103	0.3523	0.3514
2	0.5720	0.4500	0.3093	0.3150
3	0.4807	0.4050	0.2916	0.2885
4	0.4187	0.3579	0.2687	0.2619
5	0.3763	0.3271	0.2454	0.2409
6	0.3453	0.3036	0.2285	0.2217
7	0.3172	0.2822	0.2155	0.2079
8	0.2945	0.2661	0.2025	0.1947
9	0.2758	0.2512	0.1944	0.1853
10	0.2583	0.2379	0.1839	0.1769
11	0.2438	0.2261	0.1760	0.1697
12	0.2318	0.2147	0.1684	0.1614
13	0.2197	0.2052	0.1632	0.1553
14	0.2101	0.1974	0.1568	0.1493
15	0.2006	0.1896	0.1518	0.1441
Oracle Recall / Precision				
	0.3750	0.3233	0.2314	0.2219
% Perfect classifications				
	1.96%	1.50%	0.65%	0.75%

Table A10. Average recall and precision at 15 cut-off levels, recall/precision at the oracle cut-off level, and percentage of perfect classifications for four levels of specificity in the Computers & Control specialty.

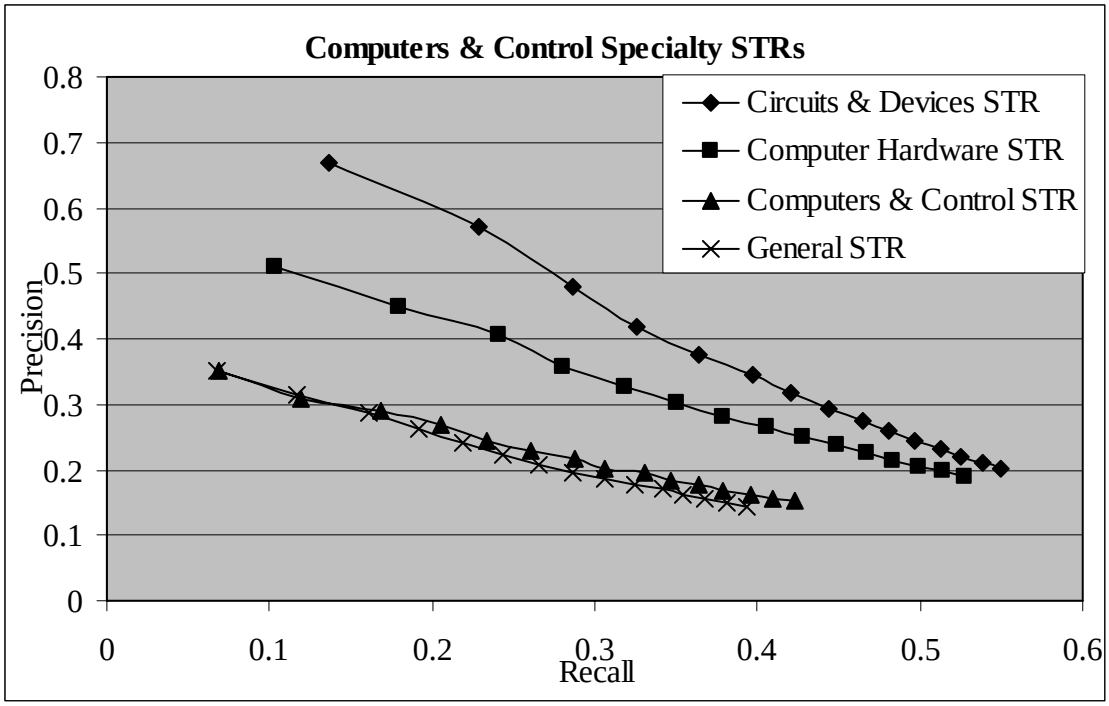

Figure A3. Recall and precision at 15 cut-off levels for four levels of specificity in the Computers & Control specialty.

