Ontology-based Document Representation for Biomedical Information Retrieval

Fabrice Camous

A dissertation submitted in fulfilment of the requirements for the award of

Doctor of Philosophy

to the

Dublin City University
School of Computing

Supervisors: Stephen Blott and Alan F. Smeaton

July 2007

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Fabrice Camous (Student Number 53148614), Dublin City University,
Date: 24/07/2007, Signature: .

Acknowledgements

First, I would like to thank my supervisors, Stephen Blott and Alan Smeaton, for giving me the opportunity to experiment with the world of academic research. Their support allowed me to complete this PhD dissertation successfully. Furthermore, I would like to thank Enterprise Ireland for financing my PhD project.

Many people were involved in this long process that started with my going back to university and ended with the completion of this dissertation. First, I would like to thank Siobhán for encouraging me to go back to university and work in research, after I had considered it for many years. Since the first day I started in DCU, I greatly appreciated the friendly atmosphere that surrounded my work in the School of Computing, especially within the CDVP group. Special thanks to Neil, Kieran, Paul Ferguson, Colm, Cathal, Hyowon, Jiamin, Paul Browne, Sandra, Georgina, Pete, Sinéad, Aidan, James, Mike, and Niall who greatly contributed to the achievement of my PhD. I am particularly grateful to Hervé for spending hours helping and advising me on my work. Very special thanks also to Ríona who made the intensity of the dissertation writing much easier to bear.

A last, I would like to thank my parents, Nicole and André, who worked very hard to provide the best environment for the successful education of their children, from primary to higher education.

Contents

Abstract

In the current era of fast sequencing of entire genomes, more data is becoming available for analysis. This data analysis, in turn, leads to an increasing amount of scientific publications. Consequently, biologists spend a considerable part of their time searching the biomedical literature. This avoids expensive experiment duplications in wet labs, and provides inspiration for new hypotheses.

Unfortunately, the fast growth of biological information, in the form of free-text, has led to a lack of standard in the naming of biological entities. As a result, different genes are referred to with the same name, or acronym, and different names refer to the same gene. The ambiguity of free-text is problematic, as the success of a search often relies on the matching of a query term with a term contained in the document representation.

Biomedical ontologies, when available, can help disambiguate the information expressed in free-text: they provide unique terms to represent concepts and therefore counterweight the occurrence of synonyms and polysems in free-text. They also contain information about the relationships between concepts. This information can be used to understand and evaluate semantic similarities between concepts.

The largest repository of biomedical research literature in the world, MEDLINE, is an entry point to biomedical information for most biologists (Hersh et al., 2004). The Medical Subject Headings (MeSH) is the controlled vocabulary used in MEDLINE to annotate the conceptual content of biomedical articles. The annotations include information about the importance of MeSH concepts in the article, and their contexts. The MeSH ontology is organized in several hierarchies that indicate the level of specificity of the MeSH concepts. This hierarchical information can be used to generate semantic similarities between concepts.

Our motivation is the improvement of MEDLINE search, as it is still a central information access point for biologists in spite of the growing availability of full journal articles on the Web. In particular, we focus on the use of the MeSH ontology to represent and retrieve biomedical articles. Although MeSH is widely used by current MELDINE search methods, we show that the information contained in MEDLINE MeSH annotations and the MeSH hierarchies is often overlooked.

We hypothesize that MeSH-based document representation can improve MEDLINE information retrieval. Specifically, our hypothesis is that the integration of information about concept relevance (from the MEDLINE annotation), and inter-concept similarities (from the MeSH hierarchies), will improve retrieval performance. We evaluate methods using such information to discriminate and compare MeSH concepts. Our methods are evaluated in the context of MEDLINE ad hoc document retrieval and document binary classifications. Our evaluations use standard datasets and metrics recently used at the Genomics track of the 2005 Text REtrieval Conference workshop.

Chapter 1

Introduction

1.1 Background

In the current era of fast sequencing of entire genomes, new experimental techniques make possible the rapid generation of high volumes of genomic data. Genes are expressed in cells through the translation of their DNA sequences into RNAs or proteins. In a single experiment, the expression of thousands of genes in cells under various conditions can be observed (Brown and Botstein, 1999). These experiments aim at finding connections between genes and connections among genes and diseases.

The analysis of this data in turn creates a high volume of genomic information in the form of research articles. The data collected from experiments can be explored and evaluated against new hypotheses. The conclusive acceptance or rejection of an hypothesis is itself likely to be reported in a scientific article.

As a consequence, biologists spend a considerable part of their time searching the research literature. The information contained in the literature helps biologists to avoid the duplication of expensive laboratory experiments on a particular gene or gene product. A recently-discovered and poorly-defined gene can be similar or related to another well-studied gene that is thoroughly described in the literature, and in a more structured way in genomic databases. What is already known about a related gene can provide guidance in the study of a new gene, and consequently reduces the amount of costly experiments in wet laboratories. Moreover, the efficient search of genomic information is crucial in allowing the generation of new hypotheses. Biomedical databases have shown to be a mine for new assumptions and a basis for new experimental work (Hearst, 1999; Blagosklonny and Pardee, 2002).

The efficient search of biomedical literature is challenged by the ambiguities that are inherent in the use of natural languages such as English (Sussna, 1993). Examples of such ambiguities are the occurrences of polysems and synonyms. The English word "bank" is a polysem as it can refer to a financial institution or the area next to a river. "Car" and "automobile" are two synonyms naming the same concept. Human readers use context and their own contextual knowledge to resolve such ambiguities. However, the automated disambiguation of information by machines is less straightforward. Indeed, the satisfactory replication in a machine of the human ability to contextualize has yet to be achieved.

Perhaps surprisingly, the scientific and domain-dependent language used in biomedical research articles does not escape the ambiguity problems associated with domain-independent literature. For example, the naming of new bio-entities, such as genes and proteins, has eluded all standardization attempts. Within the same genome, a gene can have several names and a name can refer to several genes. For example, a human gene that controls cell adhesion during immune responses and officially named SELL has fifteen different aliases in the literature (Pearson, 2001). On the other hand, the name PAP refers to five un-related human genes that are involved in different biological processes. Similar or identical genes are also given different names because they are discovered by different research teams working on different organisms. The name of a gene can relate to its role in the determination of the phenotype of an organism (colour of eyes, presence or absence of wings), or it can relate to its relationship to other genes. For example, the same gene is called "armadillo" by Drosophila geneticists because of the appearance of fly embryos when the gene is defective, and "catenin" by Mouse geneticists because it belongs to a certain gene family. Different naming practices inevitably lead to the occurrence of synonyms and polysems among gene names. Jenssen et al. (2001) reports on the identification of gene names in more than 10 million biomedical abstracts from the MEDLINE database. 2,796 gene symbols out of the 24,443 symbols identified are connected with multiples genes. Gene name abbreviations are also found to correspond to different genes or to general domain concepts (e.g. "SELL" mentioned above).

One solution to the ambiguity of gene names is to identify genes with their associated DNA sequences. The DNA sequences can be used to identify and group genes independently from the gene names associated with the sequences. The gene sequence determines the sequence of a particular protein which assumes a particular function in the activity of the cell and the life of the whole organism. Genes with identical or similar sequences across different organisms can be identified as identical or strongly related regardless of the name they were given. However, as Lord et al. (2003a) points out, sequence similarity is not enough to account for other possible relationships between genes, such as the biological process they take part in or the cellular location where they operate.

1.2 Ontologies

One solution to the ambiguity of free text is to represent information with terms drawn from biomedical or genomic ontologies. By providing unique associations between terms and concepts, ontologies reduce the occurrence of polysems and synonyms. Various ontologies are available in the biological domain. Amongst them, the Gene Ontology (GO) (Gene Ontology Consortium, 2000, 2001, 2004) is used to annotate genes and gene products across several model organism databases, and the Medical Subject Headings (MeSH)[1] are used by MEDLINE database indexers to describe the content of biomedical research articles.

Ontologies contain knowledge about the relationships that exist between con-

[1] http://www.nlm.nih.gov/mesh/

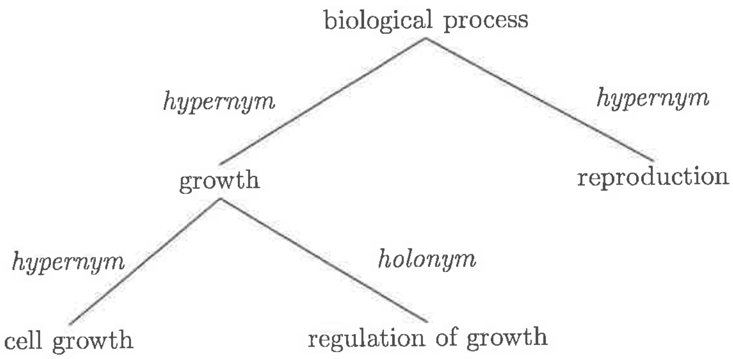


Figure 1.1: A simplified representation of the Gene Ontology "*biological process*" hierarchy.

cepts. The relationships generally include hypernymy (*is-a*), hyponymy (*has-instance*, inverse of hypernymy), holonymy (*is-part-of*), and meronymy (*has-part*, inverse of holonymy). For example, in the Gene Ontology (see Figure 1.1), the concept "*biological process*" is an hypernym of the concept "*growth*". However, the concept "*growth*" is a holonym of the concept "*regulation of growth*". The ontology is usually organized hierarchically with the most general concepts at the top and the most specific concepts at the bottom. If the hierarchy is a tree, the most general concept is called the root of the tree and the most specific concepts are called the leaves of the tree. Such trees or hierarchies are sometimes referred to as the semantic network of the ontology. The 2004 version of MeSH contains 15 hierarchies and the relationships between concepts are *narrower-than*, which includes the hyponymy and meronymy relationships, and *broader-than*, which includes the hypernymy and holonymy relationships.

While comparing a biologist's information need with a document stored in a database, an ontological representation of both the query and the document offer advantages over text-based representations. In particular, because ontologies use standard terms, they eliminate the problem of the ambiguity of free text. Consequently, the terms from a document and a query that match will correspond to identical concepts, and terms that do not match will relate to different concepts.

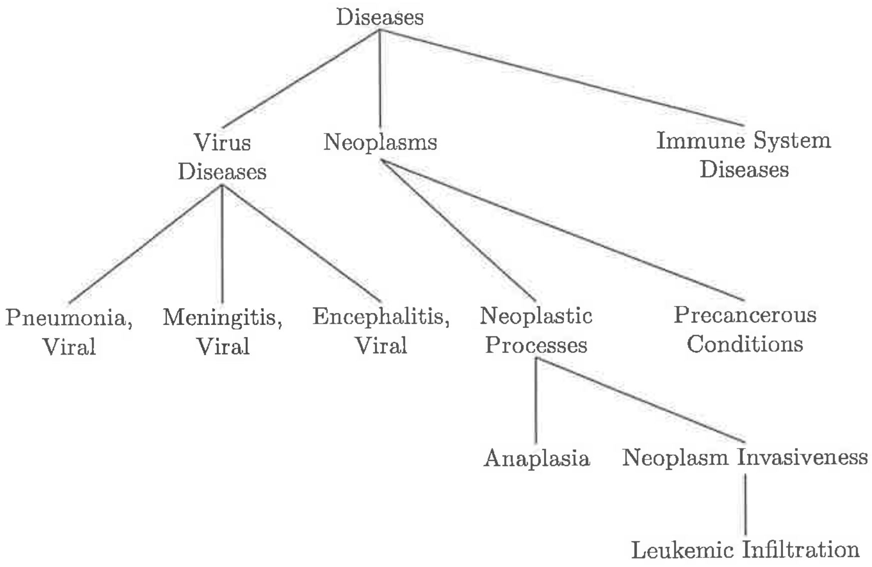


Figure 1.2: A simplified representation of the MeSH "*Diseases*" hierarchy.

However, the concepts can be different but related to a certain degree. In the MeSH ontology, the semantic network can be used to suggest the degree of relatedness between two distinct concepts. Figure 1.2 shows a simplified representation of the MeSH "*Diseases*" hierarchy. The hierarchy indicates that "*Neoplastic Processes*" and "*Precancerous conditions*" are both narrower concepts of the "*Neoplasms*" concept. This relationship suggests that the two MeSH concepts are related; more so, for example, than some other pair of concepts from distant parts of the hierarchy.

1.3 Thesis Statement

Despite the growing availability of full-text journal articles on the Web, MEDLINE remains a major entry point to biomedical research for biologists (Hersh et al., 2004, 2005). Improving the search capabilities of MEDLINE is therefore important for biomedical research. The MEDLINE database provides a MeSH-based representation of biomedical research articles. MeSH-based representations cover the conceptual content of entire articles. They are created by human indexers at the U.S. National

Library of Medicine (NLM)[2]. MeSH-based representations have been shown to be consistent across different indexers (Funk et al., 1983). This makes MeSH-based document representation, search and retrieval an attractive proposition.

1.3.1 Central Hypothesis

Our central hypothesis states that biomedical information retrieval can be improved by enhancing basic MeSH-based binary representations (only registering the presence or absence of MeSH concepts in queries or documents) with information contained in the MEDLINE MeSH fields and in the MeSH hierarchy. This central hypothesis contains two main assumptions:

- first, that non-hierarchical information about MeSH concepts can be exploited to create non-binary representations for documents and queries that increase the precision of MEDLINE retrieval (non-hierarchical hypothesis), and

- second, that hierarchical information about MeSH concepts can help us to compare concepts contained in documents and queries, and lead to better recall and precision for MEDLINE retrieval (hierarchical hypothesis).

1.3.2 Non-hierarchical Hypotheses

First, we hypothesize that corpus information and information extracted from the structure of MeSH fields can help to improve the precision of MEDLINE document retrieval. In particular:

- corpus frequency of concepts can help to evaluate the relevance of concepts in the document,

- MeSH-based representations can be improved by discriminating between concepts presented as major themes in the MeSH fields and those presented as minor themes, and

[2] http://www.nlm.nih.gov/

- associations between concepts that are found in the MeSH fields can be used to make representations more discriminative: such associations provide more context for the concepts used to describe the content of the documents.

1.3.3 Hierarchical Hypotheses

Second, we hypothesize that MeSH-based representations of documents can be further enhanced with the MeSH hierarchical information. Two methods can be used to integrate hierarchical information:

1. the hierarchy can be integrated to extend document representations, or

2. the hierarchy can be used to compare query concepts to document concepts, and concepts from different documents.

As queries and documents usually contain more than one concept, comparing documents with queries leads to the issue of combining inter-concept similarities. Combination approaches either consider all possible inter-concept similarities, or a selection of them.

MeSH concepts are located in different parts of the hierarchy that correspond to broad medical categories ("*Anatomy*", "*Diseases*", "*Chemical and Drugs*"). It can be hypothesized that concepts from different categories are too different to be compared. However, allowing concepts to be compared across categories may be beneficial for recall. Furthermore, some categories may be closer to each other than others.

The MeSH hierarchy includes nodes (concepts) and edges (relationships). It can be assumed that the edges correspond to the same semantic distances. Nevertheless, evaluating the variation of the edge semantic distances within the hierarchy can improve the accuracy of inter-concept semantic measures and increase retrieval precision.

Consequently, we develop in this dissertation secondary hypotheses regarding:

1. the combination of inter-concept semantic similarities,

2. the different parts of the MeSH hierarchy, and

3. the variation of edge distance in the MeSH hierarchy.

1.3.4 Evaluation

In this dissertation, our hypotheses are evaluated with respect to two biomedical information retrieval problems.

- The ad hoc retrieval of MEDLINE documents: we use the *trecGen2005* collection, a subset of MEDLINE containing over 4.5 million documents and 50 genomic queries with associated relevance information.

- The binary classification of MEDLINE documents for the annotiation of mouse genes with terms from the Gene Ontology: the aim is to automate the triage of MEDLINE documents likely to contain evidence to support the annotation of Mouse genes with concepts from the Gene Ontology.

These problems were chosen because MEDLINE remains a central source of information for most biologists. The first problem simulates the search of MEDLINE in relation to several genomic topics, whereas the second simulates the triage of documents for a particular topic of interest (Mouse genes).

1.3.5 Dissertation Plan

The dissertation is organized in the following manner. Chapter 2 introduces measures that use ontology hierarchies to evaluate the semantic distance or similarity between concepts. Chapter 3 presents the related research in the use of MeSH-based document representation for MEDLINE document ad hoc retrieval, classification, and clustering. Chapter 4 describes our hypotheses and methods. Chapters 5 and 6 present our experimental set-up for the evaluation of our previously formulated hypotheses. They include the description of two document retrieval evaluations organized by the Genomics track of the 2005 Text Retrieval Conference (TREC): the first task is an ad hoc retrieval task (Chapter 5), and the second involves the binary classification of MEDLINE documents (Chapter 6). Finally Chapter 7 concludes the dissertation and introduces future research directions.

Chapter 2

Network-based Semantic Measures

In this chapter we present a set of network-based semantic measures and the intuitions behind them. We point out the lack of large-scale evaluations regarding these measures, especially with the MeSH ontology. The contribution of the work described in this dissertation is partly to provide a strong evaluation framework for network-based semantic measures using the MeSH network.

Ontological representations of documents and queries often include several concepts. For example, MEDLINE records usually contain 10-12 fields containing concepts from the MeSH ontology, and queries are often expressed as a set of such concepts.

The comparison of two ontological representations involves two operations:

1. the comparison of pairs of concepts (one concept from each representation), and

2. the combination of such comparisons to produce an overall similarity measure.

Inter-concept comparisons can be derived from the semantic network underlying the ontology. A semantic network is broadly a set of nodes (representing concepts) and a set of edges, or links (representing relationships between the concepts). In particular, hierarchical relationships organize concepts in several levels of increasing specificity. For example, "*Diseases*" is a broad medical concept, whereas "*Anapla-*

sia" is more specific.

Network-based semantic measures are measures that use ontological relationships to evaluate the semantic distance or similarity between two concepts. In this chapter we give a review of network-based semantic measures. First, some background is given on WordNet, a domain-independent ontology. Most network-based semantic measures were developed and evaluated with WordNet. However, we also introduce the Medical Subject Headings (MeSH), the ontology used in this dissertation for the representation of biomedical information.

2.1 Background

2.1.1 WordNet

WordNet is a domain-independent English-language lexical reference (Miller, 1990). We briefly present WordNet in this section as some network-based semantic measures discussed later in this chapter were evaluated on it. However, we will not use WordNet in this dissertation, as an important part of our contribution is precisely to evaluate network-based semantic measures on a medical-domain-dependent ontology, MeSH, that we present in the next section.

WordNet contains nouns, verbs, adjectives and adverbs that are organized into approximately 120,000 synonym sets (synsets). The WordNet semantic network includes nine types of relationships: hypernymy (*is-a*), hyponymy (*has-instance*, inverse of hypernymy), meronymy (*component-object, member-collection, stuff-object*: 3 relations), holonymy (*object-component, collection-member, object-stuff*: 3 relations), antonymy (inverse of synonymy), and synonymy (actually intra-node, within the synsets). Most of the semantic measures discussed below concentrate on the noun part of WordNet, and hypernymy/hyponymy relationships. The noun part is the most developed and is organized into twenty-five hierarchies or primitive groups, each with a generic concept or *unique beginner* as its root (Miller, 1990). The hierarchies are connected by an abstract root node that is used by some semantic measures. WordNet version 1.7.1 contains 74,488 noun synsets and the maximum depth of the hierarchy (the distance in terms of edges between a unique beginner and a leaf node) is 14 (Devitt and Vogel, 2004).

Although WordNet was initially designed to address problems related to the domains of psychology and cognitive science, WordNet was also used to solve information management and retrieval problems, such as free-text disambiguation (Sussna, 1993), and query-document similarity calculation (Richardson and Smeaton, 1995). Moreover, the noun part of WordNet, used by most network-based measures to compare concepts, is organized in hierarchies containing relationships close to the ones used by the MeSH ontology. This suggests that network-based measures evaluated with WordNet can easily be adapted to the structure of the MeSH hierarchies.

A detailed description of WordNet is beyond the scope of the present dissertation. A full discussion on the design of WordNet and its theoretical foundations can be found in Fellbaum (1998).

2.1.2 The Medical Subject Headings (MeSH) 2004

MeSH is a biomedical controlled vocabulary developed by the U.S. National Library of Medicine (NLM)[1] since 1960. The goal of MeSH is to provide a reproducible conceptual partitioning of biomedical knowledge and information not only for indexing purposes but also for the support of information retrieval (Nelson et al., 2001). In particular, MeSH aims at providing distinctive concepts that are accessible to users and reflect the current knowledge in bio-medicine. MeSH is continually revised and updated by the MeSH staff in collaboration with indexers and experts to follow the evolution of the research literature[2].

[1] http://www.nlm.nih.gov/, last accessed: 19 January 2007

[2] http://www.nlm.nih.gov/pubs/factsheets/mesh.html

Retrieval Usefulness

From the introduction of MEDLINE in 1971, MeSH played a central part in the retrieval of articles. However, the success of a search based on MeSH is highly dependent on the user's knowledge of MeSH and the annotation practices. The usefulness of MeSH needs to be examined not from the point of view of the librarian but from the point of view of the biologist who may not know the best MeSH formulation of his/her information need.

The usefulness of MeSH for MEDLINE retrieval was examined before (Hersh et al., 1994a; Yang and Chute, 1993; Srinivasan, 1996a). A consensus seems to be that MeSH does not perform as well as free-text for searching MEDLINE. However, this is partly explained by the use of free-text terms to search the MeSH fields of records, i.e., if document are represented with MeSH concepts, queries are still represented with free-text terms. It was shown that when query free-text terms are mapped to MeSH concepts, using relevance information for example, the search of MeSH fields perform similarly as the search of free-text (Yang and Chute, 1993). Moreover, there is a substantial amount of evidence that MeSH and free-text, used in combination, improve the performance of text-only indexing. Table 2.1 shows improvements obtained by adding MeSH to free-text using various methods and various collections. The pre-retrieval and post-retrieval combinations refer to two distinct methods to combine text and MeSH representation that we will describe in details in Chapter 3. TrecGen is the collection used in this dissertation and it will also be presented in Chapter 3.

Internal Structure

MeSH 2004 includes 22,430 descriptors, 83 qualifiers, and 141,455 supplementary concepts. Descriptors are the main elements of the vocabulary. The concept covered by a descriptor can be expressed by various synonymous terms, but only one term, the MeSH *preferred term*, is used to refer to the concept. Note that this *preferred term* can be a single word (e.g. "*Neoplasms*"), a group of words (e.g. "*Neoplasm*

Table 2.1: MeSH impact on MEDLINE retrieval

Collections - Methods	Pre-retrieval combination	Post-retrieval combination
Collection from Hersh et al. (1994b) 2344 docs, 75 queries	- Srinivasan (1996a): +7.3% in 11-point average precision	- Srinivasan (1996a): +16.4% in 11-point average precision
OHSUMED collection (Hersh et al., 1994a) 348,556 docs, 101 queries	- Hersh et al. (1994a): + 9% performance improvement - Srinivasan (1996a): +6.3% in 11-point average precision	- Srinivasan (1996a): +12.9% in 11-point average precision
TrecGen collection (Hersh et al., 2004, 2005) 4.5 million docs, 50 queries	- Abdou et al. (2005): +9% in mean average precision	- Kraaij et al. (2004): +1.6% in mean average precision

Invasiveness"), or even a group of words combined with a comma (e.g. "*Pneumonia, Viral*"). Figure 2.1 shows an example of the content of a MeSH descriptor record. The MH field contains the *preferred term* to designate the concept. The MN fields indicate tree locations, effectively the positions of the descriptor in the MeSH semantic network. Definitions of some of the descriptor record's fields are given in Table 2.2.

Qualifiers are concepts that express a particular aspect of the concepts covered by descriptors. They are used to give more context and specification to descriptors. For example, the same descriptor "*Alcoholism*" can be given different contexts whether it is associated with the qualifier "*mortality*", or the qualifier "*therapy*". Qualifiers can be expressed by a single word (e.g. "*diagnosis*"), a group of words (e.g. "*radionuclide imaging*"), or a group of words combined with an ampersand (e.g. "*administration & dosage*").

Supplementary concepts are additional medical concepts that relate mainly to chemicals and drugs. Supplementary concept records contain information about their associations with descriptors. This information can be used to point to a descriptor for the conceptual description of medical information. During the annotation of biomedical articles submitted to MEDLINE, human indexers can select descriptors pointed to by supplementary concepts found in the articles. For example, if the supplementary concept "*flavophosphine*", a chemical, is found, an indexer

```
*NEWRECORD
RECTYPE
MH = Gene Library
ENTRY = DNA Libraries
ENTRY = Gene Libraries
ENTRY = Libraries, DNA
ENTRY = Libraries, Gene
ENTRY = Libraries, cDNA
ENTRY = Library, DNA
ENTRY = Library, Gene
ENTRY = Library, cDNA
ENTRY = cDNA Libraries
MN = G05.275.195
MN = G05.331.599.110.410
FX = Combinatorial Chemistry Techniques
FX = DNA, Recombinant
FX = Databases, Nucleic Acid
MH_TH = NLM (1990)
ST = T028
ST = T170
AN = do not confuse with GENOMIC LIBRARY;
do not confuse X ref GENE BANK
with BIOLOGICAL SPECIMEN BANKS
PI = Base Sequence (1978-1989)
PI = Cloning, Molecular (1980-1989)
PI = Plasmids (1978-1985)
MS = A large collection of cloned DNA fragments from a given
organism, tissue, organ, or cell type.  It may contain complete
genomic sequences (GENOMIC LIBRARY) or complementary DNA sequences,
the latter being formed from messenger RNA and lacking intron sequences.
PM = 90
HN = 90
MR = 20010725
DA = 19890515
DC = 1
UI = D015723
```

Figure 2.1: Example of a MeSH descriptor record in ASCII format

Table 2.2: Descriptor record's fields definition

AN	ANNOTATION
AQ	ALLOWABLE TOPICAL QUALIFIERS
CX	CONSIDER ALSO XREF
DA	DATE OF ENTRY
DC	DESCRIPTOR CLASS
DE	DESCRIPTOR ENTRY VERSION
DS	DESCRIPTOR SORT VERSION
DX	DATE MAJOR DESCRIPTOR ESTABLISHED
EC	ENTRY COMBINATION
FX	FORWARD CROSS REFERENCE (SEE ALSO REFERENCE)
HN	HISTORY NOTE
M##	BACKFILE POSTINGS
MH	MESH HEADING
MN	MESH TREE NUMBER
MR	MAJOR REVISION DATE
MS	MESH SCOPE NOTE
N1	CAS TYPE 1 NAME
OL	ONLINE NOTE
PA	PHARMACOLOGICAL ACTION
PI	PREVIOUS INDEXING
PM	PUBLIC MESH NOTE
PX	PRE EXPLOSION
RH	RUNNING HEAD, MESH TREE STRUCTURES
RN	CAS REGISTRY/EC NUMBER
RR	RELATED CAS REGISTRY NUMBER
UI	UNIQUE IDENTIFIER

Table 2.3: The 15 descriptor hierarchies

	descriptor hierarchy	number of elements
A:	Anatomy.	2,246
B:	Organisms.	4,386
C:	Diseases.	9,502
D:	Chemicals and Drugs.	13,911
E:	Analytical, Diagnostic and Therapeutic Techniques and Equipment.	3,242
F:	Psychiatry and Psychology.	978
G:	Biological Sciences.	2,593
H:	Physical Sciences.	596
I:	Anthropology, Education, Sociology and Social Phenomena.	526
J:	Technology and Food and Beverages.	302
K:	Humanities.	191
L:	Information Science.	405
M:	Persons.	214
N:	Health Care.	1,489
Z:	Geographic Locations.	482
	total	41,063

will be pointed to the descriptor "*Acridines*", a type of organic compound that is DNA-binding.

MeSH descriptors and qualifiers are organized into a semantic network that includes several hierarchies[3]. The MeSH hierarchies comprise fifteen descriptor hierarchies and twenty-three smaller and shallower qualifier hierarchies (some qualifier hierarchies have only one element, the root node). Tables 2.3 and 2.4 show the root concepts of the descriptor and qualifier hierarchies, respectively. The 22,430 descriptor concepts are associated with a total of 41,063 hierarchical locations (1.8 locations per descriptor on average), and the 83 qualifier concepts are associated with 99 locations (1.2 locations per qualifier on average).

The relationships within the MeSH hierarchies are of the *broader-than/narrower-than* type (Nelson et al., 2001). The *narrower-than* relationship includes the hyper-

[3] the term *hierarchy* is commonly used to refer to the main parts of the MeSH ontology, which are in fact single-rooted directed acyclic graphs

Table 2.4: The 23 qualifier hierarchies

	qualifier hierarchy	number of elements
Y01:	analysis	5
Y02:	anatomy & histology	8
Y03:	chemistry	5
Y04:	diagnosis	5
Y05:	etiology	12
Y06:	organization & administration	8
Y07:	pharmacology	10
Y08:	physiology	14
Y09:	statistics & numerical data	6
Y10:	therapeutic use	5
Y11:	therapy	9
Y19:	classification	1
Y21:	drug effects	1
Y23:	education	1
Y25:	ethics	1
Y27:	history	1
Y29:	injuries	1
Y31:	instrumentation	1
Y33:	methods	1
Y35:	pathogenicity	1
Y37:	psychology	1
Y39:	radiation effects	1
Y41:	veterinary	1
	total	99

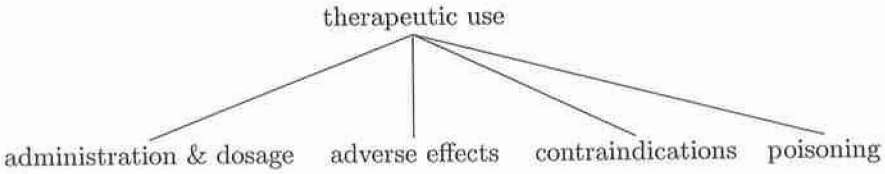


Figure 2.2: The "*therapeutic use*" qualifier hierarchy.

nymy (*is-a*) relationship and the meronymy (*part-of*) relationship. Inversely, the *broader-than* relationship includes the hyponymy (*has-instance*) relationship and the holonymy (*has-a*) relationship. A simplified version of the "*Diseases*" descriptor hierarchy is shown in Figure 1.2, and the full "*therapeutic use*" hierarchy can be seen in Figure 2.2.

Only a few network-based semantic measures were evaluated on the MeSH hierarchy, as will be shown in Section 2.3. One of the contributions of the work presented in this dissertation is the extrinsic evaluation of some network-based semantic measures on the MeSH hierarchy for MEDLINE retrieval (see Table 2.8). In the next section, various network-based semantic measures are presented.

2.2 Network-Based Similarity Measures

Network-based similarity measures are usually classified into two groups: edge-based measures and information-based measures (Budanitsky, 1999). Edge-based measures only use hierarchy information, whereas information-based measures combine concept frequency information from a corpus with hierarchy information.

In this dissertation, however, we use a different approach which classifies network-based measures into the two following categories:

1. measures that consider all hierarchy edges to correspond to the same semantic distance (simple edge weighting), and

2. measures that consider hierarchy edges to correspond to different semantic distances (complex edge weighting).

Both types of measures listed above can be applied either to the semantic comparison of two single concepts or to the comparison of groups of concepts. The semantic comparison of groups of concepts corresponds to different methods for combining inter-concept semantic comparisons.

In this section, we first introduce inter-concept measures using simple or complex edge weighting. We then describe different approaches to combine inter-concept measures in order to compare groups of concepts such as documents or queries. Finally, we present the existing evaluations available for network-based measures, and position our work in relation to these evaluations.

2.2.1 Inter-concept Measures

Simple Edge Weighting: Edge Count

The edge count approach considers that all edges between concepts in the hierarchy correspond to the same semantic distance. Figure 2.3 shows a partial representation of the MeSH network highly simplified for the sake of clarity. Note that we added an artificial "*MeSH*" node at the root in order to connect the two descriptor hierarchies "*Diseases*" and "*Chemical and Drugs*". With edge count, the semantic distance between "*Diseases*" and "*Neoplasms*" is assumed to be the same as the distance between "*Neoplasm Invasiveness*" and "*Leukemic Infiltration*" (i.e. one edge).

The assumption behind the edge count approach is that the semantic distance between two concepts is proportional to the number of edges that separate them in the hierarchy. For example, in Figure 2.3, the edge count between "*Pneumonia, Viral*" and "*Meningitis, Viral*" is 2, and the edge count between "*Pneumonia, Viral*" and "*Neoplastic Processes*" is 4. Therefore, under edge count, "*Meningitis, Viral*" is closer semantically to "*Pneumonia, Viral*" than "*Neoplastic Processes*" is.

In some hierarchies, there is more than one possible path between two concepts. In the MeSH ontology, concepts can have several parents and children. A widely-used solution to the multiple path problem is to consider only the shortest path between two concepts when calculating the semantic distance. Rada et al. (1989) uses this approach with an inter-concept measure equal to the shortest path p in the set of possible paths P between two concepts c_1 and c_2 in terms of edge count:

$$dist_{rada1}\left(c_1, c_2\right) = \min_{p \in P}\left(edge\_count_p\left(c_1, c_2\right)\right) \tag{2.1}$$

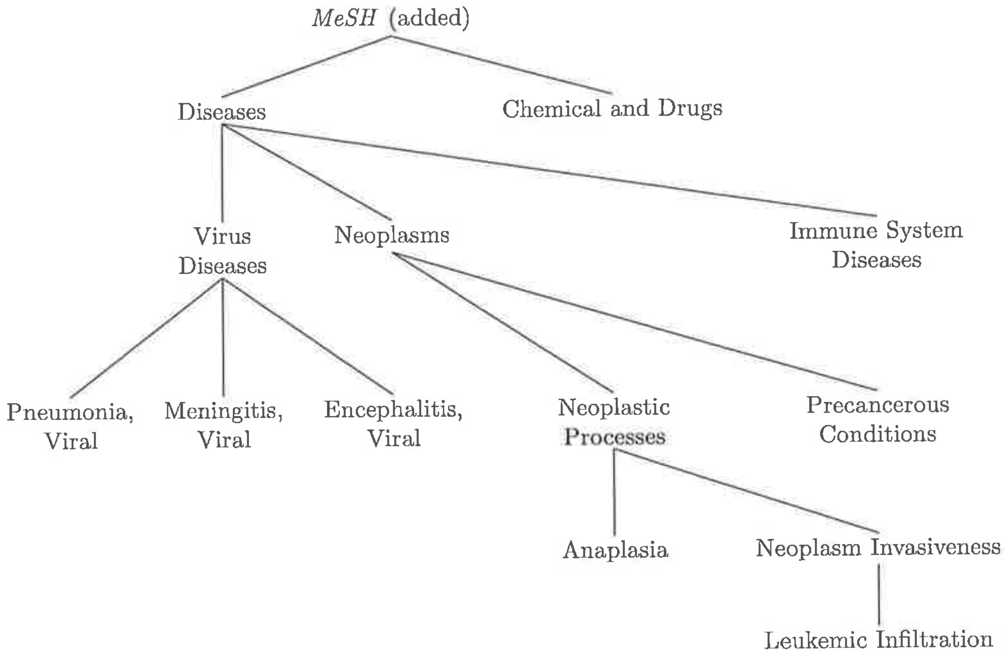


Figure 2.3: A partial representation of the MeSH hierarchy

where $edge\_count_p$ is the number of edges separating c_1 from c_2 on a path $p \in P$.

Complex Edge Weighting.

Unlike the edge count approach, the methods described in this paragraph assume that edges correspond to various semantic distances. For example, in Figure 2.3, it can be hypothesized that the edge distance between "*Anaplasia*" and "*Neoplastic Processes*" is shorter than the edge distance separating "*Virus Diseases*" from "*Diseases*". The two former concepts are located deeper in the hierarchy than the two latter. Therefore they are more specific and the edge separating them is expected to correspond to a smaller semantic distance.

We now present measures that evaluate the variation of the edge distance in the hierarchy with hierarchical parameters (hierarchy depth and/or density) and/or with corpus parameters (frequencies of the occurrences of the concepts in a corpus).

Hierarchy Information: Depth and Density. One way to evaluate edge distance variation is to exploit the variation of depth and density in the hierarchy. Sussna (1993) and Richardson and Smeaton (1995) report on observations that associate edge distance to hierarchy depth and density. In particular, the assumption is that edge distance is inversely proportional to hierarchy depth and density. The intuition behind this assumption is that:

- concepts located at high depths (high specificity) are closer semantically than concepts located at shallow depths (general concepts), and
- concepts located in dense areas of the hierarchy (detailed conceptual coverage) are closer semantically than concepts found in sparse areas (poor conceptual coverage).

The depth of a concept in a hierarchy is defined as its distance by the shortest path to the root of the hierarchy (the top concept) in terms of edge count. For example, in Figure 2.3, the depth of "*Diseases*" is 1 and the depth of "*Anaplasia*" is 4. In a hierarchy containing *broader-than/narrower-than* relationships such as the MeSH hierarchy, the depth of a concept is proportional to its specificity. Therefore "*Anaplasia*" is a more specific concept than "*Diseases*".

The density is positively correlated to the number of edges and negatively correlated to the number of nodes in the hierarchy. In graph theory, density of a graph (Preiss, 1999) is defined as:

$$\text{density} = \frac{|E|}{|N|^2}$$

where $|E|$ is the number of edges and $|N|$ the number of nodes. The hierarchy density can be calculated for each node by looking for example at the number of children of the node. The number of children for a node corresponds to the number of edges going down the hierarchy from that node. In Figure 2.3, concept "*Diseases*" has three children. In comparison, concept "*Neoplasms*" has only 2 children. Therefore

"*Neoplasms*" corresponds to a less dense area of the hierarchy than "*Diseases*" does. Edges located in dense areas are expected to correspond to smaller semantic distances than edges located in sparse areas.

Figure 2.4 gives information about the depth and density (first and second number inside the brackets, respectively) at each node of the partial MeSH network represented in Figure 2.3. To evaluate the density at a node, we use Jiang and Conrath (1997)'s method which counts the number of children of the node. According to the depth and density assumption, the edge between "*Virus Diseases*" and "*Pneumonia, Viral*" corresponds to a smaller semantic distance than the edge between "*Neoplasms*" and "*Neoplastic Processes*". Indeed, both "*Virus Diseases*" and "*Neoplasms*" have the same depth but the density is higher for "*Virus Diseases*". Similarly, the semantic distance between "*Diseases*" and "*Virus Diseases*" is expected to be higher than the distance between "*Virus Diseases*" and "*Pneumonia, Viral*". "*Diseases*" and "*Virus Diseases*" have the same density, but "*Virus Diseases*" is located deeper.

Depth Variation Integration. Some measures derive edge distance between two concepts from their respective depths, and from the depth of their lowest common ancestor (LCA) (Wu and Palmer, 1994; Ganesan et al., 2002). The LCA of two concepts is their deepest shared parent node. For example in Figure 2.4, the LCA of "*Anaplasia*" and "*Precancerous Conditions*" is "*Neoplasms*".

These measures assume that semantic distance is inversely proportional to the depths of the concepts and their LCA. The closer the concepts are to their LCA, the smaller the semantic distance is between them. This is consistent with the edge count method, as a closer LCA is equivalent to a shorter path between the concepts. In addition, the deeper the LCA for the same path length, the shorter the semantic distance between two concepts.

Wu and Palmer (1994) and Ganesan et al. (2002) use the following formula to

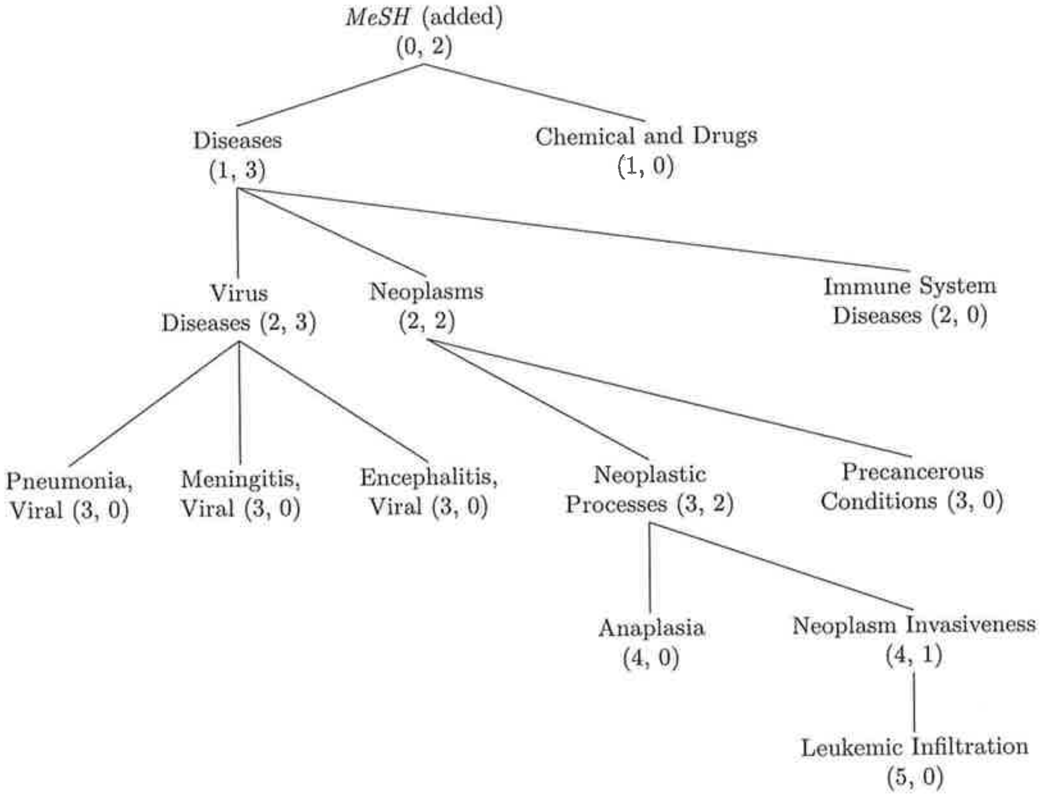


Figure 2.4: A partial MeSH hierarchy with depth and density information

measure the semantic similarity between a concept c_1 and a concept c_2:

$$sim_{wu}\left(c_1, c_2\right) = \frac{2 \times \text{depth}\left(\text{LCA}\left(c_1, c_2\right)\right)}{\text{depth}\left(c_1\right) + \text{depth}\left(c_2\right)} \tag{2.2}$$

This measure reduces edge distance as depth increases. Using Figure 2.4, we can see that:

1. sim_{wu} (*"Diseases"*, *"Virus Diseases"*) $= \frac{2\times1}{1+2} = \frac{2}{3}$, and

2. sim_{wu} (*"Virus Diseases"*, *"Pneumonia, Viral"*) $= \frac{2\times2}{2+3} = \frac{4}{5}$.

