Semantic Search as Inference

Applications in Health Informatics

Bevan R. Koopman

Semantic Search as Inference:

Applications in Health Informatics

by

Bevan Raymond Koopman

A thesis submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

School of Information Systems
Faculty of Science & Engineering

Queensland University of Technology

2014

Copyright

Statement of Original Authorship

The work contained in this thesis has not been previously submitted to meet requirements for an award at this or any other higher education institution. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made.

Signature:

Date:

To Ron and Juliet Koopman

Abstract

In this thesis, we present models for semantic search: Information Retrieval (IR) models that elicit the meaning behind the words found in documents and queries rather than simply matching keywords. This is achieved by the integration of structured domain knowledge and data-driven information retrieval methods.

The research is set within health informatics to tackle the unique challenges within this domain; specifically, how to bridge the 'semantic gap'; that is, how to overcome the mismatch between raw medical data and the way human beings interpret it. Bridging the semantic gap involves addressing two issues: *semantics*; that is, aligning the meaning or concepts behind words founds in documents and queries; and leveraging *inference*, which utilises semantics to infer relevant information.

Three semantic search models — all utilising concept-based rather than term-based representations — are developed; these include: the Bag-of-concepts model, which utilises concepts from the SNOMED CT medical ontology as its underlying representation; the Graph-based Concept Weighting model, which captures concept dependence and importance in a novel weighting function; and the core contribution of the thesis, the Graph INference model (GIN): a unified theoretical model of semantic search as inference, achieved by the integration of structured domain knowledge (ontologies) and statistical, information retrieval methods. It is the GIN that provides the necessary mechanism for *inference* to bridge the semantic gap. All three models are empirically evaluated using clinical queries and a real-world collection of clinical records taken from the TREC Medical Records Track (MedTrack).

Our evaluation shows that the use of concept-based representations in the Bag-of-concepts model leads to improved retrieval effectiveness. When concepts are combined within the Graph-based Concept Weighting model, further improvements are possible. The evaluation of GIN highlighted that its inference mechanism is suited to hard queries — those that perform poorly on a term-based system. In-depth analysis also revealed that the GIN returned many new

documents not retrieved by term-based systems and therefore never evaluated for relevance as part of the TREC MedTrack. This highlights that using standard IR test collections may underestimate the effectiveness of semantic search systems.

This work represents a significant step forward in the integration of structured domain knowledge and data-driven information retrieval methods. Furthermore, the thesis provides an understanding of inference — when and how it should be applied for effective semantic search. It shows that queries with certain characteristics benefit from inference, while others do not. The detailed investigation into the evaluation of semantic search systems shows how standard IR test collections may underestimate effectiveness of such systems and new methods of evaluation are suggested. The Graph Inference model, although developed within the medical domain, is generally defined and has implications in other areas, including web search, where an emerging research trend is to utilise structured knowledge resources for more effective semantic search.

Acknowledgements

Only one name appears on the front of this thesis but many other people have made this endeavour possible. First among them is my principle supervisor, Peter Bruza. While at the DSTC, I was inspired by discussions with Peter and, at the time, thought that if I was to do a PhD, the only person I would want to do it with was Peter. That dream is now a reality and I'm immensely grateful to Peter for the experience. I have grown as a result of his wisdom, support and insights.

I was lucky to have two associate supervisors in Laurianne Sitbon and Michael Lawley. Laurianne was especially supportive at the beginning of my PhD, when I was still finding my feet. Michael helped make my experience embedded in the Australia e-Health Research Centre (AEHRC) a very positive one and he was always supportive of my activities there.

To my pseudo-supervisor, Guido Zuccon, I also owe a great deal of thanks. From his arrival at AEHRC in 2011, he had an immediate, positive impact, always taking an interest in my research. We've enjoyed many discussions together (mainly through the small slot between our desks).

I am grateful to the QI group at QUT for creating such a diverse and interesting environment. To Mike, for letting us know "we're doing science here" and to David, Lance, Kirsty and Aneesha.

CSIRO and AEHRC supported this PhD through a top-up scholarship, funding for conference travel and payment of the medical students who provided additional relevance assessments. Most important though, the AEHRC provided a stimulating and collegial environment in which to work (not to mention a good coffee machine).

Malcolm Garrett painstakingly proofread the thesis word by word and corrected my English. The document was much improved through his efforts.

I owe many thanks my parents-in-law, Jill and Roger, who have always supported Amanda and I.

This thesis is dedicated to my parents, who have always encouraged me to

follow my dreams. The opening quote on page 8 is for them and for my brother, Dylan.

Finally, to my wife, Amanda, whose love and encouragement makes life worth living. Also, to Kai and Ella and the greatest research project of them all: parenthood.

"Twenty years from now you will be more disappointed by the things that you didn't do than by the ones you did do. So throw off the bowlines. Sail away from the safe harbor. Catch the trade winds in your sails. Explore. Dream. Discover."

— Mark Twain

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Medicine is a science of uncertainty and an art of probability.

— William Osler*

Medicine is an information-intensive field. As access to timely and relevant information is essential for effective delivery of health services, medicine is consequently dependent on information technology and, more specifically, on information retrieval (IR) systems. Much of the medical data available today is in unstructured form, namely free-text. Searching and interpreting this data presents challenges specific to the medical domain. At the core of these issues is the 'semantic gap' problem, defined as the difference between the raw medical data and the way a human being might interpret it [Patel et al., 2007]. The semantic gap might manifest as vocabulary mismatch, for example a search query of *high blood pressure* and a document containing the synonym *hypertension*, or as other associations requiring inference, for example the presence of *dialysis machine* in a patient record denoting someone suffering from *kidney disease*. These examples illustrate that highly relevant documents might have no keyword overlap with the query. The semantic gap problem is not unique to searching medical data; it is, however, accentuated to a degree that standard information retrieval approaches are rendered ineffective.

Bridging the semantic gap involves addressing two issues. The first is the the issue of *semantics*; that is, aligning the meaning or concepts behind words found in documents and queries. The second issue is leveraging *inference* to determine

*William Osler, (1849 – 1919) was a Canadian physician and one of the founding professors at John Hopkins Hospital.

the association between concepts. Two fields of research can be drawn on to address these issues: information retrieval and formal symbolic representations and reasoning (and more generally the Semantic Web). Individually, neither field fully meets the unique requirements of searching medical data. Information retrieval's dependence on term-based models and lack of implicit medical background knowledge make it susceptible to vocabulary mismatch. More importantly, current state-of-the-art information retrieval models do not support the necessary inference mechanisms required to bridge the semantic gap while symbolic ontology-based solutions using medical domain knowledge resources are too rigid, not context-specific, and do not cope well with unstructured data. However, each individual field partially addresses the requirements for effective semantic search as inference and, we argue, in combination address most requirements. Historically, there has been little overlap between the two fields, mainly because it is difficult to realise a theoretically sound, formal model that combines the two. The purpose of this thesis is to investigate and develop such a model, with the hypothesis that:

> A unified theoretical model of semantic search as inference, achieved by the integration of structured domain knowledge (ontologies) and statistical, information retrieval methods, provides the necessary mechanism for inference required for effective semantic search of medical data.

The thesis takes a two-lystep approach to addressing the above hypothesis. First, we address the problem of semantics, exploring the use of 'Bag-of-concepts' representations to overcome some of the limitations of term-based representations, specifically tackling the vocabulary mismatch problem. Secondly, we extend our Bag-of-concept model to form the Graph-Based Concept Weighting retrieval model that makes greater use of medical domain knowledge from the SNOMED CT ontology. Finally, to realise the critical requirement for inference in semantic search, we present the Graph INference model (GIN): a novel graph-based retrieval model integrating ontologies and formal information retrieval models. It is the GIN that provides the necessary mechanism for *inference* to bridge the semantic gap.

1.1 Contributions

In the development of a unified model of semantic search and evaluation of the above hypothesis, we make the following major contributions:

1. A detailed outline of the requirements for effective semantic search, identify and categorising the types of inference required to overcome the semantic gap. This contribution is specific to the medical domain but major aspects are still generally applicable.

2. The development and evaluation of concept-based representations for medical IR. Concept-based representations partially address the requirements for semantic search and demonstrate improvements in retrieval effectiveness over state-of-the-art term-based IR models.

3. The core theoretical contribution of this thesis: a unified theoretical model of semantic search as inference, which utilises a graph-based representation of a corpus comprising ontological concepts and relationships but is driven by IR probabilistic relevance estimation.

4. A three-part empirical evaluation of our retrieval models using i) TREC Medical Track test collection; ii) a novel evaluation framework developed as part of the thesis; iii) relevance assessment by medical professionals.

5. An investigation of when and why semantic search as inference succeeds and when it fails. This analysis reveals how the quality of the ontology affects retrieval performance and how the notion of conceptual relevance in an ontology differs from document/query relevance in a retrieval scenario.

In addition, the thesis provides a number of minor contributions:

1. An evaluation framework for semantic search: a method to develop a test collection of real-world medical records with associated queries and relevance judgements.

2. An analysis of the unique requirements of evaluating semantic search systems, understanding and quantifying the bias of pooling methods used in developing test collections with respect to semantic search methods.

1.2 Organisation

The thesis is organised into the following chapters:

Chapter 2 — Bridging the Semantic Gap.
This chapter details the problems in searching medical data and provides motivation for a semantic search approach. For each semantic gap issue, we detail the types of inference required to overcome the issue. The chapter finishes with a structured set of requirements for how to deal

with the semantic gap problems. These problems and requirements will be referred to throughout the thesis as each chapter attempts to address them.

Chapter 3 — Semantic Search and Medical Information Retrieval.
Literature review on semantic search. The review first covers background on symbolic representations / ontologies, followed by work on information retrieval and medical IR. The current state-of-the-art in semantic search is explained and the gap in knowledge and motivation for semantic search and inference is presented.

Chapter 4 — Bag-of-Concepts Model.
This chapter addresses the first issue of *semantics* required for a unified semantic search as inference model. We present a novel 'Bag-of-concepts' retrieval model, where queries and documents are represented as high-level concepts — taken from medical ontologies — rather than terms. This approach is reviewed in light of the semantic gap issues presented in Chapter 2 and we show how converting to higher-level concepts addresses vocabulary mismatch. Conceptual representations differ both semantically and statistically from terms. We show that it is these differences that result in an effective retrieval model using concepts. An empirical evaluation of the Bag-of-concepts model shows the effectiveness of the model compared to state-of-the-art term models, especially at improving hard queries. The chapter concludes with the finding that although the Bag-of-concepts model is effective, it addresses only some of the semantic gap problems, mainly vocabulary mismatch. This provides motivation for leveraging much deeper domain knowledge to support the necessary 'inferencing' mechanism required in semantic search.

Chapter 5 — Graph-based Concept Weighting Model.
Like bag-of-words models, the Bag-of-concepts model does not consider the innate interdependence between medical concepts (identified as one of the semantic gap issues). Thus, we extended the Bag-of-concepts model to a graph-based representation that naturally captures dependencies between concepts. In addition, we further extend previous graph-based approaches by incorporating domain knowledge that estimates the importance of a concept within the global medical domain. The incorporation of domain knowledge shows promising results and, from the previous chapter, we know that concept-based representations improve retrieval performance. These results motivated the development of a model that makes extensive use of domain knowledge. This chapter provides a link between the basic

Bag-of-concepts model of Chapter 4 and the unified model of semantic search as inference presented in Chapter 6.

Chapter 6 — Graph INference Model (GIN).
The core theoretical contribution of this thesis: a unified theoretical model of semantic search as inference, which utilises a graph-based representation of a corpus comprising ontological concepts and relationships but driven by IR probabilistic relevance estimation. We present an efficient implementation of the graph inference model using a graph traversal algorithm. Empirical evaluation of the graph inference model using the TREC Medical Records Track test collection reveals that it is not significantly more effective than our Bag-of-concepts model. However, we show that the graph inference model is effective at improving the performance of hard queries, which are more likely to require inference. Further analysis shows that the TREC MedTrack test collection is not sufficient to provide complete evaluation for semantic search systems.

Chapter 7 — Relevance Assessment and Evaluating Semantic Search.
This chapter focuses on evaluating semantic search systems. The evaluation of the GIN revealed that the model retrieved a large number of unjudged documents (those never judged by TREC assessors) and that, as a result, the retrieval effectiveness may have been *underestimated* using the TREC Medical Records Track. In this chapter, we analyse the effect that these unjudged documents have on the retrieval effectiveness estimates. This motivated the need to obtain additional relevance judgements with the aid of graduate medical students. Equipped with additional relevance judgements, we re-evaluate the Graph Inference model, showing that, indeed, the retrieval effectiveness of the GIN was underestimated. Finally, we present an alternative to the TREC-style evaluation, which uses manually coded medical records and is aimed at evaluating semantic search systems.

Chapter 8 — Discussion and Future Work.
This chapter discusses the main findings and contributions of the thesis. We discuss how each of the models proposed help to bridge the semantic gap. In doing so, we also show that the Graph Inference model provides a unified model of semantic search as inference. Furthermore, we provided an understanding of inference — when and how it should be applied for effective semantic search. We discuss the challenges for evaluating semantic search systems and how they might be overcome. Finally, the section on future work considers how the GIN can be extended and applied to other applications, including large scale web search.

Chapter 9 — Conclusion.

This chapter summarises the main conclusions and contributions of the research.

Chapter 2

Bridging the Semantic Gap

I was trying to comprehend the meaning of the words.

— Spock, Star Trek: The Final Frontier

This chapter details the requirements for effective semantic search, identifying and categorising the types of inference required to overcome the semantic gap problem. The semantic gap problem is broken down into a number of instances or sub-problems, each being detailed in the following subsections: Vocabulary Mismatch (2.1), Granularity Mismatch (2.2), Conceptual Implication (2.3), Inferences of Similarity (2.4), Negation & Family History (2.5.1), Temporality (2.5.2), Age & Gender (2.5.3), Level of Evidence (2.5.4). The analysis of the semantic gap provided in this chapter is specific to the medical domain but major aspects are still generally applicable.

To fully appreciate the effect that these issues have in a real retrieval scenario and to understand some real queries with semantic gap problems, we provide some initial results from a retrieval experiment using a state-of-the-art keyword-based retrieval system. This is done to provide concrete examples of the Semantic Gap problems and to quantify the effect that these problems have on state-of-the-art IR systems. The queries that we use are taken from the TREC Medical Records Track, a standard forum for evaluating IR systems. The chapter serves as evidence that keyword-based IR systems have limited effectiveness in searching medical data and is motivation for a semantic search and inference approach.

2.1 Vocabulary Mismatch

Vocabulary mismatch occurs when particular concepts are expressed in a number of different ways, yet have a similar underlying meaning. For example, synonyms like *boat* and *ship* are syntactically different, yet semantically very similar. In addition, there are formal and colloquial variants for terms, as well as regional differences, especially in medical natural language. Overcoming the vocabulary mismatch problem is the most common motivation and requirement for semantic search. This is also a common requirement for general IR systems and is not specific to the medical domain. However, the complexity and nature of medical language means there are often multiple variants for expressing the same concept, thus exacerbating the vocabulary mismatch problem.

Vocabulary mismatch can occur with single terms, for example, *cranium* and *skull* have similar meaning; or vocabulary mismatch can occur in multi-term phrases, such as the synonyms *heart attack* and *myocardial infarction*. Medications and pharmaceuticals are a particularly prevalent example of vocabulary mismatch — the generic name for a medication (or its active ingredient) is often synonymous with drug brand names. Acronyms and abbreviations are other instances of vocabulary mismatch; medical language makes frequent use of abbreviations. Abbreviations can be ambiguous; for example, the abbreviation *AD* may refer to *Antidepressant* or to *Alzheimer's Disease*. People can derive the correct interpretation based on the context of use but an automated system insensitive to context cannot.

The effect of vocabulary mismatch is that authors and readers might express the same information in different ways. The consequence of this, in a retrieval scenario, is that a query may have no overlapping terms with a document, yet the document could still be semantically highly relevant. A keyword-based IR system would not return these semantically relevant documents as the system returns only documents containing the query terms.

A number of approaches in IR are commonly used to address vocabulary mismatch, query expansion and pseudo relevance feedback being the most common. Here the original query is augmented (or expanded) with additional terms likely to be found in relevant documents. These additional terms can be derived in two main ways: statistically, by considering terms that co-occur highly with the query terms and semantically via the use of external resources such as thesauri, which explicitly represent term dependencies (for example, WordNet synonyms). These techniques are considered in further detail in the next chapter.

Two types of inference are required to overcome the vocabulary mismatch problem [Lancaster, 1986]. Statistical or associational inference can be employed to determine terms that are highly correlated in usage, such as synonyms.

Standard IR approaches such as query expansion take advantage of terms with highly correlated usage; these approaches are an instantiation of associational inference. In contrast, deductive inference may be used in cases where linguistic resources (such as ontologies or thesauri) describe multiple alternative terms for a concept.[1] The requirement for both association and deductive inference motivates research into a unified model that integrates structured ontologies and statistical, data-driven IR methods.

The vocabulary mismatch problem can also be described more formally. Given a query Q and document D, comprised of a sequence of terms, where each term is a string, then:

$$Q = \langle q_0, \ldots, q_n \rangle \quad D = \langle t_0, \ldots, t_n \rangle,$$

where $q_0, \ldots, q_n$ and $t_0, \ldots, t_n$ belong to common vocabulary V. Vocabulary mismatch can, therefore, be represented as:

$$q_i \notin D$$
$$t_j \in D \wedge t_j \approx q_i$$

The $\approx$ operator denotes that t_j and q_i have a similar meaning.

2.2 Granularity Mismatch

Users often formulate queries using general terms, whereas relevant documents contain specific sub-class or child concepts. For example, with the TREC MedTrack query *Patients taking atypical antipsychotics*, relevant documents would not contain the term antipsychotics; instead they would contain instances of antipsychotics, such as the drug *Clozapine* or even the brand name *Clozaril*. This problem is called granularity mismatch (sometimes referred to as specialisation / generalisation). It is another issue for information retrieval in general but even more prevalent in medical IR.

As with vocabulary mismatch, granularity mismatch is particularly prevalent when searching electronic medical records. In these records the authors provide detailed descriptions and analyses of a patient's conditions, diagnoses and treatments — they have a micro view of the information space. In contrast, users searching these documents express high-level information needs and have a macro view of the information space. As a result, the two types of users (authors and searchers) use different language to express the same inform-

[1] This is the case in the SNOMED CT medical ontology where a single concept has a 'preferred term' field and a number of 'alternative terms' descriptions for the concept.

ation. This mismatch in vocabulary renders an information retrieval system using keyword matches ineffective in searching medical data.

Overcoming granularity mismatch involves understanding when concepts are specialisations or generalisation of other concepts. Ontologies specifically attempt to address this by modelling parent-child or ISA relationships. However, an open issue is understanding when to generalise or when to specialise, as sometimes it may be appropriate to include certain parent concepts, whereas in other cases the parent may be too general.

Although ontologies encoded parent-child relationships, they do not provide a meaningful measure of distance or similarity between parent and child concepts. Some child concepts may be very similar to their parent (for example, *left kidney* is very similar to its parent *kidneys*), while other children may be quite different (for example, the child *kidney* is far less similar to its parent *organ*). Without an appropriate measure of similarity between parent and child concepts, it is difficult to determine if it is appropriate to generalise or specialise.

The ability to infer more general or more specific concepts is essential for semantic search. The inference process is typically deductive in nature: determining when one concept is a parent or child of another. However, this inference mechanism needs to include a measure of uncertainty or similarity that is lacking in hierarchical ontologies. Inference with uncertainty is the foundation of probabilistic information retrieval models that estimate a probability of relevance. This thesis proposes integrating explicit inheritance relationships from ontologies and a statistical estimation of uncertainty from IR models to address the issue of granularity mismatch.

Formally, given a query term q_i and document term t_j, granularity mismatch can be represented as:

$$q_i \notin D$$
$$t_j \in D \wedge (t_j \subset q_i \vee q_j \subset t_i),$$

where the subset operator, $\subset$, is used to denote that term t_j is a specialisation of the term q_i; that is, the possible meanings of t_j is a subset of the possible meanings of q_j.

2.3 Conceptual Implication

Although a relevant document may contain no query terms, the document may contain signs or evidence that drives a conclusion of the query. Specifically, certain terms within the document may logically infer the query terms and, by

extension, relevance of the document to the query. For example, consider the query *Kidney disease* and a document that contains the terms *Dialysis machine*. For this query, a person reading the document would deduce *Dialysis machine* $\rightarrow$ *Kidney disease*. Conceptual implication is different from vocabulary mismatch, where two concepts are expressed differently but have the same meaning and different from granularity mismatch, where one concept is general and the other is specialised. Instead, with conceptual implications the document contains evidence in the form of a concept that logically infers the conclusion of another concept.

Conceptual implication situations are particularly prevalent when deducing diseases where:

- *treatment* $\rightarrow$ *disease*: the presence of certain treatments implies that the person has a certain disease; for example certain types of chemotherapy drugs imply the presence of certain cancers.
- *organism* $\rightarrow$ *disease*: the presence of certain organisms in laboratory tests imply the disease; for example *Varicella zoster* virus $\rightarrow$ *Chicken pox*.

The required mechanism to handle conceptual implication is deductive inference. Logical deduction is the cornerstone mechanism for reasoning in ontologies [Sowa et al., 2000].

Formally, conceptual implication for semantic search can be expressed as:

$$
\begin{aligned}
& q_i \notin D \\
& t_j \in D \wedge t_j \rightarrow q_i \\
& \therefore D \rightarrow q_i,
\end{aligned}
$$

where $\rightarrow$ denotes that if t_j is present then q_i is implied.

2.4 Inferences of Similarity

While some concepts can be derived by conceptual implication, others are more associational in nature. In this case, the presence of a certain concept indicates high likelihood of another, or the two concepts are semantically similar in some way. Disease comorbidities are an example of this case; comorbidities are the presence of one disease or more in addition to a primary disease, or the effect of such additional diseases. For example, *anxiety* and *depression* are two commonly co-occurring disorders. In some cases, the two associated concepts do not just co-occur in the relevant document; they also act on each other. In such a

case, the presence of both concepts within a document does not necessarily infer relevance; the dependence between the two needs to be determined, as in the case of the TREC query *Patients treated for methicillin-resistant Staphylococcus aureus (MRSA) endocarditis*, where MRSA is a common bacterial infection and endocarditis is an infection of the heart. For this query, a document might contain both the terms *MRSA* and *endocarditis*, but MRSA is a common infection and may not be the actual cause of the endocarditis. An IR system would most likely give a high score to documents containing both MRSA and endocarditis but these document would not be relevant unless MRSA was the actual cause of the endocarditis. The mere presence of the two query terms is not enough to determine relevance.

An IR system needs to account for the innate dependence between medical concepts to be effective. The form of inference required in this case is associational. The types of relationships and associations required are typically not modelled in ontologies designed for deductive reasoning. These relationships are better suited to statistical inference mechanisms typical of data-driven IR models.

Formally, associational inference can be represented as:

$$q_i \notin D$$
$$t_j \in D \wedge t_j \sim q_i$$

where the $\sim$ denotes that t_j and q_i are strongly associated. The association metric could be implemented as a conditional probability:

$$P(t_j|q_i) > \alpha \rightarrow t_j \sim q_i$$

If the conditional probability is above some threshold α, then a strong association exists.