Cut-off level	Specialty STR	General STR	Association Weight	Prediction STR
Recall				
1	0.0846	0.0802	0.0710	0.0441
2	0.1484	0.1334	0.1359	0.0861
3	0.1845	0.1717	0.1728	0.1113
4	0.2264	0.2023	0.2057	0.1289
5	0.2507	0.2292	0.2285	0.1418
6	0.2743	0.2511	0.2471	0.1515
7	0.2959	0.2726	0.2656	0.1661
8	0.3192	0.2940	0.2862	0.1778
9	0.3396	0.3102	0.3047	0.1909
10	0.3590	0.3311	0.3265	0.2013
11	0.3722	0.3453	0.3395	0.2110
12	0.3830	0.3569	0.3487	0.2185
13	0.3994	0.3690	0.3645	0.2277
14	0.4105	0.3822	0.3771	0.2412
15	0.4192	0.3948	0.3871	0.2447
Precision				
1	0.4633	0.4467	0.4233	0.2333
2	0.4317	0.3850	0.4033	0.2233
3	0.3689	0.3422	0.3489	0.1967
4	0.3317	0.3067	0.3075	0.1733
5	0.2987	0.2780	0.2787	0.1553
6	0.2756	0.2578	0.2544	0.1411
7	0.2567	0.2405	0.2367	0.1338
8	0.2433	0.2279	0.2233	0.1263
9	0.2285	0.2133	0.2111	0.1211
10	0.2183	0.2063	0.2043	0.1160
11	0.2073	0.1970	0.1942	0.1112
12	0.1975	0.1878	0.1842	0.1064
13	0.1905	0.1795	0.1787	0.1028
14	0.1819	0.1733	0.1717	0.1002
15	0.1744	0.1673	0.1656	0.0956

Table A11. Automatic determination of the specialty in the Inspec collection for 300 queries. Recall and precision at 15 cut-off levels for Specialty STR: predicting the specialty by journal descriptor (perfect prediction); General STR: no specialty prediction; Association Weight: predicting the specialty by highest association weight (over all 3 specialties); and Prediction STR: predicting the specialty with a STR for the 3 Inspec specialties.

Appendix B. Ohsumed Experiments

	# of journals	Number of documents	Number of unique terms in titles	Number of unique Mesh headings	Number of test documents
Allergy And Immunology	4	7908	6393	2756	879
Anesthesiology	6	8097	5841	2731	898
Communicable Diseases	4	4973	5151	2412	553
Dentistry	3	3445	4378	1968	380
Dermatology	7	8087	8077	3350	900
Drug Therapy	4	4586	5750	2796	510
Education	3	3627	4270	2486	405
Emergency Medicine	5	3164	3934	2143	352
Endocrinology	4	7431	6301	2722	825
Family Practice	6	3038	3452	2140	338
Geriatrics	4	2586	3577	1698	288
Gynecology	6	8268	6728	3347	920
Health Services Research	2	3732	4303	1538	415
Hematology	2	3915	4414	1787	436
Internal Medicine	3	4257	4926	2896	474
Medical Oncology	6	7661	6780	2937	852
Nephrology	5	3138	3749	1737	349
Neurosurgery	3	3984	4414	1869	443
Nuclear Medicine	2	4014	5114	2378	446
Nursing	6	2774	3258	1574	309
Nutrition	6	5054	5198	2346	563
Ophthalmology	9	7637	6545	2566	849
Orthopedics	5	5812	4913	1631	647
Otolaryngology	7	5827	5630	2506	647
Pathology	5	5653	7056	3482	629
Pediatrics	7	8526	7114	4170	948
Psychiatry	2	2820	3080	1319	314
Radiology	3	4951	5197	2313	551
Reproduction	3	3203	3966	1691	356
Rheumatology	3	3865	4609	2150	430
Transplantation	3	7056	5516	2216	785
Urology	4	6723	5789	2223	746
Vascular Diseases	3	2661	3241	1409	296

Table B1. Ohsumed specialty collections numbers: number of journals in specialty collection, number of documents, number of unique terms, number of unique Mesh headings and number of test documents.