In the previous examples, we used the sim_{wu} to compare concepts directly connected by an edge. Additionally, the distance measure can be used to compare concepts separated by several edges in the hierarchy. In that case, hierarchy depth still influences edge distance:

1. sim_{wu} (*"Virus Diseases"*, *"Neoplasms"*) $= \frac{2\times1}{2+2} = \frac{1}{2}$, and

2. sim_{wu} (*"Neoplastic Processes"*, *"Precancerous Conditions"*) $= \frac{2\times2}{3+3} = \frac{2}{3}$,

which correspond to an average similarity between adjacent concepts of $\frac{1}{4}$ $\left(= \frac{\frac{1}{2}}{2}\right)$ in the first case, and $\frac{1}{3}$ $\left(= \frac{\frac{2}{3}}{2}\right)$ in the second. Next, we consider a measure that integrates both hierarchy depth and density to account for the variation of edge distances.

Integrating Depth and Density. Jiang and Conrath (1997) introduces a distance measure based on the assumption that edge distance is inversely proportional to hierarchy depth and density. The measure calculates the edge distance between a concept c_c and its parent concept c_p:

$$dist_{jiang1}\left(c_c, c_p\right) = \left(\beta + (1-\beta)\frac{\overline{E}}{E\left(c_p\right)}\right)\left(\frac{d\left(c_p\right)+1}{d\left(c_p\right)}\right)^{\alpha} \tag{2.3}$$

where $\alpha \geq 0$ and $0 \leq \beta \leq 1$ are the parameters that control the influence of hierarchy depth and density, respectively, on edge distance. In particular, depth

Table 2.5: Edge distance for different levels of depth and density

		depth	
		1	2
density	2	no example	1.63 (“*Neoplasms*” to “*Neoplastic Processes*”)
	3	1.45 (“*Diseases*” to “*Virus Diseases*”)	1.09 (“*Virus Diseases*” to “*Pneumonia, Viral*”)

influence increases as α increases, and density influence increases as β decreases. $E\left(c_p\right)$, the density at concept c_p, is defined as the number of children of c_p. $\overline{E}$, the average density of the hierarchy, is defined as the average number of children over the entire hierarchy. $d\left(c_p\right)$ is the depth of c_p in the hierarchy.

$dist_{jiang1}\left(c_c, c_p\right)$ decreases as depth and density increase. Figure 2.5 shows the edge distances using Formula 2.3 for the partial MeSH network represented in Figure 2.4. We set $\beta = 0$ (maximum sensitivity to density) and $\alpha = 1$ (moderate sensitivity to depth). The average density, $\overline{E}$, is equal to 2.17. Table 2.5 shows various edge distances with different levels of depth and density. Keeping density constant, increasing depth decreases edge distance. Inversely, keeping depth constant, increasing density decreases edge distance. Note that a high sensitivity to density was chosen ($\beta = 0$). This is why the highest edge distance of the network represented in Figure 2.5 is found for the highest depth value ($= 5$) but for the lowest density value ($= 1$):

$$dist_{jiang1}\left(\text{“\textit{Leukemic, Infiltration}”}, \text{“\textit{Neoplasm Invasiveness}”}\right) = 2.71$$

To calculate the semantic distance between two concepts c_1 and c_2 that do not share an edge, the weights of all the edges belonging to the shortest path between them are summed:

$$dist_{jiang2}\left(c_1, c_2\right) = \sum_{c_{ci} \in C} dist_{jiang1}\left(c_{ci}, c_{pi}\right) \tag{2.4}$$

where C is the set of concepts c_{ci} on the shortest path from c_1 to c_2 whose parents c_{pi} are also located on the shortest path. For example, in Figure 2.5:

$$\begin{aligned}
&dist_{jiang2}\left(\text{“\textit{Pneumonia, Viral}”}, \text{“\textit{Neoplastic Processes}”}\right)\\
&= 1.09 + 1.45 + 1.45 + 1.63\\
&= 5.62
\end{aligned}$$

and for an identical number of edges separating two concepts:

$$\begin{aligned}
&dist_{jiang2}\left(\text{“\textit{Diseases}”}, \text{“\textit{Leukemic Infiltration}”}\right)\\
&= 1.45 + 1.63 + 1.45 + 2.71\\
&= 7.24
\end{aligned}$$

A larger semantic distance is found because of higher depth and density levels on the shortest path from "*Diseases*" to "*Leukemic Infiltration*".

Information-based Measures. Another approach for measuring edge distance is to derive it from the distribution of concepts in a corpus. Measures using this approach are known as information-based measures. The frequencies of concepts are used to determine their information contents (ICs). The semantic distance expressed by a specific edge can be calculated by comparing the IC of the concepts sharing that edge. The assumption is that the edge distance decreases when the difference in IC between the two concepts sharing it decreases. Furthermore, differences in IC are expected to be lower for concepts located deep in the hierarchy and in dense areas.

Using information theory notions (Shannon, 1948), Resnik (1995) calculates the information content of a concept c as the negative of the log of the probability p of encountering c in the collection, based on the distribution of data in the corpus:

$$\mathrm{IC}\,(c) = -\log_2 p\,(c) \tag{2.5}$$

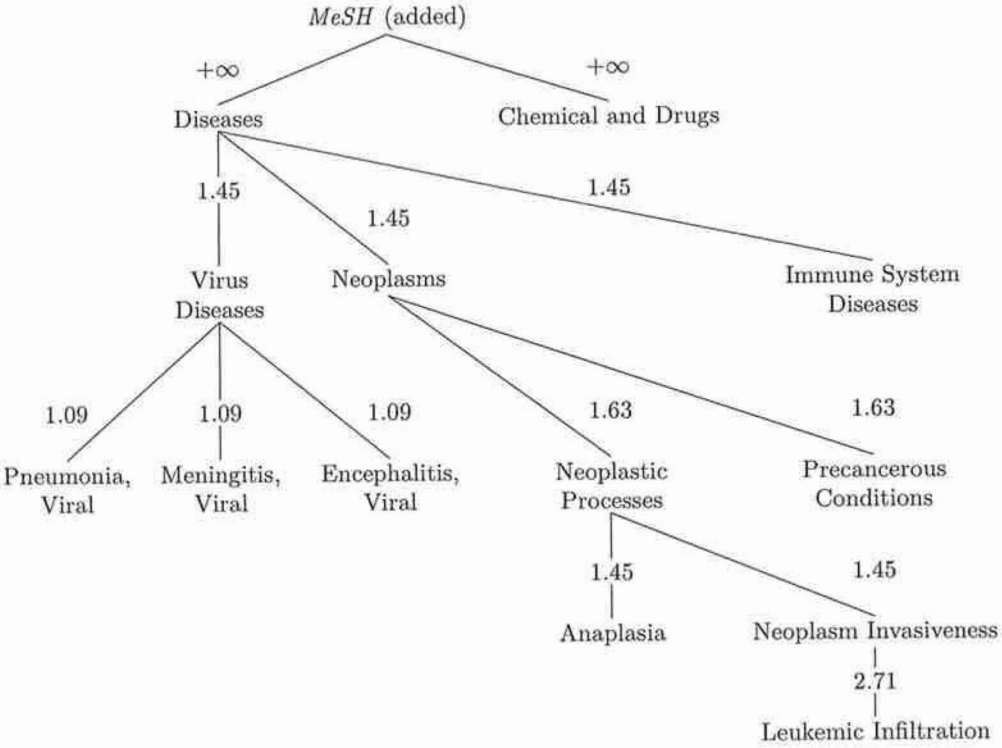


Figure 2.5: A partial MeSH hierarchy with edge weights (from network depth and density)

A specific concept is expected to have a lower probability than a general concept. Therefore, a specific concept is expected to have a higher IC than a general concept. Resnik (1995) calculates $p(c)$ as the relative frequency of concept c in the corpus. Note that an instance of a child concept is implicitly also an instance of its parents. As a consequence, for any concept c_c and its parent c_p, we have:

$$p(c_c) \leq p(c_p)$$

and hence:

$$\mathrm{IC}(c_c) \geq \mathrm{IC}(c_p)$$

As the root of the hierarchy is considered to occur in every document of the corpus, so its probability is equal to 1 and its information content is 0.

Figure 2.6 shows the concept frequencies, probabilities, and ICs calculated with Equation 2.5 (respectively, inside brackets) for the partial MeSH network representation of Figure 2.3. The frequencies used are calculated with the TrecGen04 collection, a subset of MEDLINE containing over 4.5 millions documents. The frequencies shown at each node include occurrences of all descendant nodes. Some nodes in Figure 2.6 ("*Chemical and Drugs*", "*Immune System Diseases*") do not appear in any documents in the collection. However, the figure includes only a few nodes of the MeSH hierarchy, and we can expect descendant nodes (not appearing in Figure 2.6) of these general nodes to occur in documents.

For the same level of depth and density, different IC increases are possible as we go down one level of specificity. For example, "*Anaplasia*" and "*Neoplasms Invasiveness*" are both children of "*Neoplastic Processes*" (IC (*Neoplastic Processes*) = 2.44). However, the IC increase for "*Anaplasia*" (= 9.51) is much higher than that for "*Neoplasm Invasiveness*" (= 2.46). Both concepts have the same depth and the same parent. Nonetheless, the probability of encountering "*Anaplasia*" in the corpus is lower than that of encountering "*Neoplasms Invasiveness*". Thus, the introduction of corpus information allows for further refinement of the edge distance

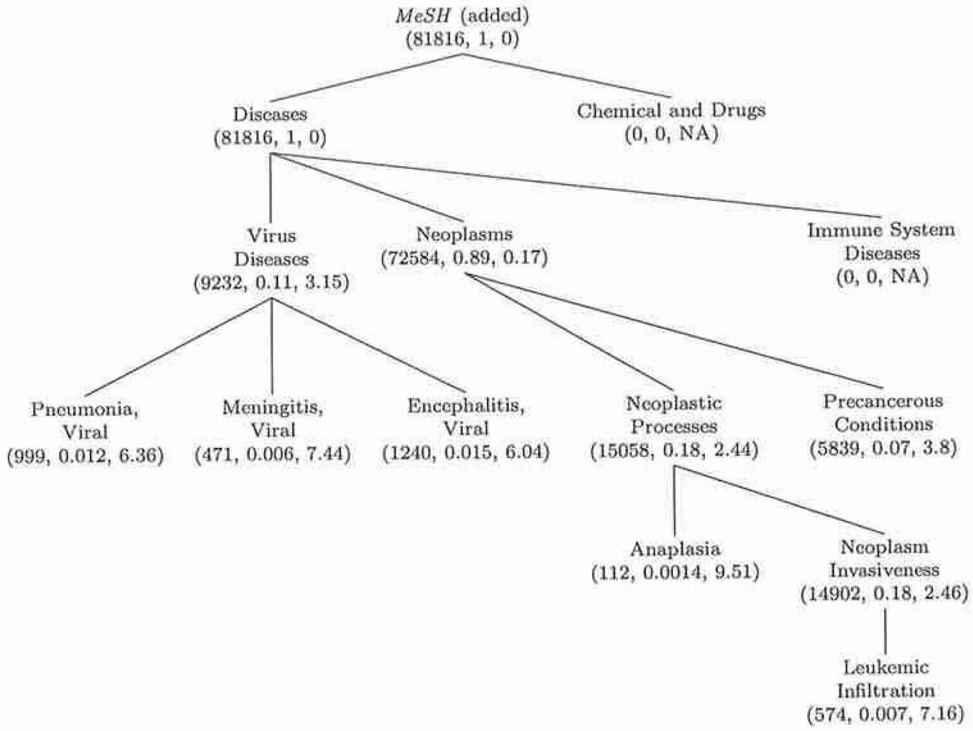


Figure 2.6: A partial representation of the MeSH hierarchy with frequencies, probabilities and information contents

calculation.

Several methods use the IC to calculate the semantic similarity between two concepts. Resnik (1995) defines the similarity between two concepts c_1 and c_2 as the IC of the concept with the highest IC contained in the set of their LCAs:

$$sim_{resnik}\left(c_1, c_2\right) = \max_i \left[-\log_2 p\left(\mathrm{LCA}_i\left(c_1, c_2\right)\right)\right] \tag{2.6}$$

where $\mathrm{LCA}_i\left(c_1, c_2\right)$ is contained in the set of LCAs of concepts c_1 and c_2. The intuition behind Formula 2.6 is that the similarity between two concepts is the extent to which they share information. This shared information is given by the IC of their LCA in the hierarchy.

When Resnik (1995)'s measure is applied to an individual edge separating a child concept c_c from a parent concept c_p, it always returns the information content of the parent concept:

$$sim_{resnik}\left(c_c, c_p\right) = -\log_2 p\left(c_p\right)$$

This is consistent with the hypothesis that edge distance is reduced when the concepts are more specific: the IC can only go up as concepts are located deeper in the hierarchy, so the edge distance can only go down.

When Resnik (1995)'s measure is used to calculate the semantic similarity between two arbitrary concepts in the hierarchy, some unsatisfactory effects are observed. First, the measure does not take in account the length of the path between two concepts. All pairs of concepts with the same LCA have the same similarity value. Figure 2.6 shows that:

1. sim_{resnik} (*"Neoplatic Processes"*, *"Precancerous Conditions"*) $= 0.17$ and

2. sim_{resnik} (*"Leukemic Infiltration"*, *"Precancerous Conditions"*) $= 0.17$,

although the edge distance is 2 for the first pair and 4 for the second. Second, Resnik (1995)'s measure is not sensitive to the IC values of the concepts being compared. For example, Figure 2.7 shows that:

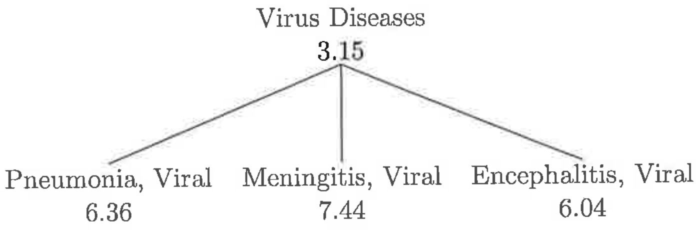


Figure 2.7: The "*Virus Diseases*" concept and its children.

1. sim_{resnik} ("*Pneumonia, Viral*", "*Meningitis, Viral*") = 3.15, and

2. sim_{resnik} ("*Pneumonia, Viral*", "*Encephalitis, Viral*") = 3.15,

as both pairs have the same LCA. However, we can see that:

$$\text{IC}\left(\text{“}\textit{Meningitis, Viral}\text{”}\right) \geq \text{IC}\left(\text{“}\textit{Encephalitis, Viral}\text{”}\right)$$

which suggests intuitively that the semantic distance between "*Pneumonia, Viral*" and "*Meningitis, Viral*" should be superior to the distance between "*Pneumonia, Viral*" and "*Encephalitis, Viral*".

In contrast with Resnik (1995)'s, other measures integrate the ICs of the concepts being compared semantically. Lin (1998) and Jiang and Conrath (1997) develop such measures.

Lin (1998)'s information-based similarity concept is based on three properties:

1. the similarity between two concepts c_1 and c_2 is positively correlated with their commonality,

2. the similarity between two concepts c_1 and c_2 is negatively correlated with the differences between them, and

3. the maximum similarity between two concepts c_1 and c_2 is obtained when c_1 and c_2 are identical.

The commonality of concepts c_1 and c_2 is defined as the IC of the concept (or set of concepts) that states their commonality. The similarity between concepts c_1 and

c_2 is defined as the ratio of the information needed to state their commonality and the information needed to fully describe them.

$$sim_{lin1}\left(c_1, c_2\right) = \frac{\mathrm{IC}\left(\mathrm{common}\left(c_1, c_2\right)\right)}{\mathrm{IC}\left(\mathrm{description}\left(c_1, c_2\right)\right)}$$

In a hierarchy, Lin (1998) defines the similarity between concepts c_1 and c_2 as:

$$sim_{lin2}\left(c_1, c_2\right) = \frac{2 \times \mathrm{IC}\left(\mathrm{LCA}\left(c_1, c_2\right)\right)}{\mathrm{IC}\left(c_1\right) + \mathrm{IC}\left(c_2\right)} \tag{2.7}$$

where sim_{lin} varies from 0 to 1.

Jiang and Conrath (1997) proposes a distance measure that is similar to Lin (1998)'s. First, the edge distance between a child concept c_c and its parent c_p is calculated with the following formula:

$$dist_{jiang3}\left(c_c, c_p\right) = -\log_2 p\left(c_c|c_p\right)$$

where $p\left(c_c|c_p\right)$ is the probability of encountering c_c in a document if c_p was already encountered in the document. This edge distance is proportional to the conditional probability of a child concept c_c given a parent concept c_p. As any instance of a child concept is implicitly an instance of its parents, similarly to Resnik (1995)'s method, this formula is equivalent to:

$$dist_{jiang3}\left(c_c, c_p\right) = -\log_2 \frac{p\left(c_c\right)}{p\left(c_p\right)}$$

or, by the rules of logarithms:

$$dist_{jiang3}\left(c_c, c_p\right) = \mathrm{IC}\left(c_c\right) - \mathrm{IC}\left(c_p\right)$$

which is the difference of information content between the child node c_c and the parent node c_p. Next, the semantic distance between any concepts c_1 and c_2 in the hierarchy is given by the sum of the edge distances from the shortest path from c_1

Table 2.6: Comparison of information-based measures

concept pair	sim_{resnik}	sim_{lin2}	$dist_{jiang4}$
Neoplastic Processes (IC = 2.44), Precancerous Conditions (IC = 3.8)	0.17	0.054	5.9
Leukemic Infiltration (IC = 7.16), Precancerous Conditions (IC = 3.8)	0.17	0.016	10.62

to c_2, which can be reduced to:

$$dist_{jiang4}\left(c_1, c_2\right) = \mathrm{IC}\left(c_1\right) + \mathrm{IC}\left(c_2\right) - 2 \times \mathrm{IC}\left(\mathrm{LCA}\left(c_1, c_2\right)\right) \tag{2.8}$$

Table 2.6 compares the results of Resnik (1995)'s, Lin (1998)'s, and Jiang and Conrath (1997)'s information-based measures for two pairs of concepts. Both concept pairs have one concept in common, "*Precancerous Conditions*", and have the same LCA, "*Neoplasms*". "*Precancerous Conditions*" is compared with "*Neoplastic Processes*" (IC = 2.44), and with "*Leukemic Infiltration*" (IC = 7.16). The difference between the two pairs comes from a higher IC value for one of the concepts. As one would expect, a higher IC value for one of the concepts leads to a decrease of sim_{lin2} and an increase of $dist_{jiang4}$. However, sim_{resnik} is unaffected as it only considers the IC of the LCA. Therefore, sim_{lin2} and $dist_{jiang4}$ are more discriminative between different pairs of concepts, and more intuitive in capturing the semantic similarities.

2.2.2 Comparing Groups of Concepts

In the previous section, the network-based measures presented focused on the similarity of pairs of concepts. However, queries and documents are usually described with several concepts. In order to compare two sets of concepts, individual inter-concept semantic comparisons need to be combined. This combination results in an inter-document measure.

We now review approaches that combine all the possible inter-concept comparisons from two sets of concepts (Azuaje et al., 2005; Rada et al., 1989; Ganesan et al., 2002), and approaches that only combine a subset of all the possible inter-concept comparisons (Azuaje et al., 2005; Ganesan et al., 2002).

All-combination Approach

This approach is based on the idea that the semantic similarity of two sets can be derived from the combination of all the possible inter-concept comparisons from the two sets.

Rada et al. (1989) defines a distance measure between two sets of concepts A and B, containing m and n concepts respectively, as the average of all the $m \times n$ inter-concept distance measures:

$$dist_{rada2}\left(A, B\right) = \frac{1}{m \times n} \times \sum_{c_i \in A, c_j \in B} dist_{rada1}\left(c_i, c_j\right) \tag{2.9}$$

where $dist_{rada1}$ is defined by Equation 2.1. This method is also used by Azuaje et al. (2005) to average inter-concept semantic comparisons. The inter-concept measures used by Azuaje et al. (2005) are sim_{resnik}, sim_{lin2}, and $dist_{jiang4}$ defined by Equations 2.6, 2.7, and 2.8, respectively.

Ganesan et al. (2002) introduces a measure called the generalized Cosine similarity measure (GCSM) that is derived from the Cosine similarity measure (CSM). In the vector space model, two documents A and B are represented respectively by feature vectors (Salton et al., 1975):

$$\vec{A} = \sum_i a_i \vec{c_i} \quad and \quad \vec{B} = \sum_i b_i \vec{c_i}$$

where c_i are the concepts contained in the vocabulary (set of all existing concepts), and a_i and b_i are their associated weights in documents A and B, respectively. The

CSM between A and B is given by:

$$sim_{CSM}\left(\vec{A},\vec{B}\right)=\frac{\vec{A}\cdot\vec{B}}{\sqrt{\vec{A}\cdot\vec{A}}\sqrt{\vec{B}\cdot\vec{B}}} \tag{2.10}$$

and the dot product between two vectors is defined by the formula

$$\vec{A}\cdot\vec{B}=\sum_i a_i b_i$$

In the traditional CSM measure, only concepts common to both documents contribute to the similarity measure. With the GCSM, the dot product between the two vectors is replaced by a weighed sum of the depth/LCA-based similarities (Paragraph 2.2.1) of all possible concept pairs and the norms of the vectors is replaced by the square root of a weighted sum of the similarities of all possible internal concept pairs for each document vector:

$$sim_{GCSM}\left(A,B\right)=\frac{\sum_{i,j} a_i b_j sim_{wu}\left(c_i,c_j\right)}{\sqrt{\sum_{i,j} a_i a_i sim_{wu}\left(c_i,c_i\right)}\sqrt{\sum_{j,j} b_j b_j sim_{wu}\left(c_j,c_j\right)}} \tag{2.11}$$

where a_i and b_j are the weights of concepts c_i and c_j contained in documents A and B, respectively, and sim_{wu} is the inter-concept similarity measure defined in Equation 2.2. With GCSM, even if A and B have no concepts in common, $sim_{GCSM}\left(A,B\right)$ can still yield a strong similarity value if the two documents contain related concepts (when $sim_{wu}\left(c_i,c_j\right)$ approaches 1).

Best-match-combination Approach

The measures described in the previous section use all inter-concept comparisons when comparing two sets of concepts. This approach is not always adequate in the case of mixed conceptual content in the sets.

Measures using all inter-concept comparisons tend to give low similarity values to documents that share concepts but also have dissimilar content. Consider three documents C, D, and E, with respective contents {"*Pneumonia, Viral*", "*Meningitis,* *Viral*"}, {"*Encephalitis, Viral*", "*Neoplasm*"}, and {"*Pneumonia, Viral*", "*Chemical and Drugs*"} extracted from the concepts of the partial MeSH network of Figure 2.3. C and D have no concept in common but their content is related and homogenously located in the "*Diseases*" area of the hierarchy. However, C and E share the same concept, "*Pneumonia, Viral*" but the content of E is more mixed than C's content, as E contains a broad concept, "*Chemical and Drugs*", located outside the "*Diseases*" area. Should D be semantically closer to C than E? The answer is not straightforward. C and D are both about diseases only. Nevertheless, C and E are both about the same disease, but in a different context.

One way to address this problem is to consider only a subset of all the possible concept pairs across the two sets. For example, Ganesan et al. (2002) and Azuaje et al. (2005) introduce measures that use the best conceptual matches between the two groups of concepts. These measures favor documents with concepts in common, such as documents C and E described above, even if they have mixed conceptual content.

Ganesan et al. (2002) defines the Optimistic Genealogy Measure (OGM) to calculate the semantic similarity between two documents A and B. OGM is an asymmetrical measure. To calculate the similarity between the set of concepts A and B, the *similarity contribution* of each concept c_i in A in relation to B needs to be calculated first. The *similarity contribution* of a concept c_i is derived from its best match c_j in document B:

$$simCon\left(c_i, B\right) = \frac{\min_j \left(\text{depth}\left(\text{LCA}\left(c_i, c_j\right)\right)\right)}{\text{depth}\left(c_j\right)}$$

Then, the OGM similarity between A and B is given by the following formula:

$$sim_{OGM}\left(A, B\right) = \frac{\sum_i a_i simCon\left(c_i, B\right)}{\sum_i a_i}$$

where a_i are the weights associated to the concepts c_i. To get a symmetrical OGM similarity between document A and B, we can simply average the OGM between A

Table 2.7: Inter-document similarities with GCSM and OGM

Similarity	GCSM	OGM
C, D	0.70	0.63
C, E	0.65	0.67

and B, and that between B and A:

$$sim_{OGMsym}\left(A,B\right)=\frac{1}{2}\left(sim_{OGM}\left(A,B\right)+sim_{OGM}\left(B,A\right)\right) \tag{2.12}$$

Azuaje et al. (2005) introduces two measures. The first is a similarity measure comparable to symmetrical OGM, as it is a combination of two asymmetrical similarity measures:

$$sim_{azu}\left(A,B\right)=\frac{1}{m+n}\left(\sum_{i}\max_{j}\left(sim\left(c_i,c_j\right)\right)+\sum_{j}\max_{i}\left(sim\left(c_i,c_j\right)\right)\right) \tag{2.13}$$

where A and B contain m and n concepts, respectively, and sim stands for any of the two similarity measures sim_{resnik}, sim_{lin2}, defined by Equations 2.6 and 2.7, respectively. The second is a distance measure:

$$dist_{azu}\left(A,B\right)=\frac{1}{m+n}\left(\sum_{i}\min_{j}\left(dist_{jiang4}\left(c_i,c_j\right)\right)+\sum_{j}\min_{i}\left(dist_{jiang4}\left(c_i,c_j\right)\right)\right) \tag{2.14}$$

where $dist_{jiang4}$ is the distance measure defined in Equation 2.8.

To illustrate the difference between measures using all inter-concept comparisons and measures using best matches, we use sim_{GCSM} (defined by Equation 2.11) and sim_{OGMsym} (defined by Equation 2.12) to calculate similarities between documents C and D, and between documents C and E. Table 2.7 shows the results. GCSM gives a higher similarity between C and D, the documents with homogenous content, whereas OGM gives a higher similarity between C and E, the documents sharing a concept, although E's content is mixed.

2.3 Evaluations

We now move to the presentation of past evaluations of the network-based semantic measures described previously. These evaluations can be either intrinsic (directly comparing the use of the network with the judgments of experts) or extrinsic (evaluating the usefulness of the network information in relation to an application). In particular, we show that some past evaluations on the MeSH network were done on a small scale, and that others were not done on the MeSH network at all.

2.3.1 Edge Count Approach

Evaluations of the edge count approach are available in the biomedical domain: Rada et al. (1989) uses the MeSH hierarchy and Caviedes and Cimino (2004) works with 3 sub-networks of the Unified Medical Language System, including MeSH.

Rada et al. (1989) evaluates the edge count approach intrinsically and extrinsically. The intrinsic evaluation concerns the use of $dist_{rada1}$ to simulate the human judgments regarding the semantic distance between two concepts. Rada et al. (1989) focuses on part of the MeSH 1986 semantic network that deals with information science related topics. The poor coverage of this part of MeSH (only 200 terms) is increased by merging it with another network, the Association of Computing Machinery's hierarchical semantic net for computer science (CRCS). The merging algorithm searches for the concepts that the two vocabularies have in common, and adds to MeSH concepts parent of child concepts that exists in CRCS. Using the MeSH network (information science part) and its augmented version (MeSH+CRCS), 12 pairs of concept are ranked according to their semantic distance with $dist_{rada1}$. This ranking is then compared to the judgments of 10 students. The students' rankings are found to significantly agree at the 0.01 level of confidence using the average Spearman's correlation coefficient. The averaged students' ranking are compared with the ranking given by $dist_{rada1}$ for each network (MeSH and MeSH+CRCS) with Spearman's rank correlation coefficient. The coefficients $\rho = 0.17$ and $\rho = 0.52$

are reported for MeSH and MeSH-CRCS, respectively. The results suggest that $dist_{rada1}$ works better when the network has a higher conceptual coverage. The 2004 version of MeSH used in this dissertation is expected to have a higher coverage than the 1986 version used in Rada et al. (1989). Moreover, the part of the MeSH network on information science is not expected to be of any relevance to us in relation to genomic topics. We would be more interested in the usefulness of $dist_{rada1}$ in genomic-related hierarchies such as "Diseases", "Chemicals and Drugs", and "Organisms".

$dist_{rada2}$ (Equation 2.9) is evaluated extrinsically by Rada et al. (1989) on the MeSH ontology. Rankings of MEDLINE documents in relation with queries by two experts are compared with rankings obtained for the same queries and documents with $dist_{rada2}$ using Spearman's rank correlation coefficient. The queries, six in total, are domain-related ("liver diseases and peritoneoscopy", "shock and endorphins") and the document, around fifty for each query, are collected via a MEDLINE search. The results show a significant correlation between the two experts' rankings, and between the experts' rankings and the rankings given by $dist_{rada2}$. The work described in Rada et al. (1989) give us some evidence on the usefulness of the MeSH hierarchy to calculate semantic distances between documents and queries. However, a large-scale evaluation, using more queries and a large collection in the context of ad hoc retrieval, is needed to confirm the value of network-based semantic measures such as $dist_{rada2}$. Such an evaluation constitutes part of the contribution of the work described in this dissertation.

More recently, Caviedes and Cimino (2004) evaluates $dist_{rada1}$ on 3 sub-networks of the Unified Medical Language System (UMLS): MeSH, ICD9CM, and SNOMED. The distance values given by each network with $dist_{rada1}$ on 55 concept pairs is compared to the averages of three expert judgments. The best correlation (0.77 with Pearson's correlation coefficient) between the distances values and the expert judgment averages is obtained with the MeSH network. Additionally, Caviedes and Cimino (2004) evaluates $dist_{rada2}$ with four clusters of two concepts each, using the same three sub-networks of the UMLS. The inter-cluster distance values are then compared to human judgments from three physicians. Strong correlations are found between network-based distance values and expert judgments for MeSH and SNOMED. Similarly to the evaluations found in Rada et al. (1989), the conclusions reached by Caviedes and Cimino (2004) regarding $dist_{rada1}$ and $dist_{rada2}$ still need to be confirmed at a larger scale. Furthermore, the concept clusters used do not correspond to the content of actual MEDLINE records. In the context of MEDLINE document retrieval, what is needed is a large-scale evaluation of $dist_{rada1}$ and $dist_{rada2}$ for the measurement of semantic distance between queries and MEDLINE MeSH-based document representations.

2.3.2 Complex Edge Weighting

Most measures assuming variation of edge distance in the hierarchy are assessed with WordNet (see section 2.1.1). Resnik (1995), Lin (1998), and Jiang and Conrath (1997) evaluate their network-based measures (sim_{resnik}, sim_{lin2}, $dist_{jiang2}$ and $dist_{jiang4}$) on the noun-part of the WordNet semantic network. The evaluations are intrinsic: the measures are compared with human judgments on a set of 30 noun pairs. Both Lin (1998) and Jiang and Conrath (1997) report slight improvements over Resnik (1995) in terms of correlation with the human judgments. However, Jiang and Conrath (1997) reports that adjustments of the measure with various values of α and β (sensitivity to network depth and density, respectively) in $dist_{jiang2}$ does not improve the results obtained with $dist_{jiang4}$, the information-based measure. This suggests that the concept frequency information of the corpus already incorporates the variation of edge distance in the network that are caused by depth and density. These intrinsic evaluations are small in scale and do not indicate whether the measures would be useful for document retrieval using a domain-dependent ontology such as MeSH.

Richardson and Smeaton (1995) assesses sim_{resnik} for the retrieval of newspapers articles. A knowledge base is build from the noun part of WordNet. Query and document words are matched to WordNet concepts with several disambiguation techniques, and sim_{resnik} is used to evaluate the semantic similarity between the concepts. The evaluation is done on 12 queries with a thousand documents each, retrieved from a collection of 742,611 text articles from newspapers with a first standard search using TF*IDF. Results show that the information-based similarity measure does not perform as well as TF*IDF. Nonetheless, they perform differently over the 12 queries, which suggest that they might be used in combination. This extrinsic evaluation is using a rather small collection and an evaluation on a larger scale is needed to confirm the results. Moreover, WordNet is a domain-independent ontology that is not tailored to satisfy the description of specific information types such as newspapers' content. In contrast, we want to evaluate network-based semantic measures on MeSH, a ontology specifically designed to describe bio-medical information.

Evaluations of sim_{resnik}, sim_{lin2}, and $dist_{jiang4}$ with domain-specific ontologies can be found. For example, the measures are used with the Gene Ontology (Gene Ontology Consortium, 2000, 2001, 2004) for gene expression correlation analysis (Wang et al., 2004, 2005) and for gene functional assessment (Azuaje et al., 2005). Moreover, they are evaluated with the Gene Ontology by Lord et al. (2003a,b) in the comparison of gene annotations with gene sequences. Additionally, Pedersen et al. (2005) compares the three measures' values in the medical ontology SNOMED-CT to human judgments. However, no evaluation of these measures is available with the MeSH network in the context of document retrieval and classification. Such an evaluation is addressed in this dissertation as an important part of our contribution.

2.4 Summary

Various measures exist that use a semantic network to evaluate the similarity between two concepts. Rada et al. (1989) uses the number of edges separating the concept nodes to evaluate the relatedness of two concepts ($dist_{rada1}$ defined by Equation 2.1). Other measures assume that the hierarchy edges do not correspond to the same semantic distance (variable edge distance). These measures calculate the edge distances with hierarchy information, such as depth and density ($dist_{jiang2}$ defined by Equation 2.3), or with corpus information (sim_{resnik}, sim_{lin2}, $dist_{jiang4}$ defined by Equations 2.6, 2.7, and 2.8, respectively).

Most measures are evaluated with WordNet, a semantic network of general English language organized with several types of relationships. Some evaluations are intrinsic and involve comparing network-based measures to human measures (Jiang and Conrath, 1997). Specific applications are also used to evaluate the benefits of the semantic network. In Sussna (1993), the measure is assessed in a word sense disambiguation application. Furthermore, Richardson and Smeaton (1995) estimates network-based measures for the retrieval of newspaper articles. Additionally, evaluations of some network-based measures are available for the biomedical and genomic domains (Rada et al., 1989; Lord et al., 2003a,b; Pedersen et al., 2005). Nevertheless, a thorough evaluation of network-based measures has yet to be done on the MeSH semantic network. Simple techniques, such as edge count, were examined in the past with MeSH (Rada et al., 1989). More complex approaches, integrating depth, density, and corpus information still need to be explored.

2.5 Contribution: Evaluation of Network-based Semantic Measures on MeSH

In this dissertation we propose an large-scale extrinsic evaluation of network-based semantic measures on the MeSH hierarchy in the context of MEDLINE document retrieval. The measures used are:

- Rada et al. (1989)'s inter-concept measure $dist_{rada1}$ as a baseline considering all links in the MeSH hierarchy to correspond to the same semantic distance, and

- Jiang and Conrath (1997)'s inter-concept measures, $dist_{jiang2}$ and $dist_{jiang4}$, to express edge distance of the MeSH relationships as a function of the depth and density of the hierarchy, and as a function of concepts' information content, respectively.

The three measures contain implicit hypotheses about the MeSH hierarchy. $dist_{rada1}$ assumes that all edges correspond to the same semantic distance. In contrast, $dist_{jiang2}$ and $dist_{jiang4}$ assume that edge distance decreases with the specificity of concepts and the density of the conceptual areas they belong to. $dist_{jiang2}$ derives specificity and density from hierarchy information, whereas $dist_{jiang4}$ derives them from corpus information. Comparing the performance of the three measures allows us to evaluate the respective hypotheses they imply. As we compare queries with documents as well as documents together, we also need to evaluate methods that compare sets of concepts by combining the comparison of individual concepts. We evaluate the following methods:

- all-combination ($dist_{rada2}$, Equation 2.9), and
- best-match-combination ($dist_{azu}$, Equation 2.14).

Finally, the evaluation comprises the following extrinsic evaluation frameworks:

- ad hoc retrieval on a subset of MEDLINE of 4.5 million documents with 50 queries and associated relevance judgements, and
- binary classification of 10,000 documents simulating expert in the triage of documents likely to give experimental support for the annotation of Mouse genes with Gene Ontology concepts.

Table 2.8 gives an overview of our contribution regarding the evaluation of network-based semantic measures on the MeSH hierarchy. In the next chapter we review related work in MEDLINE retrieval involving the use of the MeSH ontology for document representation.

Table 2.8: Contribution: Evaluation of Network-based Semantic Measures

	Past Evaluations	Our Contribution
Edge Count	Intrinsic/extrinsic evaluation on MeSH with a small set of concepts: Rada et al. (1989), Caviedes and Cimino (2004).	Large-scale extrinsic evaluation on MeSH: 1) Ad hoc retrieval on 4.5 million documents, 2) Binary classification on 12,000 documents.
Complex Measures	Intrinsic/extrinsic evaluation on: 1) WordNet: Resnik (1995), Jiang and Conrath (1997), Richardson and Smeaton (1995), 2) Gene Ontology: Wang et al. (2004, 2005), Azuaje et al. (2005), Lord et al. (2003a,b), 3) SNOMED: Pedersen et al. (2005)	Evaluation on MeSH

Chapter 3

Related Research

This chapter presents research related to the work described in this dissertation. In particular, we review the current uses of MeSH in document representation for MEDLINE document retrieval, classification, and clustering. First, some background information is given about the MEDLINE database.

3.1 The MEDLINE Database

MEDLINE, the U.S. National Library of Medicine (NLM)[1] bio-medical abstract repository, contains over 14 million reference articles from around 4,800 journals (early 2006). Approximately 400,000 new records are added to it each year (over 623,000 were added in 2006). Despite the growing availability of full-text articles on the Web, MEDLINE remains in practice a central point of access to bio-medical research (Hersh et al., 2004, 2005).

The MEDLINE record fields include text-based fields, the title and abstract fields, and ontology-based fields: the MeSH fields. Most MEDLINE records contain 10-12 MeSH fields. Some examples of textual fields and MeSH fields are shown in Figure 3.1.

In addition to the textual fields, MeSH fields are a useful source of structured and standardised information. Unlike the free-text content of the title and abstract

[1] http://www.nlm.nih.gov/, last accessed: 19 January 2007

fields, the MeSH fields unambiguously associate documents to concepts. The MeSH concepts can help us to resolve the ambiguities of free-text (see Chapter 1). In addition, MeSH concepts are assigned to the records after the examination of the entire research article by human indexers. Consequently, they can complement or add to the information contained in the title and abstract.

3.1.1 MeSH Fields Format

A MeSH field is a combination of a MeSH descriptor with zero or more MeSH qualifiers (see Section 2.1.2). Descriptors are the main conceptual vocabulary of MeSH. Qualifiers add context to the concepts described by descriptors. In Figure 3.1, "*Centrioles/*ultrastructure*" is the combination of descriptor "*Centrioles*" with qualifier "*ultrastructure*". However, "*Centrioles*" can be found in other documents associated with other qualifiers such as "*metabolism*", "*chemistry*", and "*physiology*". Each qualifier indicates a different context for the descriptor.

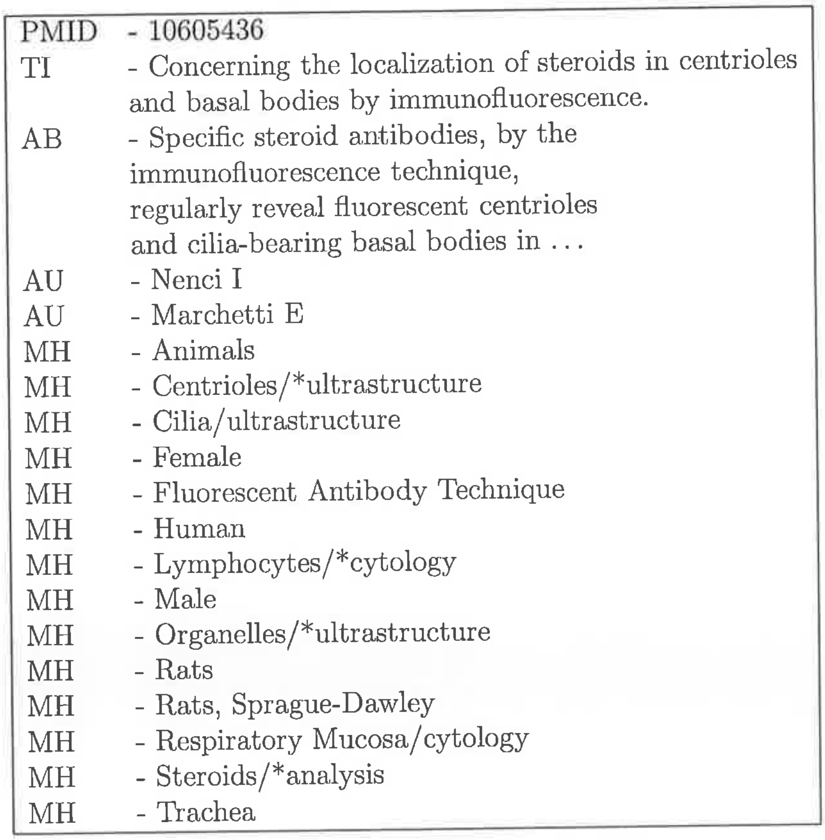
MeSH fields provides a mechanism to suggest the level of relevance of the MeSH concepts to the document. In particular, MeSH fields distinguish the major themes of an article from the minor themes. The major themes are the central concepts of a document, whereas the minor themes are peripheral concepts. A star is used to identify the major themes. Therefore the association "*Centrioles/*ultrastructure*" is a major theme of the MEDLINE record of Figure 3.1, along with "*Organelles/*ultrastructure*" and "*Steroids/*analysis*". In contrast, "*Cilia/ultrastructure*" and "*Respiratory Mucosa/cytology*" are minor themes.

The information contained in the format structure can be useful to evaluate the relevance of different documents to a query. Consider a query Q and three documents D_1, D_2, and D_3. Table 3.1 shows the content of Q and a relevant MeSH field in each of the documents. Intuitively, D_1 is more relevant to Q than D_2, as D_2 is about the same concept, "*Centrioles*", but in a different context, "*metabolism*". Moreover, D_3 is more relevant still, because the association "*Centrioles/ultrastructure*" contained in Q is a major theme in D_3, whereas it is a minor theme in D_1.