2.5 Context-specific Semantic Gap Issues

The semantic gap issues reviewed so far — vocabulary mismatch, granularity mismatch, conceptual implication and inferences of similarity — are related to the different interpretations of the terms within a document. The problems reviewed in this section still relate to interpretation of terms, but more within the context of the whole document. These problems are more specific to searching medical data, but may still affect general applications.

2.5.1 Negation and Family History

Negation and reference to family history are two unique characteristics of clinical records that affect natural language processing and search of clinical text [Chapman et al., 2001]. Commonly mentioned conditions in a patient record (e.g., *fever* or *fracture*) often appear in negated form (e.g., *denies fever*, *no fracture*). Family history details relevant hereditary conditions, for example, a patient who has a history of breast cancer in their family. From an information retrieval perspective (i.e., searching clinical documents), negation may adversely affect search effectiveness [Koopman et al., 2010; Limsopatham et al., 2012]. Traditional keyword matching IR systems denote the presence of the query terms as an indicator of relevance but do not consider situations where the terms might be explicitly negated. The situation is similar for when the term relates to a patient's family history rather than the actual patient.

Negation can be identified by certain negation identifiers: terms such as *no*, *denies*, *without*, etc. If these negation identifiers are observed, then one can conclude that the concept following them is negated. For example, if *Patient denies fever* is observed then the negation identifier *denies* indicates that the concept *fever* is negated. The same situation applies for family history with identifier terms like *father had*, *family history of*, etc. Previous research in clinical natural language processing has developed techniques for negation detection [Chapman et al., 2001].

If a negation identifier is present, then we can conclude that the following concept is negated, i.e., the conclusion is derived deductively. Therefore, deductive inference is the inference mechanism required to handle negation and family history.

2.5.2 Temporality

Temporality is another characteristic affecting search of medical data, particularly patient records. Most records contain a past medical history section that lists conditions and treatments a patient may have had in the past. Some conditions and treatments may be relevant to their current situation, while others may not affect them any more. An IR system may retrieve a patient record based on the terms found in the past medical history section, but the relevance of the record is dependent on whether the past conditions or treatments still apply to the patient or on the context of the query. To overcome this problem, the past medical history section of a document needs to be identified and handled differently from the rest of the document. This can be done in a deductive manner, similar to negation handling, i.e., by deducing that certain portions of

the document relate to past medical history.

A document representing a patient record may relate to a person's short admission to hospital spanning a few days or could relate to many months spent in treatment. The length of time covering events in the patient's record is called the document timespan and may vary considerably for different records/patients. The timespan of a document is another temporal issue affecting relevance in an IR scenario. For example, consider the TREC query *Patients treated for methicillin-resistant Staphylococcus aureus (MRSA) endocarditis*. MRSA is a bacterial infection present in many hospitals that often affects patients with low immune systems and a long hospital stay; endocarditis is an infection of the heart. MRSA is common in patients with long hospital stays and is unlikely to be the cause of the endocarditis. Thus a patient record with a long admission and containing the terms MRSA and endocarditis is unlikely to be relevant. However, a patient record containing MRSA and endocarditis, with a short admission, has a high probability of MRSA-caused endocarditis and would therefore be relevant. Many IR models use document length in the estimation of relevance: assuming the same term frequency, longer documents are less relevant than shorter documents. (In IR this is called document length normalisation.) However, in the above example, timespan normalisation, rather than document length normalisation, may provide a better means of estimating relevance. Temporality illustrates that in medical IR, relevance may be affected by many factors, some of which may not be accounted for by general retrieval models.

2.5.3 Age and Gender

When searching patient records, the age and gender of the patient can be an important determinator of relevance. Some queries have specific age or gender requirements, for example the query *Elderly women with endocarditis*. There are multiple ways to express gender (e.g., female, woman, girl); ideally the IR system would normalise these to a single form. Age can also be expressed in a number of ways (adolescent, teenage, elderly or with numeric values like 65 years-old). Again, an effective IR system needs to normalise or infer age to handle such queries effectively.

Normalising gender, e.g. *woman* $\rightarrow$ *female* and *65* $\rightarrow$ *elderly*, is a logical deduction process. Thus, deductive inference can be used to handle age and gender in medical IR.

2.5.4 Levels of Evidence

So far we have referred to patient records as a specific type of document. In reality, patient records are often comprised of a number of reports, or sub-documents, such as:

- History and examination notes, which are authored when the patient is first seen and contain past medical history and initial review of their condition.
- Laboratory tests, which include requests, results and analysis from tests and procedures, often pathology or radiology.
- Discharge summary, which is a retrospective summary of a patient's stay in hospital and recommendations for further care after they leave.

These different reports convey different information that can affect relevance in an IR scenario. For example, the initial examination may contain some initial suspected diagnosis, but the diagnosis is yet to be verified by laboratory tests. In contrast, the discharge summary is a high-quality review of findings. Thus, query terms found in the examination notes would be a less reliable indicator of relevance than query terms found in the discharge summary. Some medical IR models have begun to address this issue by treating different report types separately.

The particular report type provides evidence for concluding a relevance estimation. The required inference mechanism is therefore deductive inference.

2.6 The Semantic Gap in Effect

This chapter has highlighted the issues in searching medical data: the Semantic Gap problem. To appreciate fully the effect that this has in a real retrieval scenario and to understand some real queries that are affected by semantic gap problems, we provide some initial results from a retrieval experiment using a benchmark keyword-based retrieval system. This is done to provide concrete examples of the Semantic Gap problems and to quantify the effect that these problems have on keyword-based IR systems.

We implemented a standard keyword-based IR system and evaluated the system on a test collection of medical records. As the retrieval model, we used a Probabilistic Language Model with Dirichlet smoothing ($\mu = 20000$). Details of this retrieval model are covered in the next chapter (Section 3.3.1). For the evaluation of the retrieval model, we use the TREC Medical Records Track test

collection. The collection contains free-text electronic patients records and 81 queries specifying an information need for patients matching certain criteria, for example patients with specific diseases or receiving certain treatments. Three evaluation measures are used as part of TREC Medical Records Track: Mean Average Precision (MAP), Bpref and Precision at 10. (These measures and more details about the TREC Medical Records Track are provided in Section 4.3.1 of the next chapter). Using this standard keyword-based IR system, the retrieval results are provided in Table 2.1.

MAP	Bpref	Prec@10	Recall
0.3117	0.3891	0.4926	0.7466

Table 2.1: Retrieval results on 81 queries from TREC MedTrack (2011, 2012) using language model with Dirichlet smoothing.

The table summarises the overall retrieval results; however, we would like to focus on the performance of individual queries to understand better how the semantic gap problems may affect them. Figure 2.1, therefore, shows the performance of individual queries. The performance varies considerably between queries; in particular, the red triangles highlight those queries with the lowest performance. We hypothesise that these are queries badly affected by semantic gap problems. To understand how they manifest, we consider four queries in detail, these being provided in Table 2.2.

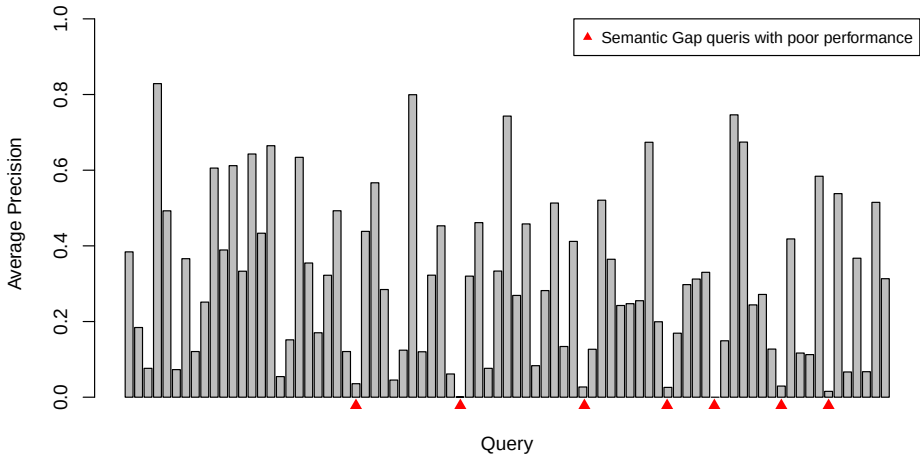


Figure 2.1: Per-query retrieval results on 81 queries from TREC MedTrack (2011, 2012) using language model with Dirichlet smoothing; ▲ shows poor performing queries, example of the Semantic Gap problem.

TREC	Query Text	Avg. Prec.	Recall
167	`Patients with AIDS who develop pancytopenia.`	0.0000	0.0000
179	`Patients taking atypical antipsychotics without a diagnosis schizophrenia or bipolar depression.`	0.0155	0.2159
174	`Elderly patients with ventilator associated pneumonia.`	0.0294	1.0000
125	`Patients co-infected with Hepatitis C and HIV.`	0.0357	0.0714

Table 2.2: Examples of queries badly affected by semantic gap problems.

Each query is characterised by one or more of the semantic gap issues outlined in this chapter:

Vocabulary mismatch. Query 167 was a typical example. The medical condition *pancytopenia* (reduction in the number of red and white blood cells) is also often referred to as *bicytopenia*. Similarly, *AIDS* could be expanded to *acquired immunodeficiency syndrome* or expressed as *HIV* or *human immunodeficiency virus*. In this case, no relevant documents were returned as all the relevant documents used the terms *bicytopenia* and *HIV* rather than the query terms *pancytopenia* and *AIDS*.

Granularity mismatch. Query 179 mentioned *antipsychotics* (the class of drugs used to treat psychosis). In detailed clinical patient records, the authors will explicitly specify the type of antipsychotic the patient is taking, rather than generally stating that they are taking antipsychotics. Therefore, many relevant documents were never retrieved because they did not contain the term *antipsychotics*, but instead specified the actual type of antipsychotic. Examples of relevant documents were those that contained *Cymbalta* and *Xanax*, both antipsychotic medications.

Conceptual implication. Query 167 specified patients affected by *pancytopenia*; however, many of the relevant documents only contained the actual causes of the pancytopenia, which include *Leukemia*, *Osteopetrosis* and *Pernicious anemia*. A qualified human reader would deduce pancytopenia from mentions of any of these causes.

Inferences of Similarity. Query 167 required patients with both *AIDS* and *pancytopenia*; many irrelevant documents contained only one of these two

disorders. Similarly, Query 125 required patients with both Hepatitis C and HIV.

Negation. Query 179 explicitly required patients *without* schizophrenia or bipolar depression. Many irrelevant documents containing these disorders were retrieved.

Temporality. Query 125 required patients with co-infections. A number of irrelevant patients were returned who had HIV and Hepatitis C infection in their history but were never co-infected with the two at the same time.

Age and Gender. Query 174 required *elderly* patients, whereas the patient records typically explicitly stated the age of the patient, for example 68 years old.

These queries provide concrete examples of the Semantic Gap problem and highlight how current benchmark keyword-based retrieval systems do not explicitly cater for the requirements of searching medical data. These poor performing, or hard queries, require a particular inference mechanism to bridge the semantic gap.

2.7 Summary

This chapter has outlined the semantic gap problem: the mismatch between the raw medical data and the way a human being might interpret it. A number of different types of semantic gap problems have been identified. For each, the required inference mechanism to overcome them is presented. These are summarised in Table 2.3. The table serves as a reference point for later chapters, each of which aims to address particular issues raised here.

In this chapter, the semantic gap is quantified in a retrieval experiment using a benchmark keyword-based retrieval system. The results show how keyword-based IR systems are limited in bridging the semantic gap. The chapter serves as motivation for investigating new IR models that utilise more semantics and inference mechanisms to overcome the semantic gap.

Semantic Gap Issue	Example	Inference required
Vocabulary mismatch: Synonyms, formal vs. colloquial terms, regional differences, abbreviations.	*Hypertension* ≈ *high blood pressure*	Associational
Granularity Mismatch: Hyponyms/hypernyms, queries use general terms, medical records use specific terms.	*Morphine* → *Opiate*	Deductive
Conceptual implication: Presence of certain terms in a medical records implies relevance to query	*Chemotherapy* → *Cancer*	Deductive
Inferences of Similarity: Causative and/or correlated.	Comorbidities, *anxiety* and *depression*.	Associational
Negation / family history	Phrases *no fever* and *mother had breast cancer*.	Deductive
Temporality	Past medical history.	Deductive
Age and gender	Terms *elderly*, *teenage*, *male*, *girl*.	Deductive
Levels of Evidence / Sub-documents	Query terms in laboratory reports vs. discharge summary.	Deductive

Table 2.3: Classification of semantic gap problems in searching medical data, including type of inference required to handle each.

Chapter 3

Semantic Search and Medical Information Retrieval

To know the road ahead, ask those coming back.

— Chinese proverb

This chapter provides background material on semantic search and medical information retrieval. It sets the thesis within the wider field of research, reviews relevant literature and identifies the gap we propose to tackle as part of a semantic search as inference approach.

3.1 Positioning of the Research

This is a multi-disciplinary thesis, drawing on a number of fields and application domains. It aims to bridge the gap between Ontologies (and more generally the Semantic Web) and Information Retrieval (IR). The motivation for combining these two different approaches is taken from cognitive science, where there are two dominating approaches to representing information: firstly, the *symbolic* approach, where cognition is seen as the manipulation of symbols and cognitive systems can be modelled as Turing machines; secondly, the *connectionist* approach, where knowledge is represented by connections between information,

statistical models and neural networks are examples of connectionist approaches. Gärdenfors argues that these two approaches, which are often seen as competing paradigms, actually "attack cognitive problems on different levels" [Gärdenfors, 1997, p. 255] and should, therefore, be seen as complementary.

Even though the two fields may be complementary, they are quite different in characteristics. Table 3.1 summarises the characteristics of these fields. It is important to note their dichotomous nature. We posit that a combination of these features is required to tackle the 'semantic gap' problem facing health informatics.

Semantic Web & Ontologies	**Information Retrieval**
Semantically 'rich'	Semantically 'shallow'
Inference by logical deduction	Statistical inference
Domain specific focus	Global focus
Heavy-weight	Light-weight
Lacks scalability	Scalability

Table 3.1: Characteristics of Semantic Web & Ontologies and Information Retrieval fields.

Terms 'information retrieval' and 'ontologies' actually cover a wide spectrum of different models and technologies that exhibit varying degrees of semantic richness; these are illustrated by the semantic spectrum in Figure 3.1. On the far right are formal representations, where information is encoded in ontologies, typically underpinned with a form of Description Logic [Frixione and Lieto, 2012]. In these types of systems the task of matching a user's query to relevant information can be viewed as logical inference and can be implemented with reasoning engines or theorem provers. Such systems utilise 'deep semantics'. In contrast, on the far left of the spectrum, concepts are simple tokenised words found in documents. Here inverted file indices capture the relationship between words and documents and term frequencies capture the relative importance of documents to queries. These representations make use of 'shallow semantics'. Moving from left to right we observe 'ankle deep semantics' [Hovy, 2001], where data may be represented in structured form, but the data may not necessarily be formally correct or complete. The structured representation may be built in an unsupervised manner (e.g., using Information Extraction methods) or constructed with the aid of human designers (e.g., incorporating taxonomies).

An important aspect to note is that a single system does not have to be based on only one point on the spectrum; an overall search solution may use different techniques along the semantic spectrum. This thesis aims to combine

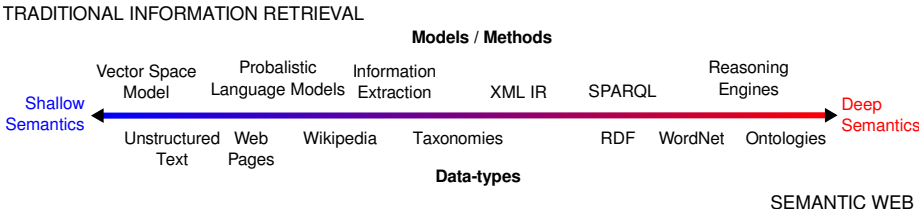


Figure 3.1: Spectrum of semantic technologies.

different methods along the semantic search spectrum; we argue that this is necessary to bridge the semantic gap in medical IR. In this chapter, deep semantic techniques are presented in Section 3.2, 'Symbolic Representations and Ontologies'. Shallow semantic techniques are presented in Section 3.3, 'Information Retrieval and Medical IR'. Work that integrates the two is presented in Section 3.4, 'Semantic Search'. Finally, we identify the gap in knowledge that we propose to address in Section 3.5, 'Semantic Search as Inference'.

3.1.1 Health Informatics

This section briefly introduces the application domain of health informatics. It is intended to contextualise the remainder of the literature review and show why health informatics is an environment where semantic search is both needed and could have significant impact.

Health informatics is a discipline at the intersection of information science, computer science and health care. It deals with the resources, devices and methods required to optimise the acquisition, storage, retrieval and use of information in health and biomedicine. Much of this information is stored in unstructured form, namely natural language. Natural language is pervasive for a number of reasons. Electronic medical records are in their infancy in many countries and those that have implemented such schemes still have enormous amounts of legacy data requiring digitisation. Additionally, as electronic medical records have been adopted using a number of different standards, interoperability between these schemes remains an open issue. Finally, medical professionals have developed sophisticated and effective natural language mechanisms to communicate with each other, for example they make extensive use to abbreviations and custom shorthand notations. As a results, they may be averse to replacing this with structured information suited to computers.

We have already introduced the 'semantic gap' as a major issue in health informatics: the mismatch between the raw medical data, such as patient records, laboratory tests or medications and the way a human (for example, a

clinician) interprets this data [Patel et al., 2007]. The ambiguity of natural language exacerbates this problem. Standardised ontologies attempt to solve this by providing an overarching semantic reference point for integrating heterogeneous data. The health informatics community has invested heavily in the development of standardised ontologies. However, as we will show, ontologies address only some of the semantic gap problems and are not well suited to dealing with natural language. An alternative approach is data-driven information retrieval, but we show that this too has limitations with respect to the semantic gap problems.[1]

Access to timely and relevant information is essential for effective delivery of health services. We deem that the challenges of dealing with this information makes semantic processing imperative. It is within this environment that a semantic search approach is required and could have significant impact.

3.2 Symbolic Representations & Ontologies

Symbolic methods involve representing concepts and relationships in a formal, structured manner. Sheth et al. [2005] define this representation as *formal semantics*, which they differentiate from *implicit semantics* such as those found in data-driven methods. Formal semantics has well defined syntactic structures and has definite semantic interpretations that make them easier for machines to process. Knowledge representation, artificial intelligence and database management are examples of research areas using formal semantics and symbolic systems. Inference is typically based on first order logic and is therefore deductive. A common realisation of symbolic information representation in information systems is the use of taxonomies and ontologies.[2] This is the underlying basis for information represented on the Semantic Web.

3.2.1 Medical Ontologies: UMLS and SNOMED CT

Two resources of medical domain knowledge are relevant to this thesis: UMLS and SNOMED CT. The Unified Medical Language System (UMLS) was developed by the U.S. National Library of Medicine and is in fact a compendium of taxonomies and ontologies in biomedical sciences. In fact, one of its ma-

[1] We use the term *data-driven* to denote approaches that are typically statistical, like those used in information retrieval. Inference in such methods can be seen as associational, in contrast to deductive inference in ontology.

[2] For the purposes of this thesis we consider a taxonomy to be a simple structured hierarchy of terms, whereas an ontology is a more expressive representation describing concepts and relationships between concepts.

jor goals is to provide a mapping between the different ontologies that make up UMLS. A concept in UMLS has a unique identifier specific to UMLS and also contains identifiers to the relevant concepts in specific ontologies. Thus, UMLS provides an overarching controlled vocabulary for medical terminologies and hence its main component is referred to as the *Metathesaurus*.

Included as part of UMLS is the SNOMED CT ontology. SNOMED CT stands for 'Systematized Nomenclature of Medicine - Clinical Terms'. It is a machine-readable collection of medical terminology covering a large range of concepts, including: disorder, procedures, organisms, body structure and pharmaceuticals [Spackman, 2008]. SNOMED CT is one of the largest domain-specific ontologies in use, with approximately 283,000 concepts, 732,000 terms and 923,000 relationships. SNOMED CT uses Description Logic as its underlying formal representation, so it is strictly a symbolic knowledge representation [Frixione and Lieto, 2012].

Concepts in SNOMED CT are represented as nodes in an acyclic graph — effectively a tree structure. Each concept has a unique identifier and a number of alternative descriptions for that concept. Concepts can be divided into a number of high-level categories, the breakdown of which is shown in the chart of Figure 3.2. SNOMED CT concepts may be defined in terms of relationships to other concepts. The most basic relationship is the inheritance, or parent-child. Thus, concepts are organised into an inheritance hierarchy. For example, Figure 3.3 shows the concept 'Viral pneumonia' as a child of 'Infectious pneumonia'. Besides inheritance, a number of other relationships can be defined between concepts. The figure shows the concept 'Viral pneumonia' has a 'Causative agent' relationship to the concept 'Virus'.

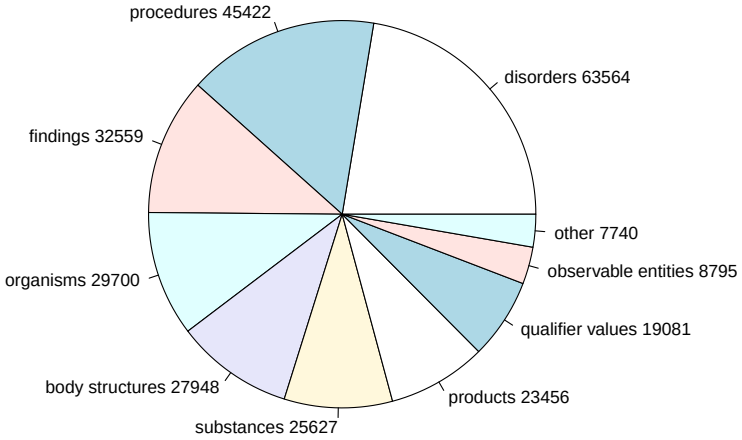


Figure 3.2: Breakdown of concept categories in the SNOMED CT ontology.

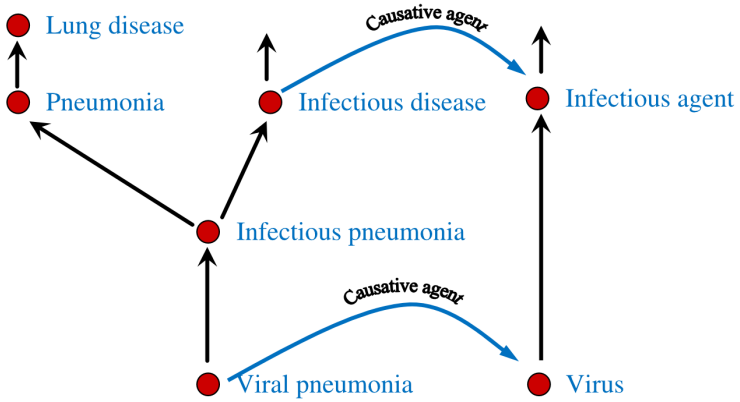


Figure 3.3: Concept hierarchy for *Viral pneumonia*.

In this thesis, we have focused on using SNOMED CT as our resource for medical knowledge. SNOMED CT covers a wide range of medical knowledge in a single, self contained resource, whereas UMLS is in fact a conglomeration of different resources, each with varying coverage. In addition, SNOMED CT has a rigorous quality control process overseen by the International Health Terminology Standards Development Organisation.[3] Finally, SNOMED CT is now mandated as the standard medical terminology in Australia and in many other counties.