Mesh / Document		Mesh / Document	
1	8.89%	7	1.18%
2	27.58%	8	0.34%
3	29.42%	9	0.06%
4	18.76%	10	0.02%
5	9.93%	11	0.01%
6	3.81%		
Average number of Mesh headings per Document: 3.11			

Table B2. Mesh heading distribution in 18,733 Ohsumed test documents.

Communicable Diseases	Gynecology	Orthopedics
Acquired Immunodeficiency Syndrome	Ultrasonography	Hip Prosthesis
Bacterial Infections	Pregnancy	Lumbar Vertebrae
Septicemia	Pregnancy Complications	Knee Prosthesis
Antibiotics	Prenatal Diagnosis	Fractures
Streptococcal Infections	Fetal Diseases	Knee Joint
Antibodies, Bacterial	Fetus	Scoliosis
Antibodies, Viral	Labor, Premature	Spinal Fusion
Meningitis	Ovarian Neoplasms	Dislocations
Disease Outbreaks	Uterine Neoplasms	Bone Neoplasms
Cytomegalic Inclusion Disease	Cervix Neoplasms	Backache
HIV Infections	Uterus	Postoperative Complications
Diarrhea	Pregnancy Complications, Infectious	Orthopedic Fixation Devices
Escherichia coli	Cesarean Section	Osteotomy
Cross Infection	Pregnancy Complications, Cardiovascular	Intervertebral Disk Displacement
Staphylococcal Infections	Pre-Eclampsia	Intervertebral Disk
Haemophilus influenzae	Placenta	Bone and Bones
Bacterial Vaccines	Amniotic Fluid	Tibia
Haemophilus Infections	Fetal Blood	Cervical Vertebrae
Staphylococcus aureus	Labor	Osteoarthritis
Opportunistic Infections	Cervix Uteri	Ligaments, Articular
Antigens, Bacterial	Pregnancy in Diabetes	Tibial Fractures
Respiratory Tract Infections	Endometriosis	Spine
Pregnancy Complications, Infectious	Pregnancy Outcome	Thoracic Vertebrae
Escherichia coli Infections	Fetal Monitoring	Femur
Mycoses	Hypertension	Shoulder Joint
Pneumonia	Menopause	Fracture Fixation, Internal
HIV	Pregnancy, Multiple	Bone Transplantation
Immunization, Passive	Postoperative Complications	Spinal Diseases
Viral Vaccines	Fetal Growth Retardation	Hip Joint
Herpes Simplex	Umbilical Arteries	Arthritis, Rheumatoid

Table B3. 30 most frequent controlled vocabulary terms in three Ohsumed specialties.

	Recall SSTR	Recall GSTR	Precision SSTR	Precision GSTR
Allergy And Immunology	0.5865	0.5434	0.2147	0.1982
Anesthesiology	0.6692	0.6095	0.2098	0.1903
Communicable Diseases	0.6572	0.5560	0.2038	0.1723
Dentistry	0.5346	0.4501	0.1524	0.1287
Dermatology	0.6486	0.5668	0.1847	0.1602
Drug Therapy	0.6384	0.5424	0.2127	0.1696
Education	0.6045	0.5893	0.1479	0.1430
Emergency Medicine	0.5824	0.5450	0.1565	0.1435
Endocrinology	0.6074	0.5526	0.2225	0.1993
Family Practice	0.5336	0.5717	0.1376	0.1435
Geriatrics	0.5574	0.5520	0.1514	0.1441
Gynecology	0.6419	0.5913	0.1886	0.1722
Health Services Research	0.5381	0.4527	0.1405	0.1190
Hematology	0.5983	0.4793	0.1959	0.1608
Internal Medicine	0.6073	0.6022	0.1643	0.1612
Medical Oncology	0.6490	0.5927	0.1788	0.1621
Nephrology	0.6506	0.6331	0.1926	0.1877
Neurosurgery	0.6773	0.5836	0.1946	0.1646
Nuclear Medicine	0.6160	0.5681	0.1870	0.1726
Nursing	0.4314	0.4099	0.1142	0.1019
Nutrition	0.6376	0.5718	0.1963	0.1732
Ophthalmology	0.6945	0.5840	0.1993	0.1658
Orthopedics	0.6992	0.6324	0.1872	0.1662
Otolaryngology	0.6213	0.5256	0.1658	0.1380
Pathology	0.5615	0.5227	0.1650	0.1502
Pediatrics	0.5974	0.5738	0.1525	0.1446
Psychiatry	0.6937	0.6225	0.1914	0.1701
Radiology	0.6215	0.5942	0.1728	0.1628
Reproduction	0.6490	0.6191	0.2006	0.1910
Rheumatology	0.6763	0.6572	0.1900	0.1830
Transplantation	0.6970	0.6558	0.2452	0.2260
Urology	0.7036	0.6674	0.1954	0.1845
Vascular Diseases	0.6730	0.6201	0.2027	0.1868