Table 3.1: An example of using information contained in the structure of MeSH fields

Q	*Centrioles/ultrastructure*
D_1	*Centrioles/ultrastructure*
D_2	*Centrioles/metabolism*
D_3	*Centrioles/*ultrastructure*

Figure 3.1: A MEDLINE record example (PMID: PubMed ID, TI: title, AB: abstract, AU: author, MH: MeSH term)



```
PMID  - 10605436
TI    - Concerning the localization of steroids in centrioles
        and basal bodies by immunofluorescence.
AB    - Specific steroid antibodies, by the
        immunofluorescence technique,
        regularly reveal fluorescent centrioles
        and cilia-bearing basal bodies in ...
AU    - Nenci I
AU    - Marchetti E
MH    - Animals
MH    - Centrioles/*ultrastructure
MH    - Cilia/ultrastructure
MH    - Female
MH    - Fluorescent Antibody Technique
MH    - Human
MH    - Lymphocytes/*cytology
MH    - Male
MH    - Organelles/*ultrastructure
MH    - Rats
MH    - Rats, Sprague-Dawley
MH    - Respiratory Mucosa/cytology
MH    - Steroids/*analysis
MH    - Trachea
```

3.1.2 MeSH Annotation Consistency

Funk et al. (1983) reported on the consistency of the annotation of MEDLINE records with MeSH concepts across NLM indexers. The evaluation relied on 760 MEDLINE records accidently indexed twice and published in 42 journals from 1974 to 1980. Hooper's measure was used to calculate the indexing consistency between two indexers, using the following equation:

$$\text{Consistency}_{\text{Hooper}} = \frac{A}{A + M + N}$$

where A is the number of terms in agreement, M the number of terms used by the first indexer but not the second, and N is the number of terms used by the second indexer but not the first. Perfect inconsistency corresponds to 0 and perfect consistency to 1. Several MeSH-based representations are evaluated for consistency using the information contained in the MeSH fields' structure (see previous section). The representations include the use of:

- "major theme" descriptors (*MH),
- descriptors (MH),
- "major theme" qualifiers (*SH),
- qualifiers (SH),
- "major theme" descriptor/qualifier associations (MH/*SM), and
- descriptor/qualifier associations (MH/SM).

Table 3.2 shows the mean Hooper's inter-indexer consistency measure over the 760 records for each representation. First, "major theme" representations (*MH, *SH, MH/*SH) give a better consistency than representations including all concepts (MH, SH, MH/SH), respectively. Second, the representations splitting the associations between descriptors and qualifiers (*MH, *SH, MH, SH) gives better consistency

Table 3.2: Inter-indexer consistency for various MeSH-based representations

MeSH-based representation	Hooper's consistency
*MH	61.1
MH	48.2
*SH	54.9
SH	48.7
MH/*SH	43.1
MH/SH	33.8

than representations maintaining the associations (MH/*SH, MH/SH), respectively. It is intuitive, and highly desirable for retrieval purposes, that the annotators should agree more on the concepts describing the major themes of the article than they should on the concepts describing the minor themes. As for the associations, the results indicate that indexers agree more on the annotation of qualifiers than on which descriptors the qualifiers should be combined with. From a retrieval point of view, the agreement results suggest that "major theme" representations will be less noisy than their all-inclusive counterparts, and will consequently favor precision over recall. Moreover, the agreements results that representations maintaining the descriptor/qualifier associations may be too specific for retrieval and may damage recall. However, low annotation agreement levels may be compensated by the use of the MeSH hierarchy, in the event of two indexers using two distinct concepts that are nevertheless closely related. Lastly, to the best of our knowledge, Funk et al. (1983)'s study is the latest published on MEDLINE indexing consistency. A more recent study including articles published over the last twenty years would be welcome in order to confirm the results presented in this section. Whether the consistency over the indexing of the 760 articles published from 1974 to 1980 is representative of collections used in our experiments remains an open question.

Table 3.3: Content of descriptor "*Epistaxis*" record

MeSH Heading	Epistaxis
Tree Number	C08.460.261
Tree Number	C09.603.261
Tree Number	C23.550.414.712
Scope Note	Bleeding from the nose.
Entry Term	Nose Bleed
Entry Term	Nosebleed
Allowable Qualifiers	BL CF CI CL CN CO DH DI DT EC EH EM EN EP ET GE HI IM ME MI MO NU PA PC PP PS PX RA RH RI RT SU TH UR US VE VI
Unique ID	D004844

3.1.3 PubMed Search

PubMed[2] is the main search interface for MEDLINE. It is developed by the National Center for Biotechnology Information (NCBI). It provides free access to MEDLINE citations and abstracts, and links to external web sites providing full-text articles.

A basic PubMed search involves the automatic mapping of a user's search terms into a Boolean query. The search is applied to several MEDLINE fields, including the author, title, abstract, and MeSH fields. During the process of automatic term mapping, a MeSH translation table is used to map the user's terms to MeSH concepts. The translation table looks for a direct match for the search term in the list of descriptors' and qualifiers' preferred terms. Then it looks at other fields, such as the ENTRY fields which contain synonyms of the preferred terms. For example, PubMed will map the term *nosebleed* from a user's query to the MeSH concept "*Epistaxis*", and retrieve documents containing that MeSH concept: the term *nosebleed* is contained in an ENTRY field of the descriptor record with preferred name "*Epistaxis*". Table 3.3 shows the content of the descriptor "*Epistaxis*" record.

PubMed search uses the MeSH hierarchy by *exploding* the MeSH concepts contained in the query. The search result will include all documents that contain narrower concepts than the original MeSH concepts. For example, consider a query

[2] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed, last accessed: 19 January 2007

Table 3.4: An example of basic PubMed Search

Q	*Neoplastic Processes*
D_1	*Anaplasia*
D_2	*Neoplasms*
D_3	*Precancerous Conditions*

Q and three documents D_1, D_2, and D_3. Table 3.4 shows the content of Q and a relevant MeSH field in each documents. Pubmed, with the *exploding* search option enabled, will retrieve D_1, as it contains a concept that is a child in the hierarchy of the query concept "*Neoplastic Processes*" (see Figure 2.3). However, D_2 and D_3 will not be retrieved although they contain a parent concept, "*Neoplasms*", and a sibling concept, "*Precancerous Conditions*", respectively, of the query concept.

Suggestions

Extending the search of a concept in documents not only to the search of all its descendants, as in PubMed, but to all related concepts in the hierarchy (parents, grandparents, siblings, cousins), is a method we want to evaluate. The difficult problem is determining how close the related concepts are to the concepts initially searched in documents. This problem is discussed in Chapter 2.

3.2 MeSH Representations in Ad Hoc Retrieval

In MEDLINE ad hoc retrieval, there are two main methods for combining MeSH-based and textual information (Srinivasan, 1996a). The first method consists of mixing the two contents to create one mixed index. This is known as pre-retrieval combination. The second method consists of building two representations for each document: a text-based representation and a MeSH-based representation. Two indices, one text-based and one MeSH-based, are queried separately, and merging techniques are used to combine the lists of documents obtained from each index. This is known as post-retrieval combination. Before presenting the two methods, we introduce the ad hoc task of the TREC 2004 and 2005 Genomics track[3], as this task is widely used for the evaluation of related work. We also describe the metrics used in the evaluations.

3.2.1 TREC 2004 and 2005 Genomics Track and the Ad Hoc Task

The Text REtrieval Conference (TREC)[4] guidelines and common evaluation procedures allow research groups from all over the world to evaluate their progress in developing and enhancing information retrieval systems. TREC has included a Genomics track since 2003.

The collection used for the TREC 2004 Genomics track ad hoc search task, TrecGen04, consists of a subset of the MEDLINE bibliographic database, and 50 topics with their associated relevance judgments (Hersh et al., 2004). In TREC terminology, a topic simply refers to a query. The subset contains ten years of completed citations from 1994 to 2003 inclusive, which amounts to a total of 4,591,008 documents. All records have a title, 75.8% contain an abstract and 99% of records contain MeSH fields. Each topic includes an ID number, a *title* field (abbreviation of information need), an *information need* field (full statement of the information need), and a *context* field (background information). Table 3.5 shows an example of a topic for the 2004 task. The number of relevant documents per topic is found in Table A.1.

The collection used for the TREC 2005 Genomics track ad hoc search task, TrecGen05, consists of the same subset of the MEDLINE as TrecGen04. However, it includes a new set of 50 topics with their associated relevance judgments (Hersh et al., 2005). 2005 topics are expressed in a particular format, or topic template, that differs from the 2004 format shown in Figure 3.5. They consist of ten instances of five distinct generic topic templates (GTTs). An example of a GTT is *find articles*

[3] http://ir.ohsu.edu/genomics/, last accessed: 19 January 2007

[4] http://trec.nist.gov/, last accessed: 19 January 2007

Table 3.5: Example of a 2004 topic

ID	5
TITLE	Protocols for isolating cell nuclei
INFORMATION NEED	Articles are relevant if they describe methods for subcellular fractionation of nuclei.
CONTEXT	Laboratory preparations can be enriched for certain kinds of proteins if the cellular compartment in which they reside is purified away from the rest of the cell contents.

Table 3.6: Description of the Generic Topic Templates

GTT#	GTT description
1	Find articles describing standard methods or protocols for doing some sort of experiment or procedure.
2	Find articles describing the role of a gene involved in a given disease.
3	Find articles describing the role of a gene in a specific biological process.
4	Find articles describing interactions (e.g., promote, suppress, inhibit, etc.) between two or more genes in the function of an organ or in a disease.
5	Find articles describing one or more mutations of a given gene and its biological impact in a given organism.

describing the role of a gene involved in a given disease. Instances of the above template replace the two generic underlined terms, gene and disease, by specific names of genes and diseases. A description of the five GTTs is shown in Table 3.6. The number of relevant documents per topic is found in Table A.4.

3.2.2 Evaluation Metrics

Precision and Recall

We now present evaluation metrics based on the concepts of precision and recall. Before defining precision and recall, we first need to define the concepts of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). Consider a retrieval system to which a particular query is submitted. TP is the set of relevant documents retrieved by the the retrieval system, whereas FN is the set of relevant documents not retrieved. FP is the set of non-relevant documents retrieved

Table 3.7: TP, FN, FP, and TN document sets

	retrieved	not retrieved
relevant	TP	FN
non-relevant	FP	TN

by the IR system, whereas TN is the set of non-relevant documents not retrieved. Table 3.7 illustrates the definitions of the 4 sets.

Precision is defined as the proportion of retrieved documents found to be relevant:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3.1}$$

and recall is defined as the proportion of relevant documents that were retrieved:

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3.2}$$

When high precision is a priority, the set of retrieved documents, TP+FP, can be limited in size. In that case, for example, precision at 5 documents retrieved (P@5) is reported, or 10, 15, 20, or more documents retrieved.

Precision and recall can be averaged over n queries Q_i:

$$\text{average precision} = \frac{1}{n} \sum_i \text{precision}\,(Q_i) \tag{3.3}$$

$$\text{average recall} = \frac{1}{n} \sum_i \text{recall}\,(Q_i) \tag{3.4}$$

where precision and recall are defined in Equations 3.1 and 3.2, respectively.

Combining Precision and Recall

If documents are retrieved in a ranked list, precision and recall can be combined by measuring precision at different levels of recall. We present three measures that use precision at different levels of recall: 11-point precision, average precision, and

R-precision.

11-point precision averages precision values at 11 standard levels of recall: 0.0, 0.1, 0.2, 0.3,.., 1.0. Consider a query for which five relevant documents are known ($R = 5$). A retrieval system retrieves ten documents for this query, as shown in Table 3.8. We notice that precision values are only directly available for five levels of recall (0.2, 0.4, 0.6, 0.8). Moreover, precision is higher at 60% recall than it is at 20% and 30% recall. This is corrected by interpolation. Interpolation consists of giving a precision value at a recall level r that correspond to the maximum precision value for all recall levels superior or equal to r. The idea behind interpolation is that precision can only decrease as the level of recall increases. Table 3.9 shows the precision values at the 11 points of recall after *interpolation* for the ranking given in Table 3.8. Note that the precision of a relevant document not retrieved is always zero. The 11-point precision is the average of the precision values at each point of recall:

$$\text{11-point precision} = \frac{1}{11}\sum_{i=1}^{11}\text{precision}_i$$

In our example, 11-point precision = 0.41. For N queries Q_j, the 11-point average precision is the average of the 11-point precision of each query:

$$\text{11-point average precision} = \frac{1}{N}\sum_{j=1}^{N}\text{11-point average precision}\,(Q_j)$$

The average precision is the average of the precision values obtained each time a relevant document D_i is retrieved. Relevant documents that are not retrieved are given zero precision values.

$$\text{average precision} = \frac{1}{R}\sum_{i=1}^{R}\text{precision}\,(D_i)$$

In the example of Table 3.8, average precision = 0.33. For N queries Q_j, the mean

Table 3.8: Example of document ranking (R=relevant, NR=non-relevant)

rank	judgment	precision	recall
1	NR	-	-
2	NR	-	-
3	R	0.33	0.2
4	NR	-	-
5	R	0.4	0.4
6	R	0.5	0.6
7	NR	-	-
8	NR	-	-
9	NR	-	-
10	R	0.4	0.8

Table 3.9: Precision at 11 standard recall points after interpolation

Precision	Recall points
0.5	0.0
0.5	0.1
0.5	0.2
0.5	0.3
0.5	0.4
0.5	0.5
0.5	0.6
0.4	0.7
0.4	0.8
0.0	0.9
0.0	1.0
0.41	11-point precision

average precision (MAP) is the mean of the average precision of each query:

$$\text{MAP} = \frac{1}{N}\sum_{j=1}^{N} \text{average precision}\,(Q_j) \tag{3.5}$$

The R-precision is the precision after R documents have been retrieved, R being the number of known relevant documents for the query. In the example of Table 3.8, $R = 5$ and the precision after 5 documents have been retrieved is 0.4.

3.2.3 Comparing Averages with Randomization Testing

When retrieval strategies are evaluated over several queries, means, such as MAP, 11-point average precision, average precision at N documents retrieved, are used to compare different strategies. The idea is to use a metric that approximates the performance of the strategies for the population of all possible queries.

The difference between two means may be due to chance with a certain probability. We need to estimate the probability of getting a certain difference between two means given that the null hypothesis (no actual difference between the two means) is be true. If the estimated probability falls below a chosen level of confidence, then the null hypothesis can be rejected and the two means are said to be different with statistical significance.

In this dissertation, we use randomization tests in order to estimate the probability that the difference between two means is due to chance. Randomization tests generate a distribution of differences between two means given that the null hypothesis is true. The number differences found equal or more extreme than the observed difference divided by the total number of differences generated gives the probability of obtained such a difference by chance.

In particular, we use an implementation of randomization testing developed by the National Institute of Standard and Technology (NIST[5]) for the TRECVID 2006

[5] http://www.nist.gov/

Workshop[6].

3.2.4 Pre-retrieval Text and MeSH Combination

Methods

The information contained in the text and MeSH fields of MEDLINE records is combined before retrieval. A mixed index is created and the original free-text query is used to search the index.

During the creation of the mixed index, text fields and MeSH fields are processed indiscriminately (Srinivasan, 1996a; Abdou et al., 2005). This means that the structure of the MeSH field described in Section 3.1 is ignored. The association between descriptors and qualifiers is broken, the distinction between major themes and minor themes is lost, and MeSH concepts represented by phrases are separated into words, as illustrated in Table 3.10. The MeSH words obtained may undergo further processing, such as stopword removal and stemming. Stopword removal aims at filtering out words that do not carry any significant information for document retrieval (Fox, 1992). They are usually identified by their high frequency in a particular collection of documents. Stemming aims at increasing search performance by producing a unique stem from all the lexical variants found in queries and documents (Frakes, 1992). In effect, stemming can be seen as method to improve recall (Kraaij and Pohlmann, 1996). For example, *computer*, *computers*, *computing*, and *computation* can all be reduced to the lexical form *comput*. After stemming, a query term such as *computer* will retrieve a document containing the term *computing*. A widely used stemming algorithm is the Porter algorithm (Porter, 1980). With the Porter algorithm, MeSH words such as *Centrioles* and *ultrastructure* are turned into *centriol* and *ultrastructur*, respectively.

A weighting scheme can be used at indexing time to reflect the relative importance of the textual and MeSH fields against each other. For example, Aronson

[6] http://www-nlpir.nist.gov/projects/trecvid/

et al. (2005) use a modified version of the SMART retrieval system (Salton, 1971) to represent text and MeSH words with different weights, 7 and 2, respectively.

Some information from the structure of MeSH fields can be integrated in the weighting process. Shin and Han (2004) uses the distinction between terms contained in MeSH major themes from terms contained in minor themes with the following weighting scheme:

$$\begin{aligned} w'_{ij} &= w_{ij} + \left(\rho + \tfrac{\rho}{4+\ln(\rho)}\right) \quad \text{if term } i \text{ is in a major theme of } D_j \\ w'_{ij} &= w_{ij} + \left(\rho - \tfrac{\rho}{4+\ln(\rho)}\right) \quad \text{if term } i \text{ is in a minor theme of } D_j \end{aligned}$$

where w_{ij} is the initial weight assigned to term i from a MeSH field of document D_j, and ρ is a parameter adjusting the sensitivity of the weighting scheme to the distinction between major and minor themes.

Evaluations

Srinivasan (1996a) describes experiments mixing text and MeSH fields before retrieval to represent documents. The MeSH field structure is discarded, MeSH phrases are broken into words (Table 3.10), and MeSH words are processed similarly to text, using stopword removal and stemming. The representation is evaluated with a collection of 2,344 MEDLINE records and 75 queries. The queries correspond to information needs expressed in short free-text statements. An increase of 7.3% (0.5169 to 0.5548) of 11-point average precision over the text-only strategy is reported. The same method is then tested on the larger OHSUMED collection (Hersh et al., 1994a). The OHSUMED collection includes 348,556 MEDLINE records, and 101 queries with their associated relevance judgements. The queries are information needs expressed with free-text. Using OHSUMED, similar improvements are reported: the 11-point average precision is increased by 6.3% (0.2316 to 0.2461) with the combination method over the method text-only baseline.

Abdou et al. (2005) assesses several vector-space and probabilistic models on the TrecGen05 collection (see Section 3.2.1). The MeSH fields' content is added to the document representation at indexing time. The 50 queries are expressed in free-text using the 5 different templates shown in Table 3.6. Adding MeSH field content is reported to increase the MAP by about 9% on average, over the nine best-performing IR models.

Shin and Han (2004) evaluated the mixed representation with the Cystic Fibrosis (CF) collection. The CF collection contains 1,239 MEDLINE records with 100 queries expressed in free-text. An improvement of 12.5% (0.279 to 0.314) in R-precision is recorded when MeSH terms are added to the document representation. Adjusting the MeSH terms' weights according to the major and minor theme distinction leads to a further 1.6% improvement in R-precision. The authors did not indicate the level of statistical significance of this improvement.

Suggestions

The pre-retrieval text and MeSH combination yields performance improvements on various documents collection. Therefore, the evaluations demonstrate the usefulness of the MeSH ontology for document retrieval. Moreover, the distinction between major themes and minor themes made by Shin and Han (2004) brings modest performance improvements but suggests that more interest may be drawn to the structure of the MeSH fields.

3.2.5 Post-retrieval Text and MeSH Combination

Method Overview

In the second method, MEDLINE text and MeSH fields are used to create two separate document representations, and hence two distinct indices. The text index is searched with free-text queries and the MeSH index is searched with MeSH queries. The results of the two searches are then combined to give a final answer to the original information need in the form of a ranked list of documents. An overview of the method is illustrated in Figure 3.2.

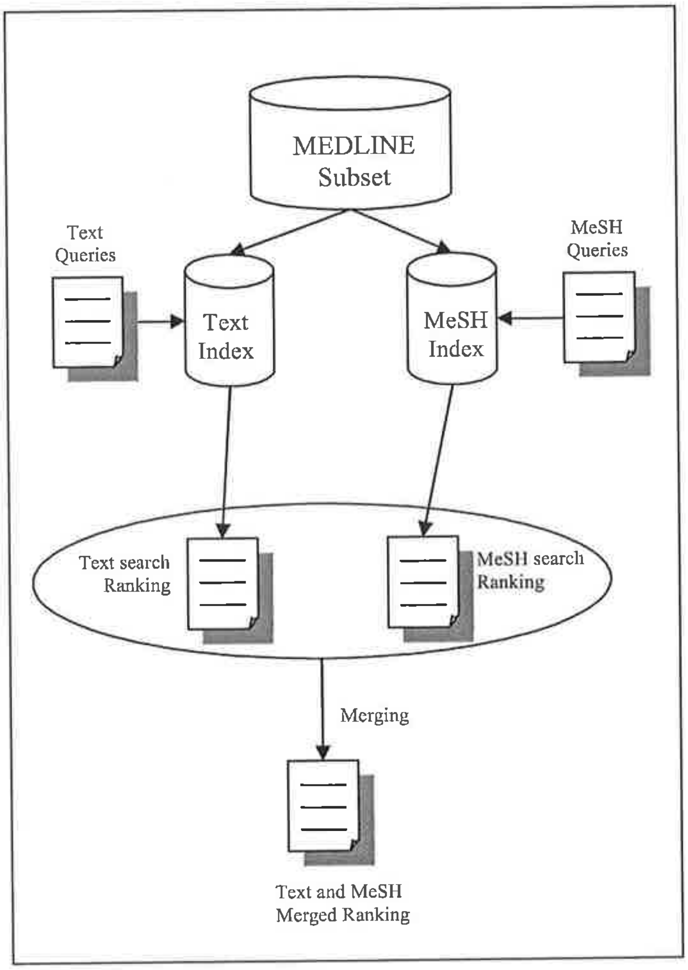


Figure 3.2: Post-retrieval text and MeSH combination overview

MeSH Query Generation

The MeSH queries can be generated either directly from the text queries by an inter-field thesaurus (Srinivasan, 1996b; Aronson and Rindflesch, 1997; Aronson, 2001; Aronson et al., 2004), or indirectly by pseudo-relevance feedback on the output of the text queries. (Srinivasan, 1996c,a; Shin et al., 2004; Kraaij et al., 2004).

Inter-field Thesauri. Inter-field thesauri can assist the generation of MeSH queries by mapping textual terms (found in the free-text fields) to MeSH concepts (found in the MeSH fields). The idea is to establish semantic similarity between free-text terms and concepts from a controlled vocabulary.

The mapping can be rule-based or knowledge-based, and can use NLP (Natural

Table 3.10: Example of MeSH field processing in Srinivasan (1996b), before stopword removal and stemming

MeSH field content	MeSH concepts
MH - Fluorescent Antibody Technique	Fluorescent Antibody Technique
MH - Rats, Sprague-Dawley	Rats Sprague-Dawley
MH - Respiratory Mucosa/cytology	Respiratory Mucosa cytology

Language Processing) techniques (Aronson and Rindflesch, 1997; Aronson, 2001; Aronson et al., 2004). It can also be based on statistical information extracted from a corpus about associations between textual terms and MeSH terms. For example, Srinivasan (1996b)'s inter-field thesaurus is based on the co-occurrence of text terms t_j and MeSH concepts c_k in MEDLINE records. Note that MeSH concepts refer here to the non-trivial words found in the MeSH fields. Table 3.10 shows how MeSH fields are processed during indexing before any stopword removal or stemming is done. Each document is represented with two vectors, a text vector and a MeSH vector. The text terms and MeSH concepts appearing in a document are given weights that are derived from their frequencies in the document (number of times they appear in the document), and from their frequencies in the collection (number of documents they appear in). The *net association strength* between a term t_j and a concept c_k is given by:

$$\text{net association}\left(t_j, c_k\right) = \sum_{i=1}^{M} w\left(t_j, D_i\right) \times w\left(c_k, D_i\right)$$

where:

- M is the number of documents D_i containing both t_j and c_k, and
- $w\left(t_j, D_i\right)$ and $w\left(c_k, D_i\right)$, the weights of t_j and c_k, respectively, in document D_i, are calculated with the *atn* and *ntn* document indexing strategies of the SMART system (Salton, 1971).

Table 3.11: *atn* and *ntn* document indexing strategies from the SMART retrieval system

	atn	*ntn*
Term Frequency	$\frac{0.5+0.5\times tf}{(max\ tf\ in\ doc)}$	*tf*
Inverse Document Frequency	$\ln\left(\frac{N}{n}\right)$	$\ln\left(\frac{N}{n}\right)$
Document Length Normalization	none	none

Table 3.11 gives a brief description of the *atn* and *ntn* document indexing strategies, where N is the number of documents in the collection, and n the number of documents containing term of interest. To generate a MeSH query from an initial text query Q_t, the concepts c_k in the thesaurus associated with at least one term in Q_t are ranked according to their *association* score with the query:

$$\text{association}\,(Q_t, c_k) = \sum_{j=1}^{r} \text{net association}\,(t_j, c_k)$$

where r is the number of terms in Q_t. Concepts are finally selected from the top of the ranked list to form the MeSH query.

Pseudo-relevance Feedback. MeSH queries can be generated by pseudo-relevance feedback (PRF) on an initial text-based search. Relevance feedback is a well known method that consists of using the result of an initial search about which a user gives relevance feedback on retrieved documents. The relevance information is then used to modify the term weights of the initial query (relevance weighting), or to add new terms (query expansion) to the query (Rocchio, 1971; Ide, 1971; Robertson and Sparck Jones, 1996). PRF is relevance feedback in the absence of relevance information from the user. Document relevance is assumed from the position in the initial ranking.

In the post-retrieval combination method, the terms extracted from the MeSH fields are not added to the original text query, but are used to create a second query to search the MeSH index, as illustrated in Figure 5.1.

Srinivasan (1996b) uses a PRF method derived from Ide (1971) to generate MeSH queries. Ide (1971) uses relevance information to modify the weights of the terms t_j contained in the old query Q_{old} and create a new query Q_{new}:

$$w\left(t_j, Q_{new}\right) = \alpha \times w\left(t_j, Q_{old}\right) + \beta \times \sum_{i=1}^{R} w\left(t_j, D_i\right) - \gamma \times w\left(t_j, D_{\mathrm{NR1}}\right) \tag{3.6}$$

where $w\left(t_j, Q_{new}\right)$ and $w\left(t_j, Q_{new}\right)$ are the weights of t_j in Q_{new} and Q_{old}, respectively, R is the number of relevant documents retrieved, and D_{NR1} is the non-relevant document retrieved at the highest rank. α determines the importance of the old query weights, whereas β and γ determine the importance of the weights derived from relevant and non-relevant documents. For MeSH query generation, there is no old query so the first term of Equation 3.6 can be dropped. Additionally, Srinivasan (1996b) assumes that all documents obtained from the initial text search are relevant, so the last term of Equation 3.6 is also dropped. Therefore, the weights of the concepts c_k in the new MeSH query Q_{MeSH} are calculated with:

$$w\left(c_k, Q_{\mathrm{MeSH}}\right) = \sum_{i=1}^{R} w\left(c_k, D_i\right) \tag{3.7}$$

where R is the number of documents assumed relevant.

Fusion

The combination of the results of a text search and a MeSH search is a special case of the fusion of various runs obtained from different representations of documents. Fusion has been used with success whilst dealing with several runs (Shaw and Fox, 1993; Belkin et al., 1995). A simple method is to sum the normalised scores obtained by the documents D_i retrieved in each ranking:

$$score_{\text{text \& MeSH}}\left(D_i\right) = score_{\text{text}}\left(D_i\right) + score_{\text{MeSH}}\left(D_i\right)$$

Note that if a document does not appear in one ranking, its score for that ranking is simply equal to zero. Moreover, weights can be given to text and MeSH document scores to reflect the performance of the ranking:

$$score_{\text{text \& MeSH}}\left(D_i\right) = \alpha \times score_{\text{text}}\left(D_i\right) + \beta \times score_{\text{MeSH}}\left(D_i\right) \tag{3.8}$$

where α and β are used to balance the two document scores against each other.

Evaluations

Srinivasan (1996b) evaluates several strategies to combine text and MeSH searches. The experiments are based on Cornell's SMART retrieval system and a collection of 2,334 MEDLINE citations with 75 queries. Three strategies are evaluated: MeSH query creation with an inter-field statistical thesaurus, MeSH query creation via PRF, and MeSH query creation combining the first two approaches. Equation 3.8 is used for text and MeSH score combination. Table 3.12 presents the results for the three strategies. The combined approach gives the best improvement with 17% over the baseline performance of 0.5169 11-point average precision over the 75 queries. The baseline corresponds to the exclusive use of text queries. However, PRF alone gives almost the same improvement with 16.4%.

Srinivasan (1996a) evaluates the post-retrieval text and MeSH combination method with MeSH query creation via PRF on the OHSUMED collection. Equation 3.8 is used again for text and MeSH score combination. Table 3.13 shows the results and the parameter settings. An improvement of 8.2% is reported over the 11-point average precision of the baseline (text-only), 0.2415.

Kraaij et al. (2004) evaluates the post-retrieval text and MeSH combination method on the TrecGen04 collection (see Section 3.2.1). Both text and MeSH

Table 3.12: Results for Srinivasan (1996b) (R=number of documents assumed relevant, T=size of MeSH query)

MeSH query generation	fusion parameters	R	T	11-point average precision	increase over baseline (text-only)
statistical thesaurus	$\alpha = 1.3$ $\beta = 1$	na	15	0.5681	+9.9%
PRF	$\alpha = 0.66$ $\beta = 1$	10	20	0.6018	+16.4%
stat. thes. + PRF	$\alpha = 0.66$ $\beta = 1$	10	35	0.6051	+17.1%

Table 3.13: Results for Srinivasan (1996a) on OHSUMED (R=number of documents assumed relevant, T=size of MeSH query)

MeSH query generation	fusion parameters	R	T	11-point average precision	increase over baseline (text-only)
PRF	$\alpha = 2$ $\beta = 1$	5	20	0.2614	+8.2%

searches are done with a retrieval engine based on generative language models (Kraaij, 2004) and uses cross-entropy between queries and documents to score relevance:

$$H(Q; D) = \sum_{w} P(w|Q) \sum_{w} \log(\lambda P(w|C) + (1 - \lambda) P(w|D))$$

where $P(w|Q)$ is the conditional probability of term w given query Q, $P(w|D)$ is the conditional probability of term w given document D, $P(w|C)$ is the relative frequency of term w in document collection C, and λ is an interpolation parameter. Text-based document representations are extracted from title and abstract fields. For MeSH document representation, MeSH fields are processed by breaking up the associations between descriptors and qualifiers, and dropping the distinction between major and minor themes. Concepts expressed as a group of words (see Section 2.1.2) are preserved by replacing blanks by underscores. However, concepts combined with a comma are split. For example, in Figure 3.1, the descriptor "*Rats,*

Table 3.14: Results for Kraaij et al. (2004)

MeSH query generation	fusion parameters	R	T	MAP	increase over baseline
PRF	$\alpha = 0.8$ $\beta = 0.2$	3	all terms in documents	0.3247	+1.6%

Sprague-Dawley" is split into two tokens, "*Rats*" and "*Sprague-Dawley*", whereas "*Fluorescent Antibody Technique*" is turned into "*Fluorescent_Antibody_Technique*". MeSH queries are generated by PRF by concatenating the terms found in the MeSH fields of the top three documents of the text search output. Equation 3.8 is used for text and MeSH score combination. Table 3.14 shows the results and the parameter settings.

Suggestions

The evaluations show that pseudo-relevance feedback is a simple and efficient solution to the generation of MeSH queries. Moreover, the post-retrieval text and MeSH combination compares well with the pre-retrieval combination in terms of improvements over text-only representation methods. The post-retrieval method is interesting as it allows us to isolate MeSH-only representations and evaluate different strategies to generate such representations. In particular, whether the information within the structure of MeSH fields is useful or not for retrieval remains an open question. More precisely, the impact of concept associations (descriptors with qualifiers), and concept discriminations (major versus minor themes) in the MeSH fields is an interesting problem to address.

3.3 MeSH Representation in Document Classification

A discussion on automated text classification and machine learning techniques used for text categorization is beyond the scope of this dissertation and can be found elsewhere (Sebastiani, 2002). Regardless of the categorization techniques used, we focus here on the use of MeSH in document representation approaches and their evaluation in the classification of MEDLINE documents. As the MeSH representations presented in this section are evaluated with the same classification task organized by the Genomics track of TREC 2004 and 2005, we first give some background about that task. We then describe the three main methods used in this task to integrate MeSH in document representation:

1. mixing text and MeSH for document representation,

2. disambiguating text with MeSH, and

3. representing documents only with MeSH.

3.3.1 TREC 2004 and 2005 Genomics Track GO Triage task

One of the tasks of the TREC 2004 and 2005 Genomics track was a biomedical document triage task for gene annotation with the Gene Ontology (GO) (Gene Ontology Consortium, 2004). GO is used by several model organism database curators in order to standardize the description of genes and gene products. The triage task simulated one of the activities of the curators of the Mouse Genome Informatics (MGI) group (Eppig et al., 2005). MGI curators manually select biomedical documents that are likely to give experimental evidence for the annotation of a gene with one or more GO terms.

For both the 2004 and 2005 GO triage tasks, the same subset of three journals from 2002 and 2003 was used. The subset contained documents that had been selected or not for providing evidence supporting GO annotation. The 5837 documents from 2002 were chosen as training documents, and the 6043 from 2003 as test documents. In 2004, the training and test sets contained 375 and 420 positive examples (documents labeled relevant), respectively. In 2005, the number of positive examples for the training and test documents was updated to 462 and 518, respectively.

The triage task was evaluated with 4 metrics: precision, recall (see Section 3.2.2), F-score, and a normalized utility measure. The F-score measure combines precision and recall, and is defined by the following formula:

$$\text{F-Score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precison} + \text{recall}} \tag{3.9}$$

The normalized utility measure is defined by:

$$\mathrm{U}_{norm} = \mathrm{U}_{raw}/\mathrm{U}_{max} \tag{3.10}$$

with U_{max} being the best possible score. U_{raw} was calculated with the following formula:

$$\mathrm{U}_{raw} = (u_r \times \mathrm{TP}) + (u_{nr} \times \mathrm{FP}) \tag{3.11}$$

where u_r is the relative utility of a relevant document and u_{nr} the relative utility of a non-relevant document. With u_{nr} set at -1 and u_r assumed positive, u_r was determined by preferred values for U_{norm} in four boundary cases: completely perfect prediction, all documents judged positive (triage everything), all documents judged negative (triage nothing), and completely imperfect prediction. With different numbers of positive examples for classification between 2004 and 2005, $u_r = 20$ and $u_r = 11$ were chosen in 2004 and 2005, respectively. Note that the goal of the task was to maximize the normalized utility and the results of the participants reflect this choice. In this dissertation, we also focus on the normalized utility. Additionally, we provide precision, recall, and F-Score values for information purposes.

Table 3.15: Text and MeSH mixing approaches in 2004 GO task

Participant	Precision	Recall	F-score	Normalized Utility
Dayanik et al. (2004)	0.1579	0.8881	0.2681	0.6512
Fujita (2004)	0.1490	0.7690	0.2496	0.5494
Darwish and Madkour (2004)	0.1510	0.719	0.2496	0.5169
Cohen et al. (2004)	0.1714	0.6571	0.2719	0.4983

For a detailed description of the task, the reader is directed to Hersh et al. (2004, 2005).

3.3.2 Mixing Text and MeSH for Document Representation

While selecting features for MEDLINE document classification, most approaches extract terms from the MeSH fields along with textual fields such as the title and the abstract of the record.

Fujita (2004) uses a classifier of soft margin linear support vector machines (SVM) and represents documents with text fields mixed with MeSH terms and gene expressions identified by a gene name tagger. Adding the MeSH terms in the feature set is reported to benefit the triage results.

Cohen et al. (2005) represents documents with text, MeSH terms, and mouse genes identified in the abstracts. Mouse genes are identified with a named entity recognition and normalization (NER+N) system. The NER+R system is dictionary-based and uses gene name information from several genomic databases including MGI (Cohen, 2005).

The MeSH terms are also mixed with the text during feature selection by Aronson et al. (2005), Cohen et al. (2004), Dayanik et al. (2004), Darwish and Madkour (2004), and Si and Kanungo (2005). Tables 3.15 and 3.16 show the results of this approach for the 2004 and 2005 GO task, respectively. Dayanik et al. (2004)'s normalized utility corresponds to the best result of the 2004 GO triage task.

Table 3.16: Text and MeSH mixing approaches in 2005 GO task

Participant	Precision	Recall	F-score	Normalized Utility
Si and Kanungo (2005)	0.1947	0.8938	0.3198	0.5577
Cohen et al. (2005)	0.2308	0.7819	0.3564	0.5449
Aronson et al. (2005)	0.3223	0.5656	0.4107	0.4575

Table 3.17: Results for approaches using MeSH to disambiguate text (2005 task only)

Participant	Precision	Recall	F-score	Normalized Utility
Niu et al. (2005)	0.2122	0.8861	0.3424	0.5870
Subramaniam et al. (2005)	0.2028	0.9015	0.3311	0.5793
Schijvenaars et al. (2005)	0.2178	0.7259	0.3351	0.4889

3.3.3 Text Disambiguation with MeSH

Other approaches use ontologies (including MeSH) in order to disambiguate and identify biological concepts in text and MeSH fields. Niu et al. (2005) uses MeSH to standardise concepts appearing in full text: synonyms are removed and terms are mapped to their standard MeSH concept. Also, in Subramaniam et al. (2005), the MeSH ontology and other mouse-specific ontologies from MGI are used to identify the set of features for classifying MEDLINE records. This concept look-up is done with the BioAnnotator tool (Subramaniam et al., 2003). BioAnnotator finds biological terms in free-text and maps them to their corresponding concepts in the chosen ontology. Similarly, Schijvenaars et al. (2005) use the Collexis indexing system[7] in order to map the document free-text terms to concepts from the MeSH vocabulary and other thesauri such as the Gene Ontology and gene name lists extracted from genomic databases. Table 3.17 gives the results of the text disambiguation approach. Note that Niu et al. (2005) and Subramaniam et al. (2005) are the highest and second highest scoring participants for the 2005 GO task, respectively.

[7] http://www.collexis.com, last accessed: 19 January 2007

Table 3.18: Results for approaches using MeSH-based document representation (2004 and 2005 task)

Participant	Precision	Recall	F-score	Normalized Utility
Seki et al. (2004)	0.1118	0.7214	0.1935	0.4348
Lee et al. (2005)	0.1873	0.8803	0.3089	0.5332

3.3.4 MeSH-based Document Representation

A final method consists of selecting document features from the MeSH fields only. Seki et al. (2004) and Lee et al. (2005) describe experiments with MeSH-only document representations.

Seki et al. (2004) describes a process that trains a naïve Bayes classifier using the MeSH features to represent documents. Additionally, a gene filter is also applied based on gene names found in the text with the YAGI[8] (Yet Another Gene Identifier) software (Settles, 2004). The results shown in Table 3.18 include the gene filter. Seki et al. (2004) does not specify how the MeSH fields are processed, i.e. whether groups of words corresponding to descriptors are split, for example (see Section 3.1).

Lee et al. (2005) used MeSH for document representation and compared this approach to other representations such as abstract-based and caption-based representations. The semantic network of the UMLS[9] (Unified Medical Language System) was also used to map terms from MeSH, free-text, captions to UMLS Semantic Types, which are broad medical concepts. An SVM classifier was used and cross-validation results on the task training documents showed the MeSH representation to give the best performance with 0.4968 in normalized utility. The second best representation was obtained by using the UMLS semantic types of the MeSH terms (0.4403). The results in Table 3.18 correspond to a combination of three representation: MeSH, captions, and UMLS semantic types associated with MeSH.

[8] http://www.cs.wisc.edu/~bsettles/yagi, last accessed: 21 January 2007

[9] http://umlsinfo.nlm.nih.gov/, last accessed: 19 January 2007

3.3.5 Suggestions

All document representation methods for the GO triage task integrated the MeSH fields. The highest-scoring methods for 2005 (Niu et al., 2005; Subramaniam et al., 2003) used MeSH to disambiguate terms contained in the text fields. Representations solely based on MeSH (Seki et al., 2004) gave lower performance. However, the information contained in the structure of the MeSH fields has not been fully examined and its use could lead to improvement of the MeSH representations. Moreover, the inter-concept relationships contained in the MeSH hierarchy can be used to extend the representation. Whether such an extension could improve the classification performance further remains an open question.

3.4 MeSH Representation and Document Clustering

MeSH-based representations have also been used for the clustering of MEDLINE records. An introduction to text clustering is beyond the scope of this dissertation and can be found elsewhere (Jain and Dubes, 1988; Jain and Murty, 1999).