3.2.2 Ontologies for Semantic Search

Having introduced ontologies and specifically the SNOMED CT medical ontology, we now analyse their applicability for semantic search. Both the advantages and limitations of ontologies for semantic search are considered. This analysis is based on both surveys of semantic search technologies [Dong et al., 2008; Mangold, 2007] and issues raised in implementations of ontology-based semantic search systems [Fang et al., 2005; Biswas et al., 2009].

Advantages of Ontologies for Semantic Search

The purpose of developing an ontology is to capture explicitly, in a standardised manner, the concepts and relationships pertaining to a particular domain. The advantages of this approach for semantic search are:

Standardisation and interoperability. Ontologies are constructed to provide an unambiguous understanding about a particular domain and this is

[3] http://www.ihtsdo.org

achieved using standardised, machine-readable languages. This has both semantic and technical advantages for their use in semantic search. From a technical perspective, this means that different systems that support the standard ontology language are able to read the ontology and process the concepts making up a given domain. The ontology (and therefore, the domain model) is decoupled from the system that acts on it — moving to a new domain simply involves moving to a new ontology. From a semantics perspective, ontologies standardise the understanding about that particular domain, thereby providing consensus on what constitutes that domain and thus reducing ambiguity. This includes standardisation around terminology (the terms used to describe different concepts). Standardisation around terminology helps to alleviate the vocabulary mismatch problem.

Inference and reasoning. Standardisation makes the ontology machine-readable, but also supports reasoning and inference. Ontologies explicitly model that given a set of axioms, certain conclusions can be inferred. For example, given the presence of *Varicella zoster virus*, one can infer the disease *Chicken pox*. Reasoning engines are tools specifically designed to make these kind of inferences. These types of inferences are important for overcoming the semantic gap problem of Conceptual Implication, where, for example, treatments or organisms logically imply certain diseases.

Explicit background knowledge. Ontologies make explicit the definition of concepts and relationships constituting a given domain. From a search perspective, these explicit definitions provide a wealth of additional information that may not be available in the data being searched but is typically understood by users. Medical records are typically authored with high level descriptions that assume substantial background knowledge that is unstated. Ontologies potentially make this implicit background knowledge explicit. By doing so, they allow inferences to be made about the information found in documents or queries. For example, ontologies make explicit the inheritance relations (parent-child); this can help alleviate the vocabulary mismatch problem.

Ontologies — specifically SNOMED CT — provide a rich resource for conceptual representation in the medical domain and hence a possible aid for semantic search. However, they have a number of limitations, which are now presented.

Limitations of Ontologies for Semantic Search

Limitations of ontologies, with respect to semantic search, stem primarily from their reliance on formal semantics. These limitations include:

Semantic similarity. Ontologies do not provide a natural means of measuring the similarity between two concepts, which we argue is crucial for semantic search. In this connection, the well known American philosopher, V. O. Quine notes:

> "... we cannot easily imagine a more familiar or fundamental notion than [semantic similarity], or a notion more ubiquitous in its application. On this score it is like the notions of logic: like identity, negation, alternation and the rest. And yet, strangely, there is something logically repugnant about it. For we are baffled when we try to relate the general notion of similarity significantly to logical terms" [Quine, 1969, p. 117].

Ontologies, based on first order logic, do not inherently represent the similarity between concepts [Gärdenfors, 2004]. In SNOMED CT, for example, there are two separate concepts, "Structure of the left knee" and "Structure of the right knee", both having the parent "Knee region structure". The left and right knee are semantically very similar and for search-related tasks the distinction is irrelevant. In contrast, "Right ventricle" and "Left ventricle" of the heart both have "Cardiac ventricle" as their parent. The distinction between the two in this case is very important as their roles are quite different. A common approach is to derive similarity by the distance between them in the ontology — these are called path-based similarity measures [Pedersen et al., 2007]. However, simple path-based measures do not naturally represent the similarity between concepts. In the above example, the left and right knee have the same path similarity as the left and right ventricle. Empirical evaluation showed that corpus-based measures are superior to path-based measures of similarity [Pedersen et al., 2007; Koopman et al., 2012b].

Uncertainty and inconsistencies. One of the advantages of symbolic systems is that they are "truth preserving"; that is, formal semantics guarantees that different systems will interpret the expressed statement in the same way — there is no ambiguity or uncertainty [Sheth et al., 2005]. The lack of uncertainty, however, is also an important limiting characteristic of these systems [Gärdenfors, 2004]. As a domain grows, it is rare to have complete agreement on a rigid conceptual model [Uschold and Gruninger,

1996; Frixione and Lieto, 2012], so it is desirable to reason with degrees of uncertainty. In addition, it may be acceptable to have contradictory statements or inconsistencies representing different views within a domain, provided that they are within different parts of the conceptual model, but ontologies typically do not deal well with such inconsistencies.

There have been some attempts to incorporate reasoning with uncertainty into ontologies and the Semantic Web. The two main approaches are probabilistic reasoning and fuzzy logic [Lukasiewicz and Straccia, 2008]. Both these approaches have the problem of how to assign prior probabilities and/or fuzzy membership functions [Sheth et al., 2005]. Also, an important open issue is the development of scalable formalisms for handling probabilistic uncertainty in ontologies [Lukasiewicz and Straccia, 2008].

Coverage. The medical domain is large and dynamic and SNOMED CT, although extensive, does not capture everything and may be lacking in areas [Dong et al., 2008]. For example, SNOMED CT captures diseases and drugs but does not specify which drug is used to treat which disease. This is a significant omission as opinions may differ on the best treatment and may change over time. Different parts of SNOMED CT are modelled with different granularity: some parts may be extremely detailed, while others may only be described at a high-level. Finally, SNOMED CT needs to be continuously updated as new medical knowledge becomes available.

Reliance on deductive reasoning. Ontologies rely on deductive reasoning as their inference mechanism. Bridging the semantic gap requires both associational and deductive reasoning, as highlighted in Chapter 2. Reliance on a single form of reasoning limits the ability to interpret medical data in different ways — similar to the way humans would. As previously mentioned, there have been attempts at incorporating uncertain inference mechanisms into ontologies, but uncertain deductive inferences still do not provide the associational inferences that we previously argued were required.

Dealing with natural language. Dealing with medical data involves interpreting natural language, a task unsuited to symbolic systems and formal semantics. Influential researcher in natural language processing, W.A. Woods, remarks:

> "...people have responded to the need for increased rigor in knowledge representation by turning to first-order logic as a semantic criterion. This is distressing, since it is already clear that

> first-order logic is insufficient to deal with many semantic problems inherent in understanding natural language as well as the semantic requirements of a reasoning system for an intelligent agent using knowledge to interact with the world." [Woods, 2004, p. 740].

A major challenge for symbolic systems is how to transform natural language into a formal representation sufficient for deductive reasoning. Automated methods are not sufficiently effective and manual methods infeasible for large amounts of data. An example of this is represented by early attempts by search engines (for example, Yahoo! and Altavista in the 1990s) to classify web pages into a taxonomy. The rate at which new webpages were added to the World Wide Web meant that this approach became unreliable and unscalable [Cohen and Widdows, 2009]. Modern search engines now adopt an automated indexing approach that includes information theory strategies. Another issue with natural language is its inherent ambiguity, both syntactically and semantically. We have already remarked on the inadequacy of symbolic systems in representing uncertainty and ambiguity.

Context Insensitive. Ontologies are constructed to capture the concepts and relationships constituting a given domain. This is typically achieved in a top-down manner: the ontological domain model is constructed first and then associated or applied to instance data. Ontologies are designed to be generally applicable and may not capture the particular characteristics of the specific data being searched. In addition, the ontology represents the view of the designers at the time, but users may have a different view of the domain and use different terms from those in the ontology [Dong et al., 2008]. In a search scenario, the top-down, designer-specific characteristics of ontologies makes them less context specific to the particular data being searched. As a consequence it can be less effective in determining the relevant information for a given query. However, the ontology may reveal well known associations that the data itself may not reveal.

Scalability / Computational Complexity. Deductive inference using ontologies is achieved using reasoning engines. For large ontologies, the tractability of such systems becomes an impediment [Mangold, 2007]. As a result, designers of ontologies and reasoning engines are forced to trade off expressiveness for tractability. Reasoning with large ontologies is computationally expensive. A semantic search system would require multiple concurrent requests with results served in a timely manner, performance

of the underlying search model being paramount to system success. A heavy-weight reasoning engine using a large ontology like SNOMED CT might not meet these requirements.

These limitations of symbolic systems are not intended to discourage their usage. It is important to point out that SNOMED CT provides a potentially very useful source of formal medical knowledge for semantic search. Instead, these limitations provide motivation for an approach that makes use of both implicit semantics with associational inference and formal semantics with deductive reasoning. The combined approach affords the possibility of exploiting the strength of both modes of inference to realise more effective semantic search.

3.3 Information Retrieval and Medical IR

Information retrieval is a wide and diverse field, as the pioneering IR researcher Gerald Salton's original general definition from the 1960s indicates:

> "Information retrieval is a field concerned with the structure, analysis, organisation, storage, searching and retrieval of information." [Salton, 1968]

This very general definition even covers the symbolic representation of information using ontologies already provided in the previous section. However, in this thesis, we adopt the standard conception of information retrieval: a user with an information need, expressed as a query, obtaining a ranked list of unstructured documents, in decreasing order of some relevance measure to the user's query. The important characteristics here are twofold: the data (documents and queries) are unstructured; and there is some measure of relevance (or uncertainty) of the document to a query. This estimation of relevance is naturally uncertain; therefore the field of information retrieval has developed a large body of knowledge around models that deal with uncertainty. These models can be considered inferential in various ways: for example, the uncertain inference that a given document is relevant to a query description or probabilistically inferring query expansion terms to augment the original query. The Probability Ranking Principle [Robertson, 1977] and Logical Uncertainty Principle [Van Rijsbergen, 1986] are two examples that illustrate the uncertain inference central to IR. Such models are in direct contrast to ontologies, where the inference is deductive. Inference under uncertainty is an important feature for semantic search — Chapter 2 has already highlighted how certain problems in medical search require associational inference, rather than deductive inference.

Besides IR's feature of inference with uncertainty, it also offers a number of other advantages. Firstly, the representation of information in IR is context specific. By representation we mean both the way the information is stored, for example in a term-document matrix and corpus statistics, and how these are used by a retrieval model, for example inferring related terms for query expansion. IR is context specific because the representation is derived from the data and, therefore, closely reflects the specific data being retrieved. If the representation is derived from the data and not handcrafted like an ontology, there is less risk of a mismatch between the designers of the model and users of the data. Deriving the representation from the data also makes the system relatively lightweight, rather than having a complex and often error prone process where designers manually construct the domain model. Many IR techniques are generally applicable, rather than domain-specific, and can therefore be applied to any domain. In contrast, with fixed, manually constructed resources such as an ontology, the ontology may need to be adapted or may not be suitable for a domain other than the one it was originally designed for. Finally, IR models are typically based on term statistics and are therefore specifically designed to work with unstructured data. As medical data is heterogeneous and much of it exists as free-text, models suited to unstructured data are naturally applicable.

Information retrieval approaches do have their limitations. The main issue for semantic search (and especially semantic search of medical data) is that IR models are dependent on terms as the representation for documents and queries. Using a term-based representation makes the model susceptible to the semantic gap problems of vocabulary and granularity mismatch. IR models are usually based on statistics from the collections used for the actual retrieval; generally no recourse is made to external sources (an exception being some IR models that derive additional statistics from external corpora [Diaz and Metzler, 2006]; these have also been applied to the medical domain [Zhu and Carterette, 2012a]). Making use of external sources is very pertinent to medical IR because medical records and the like are typically authored with high-level descriptions that assume substantial background knowledge that is unstated. Finally, uncertain inference as transacted in IR models is unsuited to the requirements of deductive inference, a mode of inference that was highlighted as being relevant to bridging the semantic gap problem.

3.3.1 Retrieval Models

Having provided a high-level definition of IR, including some advantages and limitations, we now consider some specific retrieval models. This is done for the purpose of evaluating how each may be applied to semantic search and possibly

integrated with structured domain knowledge resources such as ontologies.

Probabilistic Language Models

Most state-of-the-art information retrieval models are set within the probabilistic language modelling framework [Ponte and Croft, 1998; Hiemstra, 1998]. IR language models estimate the relevance of a document D to a query Q by the conditional probability $P(D|Q)$, where D is taken from the event space of all documents in the collection. D and Q are formed by sequences of terms drawn from a common vocabulary. Using Bayes Theorem, $P(D|Q)$ can be expressed as:

$$P(D|Q) = \frac{P(Q|D)P(D)}{P(Q)}. \tag{3.1}$$

Typically the prior probably of the query $P(Q)$ and the document $P(D)$ are assumed to be uniform. Thus, relevance of a document to a query can be estimated instead as:

$$P(D|Q) \propto P(Q|D). \tag{3.2}$$

The query Q may be made up of a number of individual terms q. If independence between query terms is assumed, as is the case with the unigram language model variant, then $P(Q|D)$ can be rewritten as:

$$P(Q|D) = \prod_{q \in Q} P(q|D). \tag{3.3}$$

For a given user's query Q, the information retrieval system returns a ranked list of documents ordered by decreasing probability of relevance, $P(Q|D)$. The estimated probabilities are often small, which can affect computers with finite precision, so the sum of logarithms is taken to produce the rank equivalent form:

$$P(Q|D) \propto \sum_{q \in Q} \log P(q|D). \tag{3.4}$$

The actual estimation of $P(q|D)$ for a given query term q can be calculated in a number of different ways. The most simple is the *Maximum Likelihood Estimate*:

$$P(q|D) = \frac{tf_{q,D}}{|D|},$$

where $tf_{q,D}$ is the term frequency of q in D, i.e., the number of occurrences of the query term q in document D and $|D|$ is the size of document D in number of terms.

Smoothing One issue with using the Maximum Likelihood Estimate is that if a document did not contain a particular query term then its estimate of $P(q|D)$ would be zero and when probabilities are multiplied in Equation 3.3, the resulting estimate for $P(Q|D)$ would also be zero. To handle such cases, *smoothing* is applied. Smoothing estimates $P(q|D)$ based on both a query term's occurrence in the document and the collection. Therefore, if the query term does not appear in the document, it will still have a non-zero probability based on its occurrence in the collection. Several smoothing methods have been proposed [Zhai, 2007]. A widely adopted smoothing method is *Dirichlet* smoothing, which combines a query term's document and collection estimates as:

$$P(q|D) = \frac{tf_{q,D} + \mu \frac{cf_q}{|C|}}{\mu + |D|}, \tag{3.5}$$

where q is the query term (which may or may not be present in the document), cf_q is the collection frequency (number of occurrences) of q, $|C|$ is the collection size (number of terms) and μ is a parameter used to control the effect of document length on the estimate. Substituting the *Dirichlet* smoothing method of Equation 3.5 into the general retrieval estimate from Equation 3.4 gives:

$$P(Q|D) \propto \sum_{q \in Q} \log \left(\frac{tf_{q,D} + \mu \frac{cf_q}{|C|}}{\mu + |D|} \right). \tag{3.6}$$

Using smoothing, an estimate of relevance can be determined for any document, even if it does not contain the query terms; all documents in the collection can be assessed for relevance. A practical limitation of this is the computation expense of assessing every document in the collection; in many cases, this may not be feasible. To overcome this issue, Azzopardi and Losada [2006] proposed a practical method of applying smoothing by first calculating the language model for an empty document: a document that contain no terms (and hence no query terms). A document model $\theta_{D_\emptyset}$ for the empty document $D^\emptyset$ that uses *Dirichlet* smoothing is:

$$\theta_{D^\emptyset} \propto \sum_{q \in Q} \frac{tf_{q,D^\emptyset} + \mu \frac{cf_q}{|C|}}{\mu + |D^\emptyset|}. \tag{3.7}$$

As the length of the empty document $D^\emptyset$ is 0 and the term frequency $tf_{q,D^\emptyset}$ is always 0, the empty document model can be simplified to:

$$\theta_{D^\emptyset} \propto \sum_{q \in Q} \frac{cf_q}{|C|}. \tag{3.8}$$

The empty document model can be calculated at indexing time by calculating

the collection statistics for each term in the vocabulary. At retrieval time, any actual document D_i being scored will first be assigned the empty document model $\theta_{D^\emptyset}$. Then, for each query term that does appear in D_i, probabilities are updated with the specific term frequency and document length statistics for D_i.

Probabilistic language models are state-of-the-art in IR and provide a formal means for modelling queries, documents and relevance estimates. Their formal foundation means extensions and adaptions can be done in a principled and formally grounded manner. They will be an important part of the unified model for semantic search proposed in this thesis.

Other Retrieval Models

Although probabilistic language models have become the state-of-the-art in information retrieval, there are other retrieval methods worthy of note. A simple, yet widely used approach, is the *tf-idf* term weighing method. The *tf* component is the term frequency, which reflects the importance of the term in the document. This is computed as the count of the term occurrences in the document. The inverse document frequency (*idf*) reflects the importance of a term in the collection. The fewer documents a term occurs in, the more discriminating the term is between documents and, therefore, the more useful it is in retrieval. Inverse document frequency is calculated as

$$idf_i = \log \frac{N}{n_i}, \tag{3.9}$$

where idf_i is the inverse document frequency for term i, N is the total number of documents in the collection and n_i is the number of documents that contain term i.

The *tf-idf* term weighting method is commonly used as part of the Vector Space Model [Salton et al., 1975]. In the Vector Space Model, documents and queries (queries can be thought of as a small document) are represented as a vector, where the elements correspond to the terms in the collection. The dimensionality of the vector is the vocabulary size of the collection. Given this geometric representation of terms and documents, measures of similarity can be developed. The most successful of these is *cosine* similarity measure [Salton et al., 1975]. Cosine similarity measures the angle between the document and query vectors; vectors are normalised so that all documents and queries vectors are of length 1. If the two vectors are identical, then the cosine angle will be 1 (the angle between them being 0). The cosine between two vectors that do not share any common terms will be 0. Cosine similarity is defined as:

$$\cos(d, q) = \frac{\sum_{i=1}^{t} d_i \cdot q_i}{\sqrt{\sum_{i=1}^{t} d_i^2 \cdot \sum_{i=1}^{t} q_i^2}}. \tag{3.10}$$

The numerator is the sum of the product of the term weights, for each matching query and document term. The denominator normalises this score by dividing by the product of the lengths of the two vectors.

Another retrieval model widely adopted in IR is the Okapi BM25 model [Robertson and Walker, 1994]. BM25 uses term frequency in the document, term frequency in the collection and document length to estimate relevance. The BM25 ranking function is:

$$\text{RSV}(D, Q) = \sum_{q \in Q} \frac{tf_{q,D}(k_1 + 1)}{tf_{q,D} + k_1(1 - b + b\frac{|D|}{|D_{\text{avg}}|})} \log \frac{|C| - df_q + 0.5}{df_q + 0.5}. \tag{3.11}$$

The left hand fraction is the term weighting component, where $tf_{q,D}$ is the term frequency of term t in document D, $|D|$ is the length of the document and $|D_{\text{avg}}|$ is the average document length. The right-hand fraction is the inverse document frequency component, where $|C|$ is the number of documents in the collection and df_q is the number of documents containing the term q. BM25 has two free parameters, b and k_1, which control the effect of term frequency and document length respectively.

It is also worth noting that the BM25 term weighing component (left-hand fraction) has been used in an alternative *tf.idf* model. This was developed by Zhai [2001] and implemented as part of the Lemur IR toolkit[4]. Thus, Lemur's *tf.idf* model encodes BM25 term weights as the components of its document vectors. Lemur's VSM also has the additional free parameters, b and k_1. This simple model was found to be the most effective in a number of experiments presented as part of this thesis; hence its mention here.

Graph-based Retrieval Models

The retrieval models reviewed so far are all bag-of-words models; that is, they do not model any term order or dependence between terms. In Chapter 2, the Semantic Gap, we identified the need to account for the innate dependence between medical concepts. Bag-of-words representations would intuitively, therefore, be limited. A number of approaches go beyond bag-of-word representations and do account for term dependence. Most common within the language modelling framework is the Markov Random Field method of Metzler and Croft [2005].

[4]Lemur is an open source IR package developed at the University of Massachusetts, Amherst and Carnegie Mellon University, available at http://www.lemurproject.org/

However, there are alternative term dependence models that are particularly relevant to semantic search, namely graph-based retrieval models.

Graph-based models have been applied in information retrieval, generally as part of connectionist approaches [Doszkocs et al., 1990]. Shifting weights between vertices in a graph is the basis for the Inference Network model of Turtle and Croft [1991] and the basis for the InQuery language used as part of Lemur. Graphs provide a convenient means of representing information for IR applications: the propagated learning and search properties of a graph provide a powerful means of identifying relevant information items (be they terms or documents) [Blanco and Lioma, 2012]. Graph-based algorithms — such as the popular PageRank algorithm [Page et al., 1999] — are examples of graph theoretic properties that can be utilised very effectively in an information retrieval scenario.

Blanco and Lioma [2012] developed a graph-based term weighting model that represents each document as a graph: vertices are terms and edges are relationships between terms. Relations may be simple co-occurrence relations within a context window, or more complex grammatical relations. The importance of a term within a document can then be estimated by the number of related terms and their importance, much in the same way PageRank estimates the importance of a page via the pages that link to it.

Graph-based representations also underly most ontologies. (Concepts in the ontology can be viewed as nodes, while relationships between concepts are edges). Certainly, this is the case for the major medical ontologies, SNOMED CT and UMLS. A graph-based representation is therefore a common feature between ontologies and the above mentioned retrieval models that aim to capture term dependence. We hypothesise that graph-based models may be very relevant to semantic search as they capture the dependencies between terms and provide a means of integrating ontologies. A graph-based representation is therefore a strong candidate for a unified model of semantic search as inference.

This section has presented some IR models of relevance to semantic search. Further detail on some models is reserved for the actual chapter where they are applied. Specifically, graph-based models are further detailed in Chapter 5 and Logic-based IR models are introduced in Chapter 6.

3.3.2 Evaluation in Information Retrieval

The ultimate goal of evaluation in information retrieval is to measure how well a user's information need is met by a ranked list of documents returned for a specific query. There is a long history of empirical evaluation in IR and robust assessments of retrieval systems is ingrained in the IR community [Cleverdon,

1991]. This section reviews some of the related work in IR evaluation. Specific semantic search evaluation issues are considered later in the chapter.

IR evaluation is based on statistical measures of retrieval effectiveness. Most measures are designed to quantify two elements of effectiveness: precision and recall [Manning et al., 2008]. Precision is a measure of what portion of the retrieved documents is relevant, or more formally:

$$\text{precision} = \frac{|D_{\text{rel}} \cap D_{\text{ret}}|}{|D_{\text{ret}}|}, \tag{3.12}$$

where D_{ret} is the set of retrieved documents and D_{rel} is the set of relevant documents. In contrast, recall is a measure of what portion of the relevant documents is retrieved, or:

$$\text{recall} = \frac{|D_{\text{rel}} \cap D_{\text{ret}}|}{|D_{\text{rel}}|}. \tag{3.13}$$

In medical IR there are different use cases requiring either the maximisation of precision or recall. A common scenario where both are required is the case of searching for patients eligible for inclusion in clinical trials [Voorhees and Tong, 2011; Voorhees and Hersh, 2012]. Clinical trials are conducted in the development of new drugs or procedures. Finding relevant patients to conduct a clinical trial can be seen essentially as a retrieval problem — the clinical trial inclusion criteria being the information need and the patient records being the document corpus. For an information need of finding patients with a rare disease, it is most important for the retrieval system to return all relevant documents (maximise recall). In this case, the user would much prefer to view many irrelevant patients than miss one of the rare relevant patients. Conversely, for a common disease, where there are a large number of relevant patients, precision is important. Users do not need all the relevant documents, but they don't wish to read irrelevant documents. Precision and recall are incorporated into a number of standard evaluation metrics. The metrics specific to the experiments and evaluation in this thesis are outlined in further detail below.