Table B4. Recall and precision at a cut-off level of 10 Mesh Headings for the individual specialty search term recommenders (SSTR) vs. the general search term recommender (GSTR) for each specialty in the Ohsumed collection.

	Oracle Recall/Precision SSTR	**Oracle Recall/Precision GSTR**
Allergy And Immunology	0.3852	0.3482
Anesthesiology	0.4684	0.3761
Communicable Diseases	0.4314	0.3249
Dentistry	0.3422	0.2383
Dermatology	0.4493	0.3291
Drug Therapy	0.4495	0.3281
Education	0.3758	0.3192
Emergency Medicine	0.3747	0.2955
Endocrinology	0.4046	0.3395
Family Practice	0.3570	0.3598
Geriatrics	0.3516	0.3152
Gynecology	0.4095	0.3255
Health Services Research	0.3404	0.2343
Hematology	0.3896	0.2977
Internal Medicine	0.4034	0.3386
Medical Oncology	0.4478	0.3401
Nephrology	0.4626	0.3789
Neurosurgery	0.4465	0.3358
Nuclear Medicine	0.4101	0.3354
Nursing	0.2687	0.2124
Nutrition	0.4211	0.3402
Ophthalmology	0.4443	0.3395
Orthopedics	0.4616	0.3467
Otolaryngology	0.4087	0.2929
Pathology	0.3409	0.2650
Pediatrics	0.3527	0.3231
Psychiatry	0.5308	0.3778
Radiology	0.4034	0.3422
Reproduction	0.4292	0.3765
Rheumatology	0.4882	0.3930
Transplantation	0.5184	0.4410
Urology	0.4749	0.3803
Vascular Diseases	0.4573	0.4172

Table B5. Recall/precision at the oracle cut-off level for the 33 specialty (SSTR) and general (GSTR) search term recommenders: for each document, the cut-off level was set at the number of original Mesh headings predicted.

	Specialty STR	General STR
Allergy And Immunology	5.23%	4.32%
Anesthesiology	12.14%	6.46%
Communicable Diseases	8.14%	5.61%
Dentistry	7.63%	3.16%
Dermatology	11.67%	6.44%
Drug Therapy	9.80%	6.67%
Education	13.33%	8.40%
Emergency Medicine	8.81%	5.40%
Endocrinology	4.48%	2.30%
Family Practice	10.65%	9.47%
Geriatrics	7.99%	6.94%
Gynecology	8.15%	4.89%
Health Services Research	7.95%	4.10%
Hematology	7.80%	3.67%
Internal Medicine	12.45%	8.86%
Medical Oncology	14.08%	8.80%
Nephrology	13.18%	6.88%
Neurosurgery	11.96%	6.32%
Nuclear Medicine	9.64%	4.71%
Nursing	7.44%	5.83%
Nutrition	8.35%	4.97%
Ophthalmology	9.42%	6.60%
Orthopedics	14.53%	7.26%
Otolaryngology	11.28%	6.34%
Pathology	5.56%	2.86%
Pediatrics	10.34%	8.86%
Psychiatry	17.83%	8.28%
Radiology	10.89%	5.26%
Reproduction	7.87%	5.62%
Rheumatology	13.02%	8.60%
Transplantation	11.21%	7.39%
Urology	14.75%	9.52%
Vascular Diseases	11.15%	9.46%

Table B6. Percentage of perfect classifications (all original Mesh headings are predicted at the oracle cut-off level) for 33 specialty and general search term recommenders.