Ontrup et al. (2003) evaluates MeSH-based document representations with a method for the automatic indexing of MEDLINE documents with MeSH concepts. The hierarchy is used to represent documents as MeSH sub-trees by expanding the descriptors contained in documents' MeSH fields up to the root of the hierarchy. Because the same descriptors may be at different locations in the hierarchy, several sub-tree representations are possible. Ontrup et al. (2003) always chooses the denser representation, i.e. the sub-trees with the lowest number of edges. Figure 3.3 shows an example of two subtrees built for two documents D_1 and D_2 containing concepts "*Anaplasia*" and "*Precancerous Conditions*", respectively. The distance $d_m(\alpha, \beta)$ between the sub-tree representations of documents α and β is calculated with the

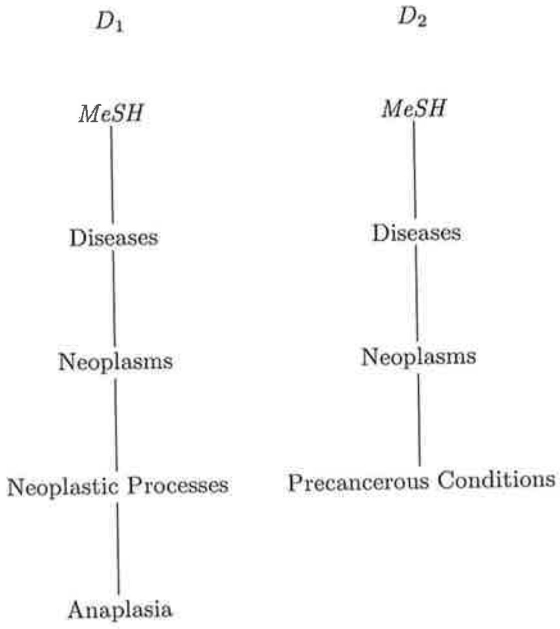


Figure 3.3: Subtree examples for documents containing "*Anaplasia*" and "*Precancerous Conditions*", respectively

following equation:

$$d_m(\alpha, \beta) = \exp^{-|B_\alpha \cap B_\beta|} + \frac{1}{1 + \exp^{-|B_\alpha \triangle B_\beta|}} \tag{3.12}$$

where B_α and B_β are the sets of branches contained in the sub-tree representations of documents α and β, respectively. $B_\alpha \triangle B_\beta = (B_\alpha \cup B_\beta)/(B_\alpha \cap B_\beta)$ is the symmetric difference between the two sub-tree representations of document α and β. For example, we can calculate $d_m(D_1, D_2)$ from Figure 3.3:

$$\begin{aligned} d_m(D_1, D_2) &= \exp^{-|B_{D_1} \cap B_{D_2}|} + \tfrac{1}{1+\exp^{-|B_{D_1} \triangle B_{D_2}|}} \\ &= \exp^{-2} + \tfrac{1}{1+\exp^{-\frac{5}{2}}} \\ &= 1.059 \end{aligned}$$

The MeSH-based representation is compared to a text-based representation in the context of automatically assigning MeSH terms to new documents. The text representations are built from the title and abstract fields of MEDLINE records. After stemming and stopword removal, the terms are assigned a TF*IDF weight. The distance $d(\alpha, \beta)$ between the text-based representations of documents α and β is derived from the Cosine similarity:

$$d(\alpha, \beta) = 1 - \cos(v^{\alpha}, v^{\beta}) \tag{3.13}$$

where v^{α} and v^{β} are the feature vectors of document α and β, respectively. Ontrup et al. (2003)'s automatic indexing method can be described in the following way. The 7,175 labeled training documents are represented separately with text and MeSH. The 5,496 unlabeled test documents are only represented with text. A total of m labeled documents ($m = 2$) are selected as nearest neighbors to an unlabeled document in terms of the text-based distance $d(\alpha, \beta)$ (Equation 3.13). Then the k_B nearest ($k_B = 4$) neighbors of the m documents are selected in terms of MeSH-based distance $d_m(\alpha, \beta)$ (Equation 3.12). The MeSH terms contained in the labeled documents (for a maximum of 10 documents) are extracted to label the unlabeled document. The results show that the text-based representation alone yields better results than the combined text and MeSH approach in terms of average precision and recall of MeSH terms.

Struble and Dharmanolla (2004) examines MeSH-based representations for the discovery of topics within the biomedical literature. The representations are built by extracting MeSH terms from the MEDLINE MeSH fields, and the documents are expressed as feature vectors. Three MeSH vectors are built for each document: one with descriptors alone, one with qualifiers alone, and one with descriptors and qualifiers. The vectors are then expanded with the MeSH hierarchy in the following way:

1. if the term was found in the MeSH fields (originally assigned), its weight is 2,
2. if a term is an ancestor of a term found in the MeSH fields (located on the path from the originally assigned term to the root of the hierarchy), its weight

Table 3.19: Expanded representations of D_1 and D_2

	D_1		D_2	
	concepts	weights	concepts	weights
Expanded Representations	Anaplasia Neoplastic Processes Neoplasms Diseases *MeSH*	2 1 1 1 1	Precancerous Conditions Neoplasms Diseases *MeSH*	2 1 1 1

is 1, and

3. in all other cases, its weight is 0.

For example, Table 3.19 shows the expanded representations of two documents D_1 and D_2 containing concepts "*Anaplasia*" and "*Precancerous Conditions*", respectively. The three MeSH-based representations are compared with a representation derived from the full-text of the articles.

Struble and Dharmanolla (2004)'s different representations are evaluated with the clustering of documents contained in two datasets: the Rat Genome Database (2713 documents), and the *Tourette's Syndrome* PubMed query result set (2241 documents). The goal of the clustering is to identify themes or topics in the biomedical literature. Struble and Dharmanolla (2004) uses AGNES, an agglomerative hierarchical clustering algorithm, with average linking. Average linking is a method that calculates the distance between two clusters with the average of the distances between the documents contained in the clusters. Principal components analysis (PCA) is used for dimension reduction. Inter-document similarities are calculated with the Cosine measure. The clustering quality is evaluated with an agglomerative coefficient a:

$$a = \text{mean}_{d \in D}\left(1 - \frac{m_d}{M}\right)$$

where m_d is the height at which document d is first merged, M is the height of the final merge, and D is the document collection. Intuitively, a larger a suggests better clustering quality when comparing datasets of a similar size. All three MeSH representation approaches yield better clustering quality than the full-text representation approach for the two datasets.

3.4.1 Suggestions

Evaluations

Ontrup et al. (2003) and Struble and Dharmanolla (2004) both used the MeSH hierarchy to extend the MeSH representations of documents. The MeSH representations worked well for the recognition of themes in the literature. However, they did not perform as well as text representations for the automatic indexing of document with MeSH concepts. This suggests that the performance of MeSH representations and their extension with the hierarchy may depend on the task they are evaluated with.

Extension Methods

In the extension methods used by Ontrup et al. (2003) and Struble and Dharmanolla (2004), we noticed that the concepts added to the initial representation were given the same importance in the representations. For example, in Formula 3.12, all the edges of the subtree have the same importance. Similarly, with Struble and Dharmanolla (2004)'s method, the added MeSH concepts are given the same weight, independently from their semantic similarity to the initial MeSH representation. However, intuitively, the importance of a concept added to a representation should depend on its semantic similarity to the initial representation. As "*Anaplasia*" is a disease, the representation of a document D_1 containing "*Anaplasia*" could be extended with the concept "*Diseases*". Nonetheless, since D_1 is not about just any diseases but a particular disease, the concept "*Diseases*" should have a lower weight than "*Anaplasia*" in the extended representation. Furthermore, the hierarchy can be used to measure the semantic distance between added concepts and the initial representation. In Figure 2.3, we can see that "*Neoplastic Processes*" is closer to "*Anaplasia*" than "*Diseases*" is. If "*Neoplastic Processes*" is added to the representation of D_1, its weight should reflect this difference. Finally, Ontrup et al. (2003) and Struble and Dharmanolla (2004) considered the ancestor concepts only during representation extension. What about the sibling concepts of "*Anaplasia*" ("*Neoplasm Invasiveness*") in Figure 2.3, the uncle concepts ("*Precancerous Conditions*"), the nephew concepts ("*Leukemic Infiltration*"), and all other related concepts in the hierarchy? This suggests examining the impact of adding all related concepts to the representations.

3.5 Summary

Most uses of the MeSH ontology for document representation make little use of the information contained in the MeSH field structure and the MeSH hierarchy relationships. Shin and Han (2004)'s approach distinguished major themes from minor themes in document representations. Following on that inspiration, more MeSH field information such as the associations between descriptors and qualifiers could be evaluated. Hierarchy relationships were used by Ontrup et al. (2003) and Struble and Dharmanolla (2004) to extend the representation of documents. Also, a basic PubMed search extends the search of a concept to all its descendent concepts in the hierarchy. These extension methods lead to interesting questions regarding the use of the hierarchy. First, should we add only ancestor concepts, descendant concepts, or concepts located anywhere in the hierarchy? Second, how can we determine the importance of the concepts added to the representation in relation to the concepts already in the representation? With the methods used by Ontrup et al. (2003), Struble and Dharmanolla (2004), and PubMed, all added concepts, ancestors or descendants, are assumed to be equally important. The opposite could be argued. Intuitively, the importance of added concepts depends on the semantic similarity between them and the initial concepts. The semantic similarity can be derived from the distances between concepts in the hierarchy.

In the next chapter, we formulate hypotheses regarding the use of information contained in the structure of MeSH fields, in the MeSH hierarchy, and in a particular corpus (TrecGen04).

Chapter 4

Hypotheses and Methods

The central hypothesis of this dissertation is that the integration of information contained in MEDLINE MeSH fields, the MeSH hierarchy, and a large corpus (such as TrecGen04) in MeSH-based representations can improve MEDLINE retrieval. In this chapter we develop this central hypothesis in further detail: we formulate several hypotheses, and describe the methods used to evaluate them. In particular, we distinguish two sets of hypotheses:

1. the first is based on the introduction of information from the MeSH fields and the TrecGen04 collection to improve MeSH-based document representations without the MeSH hierarchy (referred to below as non-hierarchical hypotheses), and

2. the second is based on the introduction of information from the MeSH hierarchy to improve MeSH-based document representations and the comparison of representations (referred to below as hierarchical hypotheses).

4.1 Non-hierarchical Hypotheses

It was previously shown in Section 3.1 that the ontological content of MEDLINE records is found in the MeSH fields. MeSH fields (see Figure 3.1) contain a combination of one descriptor with zero or more qualifiers. A star is also used to distinguish

Table 4.1: Integration of the information contained in the structure of MeSH fields

Q	*Centrioles/ultrastructure*
D_1	*Centrioles/ultrastructure* *Organelles/*ultrastructure*
D_2	*Centrioles/*metabolism* *Cilia/*ultrastructure*
D_3	*Centrioles/*ultrastructure* *Cilia/physiology*

MeSH fields containing the major themes of the document from those containing minor themes. Our intuition is that integrating information about the MeSH field structure will increase the precision of MEDLINE retrieval. Consider a query Q and three documents D_1, D_2, and D_3, the content of which is shown in Table 4.1. If we take away the distinction between major and minor themes, documents D_1 appears as relevant to Q as D_3. Both documents contain the association "*Centrioles/ultrastructure*". However, if we integrate the distinction between major and minor themes, D_3 appears to be more relevant to Q than D_1. The MeSH field structure shows that "*Centrioles/ultrastructure*" is a major theme in D_3, but only a minor theme in D_1. Consequently, the precision of retrieval is increased. Next, we take away the association between descriptors and qualifiers. Document D_2 now appears as relevant to Q as D_3 as they both contain "*Centrioles*" and "*ultrastructure*" as major themes. Once again, if we integrate the association between descriptors and qualifiers, D_3 appears to be more relevant to Q than D_1. The MeSH field structure shows that "*Centrioles*" is not associated with "*ultrastructure*" in D_2. This information helps us to increase the precision of retrieval.

Corpus information introduced in MeSH representations can also benefit the precision of MEDLINE retrieval. Corpus-based term weighting has already proven beneficial for the free-text document representation (Sparck Jones, 1972; Salton and Buckley, 1987). The motivating idea is that the presence of a term inside a document denotes an association between the two which strength and relevance varies according to the term and the document. A term may be important in one document and less so in another. Also, one term may be more important than another to distinguish between one document from another.

A well known term weighting method is the TF*IDF method that was developed with the vector space model (Salton et al., 1975) for document representation. TF*IDF uses the term frequency (TF, number of times the term appear in the document) of the term in the document and the collection frequency (CF, number of documents the term appears in) of the term. TF is positively correlated to the importance of the term in the document: the more the term appears inside the document, the more important it is expected to be for the document. However, CF is negatively correlated to the importance of the term in the document: the more the term appears in the collection, the less important it is expected to be for the document. This negative correlation for a term t is expressed in the inverse document frequency (IDF) measure:

$$\mathrm{IDF}_t = \log_2 \left(\frac{N}{\mathrm{CF}_t} \right)$$

where N is the number of documents in the collection and CF_t is the collection frequency of concept t. The TF*IDF weight for term t is then obtained by combining TF and IDF with a multiplication:

$$\mathrm{TF{*}IDF}_t = \mathrm{TF}_t \times \mathrm{IDF}_t \tag{4.1}$$

Note that TF*IDF is query-independent: it does not integrate relevance information for a particular query.

Intuitively however, the impact of TF*IDF on the MeSH representations may be limited. Unlike free-text, MeSH fields include the knowledge of the human indexer, and the knowledge contained in the MeSH ontology. The indexer already integrates the document frequency TF of free-text concepts in the article. He or she is able to identify a common concept in lexical variants and select a standard descriptor for annotation. Furthermore, the distinction between major and minor themes is available to point out the central concepts of documents. However, document frequency may be useful to distinguish qualifiers in a document. Indeed, the same qualifier can be associated with various descriptors in the same record. For example, in document D_1 of Table 4.1, "*ultrastructure*" appears twice, which suggests a preponderance of this qualifier as a context in the document. Therefore, the document frequency of qualifiers could capture their importance for the document.

The indexers are also expected not to annotate records with non-discriminative concepts. For example, indexer tend not to use general concepts such as "*Diseases*", but choose the most specific concept available to describe the content of the article (Weinberg and Cunningham, 1985). This suggests that the collection frequency of concepts will have a limited impact on retrieval precision. Nevertheless, the MeSH vocabulary includes stopwords, the check tags, which are common concepts with a high collection frequency. The full list of check tags for MeSH 2004 is given in Table 4.3. The check tags are known and can be directly excluded from representations.

Consequently, our three non-hierarchical hypotheses state that:

1. weighting concepts with corpus information (TF*IDF),
2. discriminating between major and minor themes, and
3. acknowledging associations between descriptors and qualifiers

may improve on a baseline MeSH-based document representation solely based on the presence and absence of MeSH concepts in document (binary). In particular, the improved representations are expected to result in higher precision for MEDLINE retrieval. In the case of TF*IDF, however, the impact is expected to be small, as explained above.

4.1.1 Methods

We now present the methods used to evaluate the hypotheses described above. All representations obtained are compared to a baseline that we call the binary repre-

Table 4.2: Example of binary representation

concept	weight
Animals	1
Centrioles	1
ultrastructure	1
Cilia	1
Fluorescent Antibody Technique	1
Lymphocytes	1
cytology	1
Organelles	1
Rats	1
Rats, Sprague-Dawley	1
Respiratory Mucosa	1
Steroids	1
analysis	1
Trachea	1

sentation.

Our Baseline: the Binary Representation

In our binary representation we only acknowledge the presence or absence of individual descriptors and qualifiers in the document. Associations between descriptors and qualifiers are disregarded, and so is the distinction between major and minor themes. In effect each MeSH concept is either present or absent in the document—hence the term binary. Table 4.2 shows the binary representation of the MeSH content of the document of Figure 3.1. Note that "*ultrastructure*" and "*cytology*" are only represented once although they appear in three and two MeSH fields, respectively. Moreover, some concepts, such as "*Human*", "*Female*", and "*Male*" are not represented at all as they are check tags. All other concepts not occurring in the document are given a zero weight (not shown in Table 4.2).

Corpus Information with TF*IDF Weighting

Collection frequency information is introduced with a TF*IDF weighting scheme. Descriptors always appear once in a document, so their TF is always equal to 1.

Table 4.3: List of Check Tags (CTs) for MeSH 2004

CTs list
Comparative Study
English Abstract
Female
Human
In Vitro
Male
Support, Non-U.S. Gov't
Support, U.S. Gov't, Non-P.H.S.
Support, U.S. Gov't, P.H.S.

However, as qualifiers can be combined with several descriptors in the same document, their TF value can be greater than 1. Figure 3.1 shows that "*ultrastructure*" and "*cytology*" occur three and two times in the document, respectively. Therefore $\text{TF}_{ultrastructure} = 3$ and $\text{TF}_{cytology} = 2$. Table 4.4 shows the TF*IDF representation of the MeSH content of the document of Figure 3.1. The IDF values are calculated based on the TrecGen04 collection (see Section 3.2.1).

Distinguishing Major from Minor Themes: MajMin

Here we propose a weighting scheme that distinguishes the major themes from the minor themes simply by giving the former higher weights. The starting point of this approach, which we call MajMin, is the binary representation described above. Descriptors appear only once in documents and they only represent either major or minor themes. However, qualifiers can be part of several associations in the same document, and represent both major and minor themes. In the approach described in this section, a qualifier represents a major theme in the document if it occurs at least once in a descriptor/qualifier association that represents a major theme. Table 4.5 shows a representation of the MeSH content of the document of Figure 3.1 that gives higher weight of 3 to concepts representing the major themes of the document.

Table 4.4: Example of TF*IDF representation

concept	weight
Animals	1.88
Centrioles	14.30
ultrastructure	16.55
Cilia	11.78
Fluorescent Antibody Technique	8.32
Lymphocytes	7.99
cytology	8.65
Organelles	10.59
Rats	3.72
Rats, Sprague-Dawley	5.37
Respiratory Mucosa	11.71
Steroids	9.41
analysis	3.69
Trachea	9.25

Table 4.5: Example of MajMin representation

concept	weight
Animals	1
Centrioles	3
ultrastructure	3
Cilia	1
Fluorescent Antibody Technique	1
Lymphocytes	3
cytology	3
Organelles	3
Rats	1
Rats, Sprague-Dawley	1
Respiratory Mucosa	1
Steroids	3
analysis	3
Trachea	1

Table 4.6: Example of DescQual representation

concept	weight
Animals	1
Centrioles/ultrastructure	1
Cilia/ultrastructure	1
Fluorescent Antibody Technique	1
Lymphocytes/cytology	1
Organelles/ultrastructure	1
Rats	1
Rats, Sprague-Dawley	1
Respiratory Mucosa/cytology	1
Steroids/analysis	1
Trachea	1

Associations between Descriptors and Qualifiers: DescQual

Our simple binary representation ignores the associations found in MeSH fields between descriptors and qualifiers. We expect the associations to yield a more accurate document representation as they give a specific context to the concepts described by descriptors. In this last approach, called DescQual, the associations found in the MeSH fields of the documents are retained as minimal tokens of information. Not all combinations are possible between the 22,430 descriptors and the 83 qualifiers of MeSH 2004. The total number of valid combinations is 522,928. Representations are binary: a value of 1 indicates that the combination is present in the query/document, and a value of 0 indicates that the combination is absent. Table 4.6 shows a representation of the MeSH content of the document of Figure 3.1 that keeps the associations between descriptors and qualifiers as a minimal token of information. Associations not found in the document are not represented.

4.2 Hierarchical Hypotheses

4.2.1 Main Hypothesis

Most MEDLINE records contain 10 to 12 MeSH fields. Each field contains only one descriptor and zero or more qualifiers associated with the descriptor. Without the hierarchy, the comparison of MeSH-based document representations consists of finding exact matches. Intuitively, however, based on information in the hierarchy, some concepts are more related to one another, and others less so. This suggests that hierarchy integration can lead to more refined document comparisons, and better document representations. This intuition can be explained with two examples. First, if a document contains the concept "*Neoplasms*" and the hierarchy shows (see Figure 4.1) that "*Diseases*" is a parent concept of "*Neoplasms*", we assume that the document is also about "*Diseases*" to some degree. Second, if one document contains the concept "*Neoplasms*", and another contains "*Diseases*", we can assume that the two documents are similar to some degree. In the first example, we use the hierarchy to extend the representation of documents. In the second, we use the hierarchy to compare the concepts contained in documents. Both hierarchy integrations are expected to improve the recall of MEDLINE retrieval.

Methods for Hierarchy Integration

A first approach consists of using the MeSH hierarchy to add new MeSH concepts to the original document MeSH content. A weighting scheme is used to distinguish between the original MeSH concepts and the ones derived from the extension process: the original MeSH concepts receive a weight w_o of 1, whereas the derived MeSH concepts receive a weight w_d, where $0 \leq w_d < 1$, depending upon how semantically close they are to the original MeSH representation. Each added concept is weighted independently from the others. However, each added MeSH concept is compared to the entire original MeSH representation, which typically includes 10-12 terms. As a result, the weight of the new term, w_d, is derived from the combination of several inter-concept comparisons.

A second approach consists of using the hierarchy during the comparison of two document representations. Each concept in one representation is compared to each concept in the other representation. Therefore, the comparison of two documents corresponds to the combination of several similarities of pairs of concepts.

4.2.2 Secondary Hypotheses

In this section, we develop secondary hierarchical hypotheses regarding:

1. the combination of semantic similarities of pairs of concepts,
2. the different parts of the MeSH hierarchy, and
3. the variation of edge distance in the MeSH hierarchy.

The methods used to evaluate the hypotheses are also described.

Inter-concept Similarity Combination

As documents and queries contain several concepts, evaluating the semantic distance between documents and queries involves combining semantic distances between the concepts contained in the documents.

Our intuition is that the combination strategy depends upon the method used to integrate the hierarchy information. We use two methods for hierarchy integration: the first method integrates the hierarchy information at retrieval time, and the second at indexing time.

The "retrieval time" hierarchy integration method is used for comparing two sets of concepts contained in documents and queries. Let Q be a query and D a document represented respectively by vectors ($w_1(Q)$, $w_2(Q)$, $w_i(Q)$, ..., $w_N(Q)$) and ($w_1(D)$, $w_2(D)$, $w_j(D)$, ..., $w_N(D)$), where w_i and w_j are binary weights indicating the presence or absence of concepts i or j in the query and the document, respectively. The hierarchy is used to compare the concepts contained in the document to the concepts contained in the query ($w_i(Q) = w_j(D) = 1$) when there are distinct ($i \neq j$). The similarity between identical concepts shared by the query and the document ($i = j$) is always greater than the similarity of distinct concepts compared with the hierarchy ($i \neq j$).

The "indexing time" hierarchy integration method uses the hierarhcy to extend document representations, i.e. to add more concepts to the representation with a weight reflecting its semantic distance to the initial representation. Given query Q and document D represented by the vectors defined in the previous paragraph, the concepts i and j not contained originally in Q and D ($w_i(Q) = w_j(D) = 0$), respectively, are compared individually with the set of concepts i and j already contained in Q and D ($w_i(Q) = w_j(D) = 1$), respectively. This comparison result in a score h for concepts i and j not originally contained in Q and D, respectively, such as $0 \leq h_i(Q) < 1$ and $0 \leq h_j(D) < 1$. A threshold t can be defined to control the extent of the vectors' extension, so that:

- if $h_i(Q) \geq t$ (or $h_j(D) \geq t$), then $w_i(Q) = h_i(Q)$ (or $w_j(D) = h_j(D)$), and
- if $h_i(Q) < t$ (or $h_j(D) < t$), then $w_i(Q) = 0$ (or $w_j(D) = 0$).

Q and D are then compared by comparing concepts i and j such as $i = j$ and $w_i(Q) = w_j(D) \neq 0$.

The concepts contained in MEDLINE records belong to different parts, or categories, of the MeSH ontology. The number of MeSH categories contained in MEDLINE records, 6.2 on average in the TrecGen04 collection (Hersh et al., 2004), suggests that MEDLINE records' conceptual content is rather mixed.

In Chapter 2, we presented an all-combination approach (Equation 2.9, Section 2.2.2) which considers all possible concepts pairs, and a best-match-combination approach (Equation 2.14, Section 2.2.2) which combines only best matches between concept sets. As the results presented in Table 2.7 of Section 2.2.2 suggest, all-combination penalises documents with mixed content.

When the hierarchy is used at comparison time, our assumption is that the best-match-combination approach gives better similarity evaluation measures, as this approach does not penalise the mixed content of the documents. However, when the hierarchy is used to extend document representations, each concept in turn is compared to the initial representations (which itself includes several concepts). The use of the best-match-combination approach will give high weights to concepts close to individual concepts in the initial representation. This is rather an extension of the individual concepts of the document. In contrast, we want to extend the whole content of the document by expressing it with all the other concepts not originally present. For this purpose, our intuition is that the all-combination approach will perform better than the best-match-combination approach. Consequently, our two hypotheses are:

1. best-combination performs better than all-match-combination when the hierarchy is used at comparison time, and
2. all-match-combination performs better than best-combination when the hierarchy is used to extend document representation.

The hypotheses stated above are not , to the best of our knowledge, supported by any prior evidence and are entirely based on our intuition.

Methods. A pair of measures are used to evaluate combination methods at comparison time, and another pair of measures are used to evaluate combination methods to extend document representation.

With the method integrating the hierarchy information at comparison time, $dist_{rada2}$ (Equation 2.9) and $dist_{azu}$ (Equation 2.14) are used to evaluate the all-combination and best-match-combination approaches, respectively. Note that, in Equation 2.14, $dist_{rada1}$ is used instead of $dist_{jiang4}$ in order to use the same inter-concept distance as in Equation 2.9.

With the method integrating the hierarchy to extend document representation, the semantic distance between a concept c_i and a document D containing initially

m concepts is calculated with:

$$dist_{ext1}(c_i, D) = \frac{1}{m}\sum_{j=1}^{m} dist_{rada1}(c_i, c_j) \tag{4.2}$$

to evaluate the all-combination approach, and

$$dist_{ext2}(c_i, D) = \min_{j}(dist_{rada1}(c_i, c_j)) \tag{4.3}$$

to evaluate the best-match approach.

Combined versus Distinct Hierarchies

The MeSH 2004 ontology is organized into fifteen independent descriptor hierarchies and twenty-three smaller and shallower qualifier hierarchies. To use the MeSH semantic network to evaluate the semantic distance between *all* MeSH terms found in the MeSH fields (descriptors and qualifiers), some nodes need to be added to the MeSH hierarchies. First, a "*qualifier*" node is placed over all the root nodes of the qualifier hierarchies. This node is created to compensate for the shallowness and the small size of the qualifier hierarchies. The additional "*qualifier*" node also implicitly increases the distinction (or edge count) between descriptors and qualifiers by increasing the network distance between any descriptor and any qualifier by one edge. A "*MeSH*" root node is then placed on top of all the original descriptor root nodes (the main descriptor categories) and the previously created "*qualifier*" node. Figure 4.1 illustrates the combined hierarchy.

Certain questions can be formulated about the comparison of document contents (descriptors and qualifiers) with the entire MeSH network. First of all, does it make sense to allow the comparison between concepts located in different categories? For example, can we compare a geographic locations concept to a disease? Secondly, is it appropriate to compare descriptors to qualifiers? With 22,430 distinct concepts and 41,063 network nodes, descriptors constitute the main part of the MeSH vocabulary.

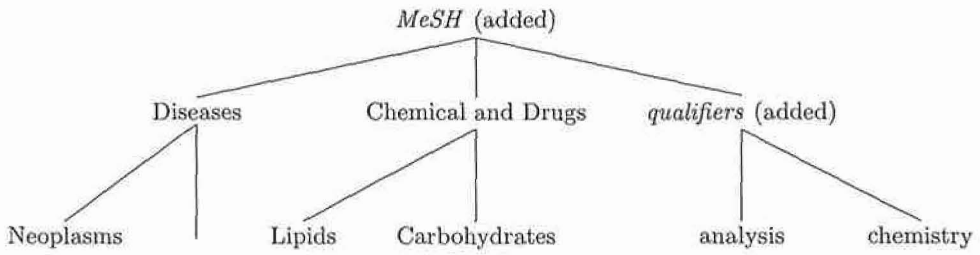


Figure 4.1: "*MeSH*" and "*qualifiers*" nodes add-up (only a few child nodes depicted for clarity)

Qualifiers are always used to provide additional contextual meaning to descriptors in the MEDLINE MeSH fields and they only include 83 distinct concepts and 99 nodes. Furthermore, qualifier hierarchies are shallow. Consequently most qualifier nodes are located only a few edges from the "*MeSH*" root node. With a semantic distance based on the minimum edge count between nodes, qualifiers will be semantically close to the general descriptor concepts located near the root.

Therefore, our two main hypotheses regarding the different parts of the MeSH hierarchy assert that:

1. comparing concepts from different MeSH categories (including the artificial "*qualifier*" category) with an added artificial "*MeSH*" root, and
2. comparing concepts from descriptor categories with concepts from qualifier categories

are expected to favor recall against precision in MEDLINE retrieval.

Methods.

Baseline. The baseline does not distinguish between categories, and uses the artificial "*qualifier*" and "*MeSH*" nodes to compare concepts from different categories. In Figure 4.1, using a simple edge count, the semantic distance of concepts from different categories or the same category can be evaluated in the same manner, for example:

$$\begin{aligned}
distance\,(\text{``}Neoplasms\text{''}, \text{``}Lipids\text{''}) &= edge\_count\,(\text{``}Neoplasms\text{''}, \text{``}Lipids\text{''}) \\
&= 4 \\
distance\,(\text{``}Neoplasms\text{''}, \text{``}analysis\text{''}) &= edge\_count\,(\text{``}Neoplasms\text{''}, \text{``}analysis\text{''}) \\
&= 4 \\
distance\,(\text{``}Lipids\text{''}, \text{``}Carbohydrates\text{''}) &= edge\_count\,(\text{``}Lipids\text{''}, \text{``}Carbohydrates\text{''}) \\
&= 2
\end{aligned}$$

Category Separation (HardSep). This method penalises the comparison of concepts from different categories by systematically giving them the maximum distance in the hierarchy, for example:

$$\begin{aligned}
distance\,(\text{``}Neoplasms\text{''}, \text{``}Lipids\text{''}) &= edge\_count\,(\text{``}Lipids, \text{``}Carbohydrates\text{''}) \\
&= 2
\end{aligned}$$

as "*Lipids*" and "*Carbohydrates*" belong to the same category, "*Chemicals and Drugs*", but:

$$\begin{aligned}
distance\,(\text{``}Neoplasms\text{''}, \text{``}Lipids\text{''}) = max\_dist \\
distance\,(\text{``}Neoplasms\text{''}, \text{``}analysis\text{''}) = max\_dist
\end{aligned}$$

as "*Neoplasms*", "*Lipids*", and "*analysis*" belong to different categories. The maximum distance in the combined MeSH hierarchy is 23 edges.

MeSH Category Ranking (SoftSep). Here we propose a softer category separation with a method that includes the ranking of the MeSH categories according to their relevance to generic genomic topics. Here, the MeSH categories refer to the categorization given to us by the organization of the MeSH vocabulary, i.e. the fifteen descriptor categories (Table 2.3) and the 23 qualifier categories (Table 2.4). We keep this categorization as is, apart from the 23 qualifier categories that we merge into one combined "*qualifier*" category. The assumption is that some categories may be less relevant than others for searching the biomedical literature. For example, we expect the "*Diseases*" category to be very relevant for a query about a possible connection between a gene and a disease. For a similar query, the "*Geographic Locations*" category might be less relevant.

One way to rank MeSH categories, thereby determining the most useful for the task at hand, is to score them according to their presence or absence in documents that were judged relevant or not relevant to generic genomic topics. A category is considered to be contained in a document, relevant or not, if at least one term from this category is present in the document. As descriptors and qualifiers can belong to several categories, a single descriptor or qualifier can increment the document frequency of several categories.

We score categories based on a relevance weight function RW introduced by Robertson and Sparck Jones (1996):

$$RW\,(\text{category}) = log\left(\frac{(r+0.5)\cdot(N-n-R+r+0.5)}{(n-r+0.5)\cdot(R-r+0.5)}\right) \qquad (4.4)$$

where r is the number of relevant documents a category occurs in, n is the number of documents the same category occurs in, R is the number of relevant documents for the query, and N is the number of documents in the collection. RW derives from a probabilistic theory of relevance weighting (Robertson and Sparck Jones, 1976). Consider the contingency table represented in Table 4.7. RW is based on the ratio between the odds of relevance for a category (ratio between the number of relevant

Table 4.7: Contingency table for relevance weighting

	doc is relevant	doc is not relevant	
category in doc	r	$n-r$	n
category not in doc	$R-r$	$N-n-R+r$	$N-n$
	R	N-R	N

documents in which it occurs, r, and the number in which it does not occur, $R-r$), and the odds of non-relevance for the same category (ratio between the number of non-relevant documents in which it occurs, $n-r$, and the number in which it does not occur, $N-n-R+r$). The constant 0.5 is added for limiting cases such as the absence of relevance information ($r = R = 0$). The probabilistic theory underlying RW uses independence assumptions (independence between terms), and ordering assumptions (ordering of documents). In particular, for RW, the distribution of categories in relevant documents is assumed independent, and so is the distribution of categories in non-relevant documents. Furthermore, the probable relevance of documents is based on both the presence and absence of categories in and from the documents.

Robertson and Sparck Jones (1996) defines the offer weights OW of terms (categories in our case):

$$OW = r \times RW \tag{4.5}$$

where r is the number of relevant a category appears in and RW is the relevance weight of the same category defined by Equation 4.4. The advantage of using the offer weight method over other PRF methods such as Rocchio's (Rocchio, 1971) is that there is no need to tune parameters over the influence of relevant documents against non-relevant ones. Instead, the offer weight score of a concept is the ratio of two odds, the odd of the concept being found in a relevant document, and the odd of the concept being found in a non-relevant document.

The method uses relevance information for the TrecGen04 and TrecGen05 collections of the TREC 2004 and 2005 ad hoc tasks, respectively (see Section 3.2.1).

Table 4.8: Category rankings obtained with relevance information of 2004 and 2005 topics

2004		2005	
1. D: 1	9. E: 0.0234	1. Q: 1	9. H: 0.0698
2. Q: 0.7866	10. F: 0.0149	2. D: 0.9569	10. M: 0.0308
3. G: 0.7818	11. M: 0.0033	3. G: 0.6603	11. L: 0.0241
4. B: 0.4764	12. J: 0.0026	4. C: 0.3912	12. N: 0.0108
5. C: 0.3051	13. Z: 0.0011	5. F: 0.3394	13. J: 0.0018
6. A: 0.3050	14. K: -0.0003	6. A: 0.2065	14. K: -0.0006
7. L: 0.0680	15. I: -0.0032	7. B: 0.2026	15. Z: -0.0053
8. H: 0.0439	16. N: -0.0075	8. E: 0.1081	16. I: -0.0079

TrecGen04 and TrecGen05 contain relevance information for 50 and 49 topics, respectively. We generate two rankings, one for each TREC year. For each topic, the offer weight of each of the 16 MeSH categories is computed with Equation 4.5. For each category, the average offer weight is calculated over the 50 and 49 topics t of TREC 2004 and 2005 respectively, and normalized based upon the highest OW average:

$$aver\_OW\left(\text{category}\right) = \frac{\left(\frac{1}{(50or49)}\sum_{t=1}^{(50or49)} OW_t\left(\text{category}\right)\right)}{\text{highest\_}OW}$$

where OW is defined in Equation 4.5. The category rankings obtained with the relevance information of the 2004 and 2005 topics are shown in Table 4.8. Q refers to the combined "*qualifier*" category, and all other capital letters refer to the 15 descriptor categories of MeSH 2004 described in Table 2.3, Section 2.1.2.

Both rankings give high scores to categories such as "*qualifiers*", "*Chemical and Drugs*", "*Biological Sciences*", "*Diseases*", "*Anatomy*" and "*Organisms*". In particular, the two categories "*qualifiers*" and "*Chemical and Drugs*" constitute the top 2 in both years. However, other categories score poorly in both rankings. On this basis, several catgories were manually selected as not relevant for generic genomic topics. Table 4.9 shows the selected categories and their scores in the 2004 and 2005 rankings. If any of the concepts being compared semantically belong

Table 4.9: Categories selected as not relevant for generic genomic topics

Categories	2004	2005
H (Physical Sciences)	0.0439	0.0698
I (Anthropology, Education, Sociology and Social Phenomena)	-0.0032	-0.0079
J (Technology and Food and Beverages)	0.0026	0.0018
K (Humanities)	-0.0003	-0.0006
L (Information Science)	0.0680	0.0241
M (Persons)	0.0033	0.0308
N (Health Care)	-0.0075	0.0108
Z (Geographic Locations)	0.0011	-0.0053

to a category contained in Table 4.9, the distance measure returns the maximum hierarchy distance. For example:

$$distance\,(\text{``}Neoplasms\text{''},\ \text{``}Lipids\text{''}) = edge\_count\,(\text{``}Neoplasms\text{''},\ \text{``}Lipids\text{''}) = 4$$
$$distance\,(\text{``}Neoplasms\text{''},\ \text{``}analysis\text{''}) = edge\_count\,(\text{``}Neoplasms\text{''},\ \text{``}analysis\text{''}) = 4$$

as "*Neoplasms*", "*Lipids*", and "*analysis*" do not belong to categories listed in Table 4.9, but

$$distance\,(\text{``}Neoplasms,\ \text{``}Coffee\text{''}) = max\_dist$$
$$distance\,(\text{``}Milk\text{''},\ \text{``}Coffee\text{''}) = max\_dist$$

as "*Coffee*" and "*Milk*" belong to a category, "*Technology and Food and Beverages*", listed in Table 4.9.

Separation between Descriptors and Qualifiers (DescQualSep). This method distinguishes descriptor concepts from qualifier concepts by also giving the maximum hierarchy distance when the semantic distance between a descriptor and a qualifier is evaluated, for example:

$$\begin{aligned}
distance\left(\text{“}Lipids\text{”},\ \text{“}Carbohydrates\text{”}\right) &= edge\_count\left(\text{“}Lipids\text{”},\ \text{“}Carbohydrates\text{”}\right)\\
&= 2\\
distance\left(\text{“}Neoplasms\text{”},\ \text{“}Lipids\text{”}\right) &= edge\_count\left(\text{“}Neoplasms\text{”},\ \text{“}Lipids\text{”}\right)\\
&= 4\\
distance\left(\text{“}chemistry\text{”},\ \text{“}analysis\text{”}\right) &= edge\_count\left(\text{“}chemistry,\ \text{“}analysis\text{”}\right)\\
&= 2
\end{aligned}$$

as the concepts being compared belong to the same category, either descriptor or qualifier, but:

$$distance\left(\text{“}Neoplasms\text{”},\ \text{“}analysis\text{”}\right) = max\_dist$$

as “*Neoplasms*” is a descriptor and “*analysis*” is a qualifier.

Edge Distance Variation

The MeSH hierarchy comprises nodes (concepts) and edges (*broader-than*/*narrower-than* relationships between concepts). Assuming that edges correspond to semantic distances between concepts, no information is explicitly available in the hierarchy about edge distances.

A simple approach is to assume that all edges in the hierarchy correspond to the same semantic distance. Intuitively, however, the creation of new and narrower concepts does not correspond to the same increase in narrowness or specificity. Therefore, the edges of the hierarchy are expected to represent various semantic distances (see Section 2.2.1).

The basic idea supporting edge distance variation is that the semantic distance between concepts is negatively correlated with the specificity of the concepts and the density of the conceptual area containing the concepts. For example, if we compare the edge connecting “*MeSH*” to “*Diseases*” with the edge connecting “*Neoplasm*

Invasiveness" to "*Leukemic Infiltration*" in Figure 2.3, we expect the semantic distance for the first edge to be higher than it is for the second. Indeed, the concepts of the second edge are more specific than the concepts of the first.

The conceptual specificity and density can be derived from the hierarchy structure (Section 2.2.1). In particular, the specificity of a concept is proportional to its depth. Additionally, the density of the conceptual area containing a concept is proportional to the number of edges connected to this concept.

The conceptual specificity and density can also be derived from the distribution of concepts in a corpus (Section 2.2.1). The idea is that both specificity and density are inversely proportional to the frequency of concepts. A specific concept is expected to appear in a limited number of documents, as few articles are published about a highly specialized area. Moreover, a high conceptual density means that many concepts are available to describe the subtleties of a research area, making each individual concept less likely to occur in publications.

Therefore we hypothesize that edge distance variation, calculated with hierarchy or corpus information, will impact positively on the precision of MEDLINE retrieval.

Methods. We evaluate our hypothesis by comparing a baseline measure that assumes edge distance to be constant with two measures that assume edge distance to vary. We calculate the variation with hierarchy and corpus information.

Our Baseline: Simple Edge Count. As a baseline, we use the edge count approach to evaluate the assumption that edge distance is constant in the MeSH hierarchy. To calculate the semantic distance between two concepts c_1 and c_2, we use $dist_{rada1}$ (Equation 2.1).

Depth and Density integration (DepthDens). A first method to calculate edge distance variation is to use a measure that is sensitive to the depth and the density of the hierarchy. To calculate the semantic distance between two concepts c_1 and c_2, we use $dist_{jiang2}$ (Equation 2.4).

Table 4.10: Overview of experimental space of the dissertation

Evaluations			Ad hoc retrieval	Binary classification
Experimental framework			Post-retrieval text & MeSH combination (Section 3.2.5)	Machine learning & classification with SVMs (SVM^{light}, Section 6.1)
Data			TrecGen05 (Section 3.2.1)	Training/test documents from 2005 GO triage task (Section 3.3.1)
Hypotheses	Non-hierarchical	TF*IDF	Section 5.1.2	Section 6.2
		MajMin	Section 5.1.2	Section 6.2
		DescQual	Section 5.1.2	Section 6.2
	Hierarchical	Integration	Section 4.2.1	Section 4.2.1
		Combination	Section 5.2.1	Section 6.3.1
		Separation	Section 5.2.2	Section 6.3.2
		Edge distance	Section 5.2.3	Section 6.3.3

Information-based Approach (InfoBased): A second method to calculate edge distance variation is to use corpus information. We use the MeSH concept distribution in the TrecGen04 collection (see Section 3.2.1) to calculate the probabilities of encountering the concepts in the collection. To calculate the semantic distance between two concepts c_1 and c_2, we use $dist_{jiang4}$ (Equation 2.8).

4.3 Hypotheses Evaluation Overview

The hypotheses presented in this chapter are evaluated in the contexts of ad hoc document retrieval and document classification in the following two chapters. Table 4.10 gives an overview of the experimental space covered.

Chapter 5

Evaluation with MEDLINE Ad Hoc Retrieval

In this chapter we evaluate the hypotheses formulated in the previous chapter in the context of MEDLINE ad hoc retrieval. First, we present our experimental set-up: the post-retrieval text and MeSH searches combination.

5.1 Post-retrieval Combination of Text and MeSH Searches

This approach, introduced in Section 3.2.5, combines the results of two searches: a text search and a MeSH search. The text queries are derived from the TREC 2005 topics described in Section 3.2.1, and the MeSH queries are obtained by pseudo-relevance feedback based on the output of the text search. An overview of the method is illustrated in Figure 5.1.