Precision at certain rank positions — for example precision at 10 — measures the number of relevant documents up to the stipulated rank position. Given a rank position n, the precision @ n is:

$$\text{precision @ n} = \frac{\sum_{i=1}^{n} \text{rel}(d_i)}{n}, \tag{3.14}$$

where $\text{rel}(d_i)$ is a function, such that $\text{rel}(d_i) = 1$ if the document d_i is relevant and $\text{rel}(d_i) = 0$ otherwise. This measure would be most appropriate when precision maximisation is important, for example the case of finding common

diseases or conditions.

Recall can also be measured at specific rank positions:

$$\text{recall @ n} = \frac{\sum_{i=1}^{n} \text{rel}(d_i)}{R}, \tag{3.15}$$

where R is the total number of relevant documents. The rank position n is often set to total number of documents returned.

Rather than have two separate measures for precision and recall, it is desirable to have a measure that encompasses both. Average precision (AP) is such a measure, calculated as:

$$\text{AP} = \frac{1}{R} \sum_{n=1}^{N} \text{P@n}, \tag{3.16}$$

where R is the number of relevant documents and N is the number of documents returned. Mean Average Precision, or MAP, is the average precision across the set of queries Q:

$$\text{MAP} = \frac{\sum_{q \in Q} \text{AP}(q)}{|Q|}. \tag{3.17}$$

MAP is a widely used measure in IR evaluation. However, it does rely on the completeness assumption: that all relevant documents within a test collection have been identified [Cleverdon, 1991]. When this assumption is violated (i.e., a substantial number of relevant documents are not assessed), then the standard evaluation measures outlined above are not robust. (More discussion on the effect of this is presented in Chapter 7.) To deal with this situation, Buckley and Voorhees [2004] introduced the *bpref* evaluation measure; bpref was specially designed to deal with incomplete relevance judgements. Bpref differs in that it considers only the documents that are explicitly assessed, whereas other measures typically assume that unjudged documents are irrelevant. Bpref is calculated as:

$$\text{bpref} = \frac{1}{|R|} \sum_{r \in R} \left(1 - \frac{|\forall n(n \in \bar{R} \wedge n < r)|}{|R|}\right), \tag{3.18}$$

where r is a document in the set of relevant documents R, n is a non-relevant document in the set of non-relevant documents $\bar{R}$, such that n occurs before r in the ranked list. Documents that have not been assessed for relevance do not affect the effectiveness measure.

TREC and the Medical Records Track

The evaluation methodology and measures outlined above are at the heart of the Text REtrieval Conference (TREC) evaluation campaign [Voorhees and Harman, 2005]. TREC aims to provide a common platform to evaluate information

retrieval systems by developing IR test collections. A test collection is made up of a corpus of documents, a set of queries (often called topics) and a set of relevance judgements provided by expert users. The document corpus and associated queries are made publicly available to teams participating in TREC. Teams use whichever retrieval method they have developed to run the queries and submit their results, in the form of a ranked list of documents, to the TREC organisers. The organisers then evaluate each team's submission according to the relevance judgements.

Early test collections contained a small number of documents; for example, the Communications of the ACM article collection (CACM) contained only 3024 documents. Such small document collections are possible to assess completely by expert judges. However, as document collections have grown — the ClueWeb collection contains 1.2 billion web documents — it has become infeasible to assess all but a small subset of documents. To deal with this issue, TREC utilised *pooling* techniques to select an appropriate subset of documents for assessment by experts. Pooling is done by taking a sample of documents for each query from each participating team. These documents are merged into a single set (called the *pool*), which is then provided to the expert assessors for judging. The intuition behind pooling is that if enough diverse systems contribute to the pool, then a representative subset of document will be assessed and the relevance judgements should not favour any particular system [Voorhees and Harman, 2005].

TREC is organised into separate sub-challenges, called Tracks, which focus on particular retrieval applications (for example the Web Track is specific to searching web documents). In 2011, TREC introduced the Medical Records Track (MedTrack), designed to "foster research on technology that allows electronic health records to be retrieved based on the semantic content of free-text fields" [Voorhees and Tong, 2011]. The document collection used in TREC MedTrack were 100,866 de-identified clinical record documents from U.S. hospitals. Topics and relevance judgments were created by medical physicians, with the topics reflecting the types of queries that might be used to identify eligible patients for inclusion in clinical trials. This test collection has been used extensively in our experiments and empirical evaluation.

The most successful teams participating in TREC MedTrack used a variety of techniques. In 2011, the best approach was by King et al. [2011], who focused on two aspects: information extraction and query expansion. For information extraction, they applied a number of NLP techniques to either reduce "uninformative content" or identify specific types of content, such as age, gender, negation or discharge diagnoses. Although handling such content provided significant improvement in retrieval performance, the approach was very specific

to the particular documents used in the MedTrack corpus, this making these approach less generally applicable. The second major approach employed was query expansion, which utilised related terms in UMLS and a number of external medical encyclopedia. The encyclopedia were treated as an external collection and query expansion terms derived using pseudo relevance feedback. Both the UMLS and encyclopedia-based query expansion results in gains in performance. Overall, King et al. [2011] applied a number of small, different techniques, each of which added some improvements in retrieval effectiveness. With respect to future directions, they remarked that concept-based indexing could be a useful technique but that further investigation was need to determine how it might be reliably applied.

In 2012, the best approach was by Zhu and Carterette [2012b], who focused on applying standard probabilistic language model approach to the task. Specifically, they first applied the Markov Random Field model of Metzler and Croft [2005] to capture term dependencies. They then investigated the effect of query expansion approaches that utilised external collections (for example, such as Wikipedia or ClueWeb). These were formally integrated into their retrieval model using the Mixture of Relevance Model proposed by Diaz and Metzler [2006]. Finally, they investigated how scoring different report types affected retrieval effectiveness; essentially tackling the Levels of Evidence semantic gap problem (Section 2.5.4). Overall, they applied a number of well known IR methods within probabilistic language modelling framework and found that each provided some improvement in retrieval effectiveness.

3.3.3 Summary — Information Retrieval and Medical IR

The notion and estimation of uncertain relevance is central to information retrieval. For semantic search, inference with uncertainty is an important requirement and IR models are therefore suited to this task. Another advantage of IR models is that the model is derived from the data. This makes the models context-specific, light-weight and well suited to dealing with natural language.

A limitation of IR models is the dependence on terms as the representation for documents and queries, making them susceptible to the vocabulary and granularity mismatch. Also, IR models do not capture background or explicit knowledge (particularly prevalent in medical data). This limits the inferences that can be made using the raw data found in documents and queries.

A number of specific retrieval models have been presented, some of which will be used as baselines for comparison of our models. Probabilistic language models are the current state of the art. Another retrieval model relevant in this thesis is graph-based retrieval. Graphs naturally capture interdependence

(between terms or concepts), identified as one of the semantic gap problems. In addition, graph-based representations also underlie most ontologies; therefore, graphs are a common feature to both ontologies and graph-based IR models. This common feature provides a means of integrating the two into a possible unified model for semantic search.

3.4 Semantic Search

Semantic search aims to retrieve documents relevant to a query, not based on just the presence of the query terms in the document, but also based on the meaning of the document and query. The focus is on deriving a higher level meaning of the queries and documents based on its content. High-level meaning might be provided by ontologies, but we have shown that pure-ontology approaches have a number of limitations. Additionally, we have outlined the limitation of standard information retrieval approaches. A hybrid approach, therefore, offers potential. A number of initiatives employ hybrid approaches and can be generally referred to as 'concept-based information retrieval'. We consider how successful previous work has been and identify the gap in knowledge that we propose to tackle as part of a semantic search as inference approach.

3.4.1 Concept-based Information Retrieval

Broadly, concept-based IR aims to make use of external knowledge sources (such as thesauri or ontologies) to provide additional background knowledge and context that may not be explicit in a document collection and users' queries. Generally, concept-based approaches fall into two categories. Most common are approaches that maintain the original term representation of documents and use a concept-based approach to improve the query representation. Previous work in medical IR most often falls into this category. The most basic approach within this category is thesaurus-based query expansion. The other category comprises far less common approaches that map the terms in documents to higher-level concepts. Retrieval is then done in 'concept space' rather than 'term space'. We review each of these categories individually.

Concept Augmented Term-based Retrieval

To start with, we review approaches that utilise concept-based representations while maintaining the original term-based representation. Important early work

in this area was done by Voorhees [1994], who investigated whether retrieval could be improved by expanding queries with WordNet synonyms. Results showed that it is very difficult to select appropriate expansion terms automatically. Human, hand-picked terms were, however, successful at improving results. Voorhees's findings also showed that performance in concept-based IR is highly dependent on the specific domain model or ontology used. General applications (those that utilise WordNet or Open Directory) struggle to outperform keyword-based systems [Voorhees, 1994; Ravindran and Gauch, 2004; Egozi et al., 2011]. As a result, concept-based IR has gained little traction in general applications. However, biomedical applications (which use domain-specific ontologies) do demonstrate consistent improvements [Zhou et al., 2007; Liu and Chu, 2007; Koopman et al., 2012a]. Research in these application areas has been more active. After Voorhees's early query expansion method, subsequent models attempted to improve the query model with concept-based representations. This was done with the aim of addressing the vocabulary mismatch problem. Query terms are normalised to concepts, the motivation being that a concept encapsulates all the lexical variants of the same term into a single entity. At retrieval time, it does not matter which term variant is used, as each variant of the term will map to the same overarching concept. Zhong and Huang [2006] successfully applied this approach to searching genomics data, although they limited the concepts to represent lexical variants of only genes in TREC Genomics Track data. Based on this initial work, there have been subsequent attempts to use concepts within probabilistic language models. Trieschnigg et al. [2010] and Trieschnigg [2010] built a query language model as a probability distribution over concepts. These approaches did demonstrate statistically significant improvements in retrieval, although with limited gains, but often the approach was very specific to the task at hand (for example, only applicable to searching genomic data).

The literature points to a critical successful factor being approaches that combine corpus-based statistics with domain knowledge. This was the finding of Stokes et al. [2008], who conducted an extensive survey on the criteria for successful query expansion. (Although specific to the genomics domain, a number of their findings can be generalised to medical IR.) Query expansion approaches that relied on only domain knowledge resources failed to provide consistent improvements in retrieval performance. However, those that augmented term-retrieval with concepts from genomics domain resources did demonstrate improvements. Methods that combine corpus-based statistics with domain knowledge were most successful. Based on this, a number of avenues have been explored that leverage more data-driven methods within a concept-based approach.

Concepts can be integrated into probabilistic language models to create a concept-based representation of the query. This is peformed pre-retrieval and therefore independent of retrieved document contents. Meij et al. [2010] and Trieschnigg [2010] extended this work by using pseudo relevance feedback to generate an updated concept-based query model. Their results showed that integrating corpus-based statistics with domain knowledge was the key component for successful query expansion.

Liu and Chu [2007] found that medical queries could be matched to a number of different scenarios, for example *treatments*, *diagnosis*, *symptoms*. The UMLS ontology provides the relevant domain knowledge about these overarching scenarios. Standard statistical query expansion methods could be applied, but then filtered based on concepts matching these specific medical scenarios. This combined statistical and ontology-based heuristic outperformed both a pure statistical and pure ontology query expansion approach. [Zhou et al., 2007] took this a step further by integrating semantic types: the higher level grouping of medical concepts into classes, for example *diseases*, *organisms*, *substances*, etc. Using concepts, semantic types and corpus statistics, they were able to derive implicit relations between concepts, which could be used for query expansion. Deriving these implicit relations represents one of the few approaches that used an *inference* mechanism within the retrieval model; this was the best approach at the TREC Genomics Track [Zhou et al., 2006]. The IR research just described was focused within the genomics domain, which is a very specific retrieval scenario. Queries are provided in the form "`Gene` (1..n) `Biological process` (1..m)" and the task is to return relevant information about the specific genes. Therefore, a number of methods are specific to this domain and cannot be applied to *ad hoc* retrieval scenarios outside this domain. They do, however, highlight that successful approaches generally utilise both domain knowledge and statistical, data-driven methods.

Pure Concept-based Retrieval

The concept-based IR literature so far falls into the category of utilising concept-based representations, while maintaining the original term-based representations. Now we consider the alternative category, which maps the terms in documents to higher-level concepts; retrieval is then done in 'concept space' rather than term space. Outside of the medical domain, a successful example of this approach is the Explicit Semantic Analysis (ESA) retrieval method [Egozi et al., 2011]. ESA is a technique that represents the meaning of texts in a high-dimensional space of concepts, where the concepts are derived from Wikipedia. Each Wikipedia article represents an individual concept and is identified

by the article title. Documents and queries are represented as concept vectors, rather than term vectors; retrieval is done by comparing these concept vectors. The motivation for ESA is that the concept representation captures 'explicit human knowledge' [Egozi et al., 2011] from Wikipedia within a data-driven (Vector Space Model) IR framework. Retrieval results using ESA show that pure concept-based approaches can be successful, especially in alleviating the vocabulary mismatch problem by representing queries and documents as higher-level concepts. Another approach that used concept-based representations was the KeyConcept system developed by Ravindran and Gauch [2004]. KeyConcept first mapped documents into a concept hierarchy; retrieval was done by combining a term-based score for the documents with a concept-based score, derived from the hierarchy of concepts. Their method is also relevant in that they utilised a combined term-concept representation. The document was scored by linearly interpolating the term and concept scores. They explored the weighting mix between terms and concepts and found the best results were obtained when both terms and concept scores were included.

Early work on developing and evaluating medical IR systems did focus on concept-based indexing and matching using UMLS; much of this research was conducted as part of the development of the SAPHIRE system [Hersh and Hickam, 1995]. The system attempted to identify concepts in both the document and the query and then match these at retrieval time. While this early work can be viewed as the first pure-concept based approach in medical IR, it was limited in scope: concepts were matched using a basic suffix striping method [Hersh et al., 1990]; the concept-matching algorithm was either Boolean matching or only considered inverse document frequency; and user's had to manually identify the most appropriate concepts for a query before documents were retrieved. In addition, the evaluation was done on medical journal abstracts, which are carefully authored and summative in nature; this is in contrast to other sources of medical data, such as patient records.

The preliminary work on mixing terms and concepts [Ravindran and Gauch, 2004] was more rigorously studied within the biomedical domain by Trieschnigg et al. [2010]. They approached the incorporation of a concept-based representation from a cross lingual perspective, which involves translating between term and concept language models; concepts were taken from either MeSH or UMLS. They experimented with a number of cross-lingual-based translation methods. The most effective translation model utilised corpus statistics derived from pseudo relevant documents. This again demonstrates the importance of including statistical methods in a concept-based approach. The approach of Trieschnigg et al. [2010] demonstrated improvements over a term baseline; however a pure concept baseline was not included.

A pure concept-based approach has largely been unexplored with the medical domain.[5] Such an approach requires the conversion of the entire document corpus to concepts derived from some domain-specific resource. If this approach is employed, a critical requirement is the coverage and quality of the domain-specific resource. In general applications, no such domain-specific resource of sufficient size and quality exists. However, the medical domain is unique in that considerable effort and resources have been expended in the development and ongoing maintenance of extensive, high quality representations of medical knowledge. In addition to a high quality domain-specific resource, a pure concept-based approach would require an accurate method to convert terms to concepts. The biomedical NLP field has tackled this problem in depth [Aronson and Lang, 2010; Liu et al., 2011; Meystre and Haug, 2006], developing tools such as the MetaMap system, which is effective at mapping free-text to UMLS concepts [Pratt and Yetisgen-Yildiz, 2003]. A contribution of this thesis is the development and evaluation of pure concept-based representations for medical IR; this is presented in the next chapter. Concept-based representations partially address the requirements for semantic search and demonstrate improvements in retrieval effectiveness over state-of-the-art term-based IR models (as shown in the next chapter). Based on this, in Chapter 4, we extend concept-based representations to incorporate inference mechanisms, which make far greater use of domain-specific knowledge, to realise semantic search as inference.

3.5 Semantic Search as Inference

Concept-based retrieval approaches show promise in medical information retrieval; they have been successful in genomics applications at least. In concept-based IR, the representation of queries and documents is augmented with higher-level concepts. This has the advantage of making the IR model less dependent on the individual terms used, thus overcoming the vocabulary mismatch problem. Concept-based IR utilises domain-specific resources (concepts from medical ontologies) and data-driven IR methods. Most concept-based IR approaches maintain the original term representation of documents and use a concept-based approach to improve the query representation. Alternatively, there have been some pure concept-based approaches (e.g., Explicit Semantic Analysis), but with little focus within the medical domain. Pure concept-based approaches have largely been unexplored. One reason for this is the lack of available means to convert from terms to concepts. However, concept identification methods (such

[5] A recent exception is work by Limsopatham et al. [2013b] which leverages similar methods to those proposed in the next chapter of this thesis.

as the MetaMap system [Aronson and Lang, 2010]) are now mature enough to achieve accurate term-concept mapping. An aim of this thesis, therefore, is developing and evaluating pure concept-based retrieval methods in detail. This is provided in the next chapter (Chapter 4).

Although concept-based representations show promise, they still only address the vocabulary mismatch problem. Chapter 2 highlighted a number of other semantic gap problems, including granularity mismatch, deductive inference and conceptual inference. In the introduction, we posit that bridging the semantic gap involves addressing two issues: *semantics*, which is aligning some meaning behind words in documents and queries; and *inference*, which is determining the association between two concepts. Concept-based IR addresses the issue of semantics by representing documents and queries with higher-level concepts. However, this lacks the necessary inference mechanism to deal with the other semantic gap problems. To address this we propose extending concept-based representations so that this inference mechanism may be realised.

Our foundation for supporting inference in concept-based representations lies in graph-based representations and graph-based retrieval models. Graphs have a number of characteristics that align with the requirements of semantic search as inference. The edges in a graph naturally capture interdependence (between terms or concepts), identified as one of the semantic gap problems. The propagation of information over a graph — such as the popular PageRank algorithm — provide a powerful means of identifying relevant information items (be they terms, concepts or documents). Importantly, graph-based representations also underly most ontologies; therefore, graphs are a common feature of both ontologies and a branch of retrieval models that also use graph-based representations. We hypothesise that graph-based representations and graph-based retrieval models provide a sound basis for a unified model of semantic search as inference. Specifically, that graph-based features and the propagation of information over a graph will provide the necessary inference mechanism needed to bridge the semantic gap. Two different graph-based retrieval models, which extend concept-based IR models, are proposed as part of this thesis. Chapter 5 presents a novel graph-based concept weighting model. Finally, our unified model of semantic search as inference is provided in Chapter 6.

This chapter has considered previous attempts at integrating ontologies and information retrieval methods, concept-based IR methods being the most relevant in this area. The requirement is for a unified model of semantic search as inference, one that combines IR methods and domain-specific resources within a single principled framework. Much of the previous work has been task-specific (for example, searching for gene-disease interactions). As such, it is *ad hoc* and often heuristic, making it difficult to extend or adapt to different applications or

domains. Some methods, like the concept-based language models, have a strong theoretical basis. However, these models are aimed at addressing the issue of *semantics* and do not tackle the issue of *inference*, which we have highlighted as essential for bridging the semantic gap.

Chapter 4

Bag-of-Concepts Model

Yet in each word some concept there must be...

— Goethe's *Faust*, Part I, Scene III

Bridging the semantic gap involves addressing two issues: *semantics* and *inference*. This chapter focuses on the issue of semantics. We present a novel 'Bag-of-concepts' retrieval model, where queries and documents are represented as high-level concepts — taken from medical ontologies — rather than terms. This approach is reviewed in light of the semantic gap issues presented in Chapter 2 and we show how converting to high-level concepts addresses vocabulary mismatch. Conceptual representations differ both semantically and statistically from term-based representations. We show that it is these differences that contribute to an effective retrieval model using concepts. An empirical evaluation of the Bag-of-concepts model using the TREC Medical Records Track shows the effectiveness of the model when compared to state-of-the-art term-based models, especially at improving hard queries.

The chapter concludes with the finding that although the Bag-of-concepts model is effective, it addresses only some of the semantic gap issues, mainly vocabulary mismatch. This provides motivation to leverage much deeper domain-knowledge to support the necessary *inference* mechanism required for semantic search.

4.1 Methods

Two separate processes are performed in the construction of the Bag-of-concepts model: first, term to concept conversion using the MetaMap system; second, concept document indexing and concept query retrieval. These processes are described separately in the subsections below.

4.1.1 Converting Terms to Concepts

As much of the medical data available is in free-text form, one of the major hurdles for using structured domain-knowledge resources such as ontologies is how to map the unstructured data to relevant entries in an ontology. To address this problem, the U.S. National Library of Medicine has developed a tool called MetaMap [Aronson and Lang, 2010] that extracts UMLS concepts from free-text; it is the state-of-the-art for medical concept identification [Pratt and Yetisgen-Yildiz, 2003]. MetaMap is widely adopted in medical NLP [Meystre and Haug, 2006; Nadkarni et al., 2011]. To understand how Metamap works, consider the example output of the MetaMap system using the input string 'heart attack and renal failure' shown in Figure 4.1.

MetaMap first analyses the input string and chunks the text into three individual phrases: "heart attack", "and" and "renal failure". We focus on the first phrase seen in Figure 4.1❶: "heart attack". For this phrase, the system produces a ranked list of possible matching candidate concepts (shown in Figure 4.1❷); in this case, there are eight candidate concepts. Included with each candidate concept is its identifier (e.g., `C0277793`), a confidence score (between 0 and 1000) and the concept's description. The highest ranking candidate(s) is/are selected from the list of candidates (shown in Fig 4.1❸). In this example, the single candidate *Heart attack (Myocardial Infarction)* (`C0027051`) is selected but in other cases multiple candidates may be selected for a single phrase (more details concerning this situation are presented later in the chapter).

Metamap performs the concept identification process through a pipeline of different sub-processes illustrated in Figure 4.2.

Firstly, the input goes through a **lexical and syntactic analysis** phase:

1. The raw text is tokenized, firstly into sentences and then into individual words. Abbreviations and acronyms are expanded to their full form.

2. For each sentence, part-of-speech tagging is performed[1].

[1] In corpus linguistics, part-of-speech tagging is the process of marking up a word as corresponding to a particular part of speech, such as noun, verb, adjective, etc.

```
|: heart attack or renal failure
|:

Phrase: "heart attack"  ❶
Meta Candidates (8):  ❷
  1000 C0027051:Heart attack (Myocardial Infarction) [Disease or Syndrome]
   861 C0018787:Heart [Body Part, Organ, or Organ Component]
   861 C0277793:Attack, NOS (Onset of illness) [Finding]
   861 C0699795:Attack (Attack device) [Medical Device]
   861 C1261512:attack (Attack behavior) [Social Behavior]
   861 C1281570:Heart (Entire heart) [Body Part, Organ, or Organ Component]
   861 C1304680:Attack (Observation of attack) [Finding]
   827 C0004063:Attacked (Assault) [Injury or Poisoning]
Meta Mapping (1000):  ❸
  1000 C0027051:Heart attack (Myocardial Infarction) [Disease or Syndrome]

Phrase: "and"

Phrase: "renal failure"
Meta Candidates (Total=6; Excluded=0; Pruned=0; Remaining=6)
  1000   C0035078:Renal Failure (Kidney Failure) [Disease or Syndrome]
  1000   C0341697:renal failure (Renal impairment) [Disease or Syndrome]
  1000   C1963154:Renal failure (Renal Failure Adverse Event) [Finding]
   861   C0022646:Renal (Kidney) [Body Part, Organ, or Organ Component]
   861   C0231174:Failure (Failure (biologic function)) [Functional Concept]
   861   C0680095:failure (Personal failure) [Individual Behavior]
Meta Mapping (1000):
  1000   C1963154:Renal failure (Renal Failure Adverse Event) [Finding]
```

Figure 4.1: MetaMap output for `heart attack or renal failure`. ❷ shows a ranked list of possible matching candidate concepts. The highest ranking candidate is shown in ❸.