Specialty Collection	Number of documents	Number of unique terms	Number of unique Mesh headings	Number of test documents
General	168,463	39,762	12,140	
Communicable Diseases	4,973	5,151	2,412	
The Journal of Infectious Diseases	2,206	3,516	1,625	244
Gynecology	8,268	6,728	3,347	
Obstetrics & Gynecology	4,776	3,282	1,691	269
Orthopedics	5,812	4,913	1,631	
Clinical Orthopaedics & Related Research	2130	3,057	993	232

Table B7. Ohsumed collection numbers for specificity experiments.

Cut-off level	Journal STR	Specialty STR	General STR
Recall			
1	0.2236	0.2192	0.1599
2	0.3567	0.3482	0.2579
3	0.4436	0.4373	0.3233
4	0.4951	0.4951	0.3898
5	0.5427	0.5409	0.4341
6	0.5749	0.5730	0.4707
7	0.5990	0.6022	0.5051
8	0.6130	0.6249	0.5245
9	0.6319	0.6452	0.5548
10	0.6411	0.6622	0.5745
Precision			
1	0.6068	0.5934	0.4396
2	0.4980	0.4875	0.3595
3	0.4230	0.4171	0.3092
4	0.3602	0.3609	0.2819
5	0.3178	0.3180	0.2539
6	0.2825	0.2831	0.2318
7	0.2540	0.2569	0.2137
8	0.2288	0.2337	0.1948
9	0.2098	0.2151	0.1838
10	0.1923	0.1994	0.1718

Table B8. Average recall and precision at 10 cut-off levels for three levels of specificity in the Ohsumed collection (averaged over 3 specialties).

Cut-off level	The Journal of Infectious Diseases STR	Communicable Diseases STR	General STR
Recall			
1	0.1976	0.1847	0.1337
2	0.3236	0.3042	0.2173
3	0.4103	0.3898	0.2733
4	0.4713	0.4623	0.3366
5	0.5233	0.5144	0.3834
6	0.5544	0.5506	0.4187
7	0.5849	0.5809	0.4601
8	0.6067	0.6035	0.4772
9	0.6227	0.6260	0.5060
10	0.6322	0.6428	0.5234
Precision			
1	0.6189	0.5820	0.4221
2	0.5205	0.4877	0.3504
3	0.4467	0.4262	0.2992
4	0.3893	0.3842	0.2787
5	0.3484	0.3443	0.2549
6	0.3081	0.3087	0.2329
7	0.2787	0.2799	0.2190
8	0.2546	0.2536	0.1988
9	0.2318	0.2345	0.1885
10	0.2127	0.2168	0.1766
Oracle Recall / Precision			
	0.4319	0.4211	0.3059
% Perfect classifications			
	6.15%	6.56%	2.87%

Table B9. Average recall and precision at 10 cut-off levels, recall/precision at the oracle cut-off level, and percentage of perfect classifications for three levels of specificity in the Communicable Diseases specialty.