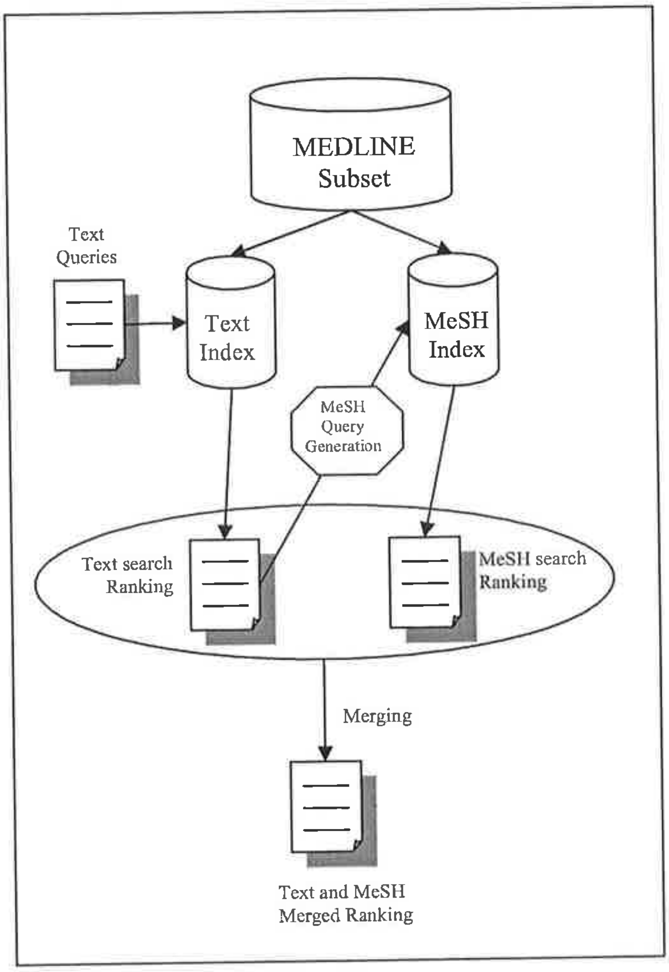


Figure 5.1: Experimental method overview

5.1.1 Text Search

Background

Probabilistic retrieval model have given the best performances for TREC 2004 and 2005 ad hoc tasks (Hersh et al., 2004, 2005). The Okapi BM25 relevance weighting function (Robertson et al., 1996) was the choice of participants with the highest results in 2004 (Fujita, 2004; Buttcher et al., 2004) and 2005 (Huang et al., 2005; Ando et al., 2005). Furthermore, Abdou et al. (2005) evaluated several probabilistic models and outlined the strong performance of models based on Divergence From Randomness (DRF). The hypothesis underlying DFR models is that terms that are informative for a particular document will occur more in that document than in other documents in the collection for which they are less informative. A full discussion on DFR models is beyond the scope of this dissertation and can be found elsewhere (Amati and Van Rijsbergen, 2002).

In this dissertation, the choice of a particular model for text-based retrieval was driven by the performance of the model in terms of MAP. High precision in the output of the text-based search is important as we assume the top documents retrieved relevant in order to generate MeSH queries. Moreover, high recall is important as the MeSH queries are then used to re-rank the documents already contained in the result set of the text-based searches. Text-based retrieval is evaluated with several models implemented by the Terrier search engine presented in the next section.

Terrier Search Engine

For text-based indexing and search, we used the Terrier search engine[1] developed at the Computing Science Department of the University of Glasgow, UK. The Terrier search engine implements several information retrieval models, including basic TF*IDF and probabilistic models such as BM25 (Robertson et al., 1996). In particular, Terrier offers a wide range of DFR models.

We experiment with three DFR models implemented by Terrier: BB2, I(n)L2, and DFR BM25. BB2 and I(n)L2 use two different randomness models, the Bose-Einstein distribution, and the inverse document frequency model, respectively. DFR BM25 is a derivation of BM25 (Robertson et al., 1996) from the DFR framework (Amati, 2003). The three DFR models use a term frequency normalization that is determined by the following formula:

$$tfn = tf \cdot \log_2(1 + c \cdot \frac{avg_l}{l}) \tag{5.1}$$

where avg_l is the average document length, l is the document length, and c is a free parameter. Terrier implementations of the TF*IDF model and the BM25 model

[1] http://ir.dcs.gla.ac.uk/terrier/, last accessed: 19 January 2007

(with Terrier standard settings: $k1 = 1.2$, $k3 = 1000$, $b = 0.75$) are also used for comparison purposes.

Indexing and Query Generation

The 4,591,008 MEDLINE documents of TrecGen05 are indexed with the Terrier search engine. The text field (title and abstract) terms are stemmed with the Porter algorithm (Porter, 1980), and MEDLINE-specific stopwords obtained from PubMed help[2] are removed from the index.

The 50 topics from TrecGen05 are instances of the five generic topic templates (GTTs) described in Section 3.2.1. An example of such an instance is *provide information about the role of the gene Interferon-beta in the disease Multiple Sclerosis.* In the following experiments, two types of queries are tested: *narratives* refer to the original 2005 text queries, and *basic narratives* refer to the queries obtained after removing from the original 2005 queries the structural terms of the GTTs. For example, for the GTT instance mentioned above, only *Interferon-beta* and *Multiple Sclerosis* are kept to generate the *basic narratives.* All other terms are common to all instances of the same GTT. Therefore, they are expected to be less discriminative for a particular topic instance. Tables A.2 and A.3 give the 2005 topics in their *narratives* and *basic narratives* format, respectively.

Results

Results are shown in Tables 5.1, 5.2, 5.3, 5.4, and 5.5. MAP and average recall are defined in Equations 3.5 and 3.4, respectively. All models yield better MAP values with the basic narratives (see Tables B.1, B.2, B.3, and B.4 for statistical significance). The DFR models (InL2, DFR BM25, BB2) give better performances in terms of MAP than TF*IDF and BM25 with the standard Terrier settings. Tables B.5, B.6, B.7 give statistical significance for the differences between the DFR

[2] http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=helppubmed.table.pubmedhelp.T42, last accessed: 19 January 2007

Table 5.1: The I(n)L2 model with various c values, basic/original narratives, and associated MAP and average recall

Basic narratives						
c value	0.5	1.0	1.5	2.0	2.5	3.0
MAP	0.2392	0.2594	0.2661	0.2680	0.2676	0.2677
Av recall	0.6827	0.7029	0.7013	0.6973	0.6976	0.6959
c value	3.5	4.0	4.5	5.0	5.5	6.0
MAP	0.2660	0.2659	0.2655	0.2644	0.2637	0.2633
Av recall	0.6948	0.6949	0.6930	0.6936	0.6927	0.6937
Narratives						
c value	0.5	1.0	1.5	2.0	2.5	3.0
MAP	0.2269	0.2457	0.2500	0.2508	0.2506	0.2500
Av recall	0.6677	0.6845	0.6795	0.6725	0.6683	0.6656
c value	3.5	4.0	4.5	5.0	5.5	6.0
MAP	0.2485	0.2480	0.2471	0.2459	0.2453	0.2447
Av recall	0.6622	0.6617	0.6596	0.6570	0.6535	0.6530

models InL2, DFR BM25, and BB2, respectively, and the TF*IDF and BM25 models. Figure 5.2 shows the MAP of the three DFR models with different c-values and Table B.8 shows the statistically significant differences between them for various values of *c*.

The DRF models perform consistently better than the TF*IDF and BM25 models in terms of MAP. Moreover, Figure 5.2 shows that BB2 reaches higher MAP values than I(n)L2 and DRF BM25 when *c* is greater than 2, although the differences are not statistically significant (see Table B.8). Consequently, the output obtained with the BB2 model, the basic narratives and a *c*-value of 5 (Equation 5.1) is selected to generate MeSH queries by pseudo-relevance feedback, and to be combined with the MeSH searches.

Comparison with Track Participant: Table 5.6 shows the best MAP results obtained by the participants to TREC 2005 Genomics track. The results indicate that our text baseline obtained with the BB2 model (0.2728 MAP) compares well with the runs submitted by the by the best participants.

Table 5.2: The DFR BM25 model with various c values, basic/original narratives, and associated MAP and average recall

Basic narratives						
c value	0.5	1.0	1.5	2.0	2.5	3.0
MAP	0.2300	0.2558	0.2622	0.2660	0.2675	0.2673
Av recall	0.6764	0.7045	0.7022	0.7021	0.6977	0.6972
c value	3.5	4.0	4.5	5.0	5.5	6.0
MAP	0.2674	0.2665	0.2658	0.2662	0.2663	0.2657
Av recall	0.6976	0.6983	0.6981	0.6998	0.6991	0.6984
Narratives						
c value	0.5	1.0	1.5	2.0	2.5	3.0
MAP	0.2222	0.2447	0.2496	0.2518	0.2524	0.2525
Av recall	0.6650	0.6876	0.6827	0.6812	0.6771	0.6647
c value	3.5	4.0	4.5	5.0	5.5	6.0
MAP	0.2519	0.2511	0.2502	0.2498	0.2492	0.2481
Av recall	0.6655	0.6660	0.6649	0.6643	0.6624	0.6594

Table 5.3: The TF*IDF model with basic/original narratives and associated MAP and average recall

Basic narratives	
MAP	0.2552
Av recall	0.6947
Narratives	
MAP	0.2412
Av recall	0.6745

Table 5.4: The BM25 model with Terrier standard settings, basic/original narratives, and associated MAP and average recall

Basic narratives	
MAP	0.2546
Av recall	0.7011
Narratives	
MAP	0.2437
Av recall	0.6833

Table 5.5: The BB2 model with various c values, basic/original narratives, and associated MAP and average recall

Basic narratives						
c value	0.5	1.0	1.5	2.0	2.5	3.0
MAP	0.2370	0.2528	0.2650	0.2688	0.2705	0.2715
Av recall	0.6656	0.6780	0.6805	0.6793	0.6837	0.6844
c value	3.5	4.0	4.5	5.0	5.5	6.0
MAP	0.2727	0.2722	0.2724	0.2728	0.2725	0.2728
Av recall	0.6837	0.6818	0.6811	0.6823	0.6819	0.6813
Narratives						
c value	0.5	1.0	1.5	2.0	2.5	3.0
MAP	0.2358	0.2490	0.2602	0.2623	0.2628	0.2631
Av recall	0.6716	0.6813	0.6839	0.6844	0.6860	0.6864
c value	3.5	4.0	4.5	5.0	5.5	6.0
MAP	0.2632	0.2624	0.2622	0.2621	0.2617	0.2615
Av recall	0.6815	0.6769	0.6772	0.6742	0.6723	0.6720

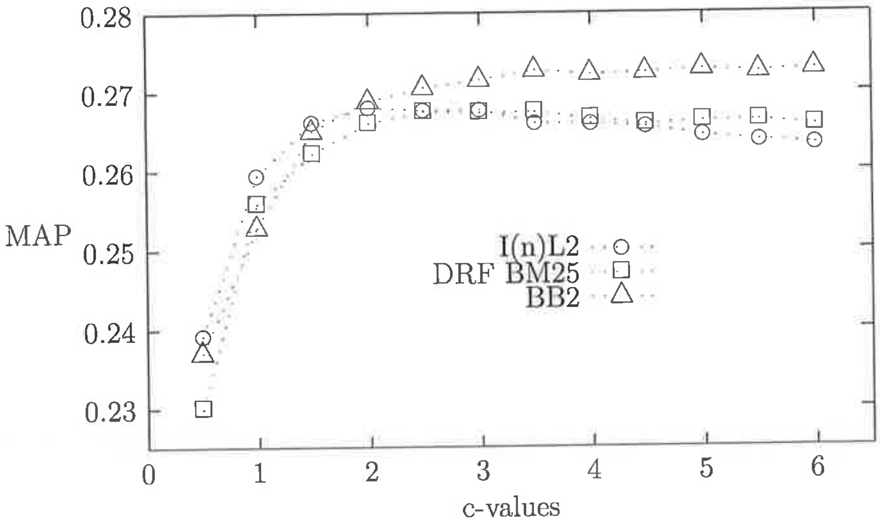


Figure 5.2: MAP for I(n)L2, DFR BM25, and BB2 for different c-values

Table 5.6: Best results for participants to TREC 2005 Genomics Track

Participants	Retrieval systems	Runs submitted

Huang et al. (2005)	0.2640 MAP
	Okapi Basic Search System (BBS):
	k1=1.4, k2=0, k3=8, b=0.55
		0.3020 MAP (best result)
		including:
		- term expansion with Acromed and LocusLink databases,
		- blank feedback with special term selection technique

Ando et al. (2005)	0.2610 MAP Lemur BM25 implementation: k1=1.2, k3=7, b=0.75	0.2883 MAP including: - structural feedback, - synonym expansion
Abdou et al. (2005)	0.2624 MAP I(n)L2 DRF model	0.2439 MAP including: - domain-specific expansion, - Rocchio expansion

5.1.2 MeSH Searches

Non-hierarchical Hypotheses Evaluation with Document/Query Representation

MeSH queries are generated with a pseudo-relevance feedback method (see Section 3.2.5). Because the queries are derived from the MeSH content of the documents retrieved during the text search, each document representation is associated with a query representation. The document representations correspond to the non-hierarchical hypotheses formulated in Chapter 4.

Pseudo-relevance Feedback Method (PRF). To generate MeSH queries, PRF is used on the text search result with the offer weight term scoring method (Robertson and Sparck Jones, 1996). MeSH terms are extracted from the MeSH fields of the top R documents of each text ranking. Each MeSH concept is then scored according to their Offer Weight score with Equation 4.5. The scored concepts are then ranked, and the top T terms are selected as the MeSH query. Suppose that two documents are retrieved by a text query ($R = 2$), that they are assumed relevant to this query, and that their MeSH representation is the one shown in Table 5.7. Note that the information of the MeSH field structure is ignored in this MeSH rep-

Table 5.7: MeSH representation of two documents obtained from a text search

document 1		document 2	
concept	weight	concept	weight
Animals	1	Animals	1
Centrioles	1	Cilia	1
ultrastructure	1	ultrastructure	1
Cilia	1	Dynein ATPase	1
Fluorescent Antibody Technique	1	Microscopy, Electron	1
Lymphocytes	1	Microtubules	1
cytology	1	Protein Conformation	1
Organelles	1	Tetrahymena pyriformis	1
Rats	1		
Rats, Sprague-Dawley	1		
Respiratory Mucosa	1		
Steroids	1		
analysis	1		
Trachea	1		

resentation: qualifiers are separated from descriptors, and the distinction between major and minor themes is dropped. Suppose that the offer weights are calculated for the concepts found in the two documents, and normalized with the highest score. A ranking of the concepts is presented in Table 5.8. Table 5.9 shows the resulting MeSH query if only five MeSH concepts are selected from the top of the ranking ($T = 5$). Note that the offer weights are used for selection only and not in the query representation.

Binary Representation. Here the documents retrieved are represented with the binary approach (Section 4.1.1). Consider two documents retrieved at the top of a ranking produced by a text search. The content of two such documents are shown in Table 5.7. If the top 2 documents only are assumed relevant ($R = 2$), as above, the content of both documents is extracted, scored with Equation 4.5, and ranked. Assume that the concept ranking returned is the one presented in Table 5.8. If we choose to keep the top 5 concepts only ($T = 5$), as above, then the concepts of the generated MeSH query will be those shown in Table 5.9.

Table 5.8: Example of Offer Weight ranking

concept	OW
Cilia	1.0
ultrastructure	0.96
Centrioles	0.94
Tetrahymena pyriformis	0.88
Fluorescent Antibody Technique	0.81
Animals	0.80
analysis	0.79
Protein Conformation	0.77
Steroids	0.75
Lymphocytes	0.73
Rats	0.72
Microscopy, Electron	0.72
cytology	0.70
Rats, Sprague-Dawley	0.69
Dynein ATPase	0.68
Organelles	0.66
Respiratory Mucosa	0.63
Trachea	0.59

Table 5.9: Example of a MeSH query obtained with PRF

concept	weight
Cilia	1
ultrastructure	1
Centrioles	1
Tetrahymena pyriformis	1
Fluorescent Antibody Technique	1

Table 5.10: TF*IDF representation of top 2 documents

document 1		document 2	
concept	weight	concept	weight
Animals	1.88	Animals	1.88
Centrioles	14.30	Cilia	11.78
ultrastructure	16.55	ultrastructure	22.07
Cilia	11.78	Dynein ATPase	12.32
Fluorescent Antibody Technique	8.32	Microscopy, Electron	6.93
Lymphocytes	7.99	Microtubules	9.55
cytology	8.65	Protein Conformation	6.72
Organelles	10.59	Tetrahymena pyriformis	14.02
Rats	3.72		
Rats, Sprague-Dawley	5.37		
Respiratory Mucosa	11.71		
Steroids	9.41		
analysis	3.69		
Trachea	9.25		

TF*IDF Representation. When documents are represented with TF*IDF, the weights of the concepts that appear several times in the documents assumed relevant need to be combined to produce the MeSH query. Consider the two documents with TF*IDF weights shown in Table 5.10. Given that only the top 2 documents are assumed relevant, that the concept ranking is shown in Table 5.8, and that only the top 5 concepts are selected, the weights of the query concepts c_q are calculated with the following formula:

$$\text{TF*IDF}_{c_q} = \frac{1}{R}\sum_{i} \text{TF*IDF}_{c_q,D_i}$$

where R is the number of documents assumed relevant, and $\text{TF*IDF}_{c_q,D_i}$ is the TF*IDF of concept c_q in document D_i. The TF*IDF weights of the selected concepts of the example are shown in Table 5.11.

MajMin Representation. The MajMin representation gives higher weights to MeSH concepts given as major themes of the document. Table 5.12 shows the content of two documents where the weight given to major theme concepts is 3.

Table 5.11: Example of TF*IDF representation of a MeSH query

concept	weight
Cilia	11.78
ultrastructure	19.31
Centrioles	7.15
Tetrahymena pyriformis	7.01
Fluorescent Antibody Technique	4.16

Suppose that the two documents are selected as relevant from a prior text search, and that only the top 5 concepts of the offer weight ranking is shown in Table 5.8 as the MeSH query. Then the MajMin weights of the query concepts c_q are calculated with the following formula:

$$\text{MajMin}_{c_q} = \frac{1}{R}\sum_i \text{MajMin}_{c_q,D_i}$$

where R is the number of documents assumed relevant, and MajMin_{c_q,D_i} is the MajMin weight of concept c_q in document D_i. The MajMin weights of the selected concepts of the example are shown in Table 5.13. We experiment with four MajMin representations, MajMin_2, 3, 4, and 5, corresponding to increasing weights given to major themes, as shown in Table 5.14.

DescQual Representation. In the DescQual representation, the associations found in MeSH fields between descriptors and qualifiers are kept as minimal tokens of information. Table 5.15 shows two documents with their DescQual representations. A zero value instead of a qualifier name indicates that the descriptor was not associated with any qualifier in the MeSH field. Assume two documents are selected as relevant from a prior text search. Their DescQual content is extracted, scored and ranked with their offer weight (Table 5.16). If only the top 5 associations are selected, the generated MeSH query is the one shown in Table 5.17.

Table 5.12: MajMin representation of top 2 documents

document 1		document 2	
concept	weight	concept	weight
Animals	1	Animals	1
Centrioles	3	Cilia	3
ultrastructure	3	ultrastructure	3
Cilia	1	Dynein ATPase	3
Fluorescent Antibody Technique	1	Microscopy, Electron	1
Lymphocytes	3	Microtubules	1
cytology	3	Protein Conformation	1
Organelles	3	Tetrahymena pyriformis	3
Rats	1		
Rats, Sprague-Dawley	1		
Respiratory Mucosa	1		
Steroids	3		
analysis	3		
Trachea	1		

Table 5.13: Example of MajMin representation of a MeSH query

concept	weight
Cilia	2
ultrastructure	3
Centrioles	1.5
Tetrahymena pyriformis	1.5
Fluorescent Antibody Technique	0.5

Table 5.14: The four MajMin representations

	weights	
	major themes	minor themes
MajMin_2	2	1
MajMin_3	3	1
MajMin_4	4	1
MajMin_5	5	1

Table 5.15: DescQual representation of top 2 documents

document 1		document 2	
concept	weight	concept	weight
Animals/0	1	Animals/0	1
Centrioles/ultrastructure	1	Cilia/ultrastructure	1
Cilia/ultrastructure	1	Dynein ATPase/ultrastructure	1
Fluorescent Antibody Technique/0	1	Microscopy, Electron/0	1
Lymphocytes/cytology	1	Microtubules/ultrastructure	1
Organelles/ultrastructure	1	Protein Conformation/0	1
Rats/0	1	Tetrahymena pyriformis/ultrastructure	1
Rats, Sprague-Dawley/0	1		
Respiratory Mucosa/cytology	1		
Steroids/analysis	1		
Trachea/0	1		

Table 5.16: Example of Offer Weight ranking with DescQual representation

rank	concept
1	Centrioles/ultrastructure
2	Cilia/ultrastructure
3	Fluorescent Antibody Technique/0
4	Tetrahymena pyriformis/ultrastructure
5	Lymphocytes/cytology
6	Organelles/ultrastructure
7	Respiratory Mucosa/cytology
8	Trachea/0
9	Dynein ATPase/ultrastructure
10	Microscopy, Electron/0
11	Microtubules/ultrastructure
12	Protein Conformation/0
13	Steroids/analysis
14	Rats/0
15	Rats, Sprague-Dawley/0
16	Animals/0

Table 5.17: Example of DescQual representation of a MeSH query

concept	weight
Centrioles/ultrastructure	1
Cilia/ultrastructure	1
Fluorescent Antibody Technique/0	1
Tetrahymena pyriformis/ultrastructure	1
Lymphocytes/cytology	1

5.1.3 Fusion Method

Our fusion method is strongly inspired by Srinivasan (1996b)'s. For each of the original text queries (*basic narratives*), we use the Terrier search engine with the BB2 model ($c = 5$) to retrieve 5000 documents. The documents are re-scored with the Cosine similarity measure (see Equation 2.10) between their MeSH vectors and the MeSH query vectors generated as described above. The scores obtained by the documents in the initial text ranking and the newly created MeSH rankings are combined with Equation 3.8 with $0 \leq \alpha \leq 1$, and $\beta = 1 - \alpha$. All documents are then re-ranked according to their new combined score, and the top 1000 are kept.

Parameter Tuning

Parameters R (number of documents assumed relevant at the top of the initial text ranking), T (number of MeSH terms kept for the MeSH queries from the top of the offer weight ranking), and α ($0 \leq \alpha \leq 1$), determine the relative importance given to text and MeSH document scores. These were tuned to optimize the MAP of the combined searches. In particular, we varied R from 5 to 20 (increments of 5), T from 5 to 30 (increments of 5), and α from 0.05 to 0.95 (increments of 0.05). The MeSH binary representation is used for the tuning. The topics of TrecGen05 are organized in 50 instances of 5 distinct GTTs (10 instances of each GTT, see Section 3.2.1). The first five instances of each GTT (25 queries) are chosen to determine the combination of R, T, and α values giving the best MAP. This combination is then tested on the remaining 5 instances of each GTTs (24 queries as topic 135 has no relevant documents).

Tables 5.18 and 5.19 show the improvement over the text searches for the training and test queries, respectively, in terms of MAP and average recall. The best improvement of MAP over the text-only searches with the training queries is obtained with R=15, T=15, and $\alpha = 0.70$. The improvement of the MAP is higher on the test queries (+11.6%) than on the training queries (+7%), which suggests that no overtraining is occurring. Interestingly, using the test queries to determine the best

Table 5.18: Parameter tuning for optimal improvement of MAP

	MAP	average recall
Text-only	0.3035	0.7140
Text and MeSH fusion with R=15 T=15 $\alpha = 0.70$	0.3247 (+7%)	0.7610

Table 5.19: Improvement over text-only baseline with test queries and tuned parameters

	MAP	average recall
Text-only	0.2408	0.0.6554
Text and MeSH fusion with R=15 T=15 $\alpha = 0.70$	0.2687 (+11.6%)	0.6818

combination of values gave the same result, i.e. R=15, T=15, and $\alpha = 0.70$. This suggests that this combination of values is consistent across topics for this method. In future work we will evaluate this combination on other topics (such as the TREC 2004 ad hoc topics).

The values R=15 and T=15 are used to generate 50 MeSH queries for all the different MeSH representations described above. The value $\alpha = 0.70$ is used to combine text and MeSH document scores with Equation 3.8.

5.1.4 Non-hierarchical Hypotheses: Results and Analysis

MeSH Searches

Table 5.20 shows the results of the different non-hierarchical approaches (binary, TF*IDF, MajMin, DescQual) for the MeSH-only searches. The best average P@10 (0.3388) (precision at ten documents retrieved) is obtained with the binary representation. The best MAP (0.1761) and the best average recall (0.6305) is obtained with the MajMin_2 representation. Tables B.9, B.10, and B.11 show the statistically significant differences between the representations in terms of average P@10, MAP, and average recall, respectively. The binary representation gets the best average P@10 with statistical significance compared to representations TFIDF, MajMin_4 and MajMin_5 only. In terms of MAP, MajMin_2 shows statistically significant differences with all other representations but binary and DescQual. Statistical significance is less pronounced for average recall as the best representation for recall, MajMin_2, only gives a significant difference with representations MajMin_4 and MajMin_5. The results over the 49 queries are contrary to our expectations: the representations integrating MeSH annotations (MajMin representations and DescQual), as well as corpus frequency information (TFIDF), do not increase the precision of the MeSH searches over the trivial binary representation. Overall, the DescQual representation is less damaging for precision than the MajMin representations, where precision decreases as the distinction weight for major themes is increased. Next, we look at precision results (P@10) for each query.

Table 5.21 shows the P@10 for all representations and all topics. The results that improve on the binary representation are in bold. TFIDF improves precision over the binary representation for 9 queries (104, 112, 114, 117, 119, 120, 142, 147, 148) and decrease precision for 20 queries (100, 105-108, 113, 116, 118, 121, 122, 129, 130-132, 134, 137-139, 141, 145). MajMin_2 improves precision over the binary representation for 11 queries (105, 112, 114, 117, 119, 120, 126, 136, 140, 142, 148) and decrease precision for 21 queries (101, 103, 106-109, 113, 116, 118 122, 124, 127, 130, 131, 134, 137-139, 141, 145, 146). MajMin_3 improves precision over the binary representation for 12 queries (105, 112, 114, 117, 119, 120, 121, 126, 136, 140, 142, 148) and decrease precision for 24 queries (101, 103, 106-109, 113, 116, 118, 122-124, 127, 128, 130-132, 134, 137-139, 141, 145, 146). MajMin_4 improves precision over the binary representation for 12 queries (100, 112, 114, 117, 119, 120, 121, 126, 136, 140, 142, 148) and decrease precision for 25 queries (101, 103, 106-109, 113, 116, 122-124, 127-132, 134, 137-139, 141, 145, 146). MajMin_5 improves precision over the binary representation for 12 queries (100, 112, 114, 117, 119, 120, 121, 126, 136, 140, 142, 148) and decrease precision for 26 queries (101, 103, 105-109, 113, 116, 118, 122-124, 127-132, 134, 137-139, 141, 145, 146). Finally, DescQual improves precision over the binary representation for 12 queries (100, 105, 114, 116-120, 127, 130, 140, 142) and decrease precision for 16 queries (106, 108, 109, 112, 122, 124, 128, 131, 132, 134, 137-139, 141, 145, 146). Each representation decreases p@10 for more queries than it increases it, compared to the binary representation. The largest difference is obtained by the MajMin_5 representation that increases p@10 for 12 queries and decreases it for 26 queries.

The results obtained for p@10 are counter-intuitive. By differentiating concepts with weights based on corpus statistics or MeSH field information, or by imposing specific association between concepts, we expected to improve precision.

A possible explanation can be found in the precision of the text searches at 15 documents retrieved. Indeed the binary MeSH queries are extracted from the 15 top documents ranked according to the free-text search of the text index. If p@15 is low, i.e. we have few relevant documents in the top 15, the MeSH concepts extracted from the top 15 may contain substantial noise. Consequently, introducing more specificity, such as specific weighting and concept associations, may damage the precision by giving importance to noisy concepts or associations.

Table 5.22 shows the Pearson correlation coefficients between the p@15 of the text searches and the p@10 of the MeSH searches for the different MeSH representations. The p@10 results for MeSH searches are well correlated with the p@15 of the text searches.

The precision of MeSH searches for all MeSH representations is linked to the precision of text searches but is the impact of MeSH representations over the MeSH binary representation also linked to the precision of the text searches? Focusing on the MajMin and DescQual representations, Table 5.23 shows the p@10 results for the MeSH searches sorted relatively to the p@15 of the text searches. The top 21 queries in the table ($0 \leq$ text p@15 < 0.33) correspond to the case where the precision of the binary representation is low presumably because of the noisy concepts introduced by pseudo-relevance feedback. Therefore, the MajMin and DescQual representations can not improve on the noisy binary representation (18 queries out of 21 are not improved). The 15 following queries ($0.33 \leq$ text p@15 < 0.66) correspond to the intermediate case where the binary representation is less noisy but noisy enough for

Table 5.20: Results for MeSH-only searches before fusion

	binary	TF*IDF	MajMin_2	MajMin_3	MajMin_4	MajMin_5	DescQual
Av. P@10	**0.3388**	0.2939	0.3102	0.2878	0.2816	0.2714	0.3102
MAP	0.1639	0.1439	**0.1761**	0.1684	0.1585	0.1507	0.1550
Av. recall	0.6256	0.6225	**0.6305**	0.6235	0.6187	0.6146	0.6158

MajMin and DescQual representations to use the wrong concepts as central themes, and the wrong associations, respectively. For 9 queries out of 15, at least 4 of the 5 MeSH representations (MajMins and DescQual) decrease precision. In the bottom part of the table, we have 13 queries for which text p@15 is greater or equal to 0.66. For these queries, we expected the binary MeSH queries generated by pseudo-relevance feedback to contain few noisy concepts. Moreover, we expect the concepts chosen as central themes in MajMin representations, and the associations generated in the DescQual representation to correspond to central themes and associations contained in relevant documents, respectively. Indeed, for 5 queries out of the 13, all 5 MeSH representations (MajMins and DescQual) increase precision over the binary representation. However, for 7 other queries, at least 4 of the 5 MeSH representations (MajMins and DescQual) decrease precision.

If the p@15 of the text searches is linked to the performance of MeSH queries generated by pseudo-relevance feedback, high levels of p@15 do not guarantee positive impact on precision for our MeSH representation policies. Looking at MajMin and DescQual representations separately, we see that it is especially the case for the MajMin representations. For 7 queries out of the bottom 13 of Table 5.23 (text p@15 $\geq$ 0.66), all MajMin representations decrease the p@10 of the binary representation. In contrast, the DescQual representation increase p@10 for 7 queries and decrease p@10 for only 4 out of the 13 queries with high p@15 levels.

To investigate the limits of the impact of text p@15 on MeSH queries, we generate MeSH queries using the document judged relevant (see Table A.4) instead of using pseudo-relevance feedback on the text searches. MeSH query generation is done by selecting the top 15 concepts according to their Offer Weight score (Equation 4.5).

Table 5.21: MeSH search P@10 results by topic for non-hierarchical representations (improvement over binary in bold)

topic	binary	TF*IDF	MajMin_2	MajMin_3	MajMin_4	MajMin_5	DescQual
100	0.4	0.2	0.4	0.4	**0.5**	**0.5**	**0.5**
101	0.1	0.1	0	0	0	0	0.1
102	0	0	0	0	0	0	0
103	0.1	0.1	0	0	0	0	0.1
104	0.1	**0.2**	0.1	0.1	0.1	0.1	0.1
105	0.6	0.5	**0.7**	**0.7**	0.6	0.5	**0.7**
106	0.4	0.2	0.1	0	0	0	0.2
107	0.3	0.2	0	0	0	0.1	0.3
108	0.8	0.5	0.4	0.5	0.5	0.5	0.5
109	0.9	0.9	0.8	0.7	0.7	0.7	0.8
110	0	0	0	0	0	0	0
111	0	0	0	0	0	0	0
112	0.1	**0.4**	**0.3**	**0.3**	**0.3**	**0.3**	0
113	0.6	0.5	0.5	0.5	0.5	0.5	0.6
114	0.7	**0.8**	**0.8**	**0.9**	**0.9**	**0.8**	**0.9**
115	0	0	0	0	0	0	0
116	0.4	0.2	0.3	0.3	0.3	0.3	**0.5**
117	0.7	**0.8**	**0.9**	**0.8**	**0.8**	**0.8**	**1**
118	0.6	0.5	0.5	0.4	0.3	0.2	**0.8**
119	0.3	**0.5**	**1**	**1**	**1**	**1**	**0.5**
120	0.9	**1**	**1**	**1**	**1**	**1**	**1**
121	0.6	0.2	0.6	**0.8**	**0.7**	**0.7**	0.6
122	0.6	0.5	0.5	0.3	0.3	0.2	0.2
123	0.3	0	0.3	0.1	0.1	0.1	0.3
124	0.9	0.9	0.7	0.6	0.4	0.4	0.5
125	0	0	0	0	0	0	0
126	0.3	0.3	**0.7**	**0.7**	**0.7**	**0.7**	0.3
127	0.1	0.1	0	0	0	0	**0.2**
128	0.3	0.3	0.3	0.2	0.1	0.1	0.1
129	0.2	0.1	0.2	0.2	0.1	0.1	0.2
130	0.5	0.4	0.4	0.4	0.4	0.3	**0.6**
131	0.8	0.6	0.6	0.4	0.4	0.4	0.6
132	0.3	0.2	0.3	0.2	0.2	0.2	0
133	0	0	0	0	0	0	0
134	0.2	0.1	0.1	0.1	0	0	0.1
136	0	0	**0.1**	**0.1**	**0.1**	**0.1**	0
137	0.3	0.2	0	0	0	0	0
138	0.3	0.2	0.1	0.1	0.1	0.1	0.2
139	0.6	0.3	0.3	0.3	0.3	0.3	0.4
140	0	0	**0.1**	**0.2**	**0.3**	**0.2**	**0.1**
141	0.3	0.2	0.2	0.1	0.1	0.1	0.2
142	0.5	**0.7**	**0.7**	**0.6**	**0.7**	**0.8**	**0.7**
143	0	0	0	0	0	0	0
144	0.1	0.1	0.1	0.1	0.1	0.1	0.1
145	0.7	0.5	0.5	0.4	0.4	0.3	0.6
146	0.7	0.7	0.5	0.5	0.6	0.6	0.6
147	0	**0.1**	0	0	0	0	0
148	0	**0.1**	**0.1**	**0.1**	**0.2**	**0.2**	0
149	0	0	0	0	0	0	0
Average	**0.3388**	0.2939	0.3102	0.2878	0.2816	0.2714	0.3102

Table 5.22: Pearson correlation coefficients between text searches p@15 and MeSH searches p@10

Representation	$r_x y$
binary	0.77
TF*IDF	0.76
MajMin_2	0.79
MajMin_3	0.79
MajMin_4	0.80
MajMin_5	0.79
DescQual	0.82

This simulates a optimal situation where pseudo-relevance feedback is only using relevant documents (P@15=1 for all queries).

Table 5.24 shows the results for all MeSH representations. The best average P@10 and MAP are obtained with the DescQual representations. The best average recall is obtained by the MajMin_2 representation. Tables B.12, B.13, and B.14 show the statistically significant differences between the representations in terms of average P@10, MAP, and average recall, respectively. The improvement of DescQual over binary in terms of average P@10 is not statistically significant, but it is in terms of MAP. Furthermore, the improvement of MajMin_2 over binary in terms of average recall is not statistically significant.

Table 5.25 focuses on the average P@10 for each query and each MeSH representations. TF*IDF increases average P@10 over binary for 12 queries (103, 109, 111, 112, 114, 117, 119, 120, 124, 128, 133, 142) and decreases P@10 for 23 queries (100, 102, 104, 105, 106, 107, 108, 115, 116, 121, 123, 126, 127, 130, 131, 137, 138, 139, 140, 141, 145, 147, 149). MajMin_2 increases average P@10 over binary for 13 queries (100, 111, 117, 119, 121, 122, 123, 127, 128, 131, 137, 139, 146) and decreases P@10 for 22 queries (101, 102, 104, 106, 107, 108, 112, 116, 118, 120, 124, 129, 130, 132, 133, 134, 138, 141, 142, 145, 147, 149). MajMin_3 increases average P@10 over binary for 13 queries (105, 111, 117, 119, 121, 122, 126, 127, 128, 131, 137, 139, 140) and decreases P@10 for 22 queries (101, 104, 106, 107, 108, 109, 110,

Table 5.23: MeSH search P@10 results by topic for non-hierarchical representations compared to the p@15 of text searches

Topics	Text p@15	binary	MajMin2	MajMin3	MajMin4	MajMin5	DescQual
102	0	0	0	0	0	0	0
110	0	0	0	0	0	0	0
115	0	0	0	0	0	0	0
125	0	0	0	0	0	0	0
136	0	0	0.1	0.1	0.1	0.1	0
143	0	0	0	0	0	0	0
147	0	0	0	0	0	0	0
148	0	0	0.1	0.1	0.2	0.2	0
101	0.0667	0.1	0	0	0	0	0.1
104	0.0667	0.1	0.1	0.1	0.1	0.1	0.1
133	0.0667	0	0	0	0	0	0
144	0.0667	0.1	0.1	0.1	0.1	0.1	0.1
149	0.0667	0	0	0	0	0	0
103	0.1333	0.1	0	0	0	0	0.1
127	0.1333	0.1	0	0	0	0	0.2
137	0.1333	0.3	0	0	0	0	0
111	0.2	0	0	0	0	0	0
123	0.2	0.3	0.3	0.1	0.1	0.1	0.3
134	0.2	0.2	0.1	0.1	0	0	0.1
138	0.2	0.3	0.1	0.1	0.1	0.1	0.2
129	0.2667	0.2	0.2	0.2	0.1	0.1	0.2
100	0.3333	0.4	0.4	0.4	0.5	0.5	0.5
106	0.3333	0.4	0.1	0	0	0	0.2
112	0.3333	0.1	0.3	0.3	0.3	0.3	0
128	0.3333	0.3	0.3	0.2	0.1	0.1	0.1
132	0.3333	0.3	0.3	0.2	0.2	0.2	0
141	0.3333	0.3	0.2	0.1	0.1	0.1	0.2
105	0.4	0.6	0.7	0.7	0.6	0.5	0.7
116	0.4	0.4	0.3	0.3	0.3	0.3	0.5
124	0.4	0.9	0.7	0.6	0.4	0.4	0.5
126	0.5333	0.3	0.7	0.7	0.7	0.7	0.3
145	0.5333	0.7	0.5	0.4	0.4	0.3	0.6
108	0.6	0.8	0.4	0.5	0.5	0.5	0.5
113	0.6	0.6	0.5	0.5	0.5	0.5	0.6
131	0.6	0.8	0.6	0.4	0.4	0.4	0.6
140	0.6	0	0.1	0.2	0.3	0.2	0.1
122	0.6667	0.6	0.5	0.3	0.3	0.2	0.2
139	0.6667	0.6	0.3	0.3	0.3	0.3	0.4
118	0.7333	0.6	0.5	0.4	0.3	0.2	0.8
142	0.8	0.5	0.7	0.6	0.7	0.8	0.7
117	0.8667	0.7	0.9	0.8	0.8	0.8	1
109	0.9333	0.9	0.8	0.7	0.7	0.7	0.8
114	0.9333	0.7	0.8	0.9	0.9	0.8	0.9
121	0.9333	0.6	0.6	0.8	0.7	0.7	0.6
146	0.9333	0.7	0.5	0.5	0.6	0.6	0.6
107	1	0.3	0	0	0	0.1	0.3
119	1	0.3	1	1	1	1	0.5
120	1	0.9	1	1	1	1	1
130	1	0.5	0.4	0.4	0.4	0.3	0.6
mean	0.4068	0.3388	0.3102	0.2878	0.2816	0.2714	0.3102

112, 116, 120, 124, 130, 132, 133, 134, 138, 141, 142, 145, 146, 147, 149). MajMin_4 increases average P@10 over binary for 11 queries (100, 111, 117, 119, 121, 126, 127, 128, 131, 137, 139) and decreases P@10 for 25 queries (101, 104, 106, 107, 108, 109, 110, 112, 114, 116, 118, 124, 130, 132, 133, 134, 136, 138, 140, 141, 143, 145, 146, 147, 149). MajMin_5 increases average P@10 over binary for 10 queries (100, 111, 113, 117, 119, 121, 127, 128, 137, 139) and decreases P@10 for 26 queries (101, 104, 105, 106, 107, 108, 109, 110, 112, 116, 118, 124, 129, 130, 133, 134, 136, 138, 140, 141, 142, 143, 145, 146, 147, 149). DescQual increases average P@10 over binary for 16 queries (105, 109, 111, 117, 119, 120, 121, 122, 123, 128, 129, 133, 136, 142, 146, 147) and decreases P@10 for 14 queries (102, 107, 108, 113, 114, 116, 126, 130, 132, 138, 139, 141, 145, 149).

The results show the difficulty of generating useful MeSH queries integrating collection concept frequency and MeSH annotation information even when documents manually judged relevant are used. In particular, the TF*IDF and MajMin representations have a negative impact on P@10 for more than 20 queries. DescQual, in contrast, increases P@10 over binary for more queries than it decreases it (16 against 14). The precision decrease obtained with TF*IDF suggests that collection document frequencies do not help to distinguish the important concepts of the relevant documents. Regarding the precision decreases obtained with the MajMin representations, the difficulty could come from the consistency with which documents relevant to the same query are annotated with similar MeSH concepts as their major themes. In this case, the results suggest that the associations used across relevant documents are more consistent than the concepts chosen as major themes. A full consistency analysis, in the line of the work done by Funk et al. (1983) needs to be done.

Text and MeSH Fusion

Table 5.26 shows the result in terms of average P@10, MAP and average recall for the MeSH representations after MeSH search outputs are combined with text

Table 5.24: Results for MeSH-only representations with MeSH queries generated with document judged relevant

	binary	TF*IDF	MajMin_2	MajMin_3	MajMin_4	MajMin_5	DescQual
Av. P@10	0.3857	0.3531	0.3673	0.3571	0.3347	0.3286	**0.3918**
MAP	0.2416	0.2208	0.2438	0.2257	0.2123	0.2054	**0.2621**
Av. recall	0.7112	0.6889	**0.7179**	0.7072	0.6966	0.6966	0.7015

search outputs. The text and MeSH binary combination (text+binary) gives the best average P@10 (0.4776), the best MAP (0.2973), and the best average recall (0.7199). All combinations of text and MeSH representations are also compared with the use of text alone ("text" in Table 5.26). Tables B.15, B.16, and B.17 show the statistically significant differences between the combinations in terms of average P@10, MAP, and average recall, respectively. The difference between the best combination for average P@10, text+binary, and the other combinations is statistically significant, except for text+TF*IDF and text+MajMin_4. Also text+binary and text+TF*IDF are the only combinations for which average P@10 is greater than text alone with statistical significance. The MAPs of combinations text+binary and text+MajMin_2 are the only ones greater than the MAP of text alone, and the MAPs of all other combinations, with statistical significance. 4 combinations (text+binary, text+TF*IDF, text+MajMin_2, text+DescQual) have average recall greater than text alone with statistical significance. The best average recall, obtained by text+binary, is greater to all other combinations but text+TF*IDF with statistical significance.