3. Finally, a syntactic analysis step is performed where terms are checked against the UMLS SPECIALIST lexicon. The lexicon contains syntactic, morphological and orthographic information for biomedical specific words. It is used to provide MetaMap with additional information about stemming or part-of-speech tagging of biomedical specific terms.

The output of this process is a number of phrases. Each phrase then goes through the following **concept mapping** process:

1. Variant generation — different variants of the terms in the phrase are identified;

2. Candidate identification — UMLS concepts matching the phrase text and its variants are identified; a candidate score (representing how well the concepts match the phrase text) is assigned to each concept;

3. Mapping construction — candidate concepts from the previous step are compiled into a ranked list of concepts (ordered by candidate score) that best match the phrase text; and optionally,

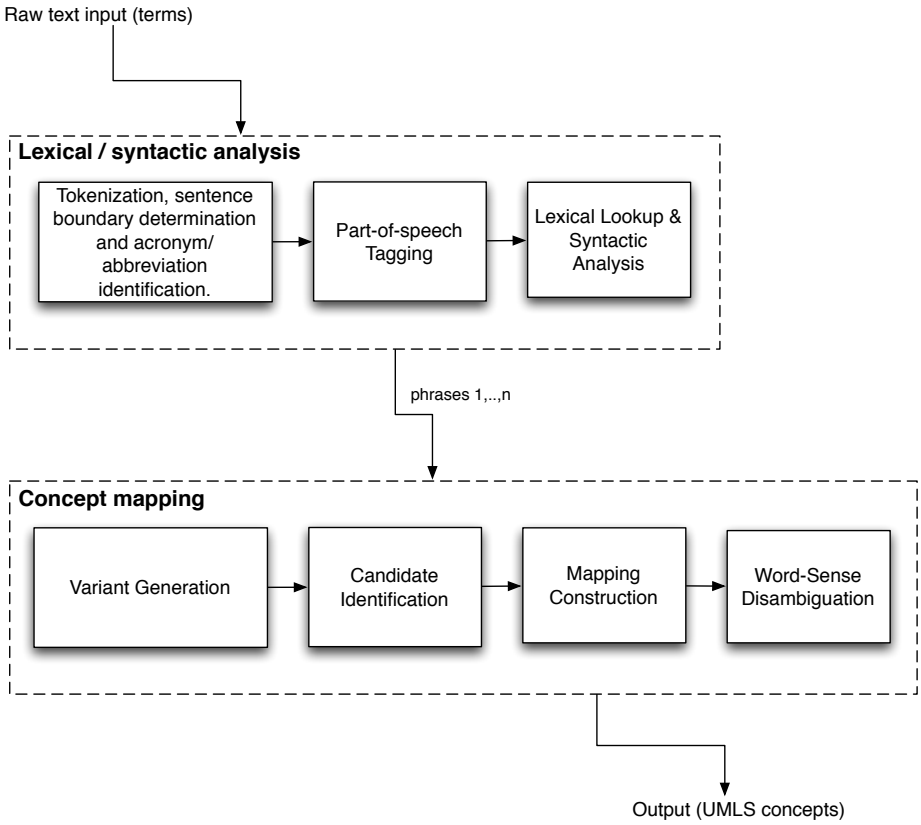


Figure 4.2: Metamap pipeline.

4. Word-sense disambiguation — concepts from the previous step are further filtered based on the semantic types of the surrounding text.[2]

The output from this process is a sequence of UMLS concepts for each phrase of input text. A more detailed example of the the conversion of term to concepts using MetaMap is provided in Appendix A.

Comparisons with human subjects have shown that MetaMap is effective in concept identification tasks (84% precision, 70% recall) [Pratt and Yetisgen-Yildiz, 2003]. Medical concept identification has been an important goal for extracting meaning from medical free-text [Hersh, 2009, p. 312]. However, much of the focus has been specifically on the concept identification task, or on categorising documents with concepts [Zheng et al., 2010; Kim et al., 2010], and less on the application of concept identification in information retrieval.

Metamap performs the important role of mapping free-text to medical concepts. It can be used to build concept-based representations of queries and

[2]Semantic types are high-level medical categories to which each concept belongs, such as *disorder*, *treatment* and *pharmaceutical*.

documents. Although a concept-based representation can be useful in itself (as our empirical evaluation will show), it also provides a means to make further use of other domain knowledge — such as the relationships between concepts — to provide inferencing capabilities, as we shall do in Chapter 6. Metamap provides a means to bootstrap the use of greater domain knowledge and is used extensively in our experiments in later chapters.

4.1.2 Indexing and Retrieving using Concepts

The previous section described the process of mapping terms to concepts. This section puts that process within the wider architecture of a Bag-of-concepts retrieval model. This architecture is illustrated in Figure 4.3. A sequence of steps is performed to develop the system:

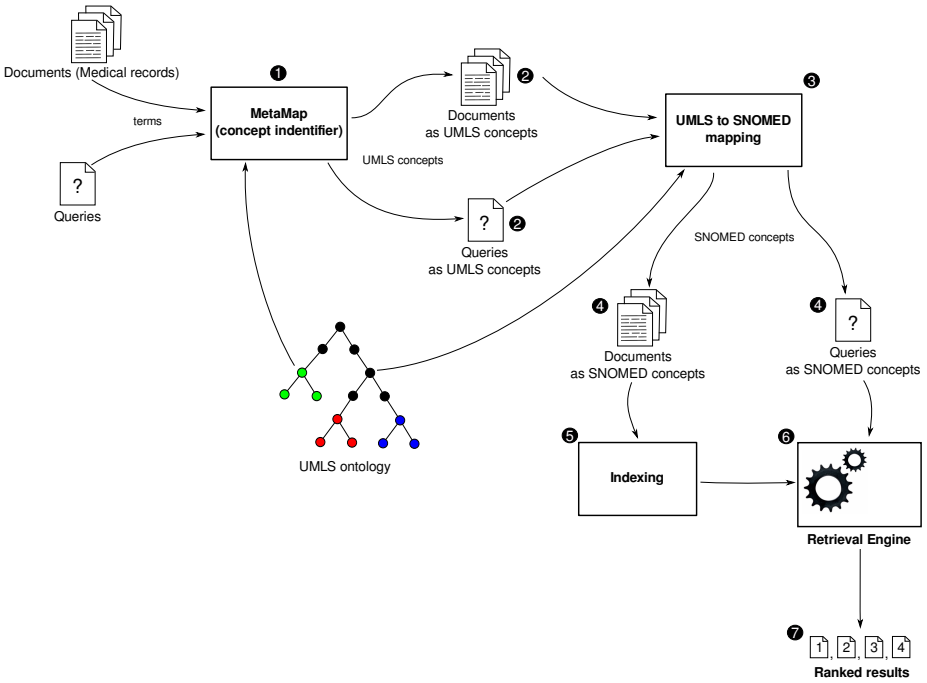


Figure 4.3: Architecture for concept-based medical information retrieval. See text for an explanation of numbered steps.

❶ The original queries and documents are fed to Metamap, which returns a sequence of UMLS concept identifiers.

❷ Each document and query is now represented as a list of UMLS concept ids (e.g. `C0027051`) rather than the original terms (e.g. `heart attack`). Documents now contain only medical concepts.

❸ The UMLS concepts are then mapped to their SNOMED CT equivalents. This mapping is provided as part of the UMLS Metathesaurus.[3]

❹ Queries and documents are now represented as a list of SNOMED CT concept ids.

❺ Documents are indexed using a standard IR search engine. The system treats the documents as a 'Bag-of-concepts'.

❻ The queries (represented as SNOMED CT concept ids) are issued to the retrieval engine.

❼ A ranked list of documents is returned and can be compared to relevance judgements to determine retrieval performance.

The figure shows that UMLS concepts are converted to SNOMED CT concepts prior to indexing and retrieval. An alternative is performing indexing and retrieval directly on UMLS documents and queries. An evaluation and discussion on the difference between these two representations is provided later in this chapter.

A number of retrieval models could be applied in the implementation of the Bag-of-concepts model. In this thesis, we focus on two state-of-the-art models: a probabilistic language model with Dirichlet smoothing and Lemur's tf-idf implementation.[4] The language model is chosen as it is widely used and state-of-the-art for keyword-based retrieval and Lemur's tf-idf model is chosen as it performs particularly well on the patient record collections used in medical IR.

A concept-based probabilistic language model can be built by applying the same method as that used for terms (covered in Section 3.3.1, Chapter 3). That is, for a concept-based query Q_c and a concept-based document D_c, both made up of one or more concepts:

$$P(Q_c|D_c) \propto \sum_{q_c \in Q_c} \log \left(\frac{cf_{q_c,D_c} + \mu \frac{Cf_{q_c}}{|C_c|}}{\mu + |D_c|} \right), \tag{4.1}$$

where cf_{q_c,D_c} is the concept frequency of query concept q_c in document D_c, Cf_{q_c} is the collection frequency (number of occurrences) of concept q_c in the collection and $|C_c|$ is the collection size in number of concepts.

[3]SNOMED CT was chosen because it covers a wide range of medical knowledge in a single, self contained resource, whereas UMLS is in fact a conglomeration of different resources, each with varying coverage. In addition, SNOMED CT is now mandated as the standard medical terminology in Australia and in many other counties.

[4]Note that Lemur's tf-idf variant uses the BM25 term weighting component.

The second retrieval model utilises Lemur's tf-idf retrieval function; the retrieval status value (RSV) for D_c under query Q_c is then:

$$\text{RSV}(D_c, Q_c) = \sum_{q_c \in Q_c} \frac{cf_{q_c,D_c}(k_1+1)}{cf_{q_c,D_c} + k_1(1-b+b\frac{|D_c|}{|D_c^{\text{avg}}|})} \log \frac{N}{n_{q_c}}, \tag{4.2}$$

where cf_{q_c,D_c} is the concept frequency within the document D_c, N is the total number of documents in the collection and n_{q_c} is the number of documents containing the query concept q_c.

These two retrieval models are used to implement a Bag-of-concepts model and evaluate the effectiveness of concept-based representations for medical IR. Before presenting an empirical evaluation of our Bag-of-concepts model, it is important to understand how a concept-based representation differs from a term-based representation and what effect these differences will have on retrieval effectiveness.

4.2 Characteristics of a Concept-based Corpus

A concept-based representation differs both semantically and statistically; we review each separately.

4.2.1 Semantics of Terms and Concepts

Firstly we consider how a concept-based representation differs semantically at a term-level. It does this in three ways: (i) by encapsulating individual terms in a single concept; (ii) by conflating term-variants to a single concept; and (iii) by expanding terms to cover multiple concepts. Each of these three is detailed below.

Term Encapsulation

MetaMap analyses the sequence of input terms and identifies relevant concepts. A single identified concept might span a number of terms: for example, the input terms *metastatic breast cancer* would be spanned by the single concept `C0278488`. Thus, the term-based representation would have three lexical units (*metastatic*, *breast* and *cancer*) but the concept-based representation would contain only the single lexical unit, `C0278488`. Mapping to concept encapsulates the entity that is "metastatic breast cancer" into the single concept `C0278488`, rather than separating it as three terms. This encapsulation of individual terms

into a single concept makes the concept explicit and distinct: there is a specific, individual entry in the index for the concept `C0278488` (*metastatic breast cancer*), rather than three separate entries in the index for each term. Term encapsulation also makes distinct concepts that share common terms; for example, the concepts *heart disease* and *liver disease*, which would otherwise be 50% similar in a term-based representation, are instead distinct in a concept-based representation. Encapsulating terms within a single concept fundamentally changes the corpus statistics of a concept-based representation; this is further explored later in Section 4.2.2.

Conflating Term-variants

Mapping to concepts encapsulates terms within a single concept but a number of different terms can map to the same concept. For example, consider the SNOMED CT concept `86406008`, which has the description *Human immunodeficiency virus infection*. This disorder can be expressed in a number of subtly different ways: as the *T-lymphotropic virus*, as the abbreviations *HIV* or *AIDS*, or as the phrase *human immunodeficiency virus*. All these variations essentially represent the same concept, `86406008` (*Human immunodeficiency virus infection*). By mapping to concepts, all these term variants for HIV map to the same concept; it does not matter which variant has been used as they all conflate to the same concept. The consequence of this in a retrieval scenario is that it does not matter how HIV has been expressed in the query or a document; each term variant conflates to a single concept and retrieval is performed by matching the single concept representing HIV.

Vocabulary mismatch was identified as the first semantic gap problem in Chapter 2 (Section 2.1) and was described as the situation where particular entities may be expressed in a number of different ways yet have a similar underlying meaning. The conflation of different term variants to a single concept specifically addresses the vocabulary mismatch problem and is, therefore, a significant benefit of a concept-based representation. Of course, this depends upon the quality of the conflation.

Concept Expansion

The previous section showed how multiple term-variants can be mapped to a single overarching concept. However, the opposite case may apply — a single term (or term phrase) may map to a number of more specific concepts. In this situation, the mapping process will produce a number of relevant concepts for a single term phrase. Consider the example of mapping the terms *esophageal reflux*

to concepts.[5] MetaMap maps *esophageal reflux* to four different SNOMED CT concepts:

- `235595009`, *Gastroesophageal reflux disease* (disorder);
- `196600005`, *Oesophagitis* (disorder);
- `47268002`, *Reflux* (finding); and
- `249496004`, *Esophageal reflux finding* (finding).

Each of these concepts covers slightly different aspects of *esophageal reflux*: the first two cover the actual disorder of *esophageal reflux* while the latter two findings indicate a positive presence of *esophageal reflux* (for example, in a laboratory test). Mapping from terms produces concepts that explicitly cover these four different aspects of *esophageal reflux* — the terms are expanded to cover a number of different concepts. This expansion mechanism has a similar effect to the query expansion process used in information retrieval that enhances the representation with other highly related terms. (In our case, a number of highly related concepts are derived from the terms.) In our Bag-of-concepts model, both queries and documents are mapped to concepts and, as a result, this concept expansion approach is applied to both. The effect on retrieval is that the model is less dependent on the particular terms used in the query or documents. For example, a query of *esophageal reflux* would map to the above four query concepts. A document that contained the term *Oesophagitis* would be retrieved as it would map to the concept `196600005` (*Oesophagitis*). In contrast, a term-based system would most likely never retrieve a document containing *oesophagitis*, as it had no overlap with the query terms *esophageal reflux* (unless a query expansion process can successfully add *oesophagitis* to the original query).

Concept expansion aids in overcoming vocabulary mismatch by making the model less dependent on the terms used in document and queries. The expanded concepts are often more specialised instantiations of the source terms (for example, the *Esophageal reflux finding* is a specific aspect of the terms *esophageal reflux*). Concept expansion therefore incorporates the specific fine-grained aspects of a higher level term description. As a result, the concept-based representation alleviates some granularity mismatch issues, identified as one of the semantic gap issues from Chapter 2 (Section 2.2).

[5] Esophageal reflux is a chronic symptom of mucosal damage caused by stomach acid coming up from the stomach into the esophagus.

4.2.2 Statistics of Terms and Concepts

This section details how the corpus statistics differ between concept and term representations and how this impacts retrieval performance. Later in the chapter, we present our empirical evaluation, showing how these differences lead to superior retrieval effectiveness.

The specific corpus that we analyse is made up of electronic patient records and is the document collection used in the TREC Medical Record Track. Each document was converted to concepts using the method described earlier in the chapter. The result of this process was three corpora, comprising terms, UMLS and SNOMED CT concepts respectively. Similarly, three sets of queries were also produced. Basic statistics of the three representations are shown in Table 4.1.

Considering first the document statistics shown in Figure 4.1(a), we observe that the average document length of concept-based documents is considerably

Representation	**#Docs**	**Average document length**	**#Vocabulary**
Terms	17,198*	2,338	218,574
UMLS	17,198*	5,417	61,302
SNOMED CT	17,198*	3,906	36,467

*100,866 original reports collapsed to patient *visit* documents.

(a) Documents statistics

Representation	**#Queries**	**Average query length**	**#Vocabulary**
Terms	82	9.01	340
UMLS	82	4.50	209
SNOMED CT	82	5.67	259

(b) Query statistics

Table 4.1: Collection statistics for three different representations (Terms, UMLS and SNOMED CT concepts) of the TREC MedTrack corpus of clinical patient records. Table 4.1(a) shows documents statistics, Table 4.1(b) shown query statistics.

larger than that of term-based documents (for both UMLS and SNOMED CT documents). This is a result of the concept expansion process, where a single term maps to multiple different concepts. UMLS documents are on average longer than SNOMED CT documents because UMLS is a much larger ontology, covering many more concepts, and is, therefore, more likely to have more concepts identified as part of the expansion process. The vocabulary size represents the number of unique terms or concepts for each representation. In this case, the term vocabulary is significantly larger than the concept vocabulary. This is because of the large number of non-medical terms that appear in the term index but are not mapped to concepts in the UMLS or SNOMED CT indices. Additionally, the term encapsulation process that converts a multi-term phrase to a single concept reduces the number of unique concepts in the UMLS and SNOMED CT indices. The UMLS ontology is larger than SNOMED CT and covers a wider variety of topics and therefore has a larger concept vocabulary than SNOMED CT.

The statistics for different query representations are shown in Table 4.1(b). On average, term-based queries are significantly longer than concept-based queries, the same trend applying for the query vocabulary size. This is because of non-medical terms — for example, "with", "for", "which", etc. — that appear in the term query but are not mapped to concepts. SNOMED CT concepts are mapped from UMLS concepts and a single UMLS concept might map to more than one SNOMED CT concept, which is why SNOMED CT queries are slightly longer (in both average query length and vocabulary size).

We have provided statistics on the average document length and vocabulary size for the three different representations, showing how they differ. Documents are longer but queries are shorter and concept vocabularies are much smaller. A core component of IR models is term-frequency. In the light of the preceding, we analyse the profile of concept frequencies in order to assess its potential impact on retrieval performance. With respect to term frequency, researchers have studied the frequency of words in natural language and shown that it obeys Zipf's law; that is, the frequency of words in a large corpus of natural language is inversely proportional to the order of their frequency of occurrence [Ha et al., 2002]. Zipf's law also applies to frequency of terms found in documents indexed by an IR system [van Rijsbergen, 1979, p. 15–16]. Furthermore, a study specifically looking at a large collection of clinical notes found that the term frequency distribution was "near-Zipfian" [Wu et al., 2012]. But does this apply to a concept-based representation? To answer this question, Figure 4.4 plots the frequency of occurrences for terms, UMLS concepts and SNOMED CT concepts found within the TREC MedTrack corpus; the y-axis shows frequency of occurrence (at log scale) for each term or concept on the x-axis and is trun-

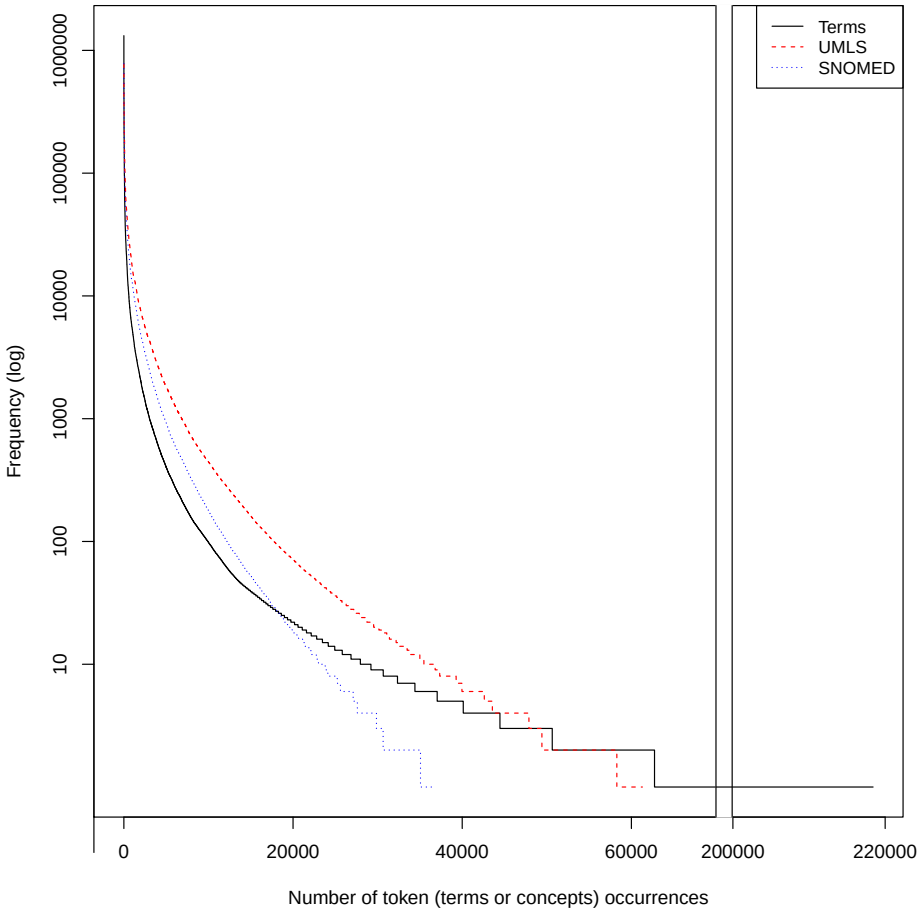


Figure 4.4: Frequency of occurrence (at log scale) for terms and concepts in the TREC MedTrack corpus; x-axis is truncated between 70,000 and 200,000 for space constraints. The term-based index follows Zipf's law: it has a small number of terms with very high frequency and a 'long tail'. Concept-based document collections do not obey Zipf's law.

cated between 70,000 and 200,000 for space constraints. The term-based index follows Zipf's law: it has a small number of terms with very highly frequency and a 'long tail' (a large number of terms that appear with low frequency).[6] In contrast, the concept-based indices do not exhibit the long tail; instead, they have only a few infrequently occurring concepts. Thus, concept-based document collections do not obey Zipf's law. Frequency statistics are important for understanding the information in text corpora [Luhn, 1958]. Therefore, a measure of word frequency is important for the purposes of information retrieval.

[6] A linear regression model using log of frequency and log of number of tokens revealed a goodness of fit of 0.9 in R-squared score; thus, making the distribution near "near-Zipfian" and in-line with Wu et al. [2012].

If the frequency of concepts in a collection differs from terms, then how does this affect retrieval using concepts? One can hypothesise that standard retrieval models, developed to use term-based frequency statistics, may not be optimal when using concept statistics, or at least standard parameter settings for these models might not apply to concepts; these questions are answered as part of our empirical evaluation.