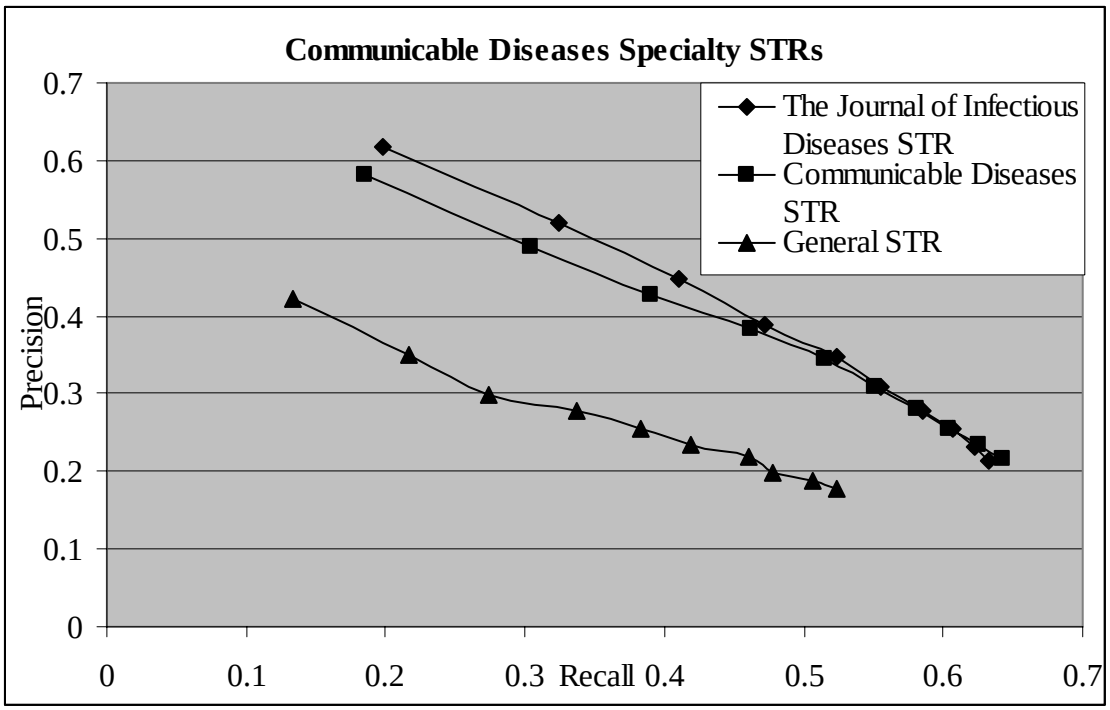

Figure B1. Recall and precision at 10 cut-off levels for three levels of specificity in the Communicable Diseases specialty.

Cut-off level	Obstetrics & Gynecology STR	Gynecology STR	General STR
Recall			
1	0.2334	0.2214	0.1562
2	0.3544	0.3356	0.2504
3	0.4340	0.4180	0.3274
4	0.4718	0.4701	0.3960
5	0.5149	0.5176	0.4394
6	0.5546	0.5448	0.4859
7	0.5790	0.5769	0.5179
8	0.5916	0.6057	0.5376
9	0.6111	0.6242	0.5722
10	0.6180	0.6494	0.5931
Precision			
1	0.6283	0.5948	0.4312
2	0.4907	0.4684	0.3532
3	0.4114	0.3941	0.3209
4	0.3411	0.3364	0.2900
5	0.2989	0.3004	0.2595
6	0.2701	0.2646	0.2404
7	0.2438	0.2432	0.2188
8	0.2189	0.2249	0.1998
9	0.2012	0.2074	0.1896
10	0.1836	0.1952	0.1777
Oracle Recall / Precision			
	0.4170	0.4022	0.3304
% Perfect classifications			
	8.55%	7.43%	4.46%

Table B10. Average recall and precision at 10 cut-off levels, recall/precision at the oracle cut-off level, and percentage of perfect classifications for three levels of specificity in the Gynecology specialty.