Overall, we observe that the text+MeSH combinations significantly improve on the use of text alone in terms of average P@10 (text+binary, text+TF*IDF), MAP (text+binary, text+MajMin_2), and average recall (text+binary, text+TF*IDF, text+MajMin_2, text+DescQual).

The comparison of the MeSH representations after combination with text reflects the observation made in the previous section: the combination text+binary gives the best results, similarly to binary on its own , followed by text+MajMin_2.

Table 5.25: Impact of MeSH representation policies (TF*IDF, MajMin, DescQual) in terms of P@10 when documents judged relevant are used to generate MeSH queries (improvement in bold)

Topics	binary	TF*IDF	MajMin2	MajMin3	MajMin4	MajMin5	DescQual
100	0.3	0.2	**0.4**	0.3	**0.4**	**0.5**	0.3
101	0.1	0.1	0	0	0	0	0.1
102	0.3	0.2	0.2	0.3	0.3	0.3	0.2
103	0	**0.1**	0	0	0	0	0
104	0.2	0.1	0.1	0.1	0.1	0.1	0.2
105	0.6	0.5	0.6	**0.7**	0.6	0.5	**0.7**
106	0.3	0.2	0.2	0.2	0.1	0.1	0.3
107	0.5	0.1	0	0	0	0	0.4
108	0.9	0.7	0.5	0.5	0.5	0.5	0.8
109	0.8	**0.9**	0.8	0.7	0.7	0.6	**0.9**
110	0.2	0.2	0.2	0	0	0	0.2
111	0.7	**1**	**0.9**	**0.9**	**0.9**	**0.9**	**0.8**
112	0.3	**0.4**	0.2	0.1	0.1	0.1	0.3
113	0.6	0.6	0.6	0.6	0.6	**0.7**	0.5
114	0.5	**0.8**	0.5	0.5	0.4	0.5	0.4
115	0.1	0	0.1	0.1	0.1	0.1	0.1
116	0.6	0.2	0.4	0.2	0.2	0.2	0.5
117	0.7	**0.8**	**0.9**	**0.8**	**0.8**	**0.8**	**0.8**
118	0.7	0.7	0.6	0.7	0.6	0.6	0.7
119	0.2	**0.5**	**1**	**1**	**1**	**1**	**0.6**
120	0.9	**1**	0.7	0.8	0.9	0.9	**1**
121	0.3	0.2	**0.6**	**0.8**	**0.7**	**0.7**	**0.7**
122	0.5	0.5	**0.7**	**0.6**	0.5	0.5	**0.6**
123	0.1	0	**0.2**	0.1	0.1	0.1	**0.3**
124	0.8	**1**	0.7	0.7	0.4	0.4	0.8
125	0	0	0	0	0	0	0
126	0.4	0.3	0.4	**0.5**	**0.5**	0.4	0.1
127	0.3	0.2	**0.4**	**0.4**	**0.4**	**0.4**	0.3
128	0	**0.2**	**0.3**	**0.3**	**0.3**	**0.3**	**0.2**
129	0.2	0.2	0.1	0.2	0.2	0.1	**0.3**
130	0.7	0.3	0.6	0.5	0.5	0.4	0.4
131	0.6	0.5	**0.7**	**0.7**	**0.7**	0.6	0.6
132	0.3	0.3	0.2	0.2	0.2	0.3	0.1
133	0.2	**0.3**	0.1	0.1	0.1	0.1	**0.3**
134	0.5	0.5	0.3	0.2	0.1	0.1	0.5
136	0.1	0.1	0.1	0.1	0	0	**0.2**
137	0.1	0	**0.2**	**0.2**	**0.2**	**0.2**	0.1
138	0.5	0.3	0.3	0.3	0.3	0.3	0.4
139	0.5	0.4	**0.8**	**0.7**	**0.7**	**0.7**	0.4
140	0.2	0.1	0.2	**0.3**	0.1	0.1	0.2
141	0.4	0	0.1	0.1	0.1	0	0.3
142	0.3	**0.7**	0.2	0.2	0.3	0.2	**0.4**
143	0.2	0.2	0.2	0.2	0.1	0.1	0.2
144	0.1	0.1	0.1	0.1	0.1	0.1	0.1
145	0.7	0.5	0.4	0.5	0.5	0.6	0.4
146	0.5	0.5	**0.7**	0.4	0.4	0.4	**0.6**
147	0.1	0	0	0	0	0	**0.2**
148	0.2	0.2	0.2	0.2	0.2	0.2	0.2
149	0.6	0.4	0.3	0.4	0.4	0.4	0.5
mean	0.3857	0.3531	0.3673	0.3571	0.3347	0.3286	0.3918

Table 5.26: Results of MeSH representations after MeSH and text searches combination

	text	text+ binary	text+ TF*IDF	text+ MajMin_2	text+ MajMin_3	text+ MajMin_4	text+ MajMin_5	text+ DescQual
Av. P@10	0.4306	**0.4776**	0.4571	0.4551	0.4531	0.4490	0.4551	0.4388
MAP	0.2728	**0.2973**	0.2813	0.2916	0.2862	0.2806	0.2774	0.2828
Av. recall	0.6823	**0.7199**	0.7099	0.7109	0.7055	0.7049	0.7010	0.7104

However, the fusion with text does not benefit the MeSH representations in the same way. For example, the average P@10 of DescQual is the same as the average P@10 of MajMin_2 (0.3102), but the average P@10 of text+DescQual (0.4388) is less than the average P@10 of text+MajMin_2 (0.4551). Similarly, the MAP of MajMin_2 (0.1761) is greater than the MAP of binary (0.1639), but the MAP of text+MajMin_2 (0.2916) is less than the MAP of text+binary (0.2973).

Focusing on binary and MajMin_2 before and after combination with text, Table 5.27 presents the MAP for each query before and after combination. The binary representation increases the average precision for 31 of the 49 topics, and MajMin_2 increases the average precision for 32 topics. Both representations benefit the performance of the same 29 topics.

A higher performance for one strategy before fusion does not always translate into a higher performance for the same strategy after fusion. For some topics (105, 114, 140, 141, 148), binary has a lower MAP than MajMin_2 but text+binary has a higher MAP than text+MajMin_2. Inversely, for other topics (101, 104, 113, 123, 129, 134), binary has a higher MAP than MajMin_2 but text+binary has a lower MAP than text+MajMin_2.

Fusion effectiveness has been associated in the past with Lee's overlap hypothesis (Lee, 1997). The hypothesis asserts that the overlap of relevant documents in the result sets must be greater than the overlap of non-relevant documents for fusion to be effective. However, recent work by Beitzel et al. (2004) contradicts the overlap hypothesis and suggests new conditions for fusion effectiveness. This study shows that fusion has limited effect if the rankings being merged already have high overlap

Table 5.27: binary and MajMin_2 representations before and after fusion with text (best result in bold)

Topics	Text	binary	majMin2	text+binary	text+majMin2
100	0.1867	0.2069	**0.2386**	0.2616	**0.3048**
101	0.0705	**0.0315**	0.0244	0.0506	**0.0601**
102	0.0114	0.0016	**0.0026**	0.0107	**0.0147**
103	0.0684	**0.0242**	0.0081	**0.0616**	0.0603
104	0.101	**0.3067**	0.304	0.2982	**0.3035**
105	0.1326	0.1943	**0.2156**	**0.2407**	0.2347
106	0.0451	**0.0513**	0.0249	**0.0735**	0.0661
107	0.501	**0.206**	0.1342	**0.5672**	0.5453
108	0.1347	**0.1647**	0.1228	**0.2397**	0.2039
109	0.6683	**0.5911**	0.5624	**0.8315**	0.8246
110	0.0035	0.0001	**0.0002**	0.0029	0.0029
111	0.1959	**0.0603**	0.0577	**0.1958**	0.18
112	0.3985	0.084	**0.2623**	0.3766	**0.4077**
113	0.5404	**0.4858**	0.4836	0.5667	**0.582**
114	0.4832	0.2527	**0.2635**	**0.5185**	0.5101
115	0.0002	0.0035	**0.005**	0.0015	**0.0033**
116	0.1706	**0.0901**	0.0817	**0.2083**	0.1792
117	0.5203	0.264	**0.3908**	0.461	**0.5186**
118	0.3827	**0.2831**	0.2023	**0.3555**	0.3262
119	0.7468	0.1696	**0.5098**	0.6842	**0.7339**
120	0.7858	0.5102	**0.6484**	0.7952	**0.7982**
121	0.8029	0.2228	**0.3618**	0.8052	**0.8097**
122	0.1919	**0.1836**	0.161	**0.2383**	0.2233
123	0.0276	**0.1119**	0.0499	0.0908	**0.0986**
124	0.1429	**0.2601**	0.2454	**0.2314**	0.2243
125	0	0	0	0	0
126	0.241	0.0458	**0.1306**	0.1276	**0.2251**
127	0.1659	**0.2656**	0.0219	**0.3305**	0.166
128	0.151	0.0462	**0.0539**	0.158	**0.1732**
129	0.1135	**0.0568**	0.0543	0.1575	**0.1726**
130	0.7609	**0.2086**	0.1232	**0.6792**	0.651
131	0.6061	**0.3619**	0.3002	**0.7512**	0.6495
132	0.1462	**0.1659**	0.1201	**0.2324**	0.2102
133	0.034	**0.0026**	0.0024	**0.0197**	0.0182
134	0.217	**0.14**	0.033	0.2395	**0.2459**
136	0.0015	0.0095	**0.3333**	0.0012	**0.0023**
137	0.0265	**0.0353**	0.0154	**0.0534**	0.0311
138	0.1982	**0.1699**	0.1167	**0.2297**	0.2113
139	0.4317	**0.3816**	0.2182	**0.4915**	0.4027
140	0.4475	0.0367	**0.0463**	**0.3041**	0.3026
141	0.3138	0.1562	**0.1588**	**0.3261**	0.303
142	0.5503	0.1834	**0.2338**	0.5601	**0.5719**
143	0.0009	0.0015	**0.0056**	0.0014	**0.003**
144	0.5	0.25	**0.5**	0.5	0.5
145	0.4044	**0.4049**	0.3337	**0.4847**	0.4519
146	0.6495	0.304	**0.4081**	0.6707	**0.6965**
147	0.0093	0.0072	**0.0085**	0.0066	**0.0075**
148	0.0565	0.0324	**0.0454**	**0.0576**	0.0566
149	0.0286	0.0027	**0.0062**	0.0162	**0.0207**
mean	0.2728	0.1639	**0.1761**	**0.2973**	0.2916

and high quality. Fusion is more likely to be effective if the rankings have minimal agreement (they contain different relevant documents) and if fusion maintains the relevant documents at a high rank. In future work we will examine these hypotheses against the effectiveness of fusing text and MeSH rankings.

Conclusion

In Chapter 4 we formulated hypotheses stating that:

1. corpus information about MeSH concepts (TF*IDF),
2. distinction between major and minor theme concepts (MajMin), and
3. higher contextual information about MeSH concepts (DescQual)

would improve the precision of MeSH-based MEDLINE retrieval. The results obtained of MeSH-only searches, and text and MeSH searches combination over 49 topics contradict our previous intuitions and do not therefore provide satisfactory support for the non-hierarchical hypotheses. However, we showed that the search performance of MeSH representations depended on the quality of the information available to generate the MeSH queries (P@15 of the text search outputs from which MeSH concepts are extracted by pseudo-relevance feedback). After the noise was reduced in the MeSH queries by generating them from documents known to be relevant, the DescQual representation was shown to have a positive impact on P@10. However, TF*IDF and MajMin representations failed to impact positively on P@10, even after the use of relevant documents for query generation. This suggested that concept frequencies in the collection were not useful to discriminate between concepts inside documents. Furthermore, we hypothesized that the failings of the MajMin representations could be explained by the lack of consistency across documents relevant to a query regarding the choice of major themes by annotators. In contrast, the success of DescQual could be explained by the consistency of descriptor/qualifier association across relevant documents. We plan to study annotation consistency in documents relevant to similar topics in future work.

In the next section we evaluate hierarchical hypotheses with the integration of hierarchy information during the comparison of queries with documents. As the binary representation is the best-performing of all the representations assessed previously, hierarchy information is used while comparing binary representations of documents and queries.

5.2 Hierarchy Integration

In this section we evaluate hierarchical hypotheses by integrating hierarchy information while comparing MeSH binary representation of documents with MeSH binary representations of queries. Our main hypothesis is that integrating hierarchy information will benefit the recall of MeSH-based MEDLINE retrieval.

5.2.1 Inter-concept Similarities Combination

Documents and queries generally contain several MeSH concepts. Therefore, comparing documents with queries involves the combination of similarities of pairs of concepts. In this section, we evaluate the hypothesis stating that the best-match-combination approach will perform better than the all-combination approach for the comparison of documents with queries (see Section 4.2.2 for intuition).

We evaluate both approaches using $dist_{rada1}$ (Equation 2.1) to compare pairs of concepts. The combined MeSH hierarchy (with added artificial nodes "*qualifier*" and "*MeSH*") is used to compare concepts from all parts of the hierarchy (see Section 4.2.2).

All-combination and best-match-combination differ in the way they combine the inter-concept distances. All-combination uses $dist_{rada2}$ (Equation 2.9), and best-match-combination uses $dist_{azu}$ (Equation 2.14, used with $dist_{rada1}$ instead of $dist_{jiang4}$). Table 5.28 describes the experimental set-up.

Intuitively, the closer a document is to a query, the better its rank. To combine the MeSH scores (distances) with text scores (similarities), semantic distances be-

Table 5.28: Experimental set-up for comparing all-combination with best-match-combination

	inter-concept measure	hierarchy separation	inter-concept measure combination
all-combination	$dist_{rada1}$	none	$dist_{rada2}$
best-match-combination	$dist_{rada1}$	none	$dist_{azu}$

tween documents and queries are turned into similarity measures with the following Equation:

$$sim\left(Q,D\right)=1-dist\left(Q,D\right)/max\_dist \tag{5.2}$$

where Q is a query, D is a document, $dist$ is either $dist_{rada2}$ or $dist_{azu}$, and max_dist, the maximum number of edges separating two concepts in the combined MeSH hierarchy, is 23.

Results

Table 5.29 shows the results for the two approaches before and after the outputs of MeSH and text searches are combined. Best-match-combination performs significantly better than all-combination in terms of MAP and average recall before and after fusion with text. Therefore, our hypothesis regarding the combination of concept-pair similarities is supported. However, contrary to our intuition, we observe that hierarchy integration fails to improve the recall over the use of non-hierarchical Cosine measure (Equation 2.10) to compare queries with documents. Before fusion, the average recall of the best-match-combination approach, 0.5848, obtained with $dist_{azu}$, is lower than the average recall obtained with the Cosine measure, 0.6256. After fusion, the average recall of the best-match-combination approach, 0.7003, is also lower than that obtained with the Cosine measure, 0.7199.

5.2.2 Hierarchy Separation

In the previous chapter (Section 4.2.2), we hypothesized that:

Table 5.29: Comparison of two approaches for inter-concept measure combination

	MeSH-only			text+MeSH		
	Av. P@10	MAP	Av. recall	Av. P@10	MAP	Av. recall
binary (no network)	0.3388	0.1639	0.6256	0.4771	0.2973	0.7199
all-combination	0.0204	0.0128	0.3344	0.4184	0.2635	0.6829
best-match-combination	0.3163	0.1401	0.5848	0.4510	0.2841	0.7003

Table 5.30: Hierarchy separation experimental set-up

	inter-concept measure	hierarchy separation	inter-concept measure combination
Baseline	$dist_{rada1}$	none	$dist_{azu}$
HardSep	$dist_{rada1}$	max_dist if 2 concepts from different categories	$dist_{azu}$
SoftSep	$dist_{rada1}$	max_dist if at least 1 of 2 concepts are in StopCat	$dist_{azu}$
DescQualSep	$dist_{rada1}$	max_dist if 1 concept is a descriptor and other a qualifier	$dist_{azu}$

1. separating the hierarchy into the different MeSH categories, or

2. separating descriptors from qualifiers,

while using the hierarchy for concept comparison, would lead to better precision for MeSH-based MEDLINE retrieval.

The first hypothesis is evaluated with two category separation methods, SoftSep and HardSep. The second hypothesis is evaluated with the DescQualSep method. The three methods are fully described in Section 4.2.2. Table 5.30 describes the experimental set-up. Before fusion with the output of text searches, distances between documents and queries are turned into similarity measures with the Equation 5.2. As $dist_{rada1}$ is used to calculate inter-concept semantic distances, the maximum distance in the combined MeSH hierarchy, max_dist, is 23 edges again.

Table 5.31: Hierarchy separation approaches

	MeSH-only			text & MeSH		
	Av. P@10	MAP	Av. recall	Av. P@10	MAP	Av. recall
Baseline	0.3163	0.1401	0.5848	0.4510	0.2841	0.7003
HardSep	0.2531	0.0781	0.4822	0.4449	0.2792	0.6918
SoftSep	0.2837	0.1227	0.5203	0.4429	0.2808	0.7016
DescQualSep	0.3061	0.1370	0.5770	0.4510	0.2842	0.7013

Results

A comparison of the baseline with the three hierarchy separation methods is presented in Table 5.31. Tables B.18, B.19, and B.20 show the statistically significant differences between the separation methods in terms of average P@10, MAP, and average recall, respectively, before fusion with text. Tables B.21, and B.22 show the statistically significant differences between the separation methods in terms of MAP and average recall, respectively, after fusion with text (differences in terms of average P@10 are not statistically significant after fusion with text) . HardSep gives the worst results before and after fusion with text, for MAP and average recall, and the worst P@10 before fusion with text . SoftSep does not decrease baseline performance as much as HardSep, even though decreases in average P@10, MAP, and average recall before fusion are statistically significant. DescQualSep, the least aggressive separation policy, also decreases the average P@10, MAP, average recall of the baseline with statistical significance before fusion with text, but it gives the best results of the separation policies.

The results in Table 5.31 show that introducing a generic concept "*MeSH*" that includes any medical concept is meaningful in order to compare concepts from different hierarchies, contrary to our intuition. Furthermore, the results for SoftSep suggest that some categories that ranked poorly with our method are still significant for genomic topics. Finally, allowing the comparison of qualifiers with descriptors in a combined MeSH hierarchy does not damage precision. It contradicts our intuition based on the fact that descriptors and qualifiers are very different vocabularies in

Table 5.32: Experimental set-up for edge distance calculation

	inter-concept measure	hierarchy separation	inter-concept measure combination
Baseline	$dist_{rada1}$	none	$dist_{azu}$
DepthDens	$dist_{jiang2}$	none	$dist_{azu}$
InfoBased	$dist_{jiang4}$	none	$dist_{azu}$

terms of size and function (see Section 2.1.2). However, the results suggest that the use of a combined MeSH hierarchy, with the addition of artificial nodes "*qualifier*" and "*MeSH*", allowing any MeSH concepts to be compared, does not retrieval precision in the conditions of our experiment.

5.2.3 Edge Distance Variation

In this section we evaluate the hypothesis, formulated in Section 4.2.2, which states that edge distance variation, calculated with hierarchy or corpus information, would impact positively on the precision of MEDLINE retrieval.

To evaluate our hypothesis, we use three methods, DepthDens, InfoBased (Section 4.2.2), and a baseline method. The first two methods, DepthDens and InfoBased, use inter-concept measures that integrate edge distance variations, $dist_{jiang2}$ and $dist_{jiang4}$, respectively. $dist_{jiang2}$ (Equation 2.4) uses hierarchy information (depth and density), whereas $dist_{jiang4}$ (Equation 2.8) uses corpus information (concept frequencies). The baseline method uses $dist_{rada1}$ (Equation 2.1), which assumes that all hierarchy edges correspond to the same distance. All three methods use the combined MeSH hierarchy to compare concepts (no separation), and combine the inter-concept distance values between queries and documents with $dist_{azu}$ (Equation 2.14). Table 5.32 describes the experimental set-up.

Before fusion with the output of text searches, distances between documents and a queries are turned into similarity measures with the Equation 5.2. Here, max_dist

corresponds to different values whether we use the baseline method, DepthDens, or InfoBased. For the baseline method using $dist_{rada1}$, $max\_dist = 23$, which is the maximum number of edges between two concepts in the combined MeSH hierarchy. For DepthDens, max_dist is different for each combination of α and β values, which set the sensitivity of $dist_{jiang2}$ to hierarchy depth and density, respectively. Calculating max_dist for every combination of β and α is computationally expensive. Instead we determine the worst case in terms of depth and density for comparing two concepts in the combined MeSH hierarchy. For depth, the worst case corresponds to the maximum distance between two concepts in the combined MeSH hierarchy, 23 edges, with one concept at depth 12, and the second at depth 11. For density, the worst case corresponds to a path between the two concepts with a minimum density, i.e. only one child per parent c_p, $E(c_p) = 1$. Note that $\overline{E}$ in Equation 2.3, the average number of children over the entire hierarchy, is approximately 3.52 in the combined MeSH hierarchy. The maximum distance for a particular combination of β and α can therefore be reduced to:

$$max\_dist\,(\beta, \alpha) = \left(\beta + (1-\beta)\,\overline{E}\right)\left(\left(\frac{13}{12}\right)^{\alpha} + 2 \times \left(\left(\frac{12}{11}\right)^{\alpha} + \left(\frac{11}{10}\right)^{\alpha} + \ldots + \left(\frac{2}{1}\right)^{\alpha}\right)\right) \qquad (5.3)$$

For InfoBased, max_dist corresponds to the worst case scenario where the LCA of two concepts c_1 and c_2 is the root node "*MeSH*", and the two concepts occur only in one document of the collection:

$$max\_dist\,(c_c, c_p) = 2 \times \log_2 p\,(\text{``}\mathit{MeSH}\text{''}) - \left(\log_2 \frac{1}{\mathrm{CF}\,(\text{``}\mathit{MeSH}\text{''})} + \log_2 \frac{1}{\mathrm{CF}\,(\text{``}\mathit{MeSH}\text{''})}\right) \approx 44.23 \qquad (5.4)$$

DepthDens is evaluated with several combinations of values of β and α: (1, 0), (0.75, 0), (0.5, 0), (0.25, 0), (1, 1), (0.75, 1), (0.5, 1), (0.25, 1), (1, 2), (0.75, 2), (0.5, 2), (0.25, 2). The (1, 0) combination is the equivalent of the baseline method (no sensitivity to hierarchy depth and density, constant edge distance in the hierarchy).

Results

Tables 5.33 and 5.34 show the results in average P@10, MAP, and average recall for the 12 (β, α) combinations before and after fusion with text, respectively. Ta-

Table 5.33: Average P@10, MAP and average recall for DepthDens before fusion with text

		α		
		0	1	2
β	1	0.3163, 0.1401, 0.5848	0.3020, 0.1313, 0.5734	0.2918, 0.1175, 0.5450
	0.75	0.3041, 0.1418, 0.5817	0.2980, 0.1298, 0.5697	0.2918, 0.1179, 0.5472
	0.5	0.2959, 0.1416, 0.5735	0.2918, 0.1379, 0.5623	0.2816, 0.1174, 0.5460
	0.25	0.2816, 0.1324, 0.5503	0.2653, 0.1276, 0.5510	0.2653, 0.1187, 0.5391

Table 5.34: Average P@10, MAP and average recall for DepthDens after fusion with text

		α		
		0	1	2
β	1	0.4510, 0.2841, 0.7003	0.4551, 0.2837, 0.7007	0.4551, 0.2834, 0.7013
	0.75	0.4347, 0.2804, 0.6929	0.4347, 0.2808, 0.6952	0.4388, 0.2803, 0.6961
	0.5	0.4347, 0.2774, 0.6901	0.4347, 0.2779, 0.6898	0.4327, 0.2778, 0.6904
	0.25	0.4367, 0.2758, 0.6856	0.4367, 0.2762, 0.6876	0.4347, 0.2763, 0.6875

bles B.23, B.24, and B.25 show the statistically significant differences between the separation methods in terms of average P@10, MAP, and average recall, respectively, before fusion with text. Tables B.26, B.27, and B.28 show the statistically significant differences between the separation methods in terms of MAP and average recall, respectively, after fusion with text.

The results show that, at a constant level of sensitivity to the hierarchy density (β value), increasing the sensitivity to the depth (α value) has a negative impact on the performance of the MeSH searches for average P@10, MAP, and average recall (Table 5.33). With $\alpha = 0$ (no sensitivity to depth), increasing the sensitivity to density (decreasing β from 1) has a positive impact at first on the MAP for the MeSH searches, but the impact becomes negative as β approaches zero. However, the differences are not statistically significant. Additionally, smaller values for β have a negative impact on the average P@10 and the average recall of the MeSH searches. This could explain the significant decrease in MAP after fusion as β decreases. As mentioned in Section 5.1.4, the effectiveness of fusion has been linked to the overlap of relevant documents in the two result sets. In particular, the results after fusion suggest that an increased sensitivity to density retrieves at high ranks relevant documents already ranked highly in the output of the text searches.

Surprisingly, the calculation of edge distances with hierarchy information damages the overall baseline retrieval performance. Notably, increasing sensitivity to depth or density does not increase precision as expected. This contradicts our assumption about the shrinking of edge distances with hierarchy depth and density.

Table 5.35 shows the results of InfoBased before and after fusion with text, in terms of MAP and average recall, in comparison the baseline method. Tables B.29 and B.30 show the statistically significant differences between the separation methods in terms of average P@10, MAP, and average recall before and after fusion with text, respectively. InfoBased performs significantly better than the baseline method with the MeSH searches, in MAP and average recall. Infobased also gives a better average P@10 than the baseline, but the difference is not statistically significant. However, after fusion with text searches, InfoBased gives lower values of average P@10, MAP and average recall than the baseline. Note that the difference after fusion between the two methods are not statistically significant. Once again we encounter a situation where a method giving good results for MeSH searches does not have the expected impact after fusion with text searches. This might be explained by the fact that InfoBased may retrieve documents that are already in the text-search result set, whereas the baseline may highly rank relevant documents not contained in the text-search result set. If we focus on the results of the MeSH searches, calculation of edges distances with corpus information improves precision of retrieval, and therefore confirms our expectations.

If we hold true the assumption that semantic distances between concepts decreases as conceptual specificity and density increase, the results obtained with DepthDens raise questions about the ability to capture this variation with MeSH hierarchy information such as depth and density. In contrast, results obtained with InfoBased suggest that the distribution of concepts in a large corpus such as Trec-

Table 5.35: Results for InfoBased before and after fusion with text

	MeSH-only			text & MeSH		
	Av. P@10	MAP	Av. recall	Av. P@10	MAP	Av. recall
baseline	0.3163	0.1401	0.5848	0.4510	0.2841	0.7003
InfoBased	0.3306	0.1645	0.6149	0.4449	0.2831	0.6968

Table 5.36: Results for MeSH searches with hierarchy integration

	no-network	baseline	DepthDens (0.75, 0)	InfoBased
Av. P@10	0.3388	0.3163	0.3041	0.3306
MAP	0.1639	0.1401	0.1418	0.1645
Av. recall	0.6256	0.5848	0.5817	0.6149

Gen05 provides useful information for the calculation of edge distances in the MeSH hierarchy.

5.2.4 Conclusion: Any Benefits from Using the Hierarchy?

Our main hierarchical hypothesis was that hierarchy integration improves the recall of MEDLINE retrieval. Table 5.36 and 5.37 show the results of the best-performing methods integrating the hierarchy for the MeSH searches, and for the combined MeSH and text searches, respectively. The no-network results corresponds to the use of the Cosine measure (Equation 2.10) to compare binary representations of documents and queries. Here, the baseline corresponds to the use of the edge count method to calculate inter-concept semantic distances (see Section 4.2.2). The re-

Table 5.37: MeSH-and-text search results for hierarchy integration

	no-network	baseline	DepthDens (0.75, 0)	InfoBased
Av. P@10	0.4776	0.4510	0.4347	0.4449
MAP	0.2973	0.2841	0.2804	0.2831
Av. recall	0.7199	0.7003	0.6929	0.6968

sults show that InfoBased is the only method integrating the hierarchy that performs comparably to no-network. Nevertheless, InfoBased is not as effective as no-network after fusion with text searches. As fusion performance was shown to be higher if the two methods retrieve different relevant documents (Beitzel et al., 2004), we can hypothesize that InfoBased retrieved more relevant documents already retrieved with text than no-network did.

Overall, the results provide little evidence to support our main hierarchical hypothesis. In the next chapter, our non-hierarchical and hierarchical hypotheses are evaluated in the context of MEDLINE document classification.

Chapter 6

Evaluation with MEDLINE Document Classification

We now proceed to evaluate the hypotheses, non-hierarchical and hierarchical, developed in Chapter 4 in the context of the binary classification of biomedical documents. First, the experimental set-up, including the metrics and classification software used, is described.

6.1 Experimental Set-up

6.1.1 TREC 2005 Genomics Track GO Triage task

The hypotheses are evaluated with the TREC 2005 Genomics track GO triage task described in Section 3.3.1. The task simulates the activity of MGI curators, i.e. the manual selection of MEDLINE documents likely to provide experimental evidence for the annotation of mouse genes with GO concepts. A total of 11,880 MEDLINE documents are provided, 5,837 to train a classifier, and 6,043 to test the resulting classifier. 462 and 518 of the training and test documents, respectively, are judged relevant for the task. These are positive examples. The rest of the documents are judged non-relevant. These are negative examples. The metrics used in the evaluation are the precision (Equation 3.1), the recall (Equation 3.2), the F-Score

(Equation 3.9), and the normalised utility U_{norm} (Equation 3.10). The parameters used to calculate the raw utility U_{raw} (Equation 3.11), $u_r = 11$ and $u_{nr} = -1$, are determined as explained in Section 3.3.1.

6.1.2 Text Categorization and SVM^{light}

A detailed discussion on text categorization and machine learning techniques is beyond the scope of this dissertation and is found elsewhere (Sebastiani, 2002). Our experiments focus on evaluating hypotheses regarding MeSH-based document representations, and do not aim at comparing several text categorization techniques.

Support vector machines (SVMs) (Vapnik, 1995) were widely used for the TREC 2004 and 2005 classification tasks. They were either used alone (Fujita, 2004; Darwish and Madkour, 2004; Cohen et al., 2004; Si and Kanungo, 2005; Niu et al., 2005; Subramaniam et al., 2005; Lee et al., 2005), or in combination with other learning methods (Aronson et al., 2005). Furthermore, they have been applied successfully to text categorization Joachims (1998). Consequently, we chose to use SVMs for the classification of MeSH-based document representations.

SVMs are based on concepts developed in Vapnik (1982). Given observations consisting of pairs of vectors x_i and binary values y_i ($y_i = 1$ indicates that x_i is a positive example, i.e. a relevant document, and $y_i = -1$ indicates that x_i is a negative example, i.e. a non-relevant document), the machine learns a family of functions $x \mapsto f(x, \alpha)$ that associates any unlabeled x (y is undetermined) to one of the two possible values of y. Each function is determined by a set α of parameter values.

The idea is to reduce the expected risk of classification error $R(\alpha)$ on the set of unlabeled vectors. First, the empirical risk $R_{emp}(\alpha)$ on l training observations is calculated:

$$R_{emp}(\alpha) = \frac{1}{2l} \sum_{i=1}^{l} |y_i - f(x_i, \alpha)|$$

$R(\alpha)$ is then bounded with probability η:

$$R(\alpha) \leq R_{emp}(\alpha) + \sqrt{\frac{h(\log(2l/h)+1) - \log(\eta/4)}{l}}$$

where h is a non-negative integer called the Vapnik Chervonenkis (VC) dimension and, the second term of the right-hand side of the equation is called the VC confidence. The VC dimension is a measure of the concept of capacity. A machine with too much capacity will tend to over-fit the training data and not generalize well. On the other hand, a machine with too little capacity will generalize too much. Also, minimizing the VC confidence will lower the upper bound of the expected risk of error.

The simplest case of SVM classification corresponds to linear machines trained on linearly separable observations x_i (See Figure 6.1). The SVM simply indentifies the hyperplane separating positive from negative examples with the largest margin, given that the training vectors x_i satisfy the following constraints:

$$x_i \cdot w + b \geq +1 \quad \text{for } y_i = +1 \tag{6.1}$$

$$x_i \cdot w + b \leq -1 \quad \text{for } y_i = -1 \tag{6.2}$$

where w is normal to the hyperplane. The vectors of positive and negative observations lying on hyperplanes $H_1 : x_i \cdot w + b = +1$ and $H_2 : x_i \cdot w + b = -1$, respectively, are called the support vectors. The margin is the distance between the two parallel hyperplanes H_1 and H_2 and is simply $2/\|w\|$. Therefore, maximizing the margin is equivalent to minimizing $\frac{1}{2}\|w\|^2$, subject to constraints 6.1 and 6.2.

When the observations x_i are not linearly separable, linear machines can still be used by introducing positive slack variables ξ_i into the constraints:

$$x_i \cdot w + b \geq +1 - \xi_i \quad \text{for } y_i = +1$$

$$x_i \cdot w + b \leq -1 + \xi_i \quad \text{for } y_i = -1$$

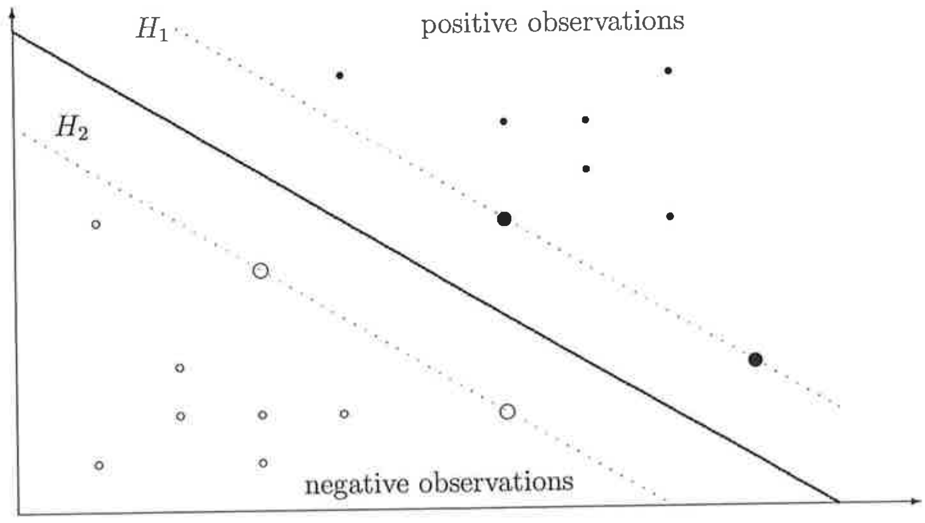


Figure 6.1: An example of linear machines trained on linearly separable data

Maximizing the margin is now equivalent to minimizing $\frac{1}{2}\|w\|^2 - C\left(\sum_i \xi_i\right)^k$. C is a user-defined parameter proportional to the cost of errors. In this dissertation we use linear machines for non-linearly separable observations to classify MeSH-based document representations. For a more detailed description of SVMs, including non-linear machines, the reader is directed to Burges (1998).

We use the SVM^{light} software which is an implementation of the Support Vector Machine method described in Joachims (1999). SVM training leads to a quadratic optimization problem and learning from large training sets can quickly become computationally expensive. The SVM^{light} implementation uses a decomposition algorithm that only uses a fixed subset of data called the working set. Therefore, a smaller quadratic problem is solved for each working set. This decomposition allows for large-scale SVM learning at lower computational cost.

We use the default settings for the learning module (svm_learn) and the classification module (svm_classify) of SVM^{light}. The automatic tuning of the software has been shown to perform well for text categorization (Joachims, 1998). Default settings include the use of a linear kernel. Moreover, parameter C, allowing a trade-off between the training error and the classifier complexity, is determined automatically from the norms of the training vectors x_i:

$$C = \left(\frac{1}{l} \sum_{i=1}^{l} \|x_i\| \right)^{-2}$$

The only modification is the setting of parameter j (the cost-factor by which training errors on positive examples out-weight errors on negative examples, the default being 1) to 11, similarly to Subramaniam et al. (2005). The j parameter allows us to tune the classifier to the difference between u_r (= 11), the relative utility of a true positive, and u_{nr} (= −1), the relative utility of a false positive (see Equation 3.11, Section 3.3.1).

6.1.3 Document Representation: Our Approach

Similarly to Seki et al. (2004) and Lee et al. (2005) (see Section 3.3.4), our approach is based on representing documents with MeSH concepts only. This approach differs from the combination of text and MeSH representations used in the ad hoc task evaluation (previous chapter). For the ad hoc task, the starting point was an information need expressed in free-text. In order to express this information need with MeSH concepts, we searched documents' textual fields and used pseudo-relevance feedback to generate a MeSH-based query. For the classification task, the starting point are relevant and non-relevant documents available to train a classifier and MeSH concepts can be used directly as classification features.

6.2 Non-hierarchical Hypotheses: Result and Analysis

Table 6.1 shows the results of the non-hierarchical approaches (binary, TF*IDF, MajMin, DescQual) described in Section 4.1. The results correspond to the performance of the SVM^{light} classifier on the test documents after learning on the training documents with the settings described above. The star sign next to the name of

Table 6.1: Classification results for test documents for various MeSH-based representations

doc rep	Precision	Recall	F-score	Normalized Utility	C (= aver$\|\|x\|\|^2$)
binary*	0.2980	0.6139	0.4013	0.4824	0.0503
TF*IDF*	0.3037	0.5212	0.3838	0.4126	0.0007
MajMin_2*	0.2922	0.5811	0.3889	0.4531	0.0265
MajMin_3*	0.2798	0.5656	0.3744	0.4333	0.0149
MajMin_4*	0.2751	0.5598	0.3690	0.4258	0.0093
MajMin_5*	0.2710	0.5598	0.3652	0.4230	0.0062
DescQual*	0.3323	0.4286	0.3744	0.3503	0.0541

a representation strategy indicates a statistically significant difference between unclassified documents and classified documents (one-tailed z-test for proportions with 0.05 significance level). However, there is no statistically significant difference between the binary representation and all the other non-hierarchical representations.

The integration of collection frequency information with the TF*IDF weighting does not overall benefit the performance of the triage task. The precision is slightly increased (+2%), but at the price of a drop in recall (-15%). As the utility function favors recall (formula 3.11), the normalized utility of the TF*IDF approach decreases by 14.5% compared to the binary representation approach. However, the increase in precision is consistent with our hypothesis.

The discrimination between major and minor themes through the MajMin method does not have a favorable impact on the overall classification performance. Both precision and recall decrease as more discrimination is introduced. Consequently, the utility is negatively affected by this discrimination policy. This contradicts our hypothesis, as discriminating between major and minor document themes was expected to favor precision.

The DescQual representation approach has a positive impact on precision (11.5% improvement on the binary approach), as expected. Unfortunately, the precision increase is accompanied by a severe drop in recall (30.2% decrease over the binary approach). Consequently, DescQual gives the weakest normalized utility of all non-hierarchical representation methods.

The results obtained with the TF*IDF and DescQual representations method confirm our non-hierarchical hypotheses regarding improved precision resulting from better discriminations between concepts within the document. Surprisingly however, the MajMin method is detrimental to both precision and recall. This suggests that the minor themes play an important part in distinguishing documents from each other. Moreover, for the three non-hierarchical representations, the gains in precision, when any, lead to higher losses in recall, which damage the task performance in terms of normalized utility. This is a consequence of the bias of the utility function towards recall.

6.3 Hierarchy Integration

The information contained in the hierarchy is used to build extended MeSH-based representations from the concepts initially found in the MeSH fields of the documents. This extension method is described in Section 4.2.1. In the evaluation, we use threshold values that correspond to the minimal score for a candidate term to be added to the representation. The threshold strategy is motivated by the reduction of features for computational cost control and noise removal.

6.3.1 Inter-concept Similarities Combination

In this section we compare the best-combination and all-match-combination approaches (Section 4.2.2) for the extension of document representations. The concepts not present in the initial representation are scored according to their semantic relevance to the concepts present in the initial representation. The semantic distance between a concept and a document D is calculated with $dist_{ext1}$ (Equation 4.2) for the all-match-combination approach, and with $dist_{ext2}$ (Equation 4.3) for the best-combination approach. The distances are turned into concept weights w_d with the

Table 6.2: Experimental set-up for comparing all-combination and best-match-combination

	inter-concept measure	hierarchy separation	inter-concept measure combination
all-combination	$dist_{rada1}$	none	$dist_{ext1}$
best-match-combination	$dist_{rada1}$	none	$dist_{ext2}$

following Equation:

$$w_d(c_i, D) = 1 - \left(dist_{ext1/2}(c_i, D) / max\_dist\right) \tag{6.3}$$

where $dist_{ext1/2}$ is either $dist_{ext1}$ or $dist_{ext2}$ depending upon the approach, and c_i is a candidate concept for extension. $dist_{rada1}$ (Equation 2.1), a simple edge count, is used to calculate inter-concept distances. To compare concepts, the combined MeSH hierarchy, with additional nodes "*qualifier*" and "*MeSH*", is used. The maximum distance *max_dist* in the combined MeSH hierarchy is 23 edges. Table 6.2 summarizes the experimental set-up.

Results

Tables 6.3 and 6.4 show results of the classification of test documents with SVM^{light} after training the classifier on the training documents for the all-combination and best-match-combination methods, respectively. For each method, a range of threshold values is reported for candidate term integration. The last row shows the result of the non-extended baseline representation approach, called binary. Next to the threshold values in brackets are the sizes of the two files containing the feature vectors of the training and test documents. The *b* letter next to the name of an extended representation threshold indicates a statistically significant difference with the binary representation (one-tailed z-test for proportions with 0.05 significance level). Both extension approaches increase recall and decrease precision as the threshold is lowered and more concepts are added. The increase in recall is consistent with our

Table 6.3: Classification results for all-combination

Threshold	Aver. Num. of concepts per Docs	recision	Recall	F-score	Normalized Utility
1.0 (binary)	20.6	0.2980	0.6139	0.4013	0.4824
0.8	20.9	0.2994	0.6139	0.4025	0.4833
0.75	45.7	0.2893	0.6467	0.3998	0.5023
0.72^b	186.8	0.2430	0.7529	0.3674	0.5397
0.7^b	500.8	0.2107	0.8784	0.3399	**0.5793**
0.68^b	1242.5	0.1941	0.8977	0.3191	0.5588
0.65^b	3406.4	0.1944	0.8514	0.3166	0.5307
0.6^b	10425.4	0.1636	0.7124	0.2660	0.3812

hypothesis regarding the effect of hierarchy integration in document representation on retrieval performance.