In summary, this section has shown how a concept-based representation differs both semantically and statistically from a term-based one. Semantically, three important mechanisms are performed when mapping term to concept: term encapsulation, term-variant conflation and concept expansion. We argue that utilising these mechanisms to produce a concept-based representation tackles some of the semantic gap problems presented in Chapter 2, specifically vocabulary and granularity mismatch. We also show how the overall statistics of a concept-based representation differs from terms. Average document length and vocabulary size differ and in addition the distribution of concepts across a collection does not obey Zipf's law. To understand how all these characteristics affect retrieval effectiveness, we now present an empirical evaluation of our Bag-of-concepts retrieval model using the TREC Medical Records Track.

4.3 Empirical Evaluation

This section presents an empirical evaluation of our Bag-of-concepts model for medical information retrieval.

4.3.1 TREC Medical Records Track Test Collection

Our evaluation uses the TREC Medical Records Track (MedTrack) [Voorhees and Tong, 2011; Voorhees and Hersh, 2012]. As this collection is also used in the evaluations of subsequent chapters, we provide some details regarding the collection.

Documents — Patient Reports & Visits

The collection contains one month of de-identified reports from multiple U.S. hospitals. There are nine types of reports: "Radiology Reports", "History and Physicals", "Consultation Reports", "Emergency Department Reports", "Progress Notes", "Discharge Summaries", "Operative Reports", "Surgical Pathology Reports" and "Cardiology Reports". In total, the collection contains 100,866 reports. A report is part of a "visit": an individual patient's single

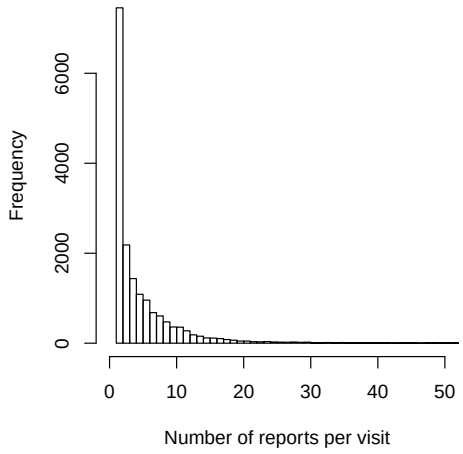


Figure 4.5: Distribution of reports per visits in the TREC Medical Records Track test collection, truncated at 50 reports per visit. Most visits contain a small number of reports (median 3 reports per visit).

admission at a hospital. Links between the same person's multiple admissions are intentionally removed for privacy as part of the de-identification process. Mapping reports to visits results in 17,198 unique visits. A single visit can represent a lengthy hospital admission and may contain many different individual reports or the admission may be short and minor with the visit comprising only a single report. Figure 4.5 shows the distributions of reports per visit. Most visits contain a small number of reports (median 3 reports per visit). The figure is truncated at 50 reports per visit with the maximum visit containing 415 reports per visit.

Queries and Relevance Judgements

Query topics represent an information need to identify cohorts of patients for clinical trails. Clinical trials are research studies involving a cohort of patients to evaluate new drugs, procedures and treatments. Researchers conducting clinical trials specify an "inclusion criteria" describing the patients required for the study. The criteria might include attributes such as diseases, treatments, age group, gender and ethnicity [Voorhees and Hersh, 2012]. A list of priority areas for conducting clinical trials is published by the U.S. Institute of Medicine [Committee on Comparative Effectiveness Research Prioritization, 2009]. These priority areas were provided to assessors to develop corresponding query topics. The assessors were physicians and students in the Oregon Health & Science University Biomedical Informatics Graduate Program and physician researchers

QueryId	Query keywords
136	Children with dental caries
102	Patients who have had a carotid endarterectomy
167	Patients with AIDS who develop pancytopenia
169	Elderly patients with subdural hematoma

Table 4.2: Example query topics from the TREC Medical Records Track test collection.

from the US National Library of Medicine. The assessors used the clinical trial priority areas to develop short inclusion criteria descriptions and these descriptions became the query topic keywords. The task for TREC MedTrack is one of ad-hoc retrieval of the free-text patient records. Several example query topics are provided in Table 4.2.

For relevance assessment, the assessors were provided with all the reports pertaining to a single patient visit and were asked to evaluate the relevance of the patient to the query topic. Thus, the unit of retrieval was a patient visit rather than an individual report document. Mapping reports to visits was left to the discretion of the teams participating in the track. TREC Medtrack 2011 contained 34 topics and 2012 contained 47 topics.[7]

Evaluation Measures

In 2011 the official evaluation measure was bpref, supplementary measures were MAP and precision @ 10. However, in 2012 the organisers used inferred measures: infNDCG as the primary measure and infAP and precision @ 10 as supplementary measures. Inferred measures required specific relevance assessments (qrels), which were not available for 2011, but bpref and precision @ 10 from 2011 could be used with 2012. It is possible to separate the evaluation into two parts, each using the evaluation measures specific to that year (34 queries for 2011 and 47 for 2012). However, it is more desirable to have a single, larger query set for statistic significance. In addition, separating the queries makes the presentation of results and discussion more cumbersome. Therefore, we combine the query sets and select bpref as the primary measure and precision @ 10 as the secondary measure. Bpref was specifically chosen because it considers only judged documents and as in MedTrack this number is small, measures that do not assume complete judgments are likely to be more reliable indicators of retrieval effectiveness. These evaluation measures were previously detailed in

[7]TREC organisers excluded topic 130 from 2011 and topics 138, 159 and 166 from 2012 due to lack of relevant visits in the corpus.

Chapter 3, Section 3.3.2.

4.3.2 Experimental Settings

This section details how the TREC Medical Records Track was utilised to evaluate our Bag-of-concepts method. As the unit of retrieval was a patient visit not an individual report document, we chose to concatenate all reports belonging to a single visit into a single visit document. This was done prior to indexing and resulted in 17,198 visit documents that were then subsequently indexed.

For the retrieval using terms, stemming was applied using the Porter stemmer and no stoplist was used.

Parameter Settings

The Bag-of-concepts model has two variants: a probabilistic language model and a tf-idf model, both models having free parameters. We have already highlighted how concept-based and term-based representations differ and have hypothesised that standard parameter settings may not apply to concept-based representations. To evaluate this hypothesis, a full sweep of the parameter space was performed separately for each of the three representations: terms, UMLS and SNOMED CT. The parameter values that maximised bpref were selected for each representation. For the language model, there is a single parameter μ, that controls the influence of document length. For the tf-idf model, there are two parameters: b controls the influence of document length and $k1$ controls the influence of term frequency. For details of the parameter sweep, refer to Table 4.3.

In addition to the above parameter sweep, we also conducted a leave-one-out cross-validation experiment: queries were repeatedly divided into ten folds, with the parameters tuned on nine folds and tested on one fold. The bpref and precision @ 10 scores were averaged across each test to give the overall

Model	Parameter	Default	Range	Increment
LM	μ	2,500	0–30,000	+1000
tf-idf	$k1$	1.2	0–40	+1
	b	0.75	0–1	+0.05

Table 4.3: Parameter selection for two model variants: language model and tf-idf. Also included is the default value for each parameter as reported in the literature.

performance. The results for cross-validation showed no significant difference to those achieved using the parameter sweep in Table 4.3. Therefore, the simplier evaluation method of the parameter sweep was favoured.

4.3.3 Results

The retrieval results for the Bag-of-concepts model and comparitive term baseline are presented in Table 4.4. Both model variants (tf-idf and language model) are presented for each representation (terms, UMLS and SNOMED); the percentages show improvements over the term baseline. The results indicate that both the concept-based approaches outperform the term baseline. UMLS demonstrates the greatest improvements of +15% in bpref and +14% in precision @ 10, while SNOMED CT shows improvements of +13% in bpref and +10% in precision @ 10. Overall, greater improvements are observed in bpref than precision @ 10. (This issue is further explored later in the discussion section.) The tf-idf model variant always exhibits superior performance over the language model.

To understand where each model was performing well, the retrieval effectiveness of individual queries is required. The plots in Figure 4.6 provide this by showing the bpref performance (y-axis) of each of the 81 queries (x-axis); queries are ordered by decreasing performance of the term baseline system. The results show that most of the gains in performance exhibited by the concept-based systems are for 'hard' queries: those that perform poorly using the term-based system. Conversely, the major losses in performance for the concept-based system are actually found in 'easy' queries: those where the term-based system exhibits good performance.

Representation	**Bpref**		**Prec@10**	
	tf-idf	LM	tf-idf	LM
Terms	0.3934	0.3917	0.4753	0.4975
UMLS	0.4513† (+15%)	0.4340† (+11%)	0.5395† (+14%)	0.5358† (+8%)
SNOMED	0.4433† (+13%)	0.4223 (+8%)	0.5235† (+10%)	0.5111 (+3%)

Table 4.4: Bag-of-concept retrieval results on TREC MedTrack using tf-idf and Language Model with Dirichlet (LM) smoothing. Percentage improvements over term baseline. † indicates statistical significance (paired t-test $p < 0.05$).

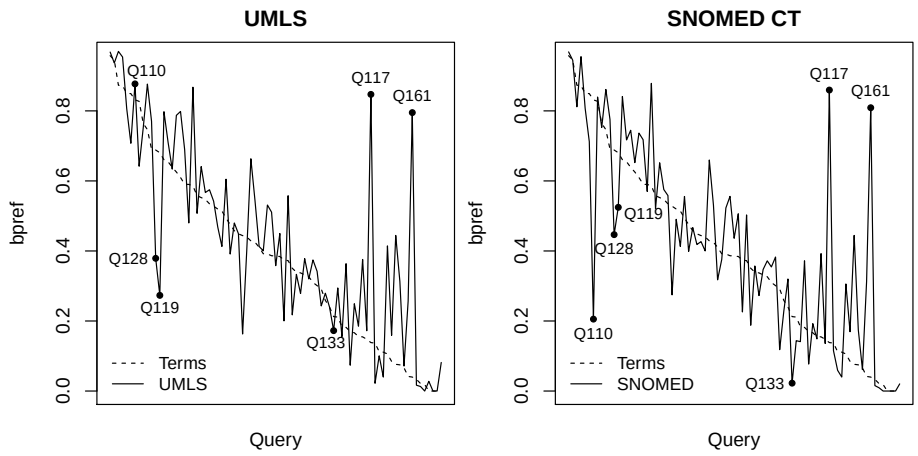


Figure 4.6: Per-query performance of UMLS and SNOMED CT concept-based systems compared to the term baseline; queries are ordered by decreasing performance of the term baseline system. Some specific queries are highlighted for further analysis in the discussion.

Parameter Sensitivity

The results from a sweep of the parameter space are provided in Table 4.5. The optimal parameter setting for bpref is shown for each representation. Also included in the table are the default settings published in the literature for each parameter [Zhai, 2007]. For all three representations, the optimal settings are significantly different from the default published for that model. However, the optimal parameter setting does not differ vastly between the term and concept-based representations.

We now examine the effect of different parameter settings on performance. Figure 4.7 shows the effect of language model's μ on bpref and precision @ 10. The greater the value of μ, the less the influence of document length, or specifically, shorter documents are less discriminating (Equation 4.1). The red vertical

	Parameter			Optimal Setting (bpref)		
Model		Influence	Default	Terms	UMLS	SNOMED
LM	μ	Doc. length effect	2,500	13,000	14,000	22,000
tf-idf	$k1$	Term freq. effect	1.2	2.9	2.1	1.5
	b	Doc. length effect	0.75	0.4	0.6	0.45

Table 4.5: Parameter selection for two model variants: language model and tf-idf.

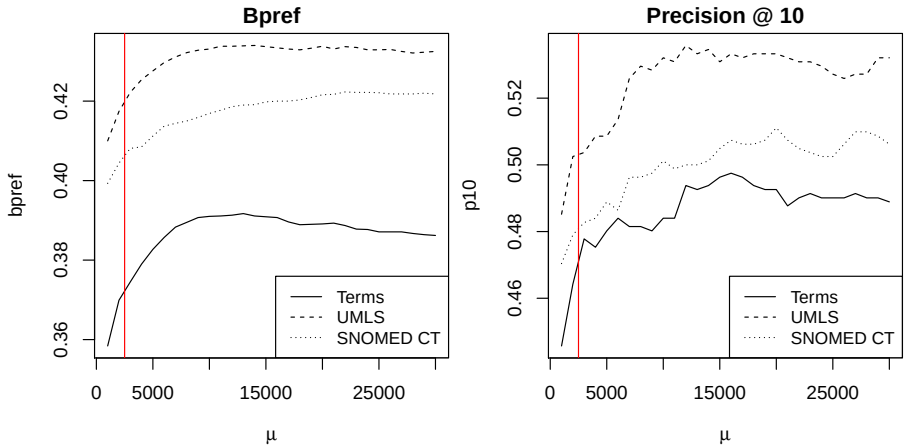


Figure 4.7: Parameter sensitivity of μ using a language model for the Bag-of-concepts model and term baseline. The greater the value of μ, the less the influence of document length. The red vertical line shows the default parameter setting reported in the literature.

line shows the default value reported in the literature. All three representation exhibit a similar trend: optimal performance is found with high values of μ and the performance stabilises for larger values. Optimal performance is achieved for values of μ much greater than the default reported in the literature (red vertical line). This means document length is not as strong an indicator of relevance for this test collection.

Figure 4.8 shows the tf-idf model's parameter sensitivity to $k1$, where the higher the value of $k1$, the greater the influence of term frequency (Equation 4.2). All three representations exhibit a similar trend: a peak is seen near the default value and thereafter a steady decline is observed. The best performance is obtained for values of $k1$ greater than the default value (red vertical line). These results show that term frequency is an important indicator of relevance for this collection.

Figure 4.9 shows the tf-idf model's parameter sensitivity to b, where the higher the value, the greater the influence of shorter documents (Equation 4.2). All three representations have a similar trend: optimal settings of b are below the default value. This indicates that shorter documents are not as strong an indicator of relevance for this test collection.

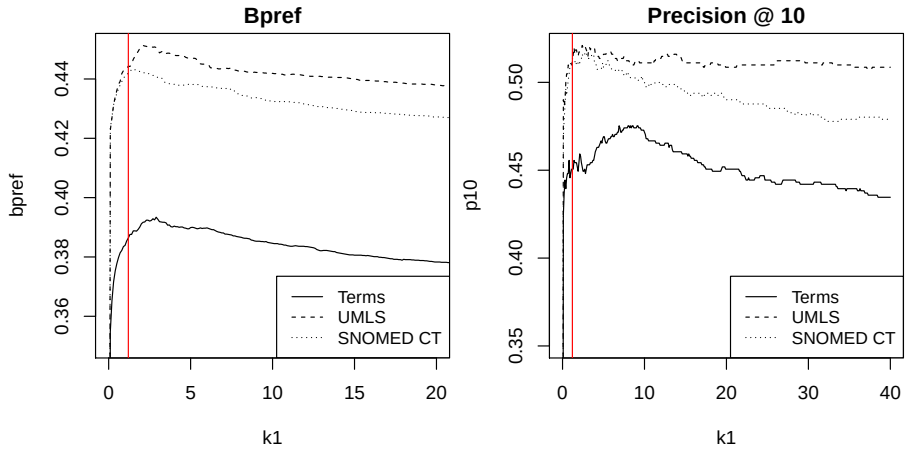


Figure 4.8: Parameter sensitivity of $k1$ using Lemur's tf-idf model for the Bag-of-concepts model and term baseline. The higher the value of $k1$, the greater the influence of term frequency. The red vertical line shows the default parameter setting reported in the literature. The value for b was fixed according to the best values reported in Table 4.5.

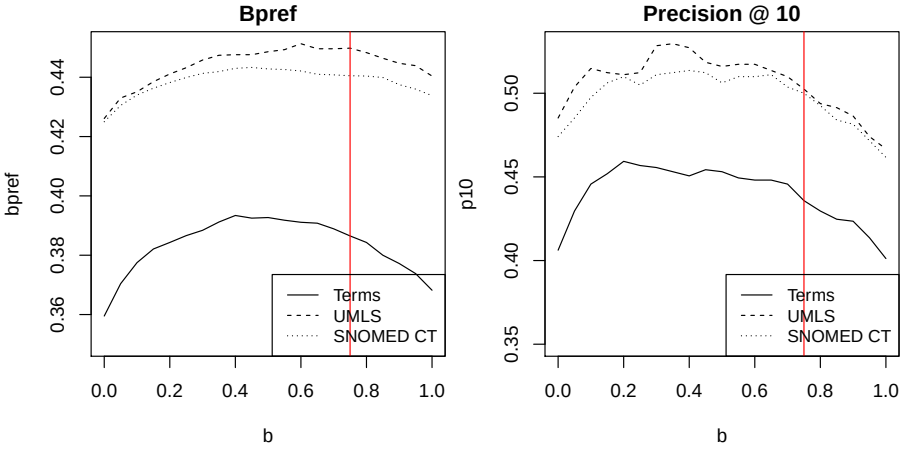


Figure 4.9: Parameter sensitivity of b using Lemur's tf-idf model for the Bag-of-concepts model and term baseline. The greater the value of b, the greater the influence of shorter documents. The red vertical line shows the default parameter setting reported in the literature. The value for $k1$ was fixed according to the best values reported in Table 4.5.

4.4 Analysis and Discussion

4.4.1 IR Models and Parameter Settings using Concepts

This chapter considers in detail how concept-based representations differed from term-based representations. Based on these differences, we conjectured that

standard IR models, or at least the default parameter settings for these models, might not directly translate to concept-based representations. We now revisit this conjecture based on the retrieval results and parameter sensitivity analysis, firstly considering the choice of retrieval model: tf-idf or language model. The concept-based representation actually exhibited a similar trend in results to the term-based representation. The best performance using a concept-based representation was always achieved with the tf-idf model (Table 4.4). For terms, by contrast, a language model was superior for precision @ 10. (Both tf-idf and LM have comparable performance in bpref.) Overall, however, the choice of retrieval model did not result in substantial differences between terms and concepts. We conclude that standard IR models — in this case tf-idf and a language model — directly translate to using a concept-based representation. Regarding the applicability of the parameter settings for terms and concepts, there was little to separate the three representations: all three followed a similar trend with respect to different parameter values of the language model's μ (Figure 4.7) and tf-idf models' $k1$ (Figure 4.8) and b (Figure 4.9). From these results we conclude that the choice of representation does not drastically affect the parameter settings.

Although the parameter settings did not differ between representations, they did differ from the default parameter values reported in the literature [Zhai, 2001, 2007]. This result highlights the specific nature of electronic patient records. Specifically, the influence of document length and term frequency. Regarding document length, shorter document length was not an influential indicator of relevance (explained by higher than default μ for the language model and lower than default b for tf-idf). This result can be explained by the fact that documents were actually a concatenation of individual reports, so their length was often determined by the number of reports in the visit. The number of reports does not, in itself, indicate that the visit was more or less relevant. In this case, a more appropriate relevance estimation would have taken into account the type of report containing the evidence. This was identified as one of the issues contributing to the semantic gap (Levels of Evidence, Section 2.5.4). Researchers have developed specific medical IR models that handle this situation. Zhu and Carterette [2013] developed a system that indexed individual reports and visits separately and a retrieval model that utilised scores from both. Another approach developed by Limsopatham et al. [2013a] grouped individual reports into departments (cardiology, radiology, emergency department); a voting model was then used that estimates the expertise of the department based on the relevance scores of its corresponding reports.

The influence of term frequency was the other characteristic that differed for electronic patient records. In this case, term frequency was a strong indicator of relevance (explained by higher than default values of $k1$ in tf-idf). A patient

was often seen by a variety of different departments and specialists and as a result their reports contained a diverse mix of content. However, much of this content was not core to the patient's main diagnoses or treatments. The content may contain past medical history, suspected diagnosis or even explicitly negated content. This type of content was covered briefly, so terms are often mentioned with low frequency. From a corpus statistic perspective, these terms appeared with low frequency in a large number of documents. However, for important aspects of the patients care, much more detailed descriptions were produced. In this case, these important terms appeared with much higher frequency and indicated the important characteristics of the patient. From a corpus statistic perspective, these terms appeared with high frequency in only a small number of documents and clearly identified these documents as potentially relevant. Thus, a retrieval model more sensitive to term frequency was able to discriminate between general characteristics or those core to the patient.

4.4.2 Gains in Hard Queries

In general, it was the hard queries (those that performed poorly on terms) that benefited the most from concept-based approaches. This was highlighted in Figure 4.6, which showed individual query performance for terms and concepts. We hypothesise that it was these queries for which the performance was most affected by the semantic gap and that the Bag-of-concepts method was effective at alleviating these issues. To understand this further, we review the specific queries 117 and 161, which were highlighted in Figure 4.6.

Query 117 contained the keywords `Patients with Post traumatic Stress Disorder` and mapped to three SNOMED CT concepts: *Patient* (`116154003`), *Posttraumatic stress disorder* (`47505003`) and *Combat fatigue* (`61157009`). This query was a typical example of the vocabulary mismatch problem and one that can be overcome using the Bag-of-concepts model. "Post traumatic Stress Disorder" can be written as one word or two: "post traumatic" or "posttraumatic", or hyphenated as "post-traumatic". In the query, it was two separate words, but a manual inspection of relevant documents revealed that it was typically expressed it as the single word, "posttraumatic". Although these documents also contained the terms "stress" and "disorder", these are very general, high frequency terms and thus were not discriminators for relevant documents. In addition, "Post traumatic Stress Disorder" is often abbreviated to "PTSD". A number of relevant documents contained only PTSD. Mapping the query to concepts also produced another concept: *Combat fatigue* (`61157009`). Posttraumatic stress disorder specific to military service is sometimes expressed as combat fatigue, especially in military care facilities. A number of documents came

from patients who were war veterans and had notes from military care facilities. In these cases, the terms "combat fatigue" were used.

Recall that mapping to concepts involved three important characteristics: term-encapsulation, conflating term-variants and concept expansion; these were detailed earlier in Section 4.2.1. When mapping to concepts all the variants — post traumatic, posttraumatic, post-traumatic and PTSD — all mapped to the single concept *Posttraumatic stress disorder* (`47505003`) and were, therefore, retrieved using the Bag-of-concepts model. This was an example of the conflating term-variants mechanism at work. Furthermore, mapping the query to concepts included the concept for Combat Fatigue. This was an example of the concept-expansion mechanism at work. Both these processes were able to overcome the vocabulary mismatch problem for this query and as a result bpref improved from 0.1012 to 0.8450 over the term baseline and precision @ 10 improved from 0.2000 to 0.8000.

The next query we review is Query 161, `Patients with adult respiratory distress syndrome`. This query was an example of the term encapsulation mechanism at work. The query was mapped to three concepts: *Patient* (`116154003`), *Adult respiratory distress syndrome* (`67782005`) and *Non-cardiogenic pulmonary edema* (`95437004`). For this query, term dependence was essential. Many documents contained the general terms "adult", "respiratory", "distress" and "syndrome", but "Adult respiratory distress syndrome" denotes a specific disease in itself. Here a term-based dependence model could be applied (for example, an n-gram language model or Markov Random Field dependence model [Metzler and Croft, 2005]), but mapping to concepts already achieved this by the term encapsulation process that mapped the query to the single concept *Adult respiratory distress syndrome* (`67782005`). Also, adult respiratory distress syndrome is also often abbreviated to ARDS. The term-encapsulation process ensured that where the abbreviation ARDS was used, it mapped to the same concept, *Adult respiratory distress syndrome* (`67782005`).

In addition, query 161 suffered from granularity mismatch, one that cannot be resolved by handling term dependence. Adult respiratory distress syndrome was often expressed as the more specific disorder "non-cardiogenic pulmonary edema". Many documents contained the latter, more specific, description rather than the general description found in the query. In the concept-based representation of the query, the concept *Non-cardiogenic pulmonary edema* was included by the concept expansion mechanism. For this query, bpref improved from 0.04 to 0.8438 and precision @ 10 from 0.1 to 1.0.