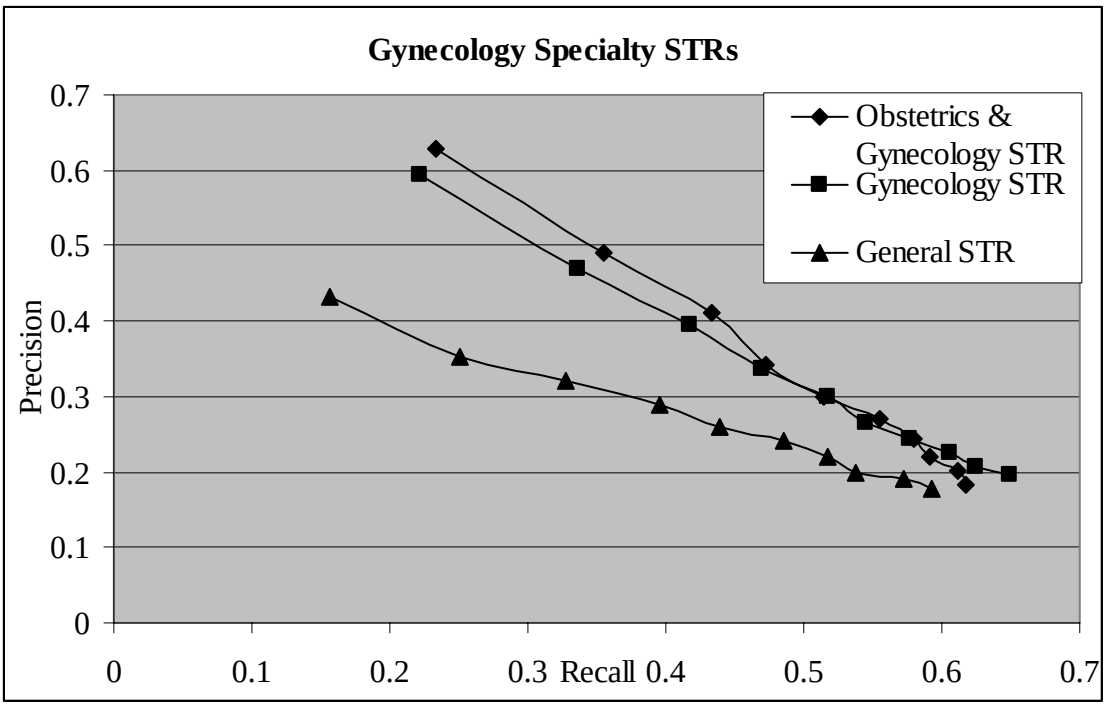

Figure B2. Recall and precision at 10 cut-off levels for three levels of specificity in the Gynecology specialty.

Cut-off level	Clinical Orthopaedics & Related Research STR	Orthopedics STR	General STR
Recall			
1	0.2396	0.2514	0.1898
2	0.3920	0.4047	0.3062
3	0.4866	0.5041	0.3691
4	0.5422	0.5529	0.4369
5	0.5898	0.5906	0.4795
6	0.6157	0.6236	0.5075
7	0.6330	0.6487	0.5373
8	0.6407	0.6655	0.5587
9	0.6620	0.6852	0.5861
10	0.6731	0.6943	0.6069
Precision			
1	0.5733	0.6034	0.4655
2	0.4828	0.5065	0.3750
3	0.4109	0.4310	0.3075
4	0.3502	0.3621	0.2769
5	0.3060	0.3095	0.2474
6	0.2694	0.2759	0.2220
7	0.2395	0.2475	0.2032
8	0.2128	0.2225	0.1859
9	0.1964	0.2035	0.1734
10	0.1806	0.1862	0.1612
Oracle Recall / Precision			
	0.4461	0.4588	0.3319
% Perfect classifications			
	12.50%	12.93%	5.60%

Table B11. Average recall and precision at 10 cut-off levels, recall/precision at the oracle cut-off level, and percentage of perfect classifications for three levels of specificity in the Orthopedics specialty.