With the all-combination approach (Table 6.3), the recall peaks at thresholds of 0.7 with 0.8784 (+43% on the binary representation). This threshold value also corresponds to the best normalized utility obtained with this strategy (0.5793, +20% on the binary representation). Lower threshold values add noisy concepts that damage both precision and recall, and hence also normalized utility.

With the best-match-combination approach (Table 6.4), more concepts are added to the representations for similar threshold values, as indicated by the file sizes. This is explained by the highest scores given to candidate concepts by $dist_{ext2}$. Threshold values smaller than 0.7 produce document representation files that are too large for the SVM^{light} software to process. With this approach, the recall is also increased but at a higher cost in precision. Consequently, the normalized utility peaks at 0.5277 (+9.4% on the binary representation).

The results confirm our intuition (Section 4.2.2) that all-combination produces better scores than best-match-combination to for the candidate concepts to representation extension.

Table 6.4: Classification results for best-match-combination

Threshold	Aver. Num. of concepts per Docs	Precision	Recall	F-score	Normalized Utility
1.0 (binary)	20.6	0.2980	0.6139	0.4013	0.4824
0.95^b	116.1	0.2657	0.6371	0.3750	0.4770
0.9^b	487.4	0.2446	0.7220	0.3654	0.5193
0.85^b	1242.9	0.2320	0.7548	0.3550	**0.5277**
0.8^b	2642.1	0.2031	0.7085	0.3157	0.4558
0.75^b	5516.7	0.1773	0.7857	0.2893	0.4542
0.7^b	11399.6	0.1631	0.9228	0.2773	0.4925

6.3.2 Hierarchy Separation

We now evaluate the hypotheses described in Section 4.2.2 regarding the separation of the MeSH hierarchy for concept comparison. Four methods are used: a baseline (no separation), HardSep (separation between all MeSH categories), SoftSep (separation between relevant and non-relevant areas of the hierarchy), and DescQualSep (separation between descriptors and qualifiers). All methods are described in detail in Section 4.2.2. The baseline corresponds to the all-combination method described above with a threshold of 0.7 (see Table 6.3). The added concepts contained in the baseline representation are then re-scored using the HardSep, SoftSep, and DescQualSep separation methods. No additional concept is added to the baseline. The distance between two concepts from separated areas of the hierarchy is set to the maximum distance max_dist within the combined hierarchy (including additional nodes "*qualifier*" and "*MeSH*"). As $dist_{rada1}$ (Equation 2.1) is used to calculate inter-concept distances, $max\_dist = 23$. $dist_{ext1}$ (Equation 4.2) is used by the four methods to combine inter-concept similarities. Distances are turned into weights with Equation 6.3. Table 6.5 describes the experimental set-up.

Table 6.5: Hierarchy separation experimental set-up

	inter-concept measure	hierarchy separation	inter-concept measure combination
Baseline	$dist_{rada1}$	none	$dist_{ext1}$ (threshold=0.7)
HardSep	$dist_{rada1}$	max_dist if 2 concepts from different categories	$dist_{ext1}$
SoftSep	$dist_{rada1}$	max_dist if at least 1 of 2 concepts are in StopCat	$dist_{ext1}$
DescQualSep	$dist_{rada1}$	max_dist if 1 concept is a descriptor and other a qualifier	$dist_{ext1}$

Results

Table 6.6 shows the results of the three separation methods for the classification of the test documents. The b letter next to the name of a separation strategy indicates a statistically significant difference with the no-separation baseline (one-tailed z-test for proportions with 0.05 significance level). All methods have a positive impact on precision and a negative impact on recall, as expected. This impact is important for HardSep as all hierarchies are separated. Precision increases by 32% and recall decreases by 30%. As a result, the normalized utility also decreases (-10.6%). SoftSep is the only method that leads to a slight improvement of normalized utility (+1.7%). This improvement is obtained with an increase in precision (+5.7%) and a decrease in recall (-1.8%) over the baseline. DesQualSep gives a similar normalised utility as the baseline, but with a higher precision (+12.1%), and a lower recall (-6.8%).

The results suggest that hierarchy separation produces weights w_d for the added concepts which favor the precision of the classification. However, given the utility function of the GO task (Equations 3.10 and 3.11) soft approaches to separation which favor recall, such as SoftSep and DescQualSep, yield better utility results.

Table 6.6: Classification results with hierarchy separation

Strategy	Precision	Recall	F-score	Normalized Utility
Baseline	0.2107	0.8784	0.3399	0.5793
HardSep^b	0.2786 (+32%)	0.6776 (-30%)	0.3948	0.5181 (-10.6%)
SoftSep^b	0.2228 (+5.7%)	0.8629 (-1.8%)	0.3542	**0.5893** (+1.7%)
DescQualSep^b	0.2363 (+12.1%)	0.8185 (-6.8%)	0.3668	0.5781 (-0.2%)

6.3.3 Edge Distance Variation

In this section we evaluate hypotheses formulated in Section 4.2.2 regarding the variation of edge distance in the MeSH hierarchy.

Following the intuition expressed in Section 2.2.1, hierarchy edges are expected to correspond to different semantic distances. Consequently, edge distances are calculated with two methods. The first one, DepthDens, uses hierarchy information: the depth and the density. The second one, InfoBased, uses corpus information which consists of the collection frequencies of concepts in a corpus. The two methods are compared to a baseline using simple edge count, hence assuming constant edge distance in the hierarchy.

All three methods, the baseline, DepthDens, and Infobased, use $dist_{ext1}$ (Equation 4.2) to calculate the semantic distance between a concept c_i, not contained in the initial representation of document D, and the initial representation of D. However, different inter-concept distance measures are used by DepthDens and InfoBased. DepthDens uses $dist_{jiang2}$ (Equation 2.4) to integrate hierarchy information, and InfoBased uses $dist_{jiang4}$ (Equation 2.8) to integrate corpus information. The three inter-concept measures, $dist_{rada1}$, $dist_{jiang2}$ and $dist_{jiang4}$, compare concepts with the combined MeSH hierarchy, which includes additional nodes "*qualifier*" and "*MeSH*". Distances are turned into weights with Equation 6.3. The maximum inter-concept distance, max_dist, is 23 edges for $dist_{rada1}$. For $dist_{jiang2}$ and $dist_{jiang4}$ however,

Table 6.7: Experimental set-up for edge distance calculation

	inter-concept measure	hierarchy separation	inter-concept measure combination
baseline	$dist_{rada1}$	none	$dist_{ext1}$ (threshold=0.7)
DepthDens	$dist_{jiang2}$	none	$dist_{ext1}$
InfoBased	$dist_{jiang4}$	none	$dist_{ext1}$

max_dist is calculated with Equations 5.3 and 5.4, respectively.

The baseline corresponds to the all-combination method with a threshold of 0.7 (see Table 6.3). The added concepts contained in the baseline representation are then re-scored using either the DepthDens or the InfoBased method. Table 6.7 summarizes up the experimental set-up.

Results

Tables 6.8, 6.9, and 6.10 respectively show precision, recall, and normalized utility results for different values of β and α used by DepthDens. Setting $\beta = 1$ and $\alpha = 0$ is the equivalent of the baseline method (no sensitivity to hierarchy depth and density, constant edge distance in the hierarchy). Decreasing β or increasing α lead to a growing influence of hierarchy density and depth, respectively.

Results in Tables 6.8 and 6.9 show that decreasing β damages precision but improves recall. Conversely, increasing α benefits precision but leads to a drop in recall. This suggests that an increased sensitivity to hierarchy depth in the determination of the added concepts' weights w_d favors precision. Therefore, the results confirm our expectations regarding the importance of hierarchy depth in the determination of edge distances. Moreover, the results support the assumption of a negative correlation between edge distance and hierarchy depth. The negative influence of β on precision, however, is surprising. It indicates that a higher sensitivity to hierarchy density in the determination of the added concepts' weights w_d damages precision.

Table 6.8: Precision of classification of test documents for different combinations of values of β and α

		α		
		0	1	2
	1	0.2107	0.2157	0.2234
β	0.75	0.1997	0.2002	0.2037
	0.5	0.1980	0.1985	0.1993
	0.25	0.1971	0.1974	0.1978

Table 6.9: Recall of classification of test documents for different combinations of values of β and α

		α		
		0	1	2
	1	0.8784	0.8668	0.8417
β	0.75	0.8880	0.8842	0.8822
	0.5	0.8938	0.8938	0.8919
	0.25	0.8938	0.8938	0.8938

Furthermore, it suggests that, contrary to our hypothesis, semantic distance between concepts is not negatively correlated to the density of the hierarchy.

The influence of β and α values on normalized utility are shown in Table 6.10. Apart from one combination of values ($\beta = 1$ and $\alpha = 1$), all combinations have a negative impact on normalized utility. However, there is no statistically significant difference between the DepthDens strategies and all the edge count baseline (one-tailed z-test for proportions with 0.05 significance level).

Table 6.11 shows the results of the InfoBased method in comparison with the

Table 6.10: Normalized utility of classification of test documents for different combinations of values of β and α

		α		
		0	1	2
	1	0.5793	0.5802	0.5756
β	0.75	0.5644	0.5630	0.5686
	0.5	0.5648	0.5656	0.5662
	0.25	0.5628	0.5635	0.5642

Table 6.11: Classification results with corpus information

Strategy	Precision	Recall	F-score	Normalized Utility
baseline	0.2107	0.8784	0.3399	0.5793
InfoBased	0.2101	0.8707	0.3385	0.5730

baseline method. There is no significant impact on precision, recall, and normalized utility (one-tailed z-test for proportions with 0.05 significance level). This suggests that the constant edge distance hypothesis gives a good approximation of $dist_{jiang4}$, a distance measure using corpus information to calculate edge distances.

6.4 Tuning of the SVM C parameter

In order to focus on comparing different MeSH-based document representations, we used the default settings of SVM^{light} for document classification in the previous evaluations. This included the automatic setting of parameter C, allowig trade-off between the training error and the classifier complexity, to the inverse of the square of the average of the norms of the training vectors x_i (see Section 6.1.2). In this section we examine the benefits of tuning C manually on the training documents using 4-fold cross-validation.

6.4.1 Non-hierarchical Representations

Table 6.12 shows the result for the non-hierarchical representations with the manual tuning of C. The *b* letter next to the name of a non-hierarchical representation indicates a statistically significant difference with the binary representation (one-tailed z-test for proportions with 0.05 significance level). First of all, the binary representation is confirmed as the representation giving the best normalized utility (0.5874) of all the non-hierarchical representations. Second, the manual tuning of C improves the normalized utility by 21.8% (0.4824 to 0.5874) over the use of the automatic setting of C. Third and last, the normalized utility obtained here by the

Table 6.12: Classification results for test documents for various MeSH-based representations with C tuning

Doc Rep.	Precision	Recall	F-score	Normalized Utility	C (4-fold cross-validation)
Binary	0.2259	0.8533	0.3572	**0.5874**	0.0085
TF*IDF^b	0.2659	0.7008	0.3856	0.5249	16e-05
MajMin_2	0.2408	0.7568	0.3653	0.5398	0.00685
MajMin_3	0.2295	0.8166	0.3583	0.5674	0.002325
MajMin_4	0.2371	0.7876	0.3644	0.5572	0.00155
MajMin_5	0.2450	0.7510	0.3694	0.5405	0.0012
DesQual	0.2251	0.8069	0.3520	0.5544	0.0096

binary representation is higher than the normalized utility (0.5793) obtained by extending the representation with the all-combination method (see Table 6.2). Note that this extension method leads to an average document size of 500.8 concepts, much higher than the initial average size of binary, which is 20.6 concepts (see Table 6.3). Furthermore, the normalized utility obtained with binary after manual tuning of C is practically the same as the normalized utility of the 2005 track (0.5870, see Table 3.17).

6.4.2 Hierarchical Representations

Next, we examine the impact of document extension on normalized utility using the all-combination method (see again Table 6.2) with the manual tuning of C. Table 6.13 shows the result for different extension threshold values. The *b* letter next to a threshold level indicates a statistically significant difference with the binary representation (one-tailed z-test for proportions with 0.05 significance level). The best normalized utility (0.5907) is obtained with a threshold value of 0.75, which correspond to an average document size of 45.7 concepts. This is an improvement over the automatic setting of C in terms of document size, as the best normalized utility was then obtained with a threshold value of 0.7 and an average document size of 500.8 concepts per document (see Table 6.3). However, the extended representa-

Table 6.13: Document extension using all-combination method with C tuning

Doc Rep.	Thres.	Aver. Num. of Concepts per Docs	Precision	Recall	F-score	Norm. Utility	C (4-fold cross-validation)
binary	1.0	20.6	0.2259	0.8533	0.3572	0.5874	0.0085
doc extension with all-combination method, no hierarchy separation	0.8	20.9	0.2291	0.8398	0.3600	0.5828	0.009
	0.75	45.7	0.2229	0.8649	0.3544	**0.5907**	0.007
	0.72	186.8	0.2270	0.8243	0.3560	0.5691	0.00725
	0.7	500.8	0.2172	0.8571	0.3466	0.5763	0.005
	0.68	1242.5	0.2065	0.8340	0.3310	0.5426	0.004
	0.65[b]	3406.4	0.1972	0.8745	0.3218	0.5509	0.002125

tion only slightly improves on the binary representation, which gave a normalized utility of 0.5874 with the manual setting of C. This suggests that document extension already corresponded to automatically determined value of C that optimized normalized utility. This could partly be explained by the influence of higher training document size on the automatic determination of C.

6.5 Conclusion

In this chapter we examined a set of hypotheses regarding the representation of MEDLINE documents with the MeSH ontology. A first set of hypotheses concerned the benefit of weighting concepts in MeSH-based document representations according to the MeSH field structure and the frequency of concepts in a corpus. A second set of hypotheses related to the impact of information contained in the MeSH hierarchy. The hypotheses were evaluated with the binary classification of MEDLINE documents. The classification task consisted of selecting articles likely to contain experimental evidence for the annotation of mouse genes with Gene Ontology concepts.

The strategies using MeSH field information did not benefit the normalised utility of the classification. Overall, the weighting methods increased precision as expected. However, they also reduced recall, which was an important factor in the utility of the evaluation task. Discriminating concepts with collection information (TF*IDF) slightly increased precision but was too costly in terms of recall to impact positively on the normalised utility. Moreover, the distinction policy between major and minor concepts had a surprising negative impact not only on recall but also on precision, contrary to our expectations. Finally, keeping the MeSH field associations between descriptors and qualifiers improved the precision but considerably damaged the recall.

The use of the hierarchy to extend MeSH-based document representations was shown to increase the recall of the classification, thus confirming our main hierarchical hypothesis. A normalized utility of 0.5793 was obtained with a simple concept selection method for document extension. The method scored candidate concepts with the average of their shortest path in terms of edge count to the concepts initially contained in the document. A score threshold of 0.7 was used for adding new concept (see Table 6.3).

Additional hypotheses regarding the separation of the MeSH hierarchy and the variation of edge distance in the hierarchy were evaluated. The hierarchy separation method based on relevance information for genomic topics was particularly useful and gave us our best result for normalized utility (0.5893). Overall, separating concepts from different areas of the hierarchy benefited precision, confirming our hypothesis. However, the evaluation of methods using hierarchy or corpus information to measure edge distance variation gave conflicting results. Increasing sensitivity to depth was shown to increase the precision of the classification, and our hypothesis, negatively correlating depth to edge distance, was confirmed. Inversely, increasing sensitivity to density was shown to increase the recall of the classification. This suggests, contrary to our hypothesis, that hierarchy density is positively correlated with edge distance in the MeSH hierarchy. In general, hierarchy information did not improve the normalized utility of the task. Finally, our method using corpus information to calculate edge distance obtained results similar to the baseline's, which assumes constant edge distance in the hierarchy. This could indicate a connection

Table 6.14: Our best method in comparison with 2005 track's best and the "*Mice*" filter

Strategy	Precision	Recall	F-score	Normalized Utility
$t = 0.6975$	0.2041	0.8745	0.3309	0.5644
2005 track's best	0.2122	0.8861	0.3424	0.5870
Mice filter	0.1889	0.9093	0.3127	0.5542

between the distribution of concepts in the corpus and the maintenance of the MeSH hierarchy. This connection, already identified and investigated in the past (Weinberg and Cunningham, 1985), remains to be examined on more recent data.

To compare our method with the best result of the 2005 GO triage task, we determine a threshold t that maximizes normalized utility by 4-fold cross-validation on the training set. The representation extension is done with the all-combination method (see Table 6.2). A value $t = 0.6975$ is found to maximize normalized utility on the training set. Table 6.14 shows the results of the classification on the test documents ($t = 0.6975$, all-combination extension method), as well as the results of the best approach for the runs submitted to the 2005 GO triage task (Niu et al., 2005). Additionally, the two strategies are compared with a simple "*Mice*" filter method. In contrast with our approach, Niu et al. (2005) used the full-text of articles for document representation (see Section 3.3.3). Feature selection involved the comparison between term frequencies in domain-specific and domain-independent corpora. The "*Mice*" filter strategy selects as relevant documents containing the MeSH descriptor "*Mice*". Its high utility result compared to the two other approaches mentioned here reflects the importance of recall in the utility of this classification task. However, our method offers a way to trade precision for recall that is domain-independent and therefore could be used with other hierarchies and other classification tasks.

Chapter 7

Conclusion and Future Work

7.1 Background

This dissertation addressed the problem of MEDLINE document retrieval. The recent discoveries in new genes, with new technologies allowing the sequencing of entire genomes, have led to the tremendous growth of the biomedical literature (see Chapter 1). Despite the increasing availability of full journal articles on the Web, the MEDLINE database remains a point of entry for biologists (Hersh et al., 2004, 2005).

We suggested that the use of a medical ontology, the Medical Subject Headings (MeSH), could help to improve the retrieval of MEDLINE documents. MeSH concepts are used by human indexers to annotate the conceptual content of MEDLINE documents. In particular, we outlined the advantages of using standard terms to name concepts in order to compensate for the ambiguities of the free text used in research articles. Moreover, the semantic network of MeSH, giving information about relationships between concepts, was shown to provide useful information to compare MeSH concepts contained in documents. Consequently, we reviewed a set of measures which use semantic networks to calculate the similarity of concepts (see Chapter 2).

Table 7.1: Contribution: MeSH-based document representation

	Previous Work	Our Contribution
Tokenization	- Only the relevant MeSH words are kept: Srinivasan (1996a). - Comma chosen as delimiter: Kraaij et al. (2004). - Descriptors and qualifiers as minimal units: Ontrup et al. (2003), Struble and Dharmanolla (2004).	- Descriptors and qualifiers are kept as minimal units. - Descriptor/qualifier associations are kept as minimal units.
Concept Weighting	Distinction between major and minor themes: Shin and Han (2004).	- A more aggressive distinction scheme is used. - The evaluation uses a larger collections for ad hoc retrieval and binary classification.
Hierarchy Integration	- Extension of query with descendent concepts: - Extension with ancestor concepts: Ontrup et al. (2003), Struble and Dharmanolla (2004). - Added concepts are given same weight as original concepts: PubMed, Ontrup et al. (2003). - Added concepts are given smaller but constant weights: Struble and Dharmanolla (2004).	- Extension with all related concepts: ancestors, descendants, siblings, cousins, etc... - Weight of added concepts depends on their semantic distance to the original document. - Hierarchy integration at indexing and retrieval time.

7.2 Thesis Statement

We showed that current uses of MeSH-based document representations made little use of the information contained in the structure of MEDLINE MeSH fields and in the MeSH semantic network or hierarchy (see Chapter 3). In particular, we suggested the use of discriminative information in the MeSH fields, as well as the use of contextual information about concepts. Table 7.1 gives an overview of our contribution regarding the use of MeSH-based document representation.

Our review of the related work regarding MeSH-based representations led to two sets of hypotheses. The first set, called non-hierarchical hypotheses, used corpus and MeSH field information to tokenize and weight MeSH concepts in MEDLINE records, and aimed at improving retrieval precision. The second set, called hierarchical hypotheses, used corpus and MeSH hierarchy information to compare concepts, and with the main aim of improving retrieval recall (see Chapter 4).

The non-hierarchical hypotheses were based on the weighting and tokenization of concepts contained in the MeSH fields of MEDLINE records. Weighting concepts provided a way to discriminate between important concepts and less important concepts for a document. The discriminative weights were either derived from the corpus distribution of the concepts (TF*IDF), or the MeSH field distinction between major and minor themes. The tokenization strategies for the MeSH fields, i.e. keeping associations or not between descriptors and qualifiers, allowed us to evaluate the impact of contextual information on precision. Overall, more discrimination between concepts and more context were expected to improve retrieval precision.

Our main hierarchical hypothesis was that the comparison of concepts with the MeSH hierarchy would improve retrieval recall. The hierarchy allows us to calculate a semantic similarity between concepts, and consequently to increase the similarity between documents containing non-identical but related concepts. The hierarchy information was integrated in two ways:

1. at comparison time, when query concepts were compared to documents concepts, and
2. at extension time, when document and query representations were extended with related concepts.

To maximize recall, the different parts of the MeSH hierarchy were combined with the addition of artificial nodes "*qualifier*" and "*MeSH*".

Secondary hierarchical hypotheses were formulated about the combination of inter-concept similarities, the separation of the MeSH hierarchy, and the semantic distances represented by the hierarchy edges.

Two methods were used for the combination of inter-concept similarities. The first, called all-combination, derived the similarity between queries and documents from the average of all possible inter-concept similarities. The second, called best-match-combination, derived the similarity between queries and documents from the best matches amongst inter-concept similarities. Our intuition was that best-match-combination would perform better than all-combination when hierarchy information was introduced at comparison time, and that all-combination would perform better than best-match-combination when hierarchy information was introduced at extension time.

Hypotheses on the separation of the MeSH hierarchy aimed at restraining the use of the hierarchy for concept comparison in order to reduce noise and increase precision. The intuition was that the main parts of the hierarchy correspond to concepts that are too different to be compared. Three approach were used. The first, called HardSep, did not allow comparison of concepts located in different MeSH categories (descriptor categories and the additional artificial "*qualifier*" category). The second, called SoftSep, used relevance information on genomic topics to determine a list of un-relevant MeSH categories. Concepts from the un-relevant categories could not be compared to concepts from other categories. The third, called DescQualSep, did not allow comparison between descriptors and qualifiers.

Hypotheses on the semantic distances represented by hierarchy edges were based on the intuition that semantic distances are negatively correlated with conceptual specificity and density. This intuition led to the calculation of hierarchy edge distances either with hierarchy information (depth and density), or with corpus information (concept distribution). The calculation of edge distance, as opposed to an approach assuming constant edge distance, was expected to reduce noise and increase retrieval precision.

7.3 Evaluations

Our hypotheses were evaluated in the context of ad hoc MEDLINE document retrieval (see Chapter 5), and in the context of MEDLINE document classification (see Chapter 6). Table 7.2 gives an overview of our evaluation framework.

The ad hoc retrieval evaluation used the TrecGen05 collection from the TREC

Table 7.2: Evaluation overview

<table>
<tr><th colspan="2">Ad hoc retrieval</th></tr>
<tr><th>Task</th><th>Experimental set-up</th></tr>
<tr><td>TrecGen05 collection:
- 50 topics.
- 4.5 million MEDLINE records
- Relevance judgments.</td><td>Post-retrieval combination:
- Inspired by Srinivasan (1996a), Kraaij et al. (2004).
- Text-based search with Terrier search engine.
- Relevance judgments.
- Fusion.</td></tr>
<tr><th colspan="2">Binary classification</th></tr>
<tr><th>Task</th><th>Experimental set-up</th></tr>
<tr><td>TREC Genomics Track GO triage task:
- 5,837 training documents.
- 6,043 test documents.
- Relevance information.</td><td>SVM^{light} (Joachims, 1999):
- Standard settings.</td></tr>
</table>

2005 Genomics track ad hoc task. The collection included 50 topics with their associated relevance judgements, and a subset of MEDLINE comprising 4,591,008 records.

The document classification evaluation used training and test documents with their respective relevance judgements from the TREC 2005 Genomics track GO (Gene Ontology) triage task. The triage task simulated one of the activities of the curators of the Mouse Genome Informatics (MGI) group (Eppig et al., 2005). MGI curators manually select biomedical documents that are likely to give experimental evidence for the annotation of mouse genes with one or more GO concepts.

7.3.1 Experimental Set-ups

For the ad hoc task, our hypotheses were evaluated with a post-retrieval combination of text and MeSH searches (see Figure 5.1). First, the text queries were derived from the 50 topics provided by the task. Second, the text queries were sent to the Terrier search engine[1]. The output of the text searches was used to generate MeSH queries by pseudo-relevance feedback. For each query, the top 5000 documents retrieved

[1] http://ir.dcs.gla.ac.uk/terrier/

with the text search were given MeSH scores derived from the comparison of their MeSH-based vectors with the MeSH query vector. The text and MeSH scores were then combined (Equation 3.8), re-ranked, and the top 1000 were kept as the final output for the combined text and MeSH searches.

For the GO triage task, we used the SVM^{light} software, an implementation of the Support Vector Machine method implemented by Joachims (1999). We used the default settings for the learning module (svm_learn) of SVM^{light}, as the automatic tuning of the software was shown to perform well for text categorization (Joachims, 1998). Default settings included the use of a linear kernel, and the automatic determination of parameter C (trade-off between the training error and the classifier complexity). However, parameter j (cost-factor by which training errors on positive examples out-weight errors on negative examples) was set to 11, to reflect the bias of the utility function (see Equation 3.11, Section 3.3.1).

7.3.2 Results

Non-hierarchical Hypotheses

Tables 5.20 and 5.26 show the results for the evaluation our non-hierarchical hypotheses with the ad hoc task. Table 6.1 show the results for the evaluation our non-hierarchical hypotheses with the classification task.

First, we examined the results of the methods aiming at weighting of concepts with corpus information (TF*IDF), and with MeSH field information (MajMin). Results showed that the TF*IDF method had a negative impact on precision for the ad hoc task, and a limited impact on precision for the classification task. This is surprising as the distribution of free-text terms in a corpus is generally useful to discriminate between relevant terms and less relevant ones. For MajMin, results showed that increasing the discrimination leads to lower precision for the ad hoc and classification task. This is also counter-intuitive as giving more importance to the central themes of documents was expected to boost precision. We concluded that the minor themes might play an important part in distinguishing documents from each other.

Second, we looked at the results of the DescQual methods that evaluated the impact of the contextual information contained in the MeSH fields. Keeping associations between descriptors and qualifiers had a moderate negative impact on precision for the ad hoc task, and a moderate positive impact on precision for the classification task. This suggests that this approach is useful for classification but less so for ad hoc retrieval.

For both the ad hoc and classification tasks, the baseline binary representation gave the best results. The best performance for the MeSH-only searches and the combined text and MeSH searches was obtained with the binary representation. Moreover, the binary representation led to the best normalized utility for the classification task.

Hierarchical Hypotheses

Tables 6.3 and 6.4 show the results for the all-combination and best-match-combination methods, respectively, for the classification task. Table 5.29 shows the results of both methods for the ad hoc task. The results support our hypothesis, which stated that all-combination performs better than best-match-combination for hierarchy integration at extension time, and that best-match-combination performs better than all-combination for hierarchy integration at comparison time.

Tables 5.31 and 6.6 show the results of the three hierarchy separation methods, HardSep, SoftSep, and DescQualSep, for the ad hoc and classification tasks, respectively. For the classification task, the results support our hypothesis, as all three separation methods increased precision. The increase in precision was highest when all MeSH categories were separated (HardSep). When only some categories judged non-relevant to genomic topics were separated (SoftSep), the increase in precision was moderate. In contrast, the results for the ad hoc task provided little support for our hypothesis regarding hierarchy separation. All three methods significantly decreased both precision and recall for the MeSH-only searches. The performance was damaged the most when all the MeSH categories were separated.

Tables 5.33 and 5.34 show the results of the DepthDens method for the ad hoc task. Tables 6.8, 6.9 and 6.10 show the results of the DepthDens method for the classification task. For the ad hoc task, moderate sensitivity to hierarchy density ($\beta = 0.75,\ 0.5$) slightly improved the precision of the MeSH-only searches. For the classification task, increasing the sensitivity to hierarchy depth ($\alpha = 1,\ 2$) increased precision and decreased recall. The last two observations were in line with our hypothesis regarding the calculation of edge distances in the hierarchy. However, increased sensitivity to hierarchy density decreased precision and increased recall for the classification task. Additionally, increased sensitivity to depth decreased the precision of the MeSH-only searches. The last two observations were in contradiction with our hypothesis. Overall, hierarchy information, such as depth and density, was shown not to be reliable for the calculation of hierarchy edge distances in search for higher retrieval precision.

Tables 5.35 and 6.11 show the results of the InfoBased method for the ad hoc and classification tasks, respectively. For the ad hoc task, InfoBased had a significant positive impact on both the precision and the recall of the MeSH-only searches. However, Infobased had no impact on performance for the classification task in comparison with the baseline, a method assuming constant hierarchy distance. The first observation suggested that corpus information is useful to calculate edge distances that favor retrieval precision, but the second suggested that assuming constant hierarchy edge distance is a good approximation of the edge distances calculated with corpus information by InfoBased.

Our main hierarchical hypothesis was that the introduction of the hierarchy would boost recall. For the classification task, the results (see Table 6.3) confirmed that using the hierarchy to extend the document representations had a positive impact on recall with the automatic tuning of the SVM^{light} classifier. For the ad hoc task however, no evidence was obtained to support our main hierarchical hypothesis.

Results (see Table 5.36) showed that even the best-performing method integrating the hierarchy, InfoBased, had little impact on the recall of a baseline method, using the Cosine measure to compare binary representations, for the MeSH-only searches. Furthermore, after combination with text searches, InfoBased yielded a significantly lower recall than the baseline (see Table 5.37).

7.4 Future Directions

7.4.1 Hierarchy Integration

In this dissertation, two methods were used to evaluate the integration of the hierarchy: integration at comparison time, and integration at extension time (see Section 4.2.1). The first method was evaluated with ad hoc retrieval, whereas the second was evaluated with document classification. Hierarchy integration was beneficial for classification but not for ad hoc retrieval, which suggests that the two integration methods need to be compared. In future work, we plan to evaluate the two methods in the same evaluation framework.

7.4.2 Fusion Effectiveness

Results showed that the effectiveness of the combination of text and MeSH searches vary according to the MeSH representation used (Tables 5.20, 5.26, 5.35, 5.36, and 5.37). Several hypotheses exist to determine the conditions leading to a successful fusion (Lee, 1997; Beitzel et al., 2004). In future work, we want to examine these hypotheses in the context of the post-retrieval text and MeSH combination method (see Section 5.1.3).

7.4.3 Growing Use of Full Text

This dissertation focused on the biomedical literature available in the MEDLINE format: the free-text is limited to titles and abstracts, although some structured information is also provided (MeSH fields, for example). The use of the MEDLINE format was motivated by the use of MEDLINE as a entry point to biomedical information, despite the growing availability of full-text article on the Web (Hersh et al., 2004). Nonetheless, the increase of full-text availability can not be ignored and the benefits of using full text for biomedical information retrieval is of great interest.

Cohen et al. (2005) already examined the impact of the use of free-text on binary classification. Free-text did not improve on the use of titles and abstract for the particular GO triage task used in this dissertation. However, it was shown to be beneficial for other binary classification task, provided that text processing techniques such as stemming and stop-words removal were used.

In 2006, the TREC Genomics[2] track moved towards the use of full-text, reflecting the interest of the biomedical information retrieval research community. In particular, the task involved the retrieval of short passages relevant to 28 topics from a collection of 162,259 full-text articles. Passage retrieval proved to be a challenge for the 2006 participants, but better results were obtained for document retrieval (Hersh et al., 2006). For the 2007 task, the same collection of full-text articles is used with new topics.

7.4.4 MeSH hierarchy

Some of our hierarchical hypotheses included separating the combined MeSH hierarchy for the comparison of concepts (see Section 4.2.2). The idea was to discriminate between categories, either assuming that they correspond to separate orthogonal conceptual areas (HardStop, DescQualSep), or by evaluating their level of relevance to our topics of interest (SoftStop). In the future we want to evaluate the relevance of each category by gradually increasing the distance between a category and the others. Similarly to the way an artificial "*qualifier*" node was added to combine the shallow qualifier categories before integrating them to the combined MeSH hierarchy, additional nodes could be added between the root node of a category and the

[2] http://ir.ohsu.edu/genomics/, last accessed on 24 May 2007.

artificial "*MeSH*" root node.

The inter-concept semantic measures used in this dissertation relied on the shortest path between two concepts in the hierarchy. This path was determined by the existence of a common ancestor node for the concepts (see Section 2.2.1). The closer the common ancestor, the shorter the path between two concepts will be. As concepts in the MeSH hierarchy can have several parents, concepts can also have common descendants. In future work we will investigate the use of common descendants to determine the shortest path between two concepts.

The use of corpus information to determine the hierarchy edge distances (InfoBased) had a positive impact on the performance of the MeSH-only ad hoc retrieval (Table 5.35). However, no impact was observed when InfoBased was used for document classification (Table 6.11). This suggested a possible connection between the concept frequencies in the corpus and the structure and building of the MeSH hierarchy. Weinberg and Cunningham (1985) shows that the process of adding more specific descriptors, although based on human intuition, depends on statistical observation of the biomedical literature and its search. Specifically, the high collection frequency of a MeSH concept triggers a process that can lead to the addition of a more specific child concept. Whether this trend would be confirmed by more recent data is an interesting subject of investigation.

7.4.5 Topic Identification

An important part of information retrieval is the identification of topics in the literature. The idea is to filter the overwhelming mass of information to select documents which are broadly relevant to one or several topics of interest. It can also be motivated by the creation of new hypotheses based on the connections discovered between several topics (Hearst, 1999; Shatkay and Feldman, 2003). Previous work has already involved the use of MeSH-based representations to identify topics in the literature (Struble and Dharmanolla, 2004; Srinivasan, 2001; Srinivasan and Rindflesch, 2002; Srinivasan, 2003). In the future we plan to evaluate the hypotheses and methods presented in this dissertation in the context of topic identification in the biomedical literature.

7.5 Papers Published

- Fabrice Camous, Stephen Blott, and Alan F. Smeaton. Ontology-based MEDLINE document Classification. To be published in proceedings of the first International Conference on Bioinformatics Research and Development (BIRD) 2007, Berlin, Germany.

- Fabrice Camous, Stephen Blott, and Alan F. Smeaton (2006). On Combining MeSH and Text Searches to Improve the Retrieval of MEDLINE documents. In proceedings of the third Confrence en Recherche d'Infomations et Applications (CORIA) 2006, Lyon, France.

- Fabrice Camous, Stephen Blott, Cathal Gurrin, Gareth J. F. Jones, and Alan F. Smeaton (2005). Structural Term Extraction for Expansion of Template-based Genomic Queries. In proceedings of the Fourteenth Text Retrieval conference (TREC 2005), Gaithersburg, MA.

- Stephen Blott, Fabrice Camous, Cathal Gurrin, Gareth J. F. Jones and Alan F. Smeaton (2005). On the use of Clustering and the MeSH Controlled Vocabulary to Improve MEDLINE Abstract Search. In proceedings of the second Confrence en Recherche d'Infomations et Applications (CORIA) 2005, Grenoble, France.

- Stephen Blott, Oisin Boydell, Fabrice Camous, Paul Ferguson, Georgina Gaughan, Cathal Gurrin, Noel Murphy, Noel O'Connor, Alan F. Smeaton, Barry Smyth, and Peter Wilkins (2004). Experiments in Terabyte Searching, Genomic Retrieval and Novelty Detection for TREC 2004. In proceedings of the Thirteenth Text Retrieval conference (TREC 2004), Gaithersburg, MA.

- Stephen Blott, Fabrice Camous, Cathal Gurrin, Gareth J. F. Jones and Alan F. Smeaton (2004). GenIRL: Genomic Information Retrieval using links. In proceedings of SIGIR 2004 Search and Discovery in Bioinformatics Workshop, Sheffield, UK.

Bibliography

Abdou, S., Savoy, J., and Ruck, P. (2005). Evaluation of Stemming, Query Expansion and Manual Indexing Approaches for the Genomic Task. In *Proceedings*, Gaithersburg, Maryland. The Fourteenth Text Retrieval Conference (TREC 2005).

Amati, G. (2003). *Probability Models for Information Retrieval based on Divergence from Randomness.* PhD thesis, Department of Computing Science, University of Glasgow, Glasgow, UK.

Amati, G. and Van Rijsbergen, C. J. (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.*, 20(4):357–389.

Ando, R., Dredze, M., and Zhang, T. (2005). TREC 2005 Genomics Track Experiments at IBM Watson. In *Proceedings*, Gaithersburg, Maryland. The Fourteenth Text Retrieval Conference (TREC 2005).

Aronson, A. R. (2001). Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. In *Proceedings*, Washington, DC. The 2001 AMIA Annual Symposium.

Aronson, A. R., Demner-Fushman, D., Humphrey, S., J. Lin, P. R., Ruiz, M., Smith, L., L.K.Tanabe, Wilbur, W., Demner-Fushman, D., Lin, J., and Liu, H. (2005). Fusion of Knowledge-Intensive and Statistical Approaches for Retrieving and Annotating Textual Genomics Documents. In *Proceedings*, Gaithersburg, Maryland. The Fourteenth Text Retrieval Conference (TREC 2005).

Aronson, A. R., Mork, J. G., Gay, C. W., Humphrey, S. M., and Rogers, W. J. (2004). The NLM indexing initiative's medical text indexer. In Fieschi, M., Coiera, E., and Li, Y.-C., editors, *Medinfo 2004*, pages 268–272, San Francisco, California. IMIA, IOS Press.

Aronson, A. R. and Rindflesch, T. C. (1997). Query Expansion Using the UMLS Metathesaurus. In *Proceedings*, pages 485–89, Nashville, TN. The 1997 AMIA Annual Fall Symposium.

Azuaje, F., Wang, H., and Bodenreider, O. (2005). Ontology-driven similarity approaches to supporting gene functional assessment. In *Proceedings*, pages 9–10, Detroit, Michigan. the ISMB'2005 SIG meeting on Bio-ontologies 2005.

Beitzel, S. M., Jensen, E. C., Chowdhury, A., Grossman, D., Frieder, O., and Goharian, N. (2004). On Fusion of Effective Retrieval Strategies in the Same Information Retrieval System. *Journal of the American Society of Information Science and Technology*, 55(10):859–868.

Belkin, N. J., Kantor, P., Fox, E. A., and Shaw, J. A. (1995). Combining the Evidence of Multiple Query Representations for Information Retrieval. *Information Processing and Management*, 31(3):431–448.

Blagosklonny, M. V. and Pardee, A. B. (2002). Conceptual biology: Unearthing the gems. *Nature*, 416(6879):373.

Brown, P. O. and Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays. *Nature Genetics*, 21:33–37.

Budanitsky, A. (1999). Lexical Semantic Relatedness and its Application in Natural Language Processing. Technical Report CSRG-390, Department of Computer Science, University of Toronto.

Burges, C. J. C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167.

Buttcher, S., Clarke, C. L. A., and Cormack, G. V. (2004). Domain-Specific Synonym Expansion and Validation for Biomedical Information Retrieval. In *Proceedings*, Gaithersburg, Maryland. The Thirteenth Text Retrieval Conference (TREC 2004).

Caviedes, J. E. and Cimino, J. J. (2004). Towards the development of a conceptual distance metric for the UMLS. *Journal of Biomedical Informatics*, 37(2):77–85.

Cohen, A. (2005). Unsupervised gene/protein named entity normalization using automatically extracted dictionaries. In *Proceedings*, Detroit, MI. BioLINK 2005 Workshop.

Cohen, A., Bhuptiraju, R., and Hersh, W. (2004). Feature generation, feature selection, classifiers, and conceptual drift for biomedical document triage. In *Proceedings*, Gaithersburg, Maryland. The Thirteenth Text Retrieval Conference (TREC 2004).

Cohen, A., Yang, J., and Hersh, W. (2005). A comparison of techniques for classification and ad hoc retrieval of biomedical documents. In *Proceedings*, Gaithersburg, Maryland. The Fourteenth Text Retrieval Conference (TREC 2005).

Darwish, K. and Madkour, A. (2004). The GUC goes to TREC 2004: using whole or partial documents for retrieval and classification in the Genomics Track. In *Proceedings*, Gaithersburg, Maryland. The Thirteenth Text Retrieval Conference (TREC 2004).

Dayanik, A., Fradkin, D., Genkin, A., Kantor, P., Lewis, D. D., Madigan, D., and Menkov, V. (2004). DIMACS at the TREC 2004 Genomics Track. In *Proceedings*, Gaithersburg, Maryland. The Thirteenth Text Retrieval Conference (TREC 2004).

Devitt, A. and Vogel, C. (2004). The Topology of WordNet: Some Metrics. In *Proceedings*, pages 106–111, Brno, Czech Republic. Second International WordNet Conference.

Eppig, J. T., Bult, C. J., Kadin, J. A., Richardson, J. E., and Blake, J. A. (2005). The Mouse Genome Database (MGD): from genes to mice—a community resource for mouse biology. *Nucleic Acids Res*, 33:471–475.

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

Fox, C. J. (1992). Lexical analysis and stoplists. In *Information Retrieval: Data Structures & Algorithms*, pages 102–130.

Frakes, W. B. (1992). Stemming algorithms. In *Information Retrieval: Data Structures & Algorithms*, pages 131–160.

Fujita, S. (2004). Revisiting again document length hypotheses - TREC 2004 Genomics Track experiments at Patolis. In *Proceedings*, Gaithersburg, Maryland. The Thirteenth Text Retrieval Conference (TREC 2004).

Funk, M. E., Reid, C. A., and McGoogan, L. S. (1983). Indexing consistency in MEDLINE. *Bull Med Libr Assoc.*, 71(2):176–183.

Ganesan, P., Garcia-Molina, H., and Widom, J. (2002). Exploiting Hierarchical Domain Structure to Compute Similarity. Technical report, Stanford University. available at http://dbpubs.stanford.edu/pub/2001-27.