4.4.3 Degradation in Easy Queries

In contrast to hard queries, concept-based IR did not improve the performance of easy queries (those that already performed well using terms). These queries were often clear and explicit, for example query 105, `Patients with dementia`. There was no ambiguity in the use of the term "dementia" and most documents containing the term were relevant. Similarly, for query 112, `Female patients with breast cancer with mastectomies during admission` where a document containing the key term *mastectomies* was typically relevant. (The other terms did not generally discriminate relevant from irrelevant documents because *mastectomy* implies breast cancer and the vast majority of mastectomies are performed on female patients.) Obviously, the semantic gap did not plague such queries. Performance gains in hard queries, but not easy queries, was a common trend uncovered in the various empirical evaluations performed in this thesis. This may be a characteristic of semantic search systems in general; further remarks on this important and usually unrecognised issue are provided in the discussion (Chapter 8).

Some queries had significantly lower performance using the Bag-of-concepts model. Queries 119, `Adult patients who presented to the emergency room with anion gap acidosis secondary to insulin dependent diabetes`, contained errors in the mapping from terms to UMLS concepts. The query was mapped to appropriate concepts but a number of documents that contained "noninsulin dependent diabetes" — and therefore were not relevant to the query — incorrectly mapped to the concept *Diabetes mellitus type 1*, which is an insulin-dependent diabetes. A large number of these irrelevant documents were retrieved by the concept-based system, thus reducing the performance on this query.

Query 128 contained the keywords `Patients admitted for hip or knee surgery who were treated with anti-coagulant medications post op`. This query was correctly mapped to concepts but still had significantly lower performance compared to the term baseline. Both term and concept models returned a similar number of relevant documents but the term baseline retrieved these documents higher in the ranked list. The concept-based model ranked highly a number of irrelevant documents pertaining to patients who had knee surgery but were not treated with anti-coagulants. This can be explained by the corpus statistics for the two query concepts *Knee joint operation* (179342005) and *Anticoagulant* (81839001). *Knee joint operation* appeared in 327 documents whereas *Anticoagulant* appeared in 1274. Thus, documents containing the rarer concept *Knee joint operation* were favoured. In contrast, the term-based query split "knee surgery" into two separate terms and both

occured frequently in the collection — more so than "anticoagulant". As a result, relevant documents that contained the rarer term "anticoagulant" were retrieved much higher in the ranking.

Another issue specifically affecting the performance of the SNOMED CT concept-based model was the mapping from UMLS to SNOMED CT concepts. UMLS to SNOMED CT mappings are provided as part of the UMLS Metathesaurus. For query 110, `Patients being discharged from the hospital on hemodialysis`, there was no mapping between the UMLS concept hemodialysis and a SNOMED CT equivalent (although the hemodialysis concept does exist in SNOMED CT). Similarly, for query 133, `Patients admitted for care who take herbal products for osteoarthritis`, there was no UMLS to SNOMED CT mapping for the UMLS concept for "herbal". As a result, both queries had poor performance on the SNOMED CT concept model. The problem of mapping between UMLS and SNOMED CT may explain why the overall SNOMED CT performance is slightly lower than UMLS. With SNOMED CT becoming the mandated standard for medical terminology, researchers are actively working on tools that directly translate free-text to SNOMED CT concepts [Suominen et al., 2013]. Such tools bypass the need to map from UMLS and avoid the problems that this can cause.

There are some limitations attached to the choice of MetaMap as a concept extraction system for clinical records, such as those used in this thesis. MetaMap was originally developed for processing biomedical literature, not clinical notes, and evaluations on the effectiveness of MetaMap have largely been done using only biomedical literature [Pratt and Yetisgen-Yildiz, 2003]. The extension of MetaMap into the clinical domain is a relatively recent phenomenon. Improvements in concept extraction in this new domain are likely to have a beneficial effect on the overall performance of the methods presented in this thesis.

Overall the Bag-of-concepts model provided improvements in retrieval effectiveness over a term baseline; greater improvements were observed in bpref than in precision @ 10. This may be explained by greater improvements in recall using concepts. Addressing the semantic gap issue of vocabulary mismatch (and to some extent granularity mismatch) mainly involved improving recall by retrieving documents not retrieved by the term-based models. Another factor explaining the difference between bpref and precision @ 10 was the effect of unjudged documents: documents never assessed for relevance by the TREC assessors. The precision @ 10 measure assumes that an unjudged document was irrelevant whereas bpref ignored these documents.[8] Unjudged documents can significantly affect the performance evaluation, especially for semantic search

[8]Details of these two evaluation measures were provided in Chapter 3, Section 3.3.2; precision @ 10 is defined in Equation 3.14 and bpref is defined in Equation 3.18.

systems; this issue is explored in detail in Chapter 7.

4.5 Summary

The empirical evaluation in this chapter has highlighted that a Bag-of-concepts model, utilising concepts defined in medical ontologies, leads to superior retrieval effectiveness. Our hypothesis was that this effectiveness stemmed from a number of specific differences between term and concept-based representations and that it was these differences that were advantageous for retrieval. Statistically, a corpus of concepts differs from one of terms. Average document length and vocabulary size differ, but also the distribution of concepts across a collection does not obey Zipf's law. However, these differences do not mean that standard IR models and parameter settings cannot be translated to a concept-based representation. More significant was the nature of the text being searched: clinical patient records. For such texts, term (or concept) frequency was shown to be an important indicator of relevance but document length was not.

It is the semantic differences between terms and concepts that lead to improvements in retrieval effectiveness. In our study, three important mechanisms influenced these semantic differences. First, term encapsulation grouped individual terms into a single concept and differentiated the concept from the individual terms comprising it. Term encapsulation naturally modeled term dependence. Second, conflating term-variants was the mechanism by which multiple term-based variants — which essentially meant the same thing — mapped to a single concept. Conflating term-variants had an important role in alleviating vocabulary mismatch. Finally, the concept expansion mechanism produced a number of different concepts for a single term or term phrase. The expanded concepts may have been more specialised instantiations of the source terms that help to address granularity mismatch.

Bridging the semantic gap involves addressing two issues: *semantics* and *inference*. Although a Bag-of-concepts system increased average performance over a term-based IR system, it mainly only addressed vocabulary mismatch. Addressing the other semantic gap issues requires inference. To support inference, a greater understanding of the dependence between concepts is required. Some of this dependence information is provided in medical ontologies in the form of explicit relationships between concepts; other information can be derived from co-occurrence statistics. The next chapter extends the Bag-of-concepts model to capture some of the dependencies that exist between concepts. This is realised using both co-occurrence statistics and by leveraging more domain knowledge in the form of explicit concept relationships from the SNOMED CT ontology.

Chapter 5

Graph-based Concept Weighting Model

"There is an old saying," said Erdős. "Non numerantur, sed ponderantur."
(They are not counted but weighed).

— Paul Hoffman, The Man Who Loved Only Numbers:
The Story of Paul Erdős* and the Search for Mathematical Truth

This chapter extends the Bag-of-concepts model to account for the innate dependencies that exist between medical concepts. We propose a retrieval model that integrates the Bag-of-concepts model with previous work on graph-based term weighting. In addition, we propose a novel concept weighting method that incorporates the importance of the concept within the global medical domain (rather than just a single corpus). This weighting method is achieved by incorporating domain knowledge from the SNOMED CT ontology into the retrieval function. An empirical evaluation demonstrates the effectiveness of our graph-based concept weighting model over both term and concept baselines. The improvements in retrieval effectiveness by incorporating domain knowledge are promising and motivate a model that makes far more use of domain knowledge.

*Paul Erdős (1913 – 1996) was an Hungarian mathematician who made considerable contributions in graph theory and probability theory.

5.1 Motivation

One of the semantic gap issues introduced in Chapter 2 was Inference of Similarity (Section 2.4), which included the need to account for the innate *dependence* between medical concepts. This requirement is important when the query expresses multiple constraints that all have to be met within a document for it to be relevant. For example, in the query `Patients who present to the hospital with episodes of acute loss of vision secondary to glaucoma`, relevance depends on the vision loss caused by the glaucoma and not as a result of some other condition. Thus, the mere presence of *acute loss of vision* and *glaucoma* within a document does not necessarily indicate relevance; instead, the dependence between the two concepts needs to be determined.

Many IR models represent documents as bag-of-words; that is, the representation does not consider word order or dependence between terms. Some approaches go beyond bag-of-word representations and do account for term dependence. Most common within the language modelling framework is the Markov random field method of Metzler and Croft [2005]. However, graph-based retrieval models can also capture term dependence and are effective in empirical evaluations [Blanco and Lioma, 2012]. In addition, graph-based retrieval models have a number of characteristics attractive for semantic search: the propagated learning and search properties of a graph provide a powerful means of identifying important or relevant information items (be they terms, concepts or documents) [Turtle and Croft, 1991; Blanco and Lioma, 2012]. The popular PageRank algorithm [Page et al., 1999] is a prominent example of this class of algorithm and is one practical method to identify these important information items.

The previous chapter showed that the Bag-of-concepts model is effective when compared to term-based models. In addition, a graph-based model has a number of characteristics attractive for semantic search. Therefore, we provide a novel model that integrates both Bag-of-concepts and graph-based models. This new model also provides a means of incorporating more domain knowledge in the form of a measure of importance for a concept with the medical domain, which proves to be an effective indictor of relevance.

5.2 Graph-based Term Weighting

Blanco and Lioma [2012] developed a graph-based term weighting model that represents each document as a graph: vertices are terms and edges are relationships between terms. Relationships may be defined by simple co-occurrence of

terms within a context window or based on grammatical relationships between terms (for example, verb-noun or adverb-verb relationships). Using this method, a term is represented as node within the document graph and is connected to one or more other terms (nodes). The importance of a term within a document can then be estimated by the number of neighbouring terms and the importance of the neighbours. This measures importance in the same way PageRank estimates the importance of a page via the pages that link to it.[1]

We hypothesise that this graph-based term weighting model, adapted to a concept representation of documents, might be a powerful tool for medical IR as it would capture the dependencies between concepts found in medical free-text. The remainder of this section provides an explanation of the original graph-based model. In the next section we show how this model can be integrated with our Bag-of-concepts model.

In Blanco & Lioma's graph-based term weighting model, a term i in a document is represented by the vertex (or node) v_i. A vertex is connected to other vertices and $\mathcal{V}(v_i)$ denotes the set of vertices connected to v_i. The weight of v_i within a document is initially set to 1 and the following function is applied for several iterations:

$$S(v_i) = (1 - \phi) + \phi * \sum_{v_j \in \mathcal{V}(v_i)} \frac{S(v_j)}{|\mathcal{V}(v_j)|} \quad (0 \leq \phi \leq 1), \tag{5.1}$$

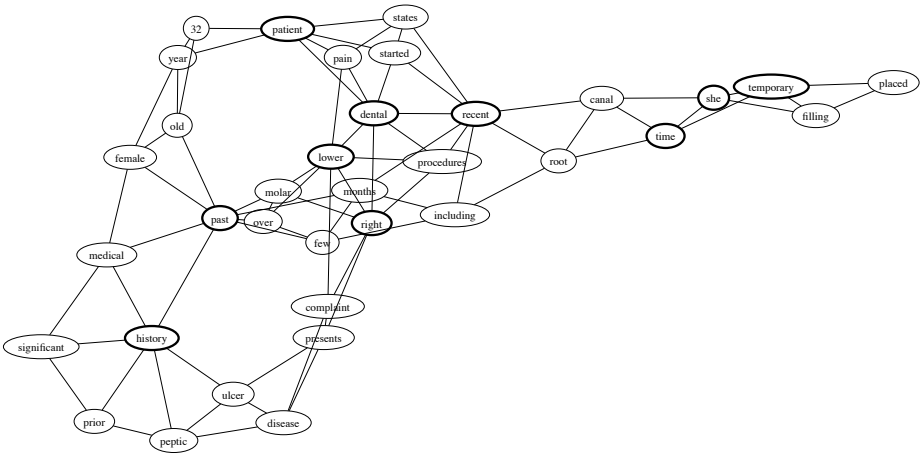
where ϕ is the damping factor that controls "vote recycling" from the original PageRank algorithm [Page et al., 1999]. Blanco and Lioma [2012] showed that only a small number of iterations (< 50) is required to obtain convergence.

Next, we present an example of the graph produced when the above method is applied to a small sample document of medical text; this is done to highlight some of the characteristics of a graph-based representation. A sample medical text document is shown in Figure 5.1(a) and the corresponding graph constructed from this document (using a context window of $N = 3$ terms) is shown in Figure 5.1(b). The vertex scoring algorithm of Equation 5.1 is applied to each vertex and the ten vertices with the highest score are bold highlighted; these include the terms `dental`, `patient` and a number of temporal terms (`history`, `past`, `time` and `recent`). The terms with higher scores provide an indication of the important terms appearing in this document. The next section shows how this information is included into a retrieval model.

[1] PageRank is a link analysis algorithm used by the Google web search engine to measure the relative importance of a webpage based on a hyperlinked set of documents.

```
"The patient is a 32-year-old female with a past medical history
significant for a prior history of peptic ulcer disease who
presents with a complaint of right lower dental pain.  The
patient states that she was started on recent dental procedures,
on a right lower molar, over the past few months, including a
recent root canal, at which time she had a temporary filling
placed."
```

(a) Sample medical text document.



(b) Term-based graph of the sample medical text document; stop words removed.

Figure 5.1: Resulting term graph built from the above medical document. Built using co-occurrence window $N = 3$. Bolded nodes indicate the 10 terms with greatest score within the document (according to Equation 5.1).

Retrieval Function

The graph-based vertex score of Equation 5.1 is now integrated into a retrieval function that estimates the relevance between a document and a query:

$$R(d, q) = \sum_{t \in q} w(t, q) * w(t, d), \tag{5.2}$$

where $w(t, q)$ is the weight of the term in query. This is often uniform for ad-hoc queries; thus $w(t, q) = 1$. The second component, $w(t, d)$, is the weight of the term in the document. The graph-based score provides a means of estimating $w(t, d)$:

$$w(t, d) = idf(t) * S(v_t), \tag{5.3}$$

where $S(v_t)$ is the vertex score from Equation 5.1 for term t and $idf(t)$ is the inverse document frequency of the term. The general retrieval function from

Equation 5.2 can be expressed as:

$$R(d,q) = \sum_{t \in q} idf(t) * S(v_t). \tag{5.4}$$

The *idf* component provides a measure of the importance of the term in the collection, while the PageRank score provides a measure of the importance of the term in the document.

In the next section we apply the graph-based term weighing method to use concepts from the Bag-of-concepts model.

5.3 Graph-based Concept Weighting

Building a graph of concepts is performed in the same way as building a graph of terms: a context window of fixed length is moved across a document and concepts that co-occur within the context window are connected via an edge in a graph of concepts. Although the process of creating the graph for terms and concepts is the same, the resulting graph itself can differ significantly for the concepts. To demonstrate this, we revisit the sample document and resulting graph from Figure 5.1. Converting the same document to concepts and constructing the graph results in the graph shown in Figure 5.2. The concepts are identified by their concept *id* in both the document and the graph but we also include their description in parentheses to make the example readable. The PageRank function from Equation 5.1 is applied and the 10 vertices with the highest scores are highlighted.

There are a number of differences between the term and concept graphs. First, the concept graph is much larger: there are many more concepts than terms. This is a result of the concept expansion mechanism, where a single term can map to multiple concepts. However, multiple terms also map to a single concept. For example, the phrase *Peptic ulcer disease* maps to the single concept `C0030920`. This is the term-encapsulation mechanism at work.

Both the term and concept graphs contain similar high score items: `dental` appears in both, as do `patient` and temporal items like `history`, `year`, `recent` and `time`. The one major difference, however, is the concept `Peptic Ulcer`, which appears in the concept graph but not in the term graph. The reason for this is twofold: firstly, when converting to concepts, the n-gram `peptic ulcer` from the original text maps to the single concept `c0030920` (a result of the term-encapsulation mechanism); secondly, when represented in graph form, the concept is highly connected and therefore receives a high score. The high score for `Peptic Ulcer` reveals it as an important concept within the concept graph

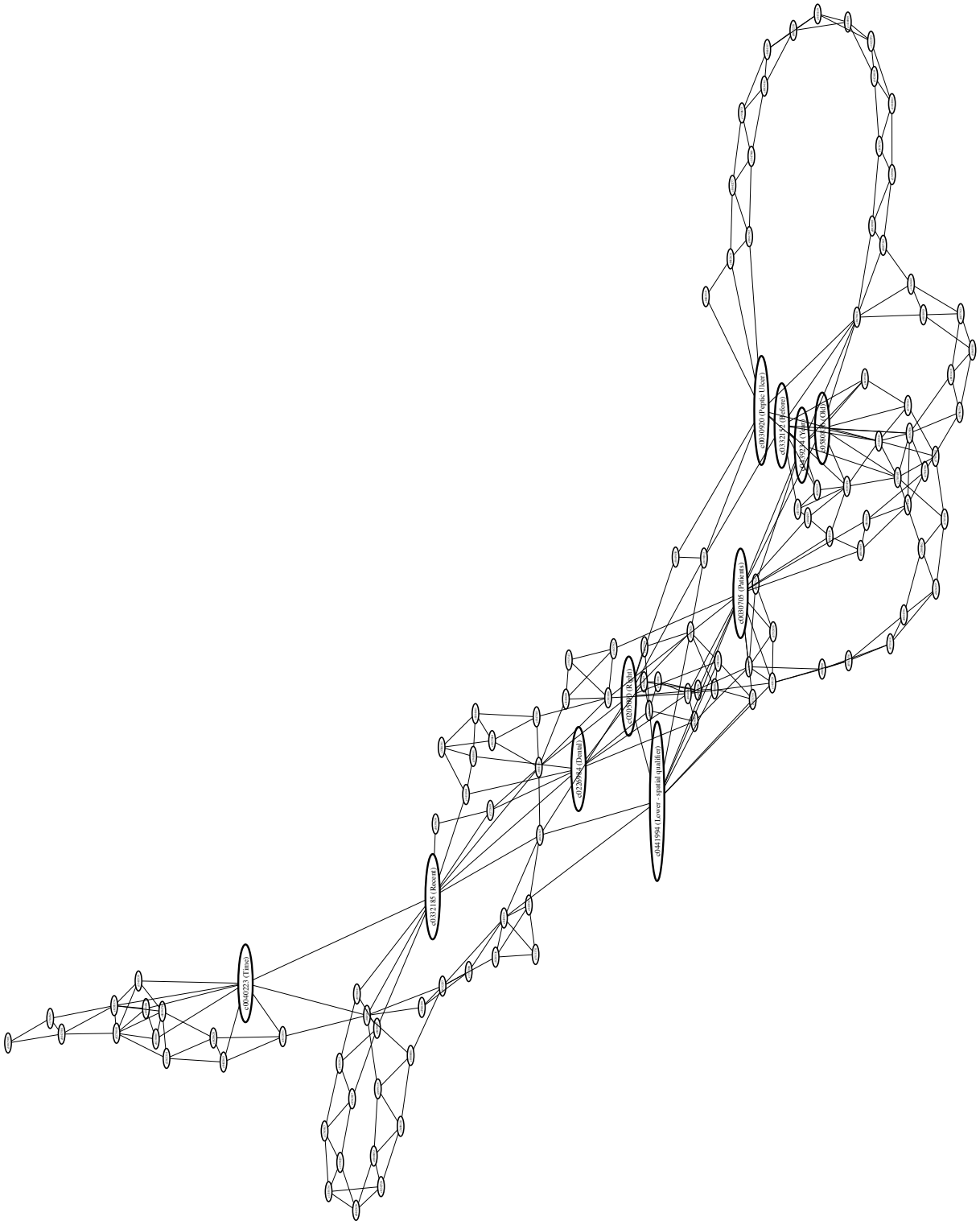


Figure 5.2: Resulting concept graph built from the medical document from Figure 5.1(a). Built using co-occurrence window $N = 3$. Bolded nodes indicate the 10 concepts with greatest score within the document (according to Equation 5.1).

(and therefore this document) and is a feature not present in the term graph.

5.3.1 Concept Retrieval Function

The same retrieval function used for terms can be applied to concepts. The original term weighting function from Equation 5.3 is modified to weight a concept c within document d_c as:

$$w(c, d_c) = idf(c) * S(v_c). \tag{5.5}$$

Then the original retrieval function is modified to:

$$R(d_c, q_c) = \sum_{c \in q_c} idf(c) * S(v_c), \tag{5.6}$$

where d_c is the document converted to concepts and q_c is the query converted to concepts.

5.3.2 Incorporating Domain Knowledge

The concepts in our concept-based graph model are taken from the SNOMED CT medical ontology. SNOMED CT also defines explicit relationships between concepts: for example the *HIV* virus concept is related to the *AIDS* disease concept. SNOMED CT can therefore also be modelled as a graph: concepts are vertices and concept relationships are edges. The number of relationships a concept has can be an indicator of the importance of the concept within the medical domain. Consider the example of the concept *Asthma*, which is related to a total of fifty other concepts, a subset of which is shown in Figure 5.3.

Concepts important to the medical domain, concepts such as diseases and treatments, are carefully modelled by the designers of SNOMED CT and contain detailed relationships to other concepts. In contrast, concepts that are peripheral to the medical domain are only broadly defined and typically contain only a small number of relationships. In contrast to the *Asthma* example, SNOMED CT defines the concept *Dog*, which is only related to five other concepts, reflecting that this concept is perhaps of lesser importance to the medical domain.

Identifying the important concepts within the medical domain may provide an indication of what users may be interested in when searching medical documents. We would like to include this indication of importance within the medical domain into our graph-based concept weighting model. Currently, the concept weighting scheme is based on the number of related concepts within the graph

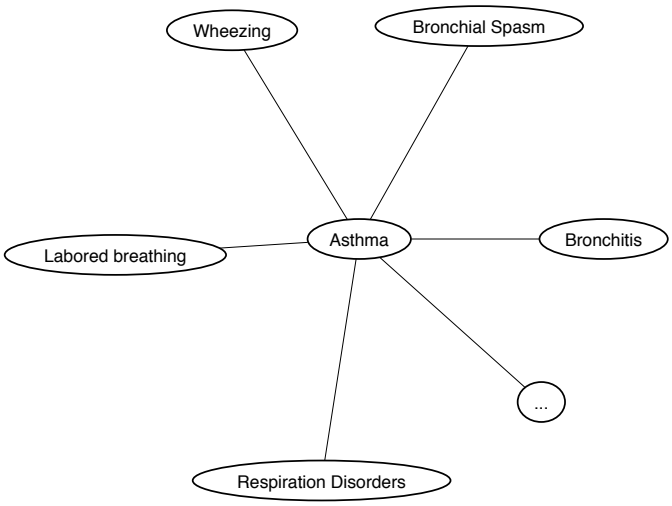


Figure 5.3: The concept *Asthma* is related to fifty other concepts in the SNOMED CT ontology. This provides an indication of its importance within the medical domain.

built for a single document. This method captures the importance of a concept within a document but does not consider the importance of a concept within the wider medical domain. The original concept weight can be adjusted by the number of related concepts within the SNOMED CT ontology which represents the 'background' importance of the concept within the medical domain. The weighting function $w(c, d_c)$ of Equation 5.5 can then be augmented as

$$w(c, d_c) = S(v_i) * idf(c) * \log(|\mathcal{V}_s(c)|), \tag{5.7}$$

where $\mathcal{V}_s(c)$ is the set of edges adjacent to concept c in the SNOMED CT ontology graph. A concept's weight is therefore adjusted based on its background weight within the medical domain, similar to the way background smoothing is applied in language models based on a term's frequency within the corpus. The logarithmic scaled value was chosen to dampen the effect of concepts with a very large number of related concepts. Using a logarithmic scaled value proved more effective than just weighting using $|\mathcal{V}_s(c)|$. Also, multiplying the value was more effective than a linear combination.