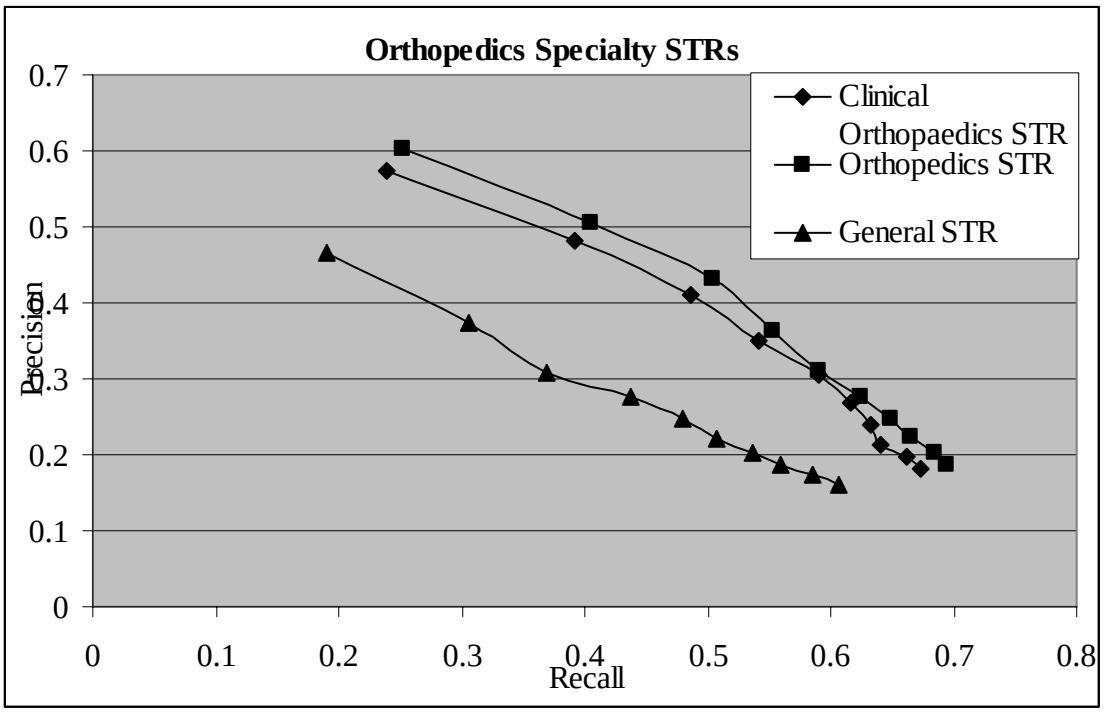

Figure B3. Recall and precision at 10 cut-off levels for three levels of specificity in the Orthopedics specialty.

Cut-off level	Specialty STR	General STR	Association Weight	Prediction STR
Recall				
1	0.2126	0.1610	0.1599	0.1443
2	0.3306	0.2484	0.2410	0.2204
3	0.4117	0.3374	0.2963	0.2715
4	0.4674	0.3853	0.3381	0.3064
5	0.4936	0.4266	0.3695	0.3272
6	0.5194	0.4556	0.3834	0.3494
7	0.5439	0.4842	0.4050	0.3718
8	0.5592	0.5073	0.4192	0.3818
9	0.5793	0.5185	0.4299	0.3930
10	0.5945	0.5312	0.4374	0.4047
Precision				
1	0.5394	0.4091	0.4121	0.3667
2	0.4379	0.3379	0.3167	0.2894
3	0.3717	0.3040	0.2616	0.2434
4	0.3220	0.2598	0.2265	0.2106
5	0.2752	0.2333	0.2000	0.1824
6	0.2434	0.2091	0.1742	0.1641
7	0.2186	0.1900	0.1576	0.1489
8	0.1973	0.1746	0.1436	0.1345
9	0.1818	0.1599	0.1317	0.1232
10	0.1688	0.1482	0.1212	0.1148

Table B12. Automatic determination of the specialty in the Ohsumed collection for 330 queries. Recall and precision at 10 cut-off levels for Specialty STR: predicting the specialty by journal descriptor (perfect prediction); General STR: no specialty prediction; Association Weight: predicting the specialty by highest association weight (over all 33 specialties); and Prediction STR: predicting the specialty with a STR for the 33 Ohsumed specialties.