Gene Ontology Consortium (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29.

Gene Ontology Consortium (2001). Creating the gene ontology resource: design and implementation. *Genome Research*, 11(8):1425–33.

Gene Ontology Consortium (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, 32. Database issue D258-D261.

Hearst, M. A. (1999). Untangling Text Data Mining. In *Proceedings*, University of Maryland. ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics.

Hersh, W., Buckley, C., Leone, T. J., and Hickam, D. (1994a). OHSUMED: an interactive retrieval evaluation and new large test collection for research. In *Proceedings*, pages 192–201, Dublin, Ireland. The 17th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '94).

Hersh, W., Cohen, A. M., Roberts, P., and Rekapalli, H. K. (2006). Trec 2006 genomics track overview. In *Proceedings*, Gaithersburg, Maryland. The Fifteenth Text Retrieval Conference (TREC 2006).

Hersh, W. R., Bhuptiraju, R. T., Ross, L., Johnson, P., Cohen, A. M., and Kraemer, D. F. (2004). TREC 2004 Genomics Track Overview. In *Proceedings*, Gaithersburg, Maryland. The Thirteenth Text Retrieval Conference (TREC 2004).

Hersh, W. R., Cohen, A. M., Yang, J., Bhuptiraju, R. T., Roberts, P. M., and Hearst, M. A. (2005). TREC 2005 Genomics Track Overview. In *Proceedings*, Gaithersburg, Maryland. The Fourteenth Text Retrieval Conference (TREC 2005).

Hersh, W. R., Hickam, D. H., Haynes, R. B., and McKibbon, K. A. (1994b). A performance and failure analysis of SAPHIRE with a MEDLINE test collection. *Journal of the American Medical Informatics Association*, 1(1):51–60.

Huang, X., Zhong, M., and Si, L. (2005). York University at TREC 2005: Genomics Track. In *Proceedings*, Gaithersburg, Maryland. The Fourteenth Text Retrieval Conference (TREC 2005).

Ide, E. (1971). *New experiments in relevance feedback*, chapter 16, pages 337–354. Prentice-Hall, Englewood Cliffs, NJ.

Jain, A. K. and Dubes, R. C. (1988). *Algorithms for clustering data.* Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Jain, A. K. and Murty, M. N. (1999). Data Clustering: A Review. *ACM Computing Surveys*, 3(31):264–323.

Jenssen, T., Laegreid, A., Komorowski, J., and Hovig, E. (2001). A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, 28(1):21–28.

Jiang, J. J. and Conrath, D. W. (1997). Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In *Proceedings*, Taipei, Taiwan. ROCLING X.

Joachims, T. (1998). Text categorization with support vector machines: learning with many relevant features. In *Proceedings*, pages 137–142, Chemnitz, DE.

Joachims, T. (1999). Making large-scale support vector machine learning practical. In *Advances in kernel methods: support vector learning*, pages 169–184. MIT Press, Cambridge, MA, USA.

Kraaij, W. (2004). *Variations on Language Modeling for Information Retrieval.* PhD thesis, University of Twente, Twente, Netherlands.

Kraaij, W. and Pohlmann, R. (1996). Viewing Stemming as Recall Enhancement. In *Proceedings*, pages 40–48, Zurich, Switzerland. The 19th International Conference on Research and Development in Information Retrieval (SIGIR '96).

Kraaij, W., Weeber, M., Raaijmakers, S., and Jelier, R. (2004). MeSH based feedback, concept recognition and stacked classification for curation tasks. In *Proceedings*, Gaithersburg, Maryland. The Thirteenth Text Retrieval Conference (TREC 2004).

Lee, C., Hou, W.-J., and Chen, H.-H. (2005). Identifying Relevant Full-Text Articles for Database Curation. In *Proceedings*, Gaithersburg, Maryland. The Fourteenth Text Retrieval Conference (TREC 2005).

Lee, J. H. (1997). Analyses of multiple evidence combination. In *Proceedings*, pages 267–276, Philadelphia, PA. The 20th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '97).

Lin, D. (1998). An Information-Theoretic Definition of Similarity. In *Proceedings*, Madison, Wisconsin. The 15th International Conf. on Machine Learning (ICML '98).

Lord, P. W., Stevens, R. D., Brass, A., and Goble, C. A. (2003a). Investigating similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, 19(10):1275–83.

Lord, P. W., Stevens, R. D., Brass, A., and Goble, C. A. (2003b). Semantic similarity measures as tools for exploring the Gene Ontology. In *Proceedings*, Lihue, Hawaii. Pacific Symposium of Biocomputing.

Miller, G. A. (1990). Nouns in WordNet: A Lexical Inheritance System. *International Journal of Lexicography*, 3(4):245–264.

Nelson, S. J., Johnston, D., and Humphreys, B. L. (2001). *Relationships in Medical Subject Headings*, chapter 11, pages 171–184. Kluwer Academic Publishers, New York. Book title: Relationships in the organization of knowledge.

Niu, J., Sun, L., Lou, L., Deng, F., Lin, C., Zheng, H., and Huang, X. (2005). WIM at TREC 2005. In *Proceedings*, Gaithersburg, Maryland. The Fourteenth Text Retrieval Conference (TREC 2005).

Ontrup, J., Nattkemper, T., Gerstung, O., and Ritter, H. (2003). A MeSH Term based Distance Measure for Document Retrieval and Labeling Assistance. In *Proceedings*, Cancun, Mexico. 25th Annual Int. Conf. of the IEEE Engineering in Med. and Biol. Soc. (EMBC2003).

Pearson, H. (2001). Biology's Name Game. *Nature*, 411(6838):631–2.

Pedersen, T., Pakhomov, S., and Patwardhan, S. (2005). Measures of Semantic Similarity and Relatedness in the Medical Domain. Technical report, University of Minnesota Digital Technology Center. Research Report DTC 2005/12.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

Preiss, B. R. (1999). *Data structures and algorithms with object-oriented design patterns in C++*. John Wiley & Sons, Inc., New York, NY, USA.

Rada, R., Mili, H., Bicknell, E., and Blettner, M. (1989). Development and application of a metric on semantic nets. *IEEE Transaction on Systems, Man, and Cybernetics*, 19(1):17–30.

Resnik, P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings*, pages 448–453, Hyberabad, India. The 14th International Joint Conference on Artificial Intelligence.

Richardson, R. and Smeaton, A. F. (1995). Using Wordnet in a Knowledge-Based Approach to Information Retrieval. Working paper CA-0395, School of Computer Applications, Dublin City University, Dublin.

Robertson, S. and Sparck Jones, K. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27:129–46.

Robertson, S. E. and Sparck Jones, K. (1996). Simple, proven approaches to text retrieval. Technical report, Computer Laboratory, University of Cambridge. Technical report 356.

Robertson, S. E., Walker, S., Hancock-Beaulieu, M., and Gatford, M. (1996). Okapi at TREC-3. In *Proceedings*, Gaithersburg, Maryland. The Third Text Retrieval Conference (TREC 1996).

Rocchio, J. J. (1971). *Relevance feedback in information retrieval*, chapter 14, pages 313–323. Prentice-Hall, Englewood Cliffs, NJ.

Salton, G. (1971). *The SMART Retrieval System—Experiments in Automatic Document Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Salton, G. and Buckley, C. (1987). Term weighting approaches in automatic text retrieval. Technical report, Cornell University, NY.

Salton, G., Wong, A., and Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of ACM*, 18:613–620.

Schijvenaars, B., Schuemie, M., van Mulligen, E., Weeber, M., Jelier, R., Mons, B., Kors, J., and Kraaij, W. (2005). Notebook Paper TREC 2005 Genomics Track: A Concept-Based Approach to Text Categorization. In *Proceedings*, Gaithersburg, Maryland. The Fourteenth Text Retrieval Conference (TREC 2005).

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47.

Seki, K., Costello, J. C., Singan, V. R., and Mostafa, J. (2004). TREC 2004 Genomics Track experiments at IUB. In *Proceedings*, Gaithersburg, Maryland. The Thirteenth Text Retrieval Conference (TREC 2004).

Settles, B. (2004). Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets. In *Proceedings*, Geneva, Switzerland. The COLING 2004 International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA).

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 and 623–656.

Shatkay, H. and Feldman, R. (2003). Mining the Biomedical Literature in the Genomic Era: An Overview. *Journal of Computational Biology*, 10(6):821–55.

Shaw, J. A. and Fox, E. A. (1993). Combination of Multiple Searches. In *Proceedings*, Gaithersburg, Maryland. The Second Text Retrieval Conference (TREC 1993).

Shin, K. and Han, S.-Y. (2004). Improving information retrieval in medline by modulating mesh term weights. In *Proceedings*, pages 388–394, Manchester, UK. The 9th International Conference on Applications of Natural Languages to Information Systems, (NLDB 2004).

Shin, K., Han, S.-Y., Gelbukh, A., and Park, J. (2004). Advanced Relevance Feedback Query Expansion Strategy for Information Retrieval in MEDLINE. In *Lecture Notes in Computer Science*, pages 425–431, Puebla, Mexico. The 9th Iberoamerican Congress on Pattern Recognition, Springer-Verlag.

Si, L. and Kanungo, T. (2005). Thresholding Strategies for Text Classifiers: TREC 2005 Biomedical Triage Task Experiments. In *Proceedings*, Gaithersburg, Maryland. The Fourteenth Text Retrieval Conference (TREC 2005).

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

Srinivasan, P. (1996a). Optimal document-indexing vocabulary for MEDLINE. *Information Processing and Management*, 32(5):503–514.

Srinivasan, P. (1996b). Query Expansion and MEDLINE. *Information Processing and Management*, 32(4):431–443.

Srinivasan, P. (1996c). Retrieval Feedback in MEDLINE. *Journal of the American Medical Informatics Association*, 3(2):157–167.

Srinivasan, P. (2001). MeSHmap: a text mining tool for MEDLINE. In *Proceedings*, pages 642–646, Washington, DC. AMIA 2001 Symposium.

Srinivasan, P. (2003). Text mining: Generating hypotheses from MEDLINE. *Journal of the American Society for Information Science and Technology*, 55(5):396–413.

Srinivasan, P. and Rindflesch, T. (2002). Exploring Text Mining from MEDLINE. In *Proceedings*, pages 722–6, San Antonio, TX. AMIA 2002 Symposium.

Struble, C. A. and Dharmanolla, C. (2004). Clustering MeSH Representations of Biomedical Literature. In *Proceedings*, pages 41–47, Boston, MA. BioLINK 2004.

Subramaniam, L., Mukherjea, S., Kankar, P., Srivastava, B., Batra, V. S., Kamesam, P. V., and Kothari, R. (2003). Information extraction from biomedical literature: methodology, evaluation and an application. In *Proceedings*, pages 410–417, New Orleans, LA, USA. The twelfth international conference on Information and knowledge management.

Subramaniam, L., Mukherjea, S., and Punjani, D. (2005). Biomedical Document Triage: Automatic Classification Exploiting Category Specific Knowledge. In *Proceedings*, Gaithersburg, Maryland. The Fourteenth Text Retrieval Conference (TREC 2005).

Sussna, M. (1993). Word sense disambiguation for free-text indexing using a massive semantic network. In *Proceedings*, pages 67–74, Washington, D.C., United States. The second international conference on Information and knowledge management.

Vapnik, V. (1982). *Estimation of Dependences Based on Empirical Data: Springer Series in Statistics (Springer Series in Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA.

Wang, H., Azuaje, F., and Bodenreider, O. (2005). An ontology-driven clustering method for supporting gene expression analysis. In *Proceedings*, pages 389–394, Jordanstown, NI. 18th IEEE Symposium on Computer-Based Medical Systems,.

Wang, H., Azuaje, F., Bodenreider, O., and Dopazo, J. (2004). Expression correlation and gene ontology-based similarity: an assessment of quantitative relationships. In *Proceedings*, pages 25–31, La Jolla, CA, USA. IEEE 2004 Symposium on Computational Intelligence in Bioinformatics and Computational Biology.

Weinberg, B. H. and Cunningham, J. A. (1985). The Relationship Between Term Specificity in MeSH and Online Postings in MEDLINE. *Bulletin of Medical Library Association*, 73(4):365–372.

Wu, Z. and Palmer, M. (1994). Verb Semantics and lexical selection. In *Proceedings*, pages 133–8, Las Cruces, New Mexico. The 32nd Annual Meeting of the Association for Computational Linguistics.

Yang, Y. and Chute, C. G. (1993). Words or concepts: the features of indexing units and their optimal use in information retrieval. In *Proceedings*, pages 685–9. Proc Annu Symp Comput Appl Med Care.

Appendix A

Relevance Information for the Ad Hoc Task

Table A.1: Relevant documents for 2004 topics

topic	total judgements	non-relevant docs	relevant docs
1	879	800	79
2	1264	1163	101
3	1189	1008	181
4	1170	1140	30
5	1171	1147	24
6	787	693	94
7	730	615	115
8	938	777	161
9	593	478	115
10	1126	1122	4
11	742	631	111
12	810	554	256
13	1118	1094	24
14	948	927	21
15	1111	1021	90
16	1078	931	147
17	1150	1147	3
18	1392	1391	1
19	1135	1134	1
20	814	698	116
21	676	596	80
22	1085	875	210
23	915	757	158
24	952	926	26
25	1142	1110	32
26	792	745	47
27	755	726	29
28	836	823	13
29	756	713	43
30	1082	917	165
31	877	739	138
32	1107	611	496
33	812	748	64
34	778	747	31
35	717	446	271
36	676	422	254
37	476	327	149
38	1165	742	423
39	1350	1033	317
40	1168	891	277
41	880	298	582
42	1005	308	697
43	739	544	195
44	1224	575	649
45	1139	983	156
46	742	545	197
47	1450	1085	365
48	1121	966	155
49	1100	1027	73
50	1091	789	302
Total	48753	40485	8268

Table A.2: 2005 topics as *narratives*

topic	*narratives*
	Describe the procedure or methods for:
100	- how to "open up" a cell through a process called "electroporation."
101	- exact reactions that take place when you do glutathione S-transferase (GST) cleavage during affinity chromatography.
102	- different quantities of different components to use when pouring a gel to make it more or less porous.
103	- green fluorescent protein (GFP) tagged proteins to do experiments with tagged proteins.
104	- how to do a microsomal budding assay, i.e., budding of vesicles from microsomes in vitro.
105	- purification of rat IgM.
106	- chromatin IP (Immuno Precipitations) to isolate proteins that are bound to DNA in order to precipitate the proteins out of the DNA.
107	- normalization procedures that are used for microarray data.
108	- identifying in vivo protein-protein interactions in time and space in the living cell.
109	- fluorogenic 5'-nuclease assay.
	Provide information about the role of the gene:
110	- Interferon-beta in the disease Multiple Sclerosis.
111	- PRNP in the disease Mad Cow Disease.
112	- IDE gene in the disease Alzheimer's Disease.
113	- MMS2 in the disease Cancer.
114	- APC (adenomatous polyposis coli) in the disease Colon Cancer.
115	- Nurr-77 in the disease Parkinson's Disease.
116	- Insulin receptor gene in the disease Cancer.
117	- Apolipoprotein E (ApoE) in the disease Alzheimer's Disease.
118	- Transforming growth factor-beta1 (TGF-beta1) in the disease Cerebral Amyloid Angiopathy (CAA).
119	GSTM1 in the disease Breast Cancer.
	Provide information on the role of the gene:
120	- nucleoside diphosphate kinase (NM23) in the process of tumor progression.
121	- BARD1 in the process of BRCA1 regulation.
122	- APC (adenomatous polyposis coli) in the process of actin assembly.
123	- COP2 in the process of transport of CFTR out of the endoplasmic reticulum.
124	- casein kinase II in the process of ribosome assembly.
125	- Nurr-77 in the process of preventing auto-immunity by deleting reactive T-cells before they migrate to the spleen or the lymph nodes.
126	- P53 in the process of apoptosis.
127	- alpha7 nicotinic receptor subunit gene in the process of ethanol metabolism.
128	- gamma-aminobutyric acid receptors (GABABRs) in the process of inhibitory synaptic transmission.
129	- Interferon-beta in the process of viral entry into host cell.
	Provide information about the genes:
130	- BRCA1 regulation of ubiquitin in cancer.
131	- L1 and L2 in the HPV11 virus in the role of L2 in the viral capsid.
132	- APC (adenomatous polyposis coli) and wnt in colon cancer.
133	- phospholipase A2 (PLA2) and SAR1 in Endoplasmic reticulum transport (i.e. vesicle budding from the ER).
134	- CFTR and Sec61 in degradation of CFTRwhich leads to cystic fibrosis.
135	- Bop and Pes in cell growth.
136	- alpha7 nicotinic receptor gene and ApoE gene in the neurotoxic effects of ethanol.
137	- Insulin-like GF and insulin receptor gene in the function in skin.
138	- HNF4 and COUP-TF I in the suppression in the function of the liver.
139	- Ret and GDNF in kidney development.
	Provide information about:
140	- BRCA1 185delAG mutation and its/their role in ovarian cancer.
141	- Huntingtin mutations and its/their role in Huntington's Disease.
142	- Sonic hedgehog mutations and its/their role in developmental disorders.
143	- Mutations of NM23 and its/their impact on tracheal development.
144	- Mutations in metazoan Pes and its/their effect on cell growth.
145	- Mutations of hypocretin receptor 2 and its/their role in narcolepsy.
146	- Mutations of presenilin-1 gene and its/their biological impact in Alzheimer's disease.
147	- Mutations of alpha7 nAChR gene and its/their biological impact in alcoholism.
148	- Mutation of familial hemiplegic migraine type 1 (FHM1) and its/their neuronal Ca2+ influx in hippocampal neurons.
149	- Mutations of the alpha 4-GABAA receptor and its/their impact on behavior.

Table A.3: 2005 topics as *basic narratives*

topic	*narratives*
100	- how to "open up" a cell through a process called "electroporation."
101	- exact reactions that take place when you do glutathione S-transferase (GST) cleavage during affinity chromatography.
102	- different quantities of different components to use when pouring a gel to make it more or less porous.
103	- green fluorescent protein (GFP) tagged proteins to do experiments with tagged proteins.
104	- how to do a microsomal budding assay, i.e., budding of vesicles from microsomes in vitro.
105	- purification of rat IgM.
106	- chromatin IP (Immuno Precipitations) to isolate proteins that are bound to DNA in order to precipitate the proteins out of the DNA.
107	- normalization procedures that are used for microarray data.
108	- identifying in vivo protein-protein interactions in time and space in the living cell.
109	- fluorogenic 5'-nuclease assay.
110	- Interferon-beta Multiple Sclerosis.
111	- PRNP Mad Cow Disease.
112	- IDE gene Alzheimer's Disease.
113	- MMS2 Cancer.
114	- APC (adenomatous polyposis coli) Colon Cancer.
115	- Nurr-77 Parkinson's Disease.
116	- Insulin receptor gene Cancer.
117	- Apolipoprotein E (ApoE) Alzheimer's Disease.
118	- Transforming growth factor-beta1 (TGF-beta1) Cerebral Amyloid Angiopathy (CAA).
119	GSTM1 in the disease Breast Cancer.
120	- nucleoside diphosphate kinase (NM23) in the process of tumor progression.
121	- BARD1 in the process of BRCA1 regulation.
122	- APC (adenomatous polyposis coli) in the process of actin assembly.
123	- COP2 in the process of transport of CFTR out of the endoplasmic reticulum.
124	- casein kinase II in the process of ribosome assembly.
125	- Nurr-77 in the process of preventing auto-immunity by deleting reactive T-cells before they migrate to the spleen or the lymph nodes.
126	- P53 in the process of apoptosis.
127	- alpha7 nicotinic receptor subunit gene in the process of ethanol metabolism.
128	- gamma-aminobutyric acid receptors (GABABRs) in the process of inhibitory synaptic transmission.
129	- Interferon-beta in the process of viral entry into host cell.
130	- BRCA1 regulation of ubiquitin in cancer.
131	- L1 and L2 in the HPV11 virus in the role of L2 in the viral capsid.
132	- APC (adenomatous polyposis coli) and wnt in colon cancer.
133	- phospholipase A2 (PLA2) and SAR1 in Endoplasmic reticulum transport (i.e. vesicle budding from the ER).
134	- CFTR and Sec61 in degradation of CFTRwhich leads to cystic fibrosis.
135	- Bop and Pes in cell growth.
136	- alpha7 nicotinic receptor gene and ApoE gene in the neurotoxic effects of ethanol.
137	- Insulin-like GF and insulin receptor gene in the function in skin.
138	- HNF4 and COUP-TF I in the suppression in the function of the liver.
139	- Ret and GDNF in kidney development.
140	- BRCA1 185delAG mutation role in ovarian cancer.
141	- Huntingtin mutations role in Huntington's Disease.
142	- Sonic hedgehog mutations role in developmental disorders.
143	- Mutations of NM23 impact on tracheal development.
144	- Mutations in metazoan Pes effect on cell growth.
145	- Mutations of hypocretin receptor 2 role in narcolepsy.
146	- Mutations of presenilin-1 gene biological impact in Alzheimer's disease.
147	- Mutations of alpha7 nAChR gene biological impact in alcoholism.
148	- Mutation of familial hemiplegic migraine type 1 (FHM1) neuronal Ca2+ influx in hippocampal neurons.
149	- Mutations of the alpha 4-GABAA receptor impact on behavior.

Table A.4: Relevant documents for 2005 topics

topic	total judgements	non-relevant docs	relevant docs
100	704	630	74
101	651	631	20
102	1164	1154	10
103	701	676	25
104	629	625	4
105	1133	1044	89
106	1230	1061	169
107	484	294	190
108	1092	889	203
109	389	210	179
110	934	918	16
111	675	473	202
112	870	859	11
113	1356	1342	14
114	754	375	379
115	1350	1335	15
116	1265	1179	86
117	1094	385	709
118	937	905	32
119	589	528	61
120	527	182	345
121	422	380	42
122	871	815	56
123	1029	992	37
124	752	691	61
125	1202	1191	11
126	1320	1013	307
127	841	837	4
128	954	880	74
129	987	949	38
130	813	781	32
131	431	389	42
132	531	501	30
133	523	518	5
134	732	721	11
135	1057	1057	0
136	853	850	3
137	1129	1078	51
138	501	489	12
139	380	345	35
140	395	366	29
141	520	439	81
142	528	257	271
143	902	898	4
144	1212	1210	2
145	288	256	32
146	825	388	437
147	659	649	10
148	536	525	11
149	1294	1271	23
Total	41015	36431	4584

Appendix B

Randomization Tests for the Ad Hoc Task

Table B.1: InL2 runs, basic narratives vs. narratives. Number of iteration: 100000, significance level: 0.05

c values	Significant Difference
0.5	Basic narratives > Narratives
1.0	Basic narratives > Narratives
1.5	Basic narratives > Narratives
2.0	Basic narratives > Narratives
2.5	Basic narratives > Narratives
3.0	Basic narratives > Narratives
3.5	Basic narratives > Narratives
4.0	Basic narratives > Narratives
4.5	Basic narratives > Narratives
5.0	Basic narratives > Narratives
5.5	Basic narratives > Narratives
6.0	Basic narratives > Narratives

Table B.2: DRF runs, basic narratives vs. narratives. Number of iteration: 100000, significance level: 0.05

c values	Significant Difference
0.5	none
1.0	none
1.5	Basic narratives > Narratives
2.0	Basic narratives > Narratives
2.5	Basic narratives > Narratives
3.0	Basic narratives > Narratives
3.5	Basic narratives > Narratives
4.0	Basic narratives > Narratives
4.5	Basic narratives > Narratives
5.0	Basic narratives > Narratives
5.5	Basic narratives > Narratives
6.0	Basic narratives > Narratives

Table B.3: TFIDF and BM25 runs, basic narratives vs. narratives. Number of iteration: 100000, significance level: 0.05

Model	Significant Difference
TFIDF	none
BM25	Basic narratives > Narratives

Table B.4: BB2 runs, basic narratives vs. narratives. Number of iteration: 100000, significance level: 0.05

c values	Significant Difference
0.5	none
1.0	none
1.5	none
2.0	none
2.5	none
3.0	none
3.5	none
4.0	none
4.5	none
5.0	none
5.5	none
6.0	none

Table B.5: InL2 vs. TFIDF and BM25 (basic narratives. Number of iteration: 100000, significance level: 0.05

c values	Significant Difference
0.5	BM25 > InL2 TFIDF > InL2
1.0	InL2 > BM25 InL2 > TFIDF
1.5	InL2 > BM25 InL2 > TFIDF
2.0	InL2 > BM25 InL2 > TFIDF
2.5	InL2 > BM25 InL2 > TFIDF
3.0	InL2 > BM25 InL2 > TFIDF
3.5	InL2 > BM25 InL2 > TFIDF
4.0	InL2 > BM25
4.5	none
5.0	none
5.5	none
6.0	none

Table B.6: DFR BM25 vs. TFIDF and BM25 (basic narratives. Number of iteration: 100000, significance level: 0.05

c values	Significant Difference
0.5	BM25 > DFR BM25
	TFIDF > DFR BM25
1.0	DFR BM25 > BM25
	DFR BM25 > TFIDF
1.5	DFR BM25 > BM25
	DFR BM25 > TFIDF
2.0	DFR BM25 > BM25
	DFR BM25 > TFIDF
2.5	DFR BM25 > BM25
	DFR BM25 > TFIDF
3.0	DFR BM25 > BM25
	DFR BM25 > TFIDF
3.5	DFR BM25 > BM25
	DFR BM25 > TFIDF
4.0	DFR BM25 > BM25
	DFR BM25 > TFIDF
5.0	DFR BM25 > BM25
	DFR BM25 > TFIDF
5.5	DFR BM25 > BM25
6.0	none

Table B.7: BB2 vs. TFIDF and BM25 (basic narratives. Number of iteration: 100000, significance level: 0.05

c values	Significant Difference
0.5	BM25 > BB2
	TFIDF > BB2
1.0	none
1.5	none
2.0	BB2 > BM25
	BB2 > TFIDF
2.5	BB2 > BM25
	BB2 > TFIDF
3.0	BB2 > BM25
	BB2 > TFIDF
3.5	BB2 > BM25
	BB2 > TFIDF
4.0	BB2 > BM25
	BB2 > TFIDF
5.0	BB2 > BM25
	BB2 > TFIDF
5.5	BB2 > BM25
	BB2 > TFIDF
6.0	BB2 > BM25
	BB2 > TFIDF

Table B.8: Comparison of BB2, InL2, DRF BM25 models with basic narratives. Number of iteration: 100000, significance level: 0.05

c values	Significant Difference
0.5	InL2 > DFR BM25
1.0	InL2 > DFR BM25
1.5	InL2 > DFR BM25
2.0	InL2 > DFR BM25
2.5	none
3.0	none
3.5	none
4.0	none
4.5	none
5.0	none
5.5	DFR BM25 > InL2
6.0	DFR BM25 > InL2

Table B.9: MeSH representations (av. P@10). Number of iteration: 100000, significance level: 0.05

Significant Difference
binary > TFIDF
binary > majMin4
binary > majMin5
majMin2 > majMin3
majMin2 > majMin4
majMin2 > majMin5

Table B.10: MeSH representations (MAP). Number of iteration: 100000, significance level: 0.05

Significant Difference
binary > TFIDF
majMin2 > TFIDF
majMin2 > majMin3
majMin2 > majMin4
majMin2 > majMin5
majMin3 > majMin4
majMin3 > majMin5
majMin4 > majMin5

Table B.11: MeSH representations (av. recall). Number of iteration: 100000, significance level: 0.05

Significant Difference
majMin2 > majMin4
majMin2 > majMin5
majMin3 > majMin4
majMin3 > majMin5
majMin4 > majMin5

Table B.12: MeSH representations (av. P@10). MeSH queries generated from rel. docs. Number of iteration: 100000, significance level: 0.05

Significant Difference
binary > majMin5
descQual > TFIDF
majMin2 > majMin4
majMin2 > majMin5
majMin3 > majMin4
majMin3 > majMin5
descQual > majMin4
descQual > majMin5

Table B.13: MeSH representations (MAP). MeSH queries generated from rel. docs. Number of iteration: 100000, significance level: 0.05

Significant Difference
binary > TFIDF
binary > majMin4
binary > majMin5
descQual > binary
descQual > TFIDF
majMin2 > majMin3
majMin2 > majMin4
majMin2 > majMin5
majMin3 > majMin4
majMin3 > majMin5
descQual > majMin3
majMin4 > majMin5
descQual > majMin4
descQual > majMin5

Table B.14: MeSH representations (av. recall). MeSH queries generated from rel. docs. Number of iteration: 100000, significance level: 0.05

Significant Difference
binary > TFIDF
majMin2 > TFIDF
majMin3 > TFIDF
majMin2 > majMin3
majMin2 > majMin4
majMin2 > majMin5
majMin3 > majMin4
majMin3 > majMin5

Table B.15: Text alone and text+MeSH combinations (av. P@10). Number of iteration: 100000, significance level: 0.05

Significant Difference
text+TFIDF > text
text+binary > text
text+TFIDF > text+descQual
text+binary > text+descQual
text+binary > text+majMin2
text+binary > text+majMin3
text+binary > text+majMin4

Table B.16: Text alone and text+MeSH combinations (MAP). Number of iteration: 100000, significance level: 0.05

Significant Difference
text+binary > text
text+majMin2 > text
text+binary > text+TFIDF
text+binary > text+majMin3
text+binary > text+majMin4
text+binary > text+majMin5
text+binary > text+descQual
text+majMin2 > text+TFIDF
text+majMin2 > text+majMin3
text+majMin2 > text+majMin4
text+majMin2 > text+majMin5
text+majMin2 > text+descQual
text+majMin3 > text+majMin4
text+majMin3 > text+majMin5
text+majMin4 > text+majMin5

Table B.17: Text alone and text+MeSH combinations (av. recall). Num. of iteration: 100000, signif. level: 0.05

Significant Difference
text+TFIDF > text
text+binary > text
text+descQual > text
text+majMin2 > text
text+binary > text+descQual
text+binary > text+majMin2
text+binary > text+majMin3
text+binary > text+majMin4
text+binary > text+majMin5
text+majMin2 > text+majMin3
text+majMin2 > text+majMin4
text+majMin2 > text+majMin5
text+majMin3 > text+majMin5
text+majMin4 > text+majMin5

Table B.18: Separation policies versus baseline (av. P@10), MeSH alone. Num. of iteration: 100000, signif. level: 0.05

Significant Difference
baseline > descQualSep
descQualSep > hardSep
baseline > hardSep
baseline > softSep
softSep > hardSep

Table B.19: Separation policies versus baseline (MAP), MeSH alone. Num. of iteration: 100000, signif. level: 0.05

Significant Difference
baseline > descQualSep
baseline > HardSep
baseline > SoftSep
descQualSep > HardSep
descQualSep > SoftSep
SoftSep > HardSep

Table B.20: Separation policies versus baseline (av. recall), MeSH alone. Number of iteration: 100000, significance level: 0.05

Significant Difference
baseline > descQualSep
descQualSep > hardSep
descQualSep > softSep
baseline > hardSep
baseline > softSep
softSep > hardSep

Table B.21: Separation policies versus baseline (MAP), text+MeSH combinations. Number of iteration: 100000, significance level: 0.05

Significant Difference
text+baseline > text+hardSep
text+baseline > text+SoftSep
text+descQualSep > text+hardSep
text+descQualSep > text+SoftSep

Table B.22: Separation policies versus baseline (av. recall), text+MeSH combinations. Number of iteration: 100000, significance level: 0.05

Significant Difference
text+descQualSep > text+hardSep
text+baseline > text+hardSep
text+softSep > text+hardSep

Table B.23: DepthDens runs vs. baseline (av. P@10), MeSH alone. Number of iteration: 100000, significance level: 0.05

Significant Difference	Significant Difference
baseline (1, 0) > 0.75, 0	0.75, 0 > 0.25, 2
baseline (1, 0) > 0.5, 0	0.5, 0 > 0.25, 1
baseline (1, 0) > 0.25, 0	0.5, 0 > 0.25, 2
baseline (1, 0) > 1, 1	1, 1 > 0.25, 1
baseline (1, 0) > 0.75, 1	1, 1 > 0.5, 2
baseline (1, 0) > 0.5, 1	1, 1 > 0.25, 2
baseline (1, 0) > 0.25, 1	0.75, 1 > 0.25, 1
baseline (1, 0) > 1, 2	0.75, 1 > 0.5, 2
baseline (1, 0) > 0.75, 2	0.75, 1 > 0.25, 2
baseline (1, 0) > 0.5, 2	0.5, 1 > 0.25, 1
baseline (1, 0) > 0.25, 2	0.5, 1 > 0.25, 2
0.75, 0 > 0.25, 0	1, 2 > 0.25, 1
0.75, 0 > 0.25, 1	0.75, 2 > 0.25, 1
0.75, 0 > 0.5, 2	0.75, 2 > 0.25, 2

Table B.24: DepthDens runs vs. baseline (MAP), MeSH alone. Number of iteration: 100000, significance level: 0.05

Significant Difference	Significant Difference
baseline (1, 0) > 1, 1	0.5, 0 > 0.5, 2
baseline (1, 0) > 0.75, 1	0.5, 0 > 0.25, 2
baseline (1, 0) > 0.25, 1	0.25, 0 > 0.25, 1
baseline (1, 0) > 1, 2	0.25, 0 > 0.5, 2
baseline (1, 0) > 0.75, 2	0.25, 0 > 0.25, 2
baseline (1, 0) > 0.5, 2	1, 1 > 1, 2
baseline (1, 0) > 0.25, 2	1, 1 > 0.75, 2
0.75, 0 > 1, 1	1, 1 > 0.5, 2
0.75, 0 > 0.75, 1	1, 1 > 0.25, 2
0.75, 0 > 0.25, 1	0.75, 1 > 1, 2
0.75, 0 > 1, 2	0.75, 1 > 0.75, 2
0.75, 0 > 0.75, 2	0.75, 1 > 0.5, 2
0.75, 0 > 0.5, 2	0.75, 1 > 0.25, 2
0.75, 0 > 0.25, 2	0.5, 1 > 0.25, 1
0.5, 0 > 0.25, 0	0.5, 1 > 1, 2
0.5, 0 > 0.75, 1	0.5, 1 > 0.75, 2
0.5, 0 > 0.25, 1	0.5, 1 > 0.5, 2
0.5, 0 > 1, 2	0.5, 1 > 0.25, 2
0.5, 0 > 0.75, 2	

Table B.25: DepthDens runs vs. baseline (av. recall), MeSH alone. Number of iteration: 100000, significance level: 0.05

Significant Difference	Significant Difference
baseline (1, 0) > 0.5, 0	0.5, 0 > 0.75, 2
baseline (1, 0) > 0.25, 0	0.5, 0 > 0.5, 2
baseline (1, 0) > 1, 1	0.5, 0 > 0.25, 2
baseline (1, 0) > 0.75, 1	1, 1 > 0.25, 0
baseline (1, 0) > 0.5, 1	0.75, 1 > 0.25, 0
baseline (1, 0) > 0.25, 1	1, 1 > 0.75, 1
baseline (1, 0) > 1, 2	1, 1 > 0.5, 1
baseline (1, 0) > 0.75, 2	1, 1 > 0.25, 1
baseline (1, 0) > 0.5, 2	1, 1 > 1, 2
baseline (1, 0) > 0.25, 2	1, 1 > 0.75, 2
0.75, 0 > 0.5, 0	1, 1 > 0.5, 2
0.75, 0 > 0.25, 0	1, 1 > 0.25, 2
0.75, 0 > 1, 1	0.75, 1 > 0.5, 1
0.75, 0 > 0.75, 1	0.75, 1 > 0.25, 1
0.75, 0 > 0.5, 1	0.75, 1 > 1, 2
0.75, 0 > 0.25, 1	0.75, 1 > 0.75, 2
0.75, 0 > 1, 2	0.75, 1 > 0.5, 2
0.75, 0 > 0.75, 2	0.75, 1 > 0.25, 2
0.75, 0 > 0.5, 2	0.5, 1 > 0.25, 1
0.75, 0 > 0.25, 2	0.5, 1 > 1, 2
0.5, 0 > 0.25, 0	0.5, 1 > 0.75, 2
0.5, 0 > 0.5, 1	0.5, 1 > 0.5, 2
0.5, 0 > 0.25, 1	0.5, 1 > 0.25, 2
0.5, 0 > 1, 2	0.25, 1 > 0.25, 2

Table B.26: DepthDens runs vs. baseline (av. P@10), text+MeSH. Number of iteration: 100000, significance level: 0.05

Significant Difference	Significant Difference
baseline (1, 0) > 0.75, 0	1, 2 > 0.25, 0
baseline (1, 0) > 0.5, 0	1, 1 > 0.75, 1
baseline (1, 0) > 0.25, 0	1, 1 > 0.5, 1
baseline (1, 0) > 0.75, 1	1, 1 > 0.25, 1
baseline (1, 0) > 0.5, 1	1, 1 > 0.75, 2
baseline (1, 0) > 0.25, 1	1, 1 > 0.5, 2
baseline (1, 0) > 0.75, 2	1, 1 > 0.25, 2
baseline (1, 0) > 0.5, 2	1, 2 > 0.75, 1
baseline (1, 0) > 0.25, 2	1, 2 > 0.5, 1
1, 1 > 0.75, 0	1, 2 > 0.25, 1
1, 2 > 0.75, 0	1, 2 > 0.75, 2
1, 1 > 0.5, 0	1, 2 > 0.5, 2
1, 2 > 0.5, 0	1, 2 > 0.25, 2
1, 1 > 0.25, 0	

Table B.27: DepthDens runs vs. baseline (MAP), text+MeSH. Number of iteration: 100000, significance level: 0.05

Significant Difference	Significant Difference
baseline (1, 0) > 0.75, 0	1, 2 > 0.25, 0
baseline (1, 0) > 0.5, 0	0.75, 2 > 0.25, 0
baseline (1, 0) > 0.25, 0	0.5, 2 > 0.25, 0
baseline (1, 0) > 0.75, 1	0.25, 2 > 0.25, 0
baseline (1, 0) > 0.5, 1	1, 1 > 0.75, 1
baseline (1, 0) > 0.25, 1	1, 1 > 0.5, 1
baseline (1, 0) > 0.75, 2	1, 1 > 0.25, 1
baseline (1, 0) > 0.5, 2	1, 1 > 0.75, 2
baseline (1, 0) > 0.25, 2	1, 1 > 0.5, 2
0.75, 0 > 0.5, 0	1, 1 > 0.25, 2
0.75, 0 > 0.25, 0	0.75, 1 > 0.5, 1
1, 1 > 0.75, 0	0.75, 1 > 0.25, 1
0.75, 0 > 0.5, 1	0.75, 1 > 0.5, 2
0.75, 0 > 0.25, 1	0.75, 1 > 0.25, 2
0.75, 0 > 0.5, 2	0.5, 1 > 0.25, 1
0.75, 0 > 0.25, 2	1, 2 > 0.5, 1
0.5, 0 > 0.25, 0	0.75, 2 > 0.5, 1
1, 1 > 0.5, 0	0.5, 1 > 0.25, 2
0.75, 1 > 0.5, 0	1, 2 > 0.25, 1
0.5, 1 > 0.5, 0	0.75, 2 > 0.25, 1
0.5, 0 > 0.25, 1	0.5, 2 > 0.25, 1
1, 2 > 0.5, 0	1, 2 > 0.75, 2
0.75, 2 > 0.5, 0	1, 2 > 0.5, 2
0.5, 0 > 0.25, 2	1, 2 > 0.25, 2
1, 1 > 0.25, 0	0.75, 2 > 0.5, 2
0.75, 1 > 0.25, 0	0.75, 2 > 0.25, 2
0.5, 1 > 0.25, 0	0.5, 2 > 0.25, 2
0.25, 1 > 0.25, 0	

Table B.28: DepthDens runs vs. baseline (av. recall), text+MeSH. Number of iteration: 100000, significance level: 0.05

Significant Difference	Significant Difference
baseline (1, 0) > 0.75, 0	1, 2 > 0.25, 0
baseline (1, 0) > 0.25, 0	0.75, 2 > 0.25, 0
baseline (1, 0) > 0.75, 1	0.5, 2 > 0.25, 0
baseline (1, 0) > 0.5, 1	1, 1 > 0.75, 1
baseline (1, 0) > 0.25, 1	1, 1 > 0.5, 1
baseline (1, 0) > 0.75, 2	1, 1 > 0.25, 1
baseline (1, 0) > 0.5, 2	1, 1 > 0.75, 2
baseline (1, 0) > 0.25, 2	1, 1 > 0.5, 2
0.75, 0 > 0.5, 0	1, 1 > 0.25, 2
0.75, 0 > 0.25, 0	0.75, 1 > 0.5, 1
1, 1 > 0.75, 0	0.75, 1 > 0.25, 1
0.75, 0 > 0.5, 1	1, 2 > 0.75, 1
0.75, 0 > 0.25, 1	0.75, 1 > 0.5, 2
1, 2 > 0.75, 0	0.75, 1 > 0.25, 2
0.75, 0 > 0.5, 2	1, 2 > 0.5, 1
0.75, 0 > 0.25, 2	0.75, 2 > 0.5, 1
0.5, 0 > 0.25, 0	0.5, 1 > 0.25, 2
1, 1 > 0.5, 0	1, 2 > 0.25, 1
0.75, 1 > 0.5, 0	0.75, 2 > 0.25, 1
0.5, 0 > 0.25, 1	0.5, 2 > 0.25, 1
1, 2 > 0.5, 0	1, 2 > 0.75, 2
0.75, 2 > 0.5, 0	1, 2 > 0.5, 2
0.5, 0 > 0.25, 2	1, 2 > 0.25, 2
1, 1 > 0.25, 0	0.75, 2 > 0.5, 2
0.75, 1 > 0.25, 0	0.75, 2 > 0.25, 2
0.5, 1 > 0.25, 0	0.5, 2 > 0.25, 2

Table B.29: InfoBased versus baseline (av. P@10, MAP, av. recall), MeSH alone. Number of iteration: 100000, significance level: 0.05

Measure	Significant Difference
average P@10	none
MAP	InfoBased > baseline
average recall	InfoBased > baseline

Table B.30: InfoBased versus baseline (av. P@10, MAP, av. recall), text+MeSH alone. Number of iteration: 100000, significance level: 0.05

Measure	Significant Difference
average P@10	none
MAP	none
average recall	none