Now the weighting function contains three measures of importance: 1) the PageRank score, which represents the importance of the concept with the document; 2) the *idf*, which represents the importance of the concept within the collection; and 3) the number of edges in SNOMED CT, which represents the im-

portance of the concept within SNOMED CT. The weighting using SNOMED CT is independent of the document corpus and utilises a global measure of importance for the concept within the medical domain.

The graph-based concept weighting method described here has a number of similarities with the MEDRank system [Herskovic et al., 2011], aimed at automatically indexing biomedical articles. Using MEDLINE abstracts, MedRank first mapped the terms to concepts and then built a concept graph similar to that described in this chapter. Relationships between concepts were either determined by co-occurrence within a window (as we do) or via an external relationships database. Concepts were then ranked by decreasing PageRank score and the top k concepts chosen as the indexing labels to apply to the MEDLINE abstract. Although the concept graph and use of PageRank is similar to our method, there are some key differences. Firstly, the method was applied to a different task: MedRank produces a ranking of concepts based on a single document (abstract), instead our method produces a ranking of documents based on a set of concepts in a query. Secondly, MEDRank was developed to index journal abstracts, which differ both in length and in nature to detailed clinical records such as those in TREC MedTrack. Finally, our retrieval function uses term frequency, PageRank score and the incorporation of domain knowledge (importance of the concept within SNOMED CT) to weight a document, whereas MEDRank uses only the PageRank score.

5.4 Empirical Evaluation

This section contains the evaluation of our graph-based concept weighting model and includes our experimental setup, evaluation methodology and retrieval results.

5.4.1 Experimental Setup

The TREC Medical Records Track was adopted as the test collection. Details of this test collection were introduced in Chapter 4, Section 4.3.1. A number of baselines were implemented for comparison:

terms-tfidf: This baseline was a state-of-the-art bag-of-words model. The results from Chapter 4 showed that tf-idf demonstrated the best performance over a Language Model with Dirichlet smoothing. Therefore, Lemur's tf-idf variant from Chapter 4 was adopted for this experiment. The parameters $k1$ and b were selected based on the setting that maximised bpref

($k1 = 2.9$ and $b = 0.4$). This strong tf-idf tuned baseline is denoted terms-tfidf.

terms-graph: This baseline was an implementation of Blanco & Lioma's graph weighting method and applied to terms. The damping factor parameter ϕ from Equation 5.1 was set to 0.85 according to the findings of Blanco and Lioma [2012]. Similarly, the number of iterations and the context window size were set at 20 and 10 respectively, in line with Blanco & Lioma. This baseline is denoted terms-graph.

concepts-tfidf: This baseline was the Bag-of-concepts model from Chapter 4 using Lemur's tf-idf retrieval function. The parameters, $k1$ and b, were selected based on the setting that maximised bpref ($b = 0.75$ and $k1 = 1.5$). This tuned baseline is denoted concepts-tfidf.

The above baselines were compared against two of our proposed retrieval models:

concepts-graph: This model was the graph-based weighting method applied to concepts, as described in Section 5.3.1. The same parameter settings as terms-graph (ϕ, the number iterations and the context window size) were adopted. This model is denoted concepts-graph.

concepts-graph-snomed: This model extended the concepts-graph model by the incorporation of domain knowledge, as described in Section 5.3.2 (maintaining the same parameter settings as those used for concepts-graph). This model is denoted concepts-graph-snomed.

Evaluation was performed using the 81 topics from the TREC MedTrack collection (2011 and 2012). Retrieval results were evaluated using bpref and precision @ 10.

5.4.2 Results

The retrieval results of the three baselines and the two graph-based concept models are reported in Table 5.1.

Comparing the bag-of-words (terms-tfidf) and Bag-of-concepts (concepts-tfidf) models, the concept-based representation demonstrated improved performance. (This was the finding in Chapter 4.) However, the effect of graph-based weighting on terms (comparing terms-tfidf and terms-graph) exhibited degraded performance in relation to the baseline, although, when concepts were used to construct the graph (comparing concepts-tfidf and concepts-graph), performance improved. The incorporation of domain knowledge using SNOMED CT

Run	Bpref	Prec@10
terms-tfidf	0.3827	0.4740
concepts-tfidf	0.4147	0.4988
terms-graph	0.3525	0.4358
concepts-graph	0.4279 (+12%)	0.5086 (+7%)
concepts-graph-snomed^{t,c,g}	**0.4404** (+15%)	**0.5123**^{g} (+8%)

Table 5.1: Retrieval results on TREC MedTrack using both term and concept representations and after applying graph-based weighting and incorporation of domain knowledge. Percentage improvement shown over terms-graph. Statistic significance (paired t-test, $p < 0.05$) over t=terms-tfidf, c=concepts-tfidf, g=terms-graph.

(concepts-graph-snomed) provided additional improvements over concepts-graph in both bpref and precision. Analysis of results is presented in the next section.

Statistical significance using paired t-test was not found for any of the above results. The test collection contained only 81 query topics; van Rijsbergen comments that paired t-test may not reliably indicate statistical significance with small query sets [van Rijsbergen, 1979]. Ideally, a larger query set or additional test collections would have been used; however, the medical domain does not currently have the diversity of evaluation resources available to other domains.

5.5 Analysis and Discussion

First, we analysise the effect that graph-based weighting has on retrieval effectiveness using terms. When comparing the terms-tfidf and terms-graph baselines, we observed that the use of graph weighting actually degraded retrieval performance by 8%. This result is contrary to the findings of Blanco and Lioma [2012], who reported improvements in both bpref and precision @ 10 using the graph model on a number of test collections (over both tf-idf and BM25 baselines). In this study, the corpora used comprised newswire articles, web documents and blogs. The graph-based term weighting method may not be as suited to the peculiarities of medical documents; further analysis would be required to fully understand the reason for this.

In contrast to using terms, applying graph-based weighting to concepts did improve performance. The concepts-graph model showed improvements over both the terms-tfidf and concepts-tfidf baselines, more so in bpref, where

concepts-graph exhibited a 12% improvement in bpref over the tuned terms-tfidf baseline and a 3% improvement in bpref over the tuned concept-tfidf baseline. Graph-based weighting was effective when using concepts, but not when using terms. We hypothesise that this was due to the term-encapsulation mechanism, which encapsulates important medical n-grams as a single vertex in the graph (such as the Peptic Ulcer example from the concept graph of Figure 5.2). In contrast, the term-based graph did not encode these n-grams; instead, the two terms were split as separate vertices, both receiving a lower weight.

Overall, both the graph-based concept weighting methods (concepts-graph and concepts-graph-snomed) outperformed the other three baselines in both bpref and precision @ 10. When comparing concept-graph-snomed to concept-graph, the incorporation of domain knowledge using SNOMED CT into the weighting provided an improvement in both bpref (4%) and precision (2%). Although the overall performance after incorporating domain knowledge is not considerably higher, the method promoted additional robustness across the query set. To illustrate this, Figure 5.4 shows the number of queries exhibiting change in bpref over the terms-graph baseline for both concept graph models. The histogram shows that concept-graph-snomed tended to make small variations (gains and losses) to a larger number of queries, whereas the concepts-graph had larger variations on a smaller number of queries. The former (small gains on many queries) indicates increased robustness and is more desirable for the general applicability of the model. In summary, both graph concept models demonstrated encouraging potential to benefit some queries substantially. Further study is needed to enhance this aspect.

We now consider some interesting characteristics of the incorporation of domain knowledge. From Equation 5.7, the weighting of concept c was dependent on the logarithm of the number of edges adjacent to c in the SNOMED CT graph. Note that when a concept had only one adjacent edge in the SNOMED CT graph, then the weight w_b of query concept c for document d is zero ($\log |\mathcal{V}(c)| = \log 1 = 0$). In practice, this meant that query concepts that contained only one edge in SNOMED CT were essentially ignored (their weight always being zero). Intuitively, this seems an undesirable characteristic that could have led to significant degradation in performance. To understand the extent of this characteristic and how it actually affected performance we first consider how many queries contained concepts with only one edge in SNOMED CT (and therefore had scores of zero). The 81 test queries contained 1072 concepts in total; of these a total of 279 (26%) had only one edge in the SNOMED CT graph and were therefore ignored. Intuitively, ignoring so many concepts in the query set would have a drastic effect on retrieval performance; however, empirical results showed the contrary. This is confirmed by Figure 5.5, which compares the change in bpref,

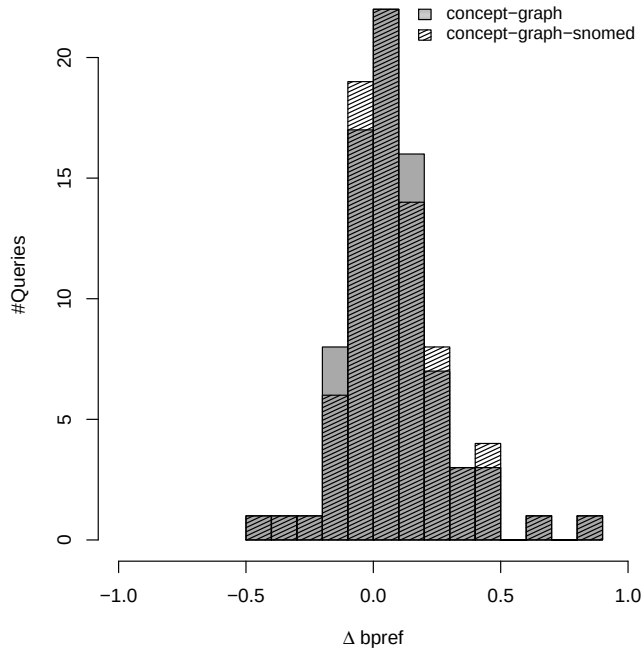


Figure 5.4: Histogram showing #queries exhibiting change in bpref over term-graph for both concept graph models. Results show concepts-graph-snomed tends to make more small improvements to many queries — an indicator of increased robustness.

after applying the SNOMED CT weighting, against the percentage of concepts excluded within the given query (i.e., where $|\mathcal{V}(c)| = 1$). Points on the far right of the x-axis indicate queries where many concepts have been excluded. The figure shows that every query had at least one concept excluded after applying the SNOMED CT weighting. For some queries, a large proportion of the concepts were excluded (far right of the x-axis); however, these queries still exhibited positive changes in bpref. These queries contained a large number of concepts that were deemed as peripheral to the medical domain. Thus, when they were excluded, performance improved. Rather than completely exclude concepts, we performed additional experiments with alternative approaches that simply assigned a logarithmic scaled weight (e.g., $1 + log(|\mathcal{V}_s(c)|)$ or $log(1 + |\mathcal{V}_s(c)|)$), but these methods never performed as well compared to when query concepts with only one adjacent edge in SNOMED CT were completely excluded. We conclude that a concept's lack of connectedness to other concepts (i.e., having only one edge) indicated that the concept provided no additional information for the query in a retrieval scenario and, in fact, the concept may have been misleading and a cause of query drift, the consequence of which was degraded

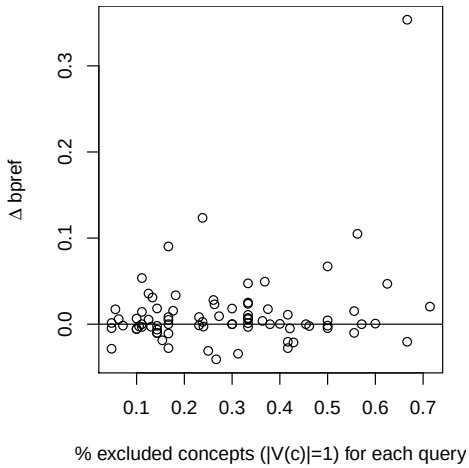


Figure 5.5: The Δbpref when excluding query concepts with only one edge in the SNOMED CT graph. x-axis indicates the percentage of concepts for a given query where $|\mathcal{V}_s(c)| = 1$ (and are therefore excluded).

performance.

The exclusion of certain concepts based on the SNOMED CT connectedness was in effect a form of *query reduction*. Previous work in information retrieval has considered query reduction methods [Kumaran and Carvalho, 2009; Bendersky and Croft, 2008], the motivation being that finding an ideal subset of query terms can result in substantial performance gains. Kumaran and Carvalho [2009] adopted a learning-to-rank approach that used statistical predictors (such as IDF, tf, Mutual Information and Query Clarity) to find an optimal query subset — they found an upper bound of 30% increase in performance, but their predictors provided only an 8% increase. Bendersky and Croft [2008] made use of corpus based statistics (such as IDF) and corpus independent indicators (such as Google n-grams[2]) to identify and weight 'key concepts' within the query. This study showed improvements in average retrieval effectiveness but found no robust feature across different test collections. In our case, we have shown that the use of a concept's connectedness in the SNOMED CT ontology provided an indicator of importance; in practice, providing a useful feature for the implementation of an implicit query reduction method. Unlike previous approaches, our method used only one feature and avoided the use of heavyweight machine learning to find an optimum feature combination; that is, no additional parameters were introduced. An interesting avenue of future work from this study is to consider query reduction specific to medical information retrieval, especially given the rich amount of domain knowledge available in

[2]Google n-grams charts the yearly count of selected n-grams found in books digitized by Google.

resources such as SNOMED CT.

Finally, the findings of this study are applicable outside of the medical domain, specifically the incorporation of domain knowledge representing the importance of a concept outside of the corpus being indexed. We used connectedness in SNOMED CT as the indicator of importance. The alternative weighting could be based on the connectedness within any other resource represented as a graph, including domain specific resources or general resources like WordNet.

5.6 Summary

This chapter presents a graph-based method to weight medical concepts found in documents for the purpose of medical IR. Existing graph-based term weighting methods were adapted and applied to concepts; a concept's weight was based on its PageRank score within the document. In addition, we presented a novel method for the incorporation of domain knowledge representing the importance of a concept within the wider medical domain (not just the corpus itself). This method had an interesting characteristic of excluding a large number of query concepts, resulting in a form of query reduction, which in turn led to improvements in performance.

Graph-based representations were chosen over bag-of-words representations because they can capture the relationships that exist between concepts. In relation to the challenge of bridging the semantic gap, the concept-based representation was used to overcome vocabulary mismatch and the graph-based representation was used to capture the innate dependence between medical concepts, which was a characteristic of the Inference of Similarity semantic gap issue.

The empirical evaluation using a number of strong baselines showed that our graph-based concept weighting method demonstrates superior retrieval performance. In particular, the use of additional domain knowledge in the form of the connectedness in SNOMED CT, although a simple measure, yielded promising results. This measure highlights just one of potentially many useful features from domain knowledge resources that could be exploited within a data-driven IR approach. However, the feature that we used captures only the number of relationships pertaining to a concept. Considerable additional information from SNOMED CT regarding a concept could potentially be utilised, including other concepts that it is connected to and the type of relationship connecting concepts. We hypothesise that this additional information is required to underpin the inference mechanisms necessary to bridge the semantic gap. The following chapter presents a retrieval model that makes extensive use of domain knowledge; this represents a unified model of semantic search as inference. The

foundation of the model is a graph-based representation of a corpus comprising ontological concepts and relationships but driven by IR probabilistic relevance estimation.

Chapter 6

Graph Inference Model

Vielleicht noch mehr als der Berührung der Menschheit mit der Natur verdankt die Graphentheorie der Berhrung der Menschen untereinander.

Perhaps even more than to the contact between mankind and nature, graph theory owes (its existence) to the contact of human beings between each other.

— Dénes König*

This chapter presents a unified model of semantic search as inference — the Graph INference model (GIN). The model utilises a graph-based representation of a corpus comprising concepts and relationships taken from a domain knowledge resource, but the model is driven by IR-based probabilistic relevance estimation. A concept-based representation, like that of the Bag-of-concepts model, is employed; however, this is integrated into a novel graph-based representation of a corpus. This graph-based representation uses background domain knowledge as the underlying structure, on top of which documents are represented. The theoretical foundations for the GIN are intuitively inspired by logic-based IR, where retrieval is modelled as a process of logical inference. In the GIN, the retrieval inference mechanism is realised as a traversal over the graph structure, from the query nodes to the document nodes.

*Dénes König (1884 – 1944) was a Hungarian mathematician who wrote the first textbook on the field of graph theory.

6.1 Background

The GIN rests on two areas of related work: firstly, logic-based IR, in which the retrieval process is modelled as one of logical inference; secondly, measures of semantic similarity, which, we will show, relate to the Logical Uncertainly Principle in logic-based IR and are an essential component of the Graph Inference model presented later in the chapter.

6.1.1 Logic-based Information Retrieval

Logic-based IR is an area of research that models the retrieval process as one of a non-classical implication, denoted $d \rightarrow q$, rather than as the traditional matching function between document d and query q. Owing to uncertainties in both query and document representations, it is usually the case that the query q cannot be inferred from the document d; therefore $P(d \rightarrow q)$ is evaluated instead, where P is a probability estimating the strength of the implication.

Fundamental to logic-based IR is the Logical Uncertainty Principle [van Rijsbergen, 2000], which provides a means of evaluating $P(d \rightarrow q)$. The Logical Uncertainty Principle states that if $d \rightarrow q$ cannot be immediately evaluated (as is often the case in IR where partial relevance exists), then additional information is added to d resulting in a document d', such that $d' \rightarrow q$ is true. The measure of the uncertainty is determined by the amount of information that needs to be added to d to allow $d' \rightarrow q$ to be true.

Following on from initial work by Van Rijsbergen [1986], Nie [1989] described the uncertainty of implication as the distance or *effort* required to alter d to d', formally:

$$P(d \rightarrow q) \propto \frac{1}{\epsilon(d, d')},$$

where the function $\epsilon(d, d')$ measures the effort (or alternatively, distance) to move from d to d'. The effort is further described as a sequence of changes, starting from d and finishing at d', thus:

$$\epsilon(d, d') = \sum_{d_i \in \langle d, \ldots, d_{i-1}, d_i, \ldots, d' \rangle} \epsilon(d_{i-1}, d_i). \tag{6.1}$$

The measure of effort $\epsilon(d, d')$ can be considered as inverse to a measure of similarity — the more similar two documents, the less the effort or distance between them. In this way, the uncertainty of the implication can be determined

by a sequence of similarity estimations:

$$\begin{aligned} P(d \rightarrow q) &\propto \delta(d, d') \\ &\propto \bigotimes_i \delta(d_{i-1}, d_i) \end{aligned} \tag{6.2}$$

where, assuming a sequence of transitions $d, \ldots, d_i, \ldots, d', i \geq 0$, the function $\delta(d, d')$ measures the *similarity* between d and d'. The $\bigotimes$ operator determines how individual similarity measures are combined. The actual implementation of both the $\bigotimes$ operator and the similarity function $\delta(d, d')$ are intentionally unspecified so that the model remains abstract and, therefore, can be instantiated in a way that best suits the particular application. To describe the sequence of transitions from d to d', Nie used a graph analogy and presents the illustration shown in Figure 6.1, showing the sequence of changes as a traversal over a graph.

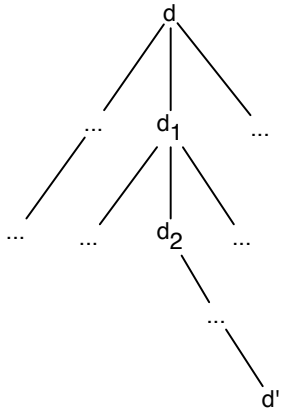


Figure 6.1: A graph analogy of the Logical Uncertainty Principle, described by Nie [1989] as the sequence of transitions from d to d'.

The literature on logic-based IR is primarily theoretical in nature and usually does not report large scale evaluations (an exception being the Logical Imaging approach of Crestani [1998]). However, logic-based IR provides a number of aspects particularly pertinent to this thesis. Firstly, this thesis argues that semantic search requires inference — and logic-based IR models the retrieval process as a process of logical inference. Secondly, the Logical Uncertainty Principle incorporates some measure of effort — and measures of effort have often been modelled in the literature by means of measures of similarity. This directly addresses the requirements from the semantic gap problem of Inference of Similarity (Section 2.4). Finally, the graph analogy presented by Nie [1989] provides an intuition for instantiating a retrieval model that incorporates an

inference mechanism. It aligns well with the focus on graph-based retrieval from the previous chapter. Logic-based IR provides the theoretical foundations for our unified model of semantic search as inference.

6.1.2 Semantic Similarity

The previous section showed that the similarity measure is a key component of logic-based IR. The GIN presented in this chapter will make extensive use of similarity measures, so it is worth considering the choice of measure here. In logic-based IR, similarity is directly related to distances between 'possible worlds'. The reason for this flows directly from the Logical Imaging [Crestani, 1998; Zuccon et al., 2009]. The generic form of imaging is summarised as follows: If $x \rightarrow y$ does not go through at a world w, then the implication at a neighbouring world w' is evaluated. If the implication holds at this world, then the probability of the implication holding at the original world w is inversely proportional to the distance between these worlds, or in other words, proportional to the similarity between these worlds. In IR there are typically two choices for worlds: documents or terms. Therefore, similarity between such worlds can be operationalised by semantic similarity. This then goes beyond previous work in logic-based IR by equating a concept to a "world".

In the literature, semantic similarity between two terms or concepts is usually calculated in one of two ways: path-based or corpus-based. Path-based measures use external resources such as ontologies (similarity being inversely proportional to the length of the path between two concepts in the thesaurus). Path-based measures are dependent on only the external thesauri; they do not derive any measure of similarity from the corpus in which they occur. In contrast, corpus-based measures make use of only corpus statistics to derive the measure of similarity. A comparison between path-based and corpus-based measures in the biomedical domain by Pedersen et al. [2007] showed that a corpus-based measure correlated most strongly with human judged similarity measures provided by medical professionals. Based on this finding, we evaluated a number of different corpus-based measures of semantic similarity to determine which correlated most strongly with human-judged similarity of medical concepts. The measures evaluated included Random Indexing, Latent Semantic Analysis, Hyperspace Analogue to Language, Document Vector Cosine similarity, Positive Pointwise Mutual Information, Cross Entropy Reduction, Language Model with Jensen-Shannon Divergence and Latent Dirichlet Allocation. Full details of the evaluation of these different measures are provided in Appendix B. The findings from this study highlighted the effectiveness and robustness of the Document Vector Cosine similarity measure; that is, the cosine angle between two terms or

concepts represented by their document vectors and weighted with tf-idf. Based on this finding, the Document Vector Cosine similarity measure will be adopted later in the chapter as the similarity measure in the GIN.

6.2 Graph Inference Model Theory

This section presents the theoretical aspects of the GIN. The model is described independent of its application in medical IR; this is intentional to emphasise the general applicability of the model. Implementation specific aspects are left until Section 6.3.

6.2.1 Information Units and Relationships

This section defines the basic elements that make up our graph-based representation of queries and documents. Firstly, we define an Information Unit.

Definition 1 *Let* $\mathbb{U}$ *denote a non-empty set of Information Units.*

An Information Unit ($u \in \mathbb{U}$) is an abstract and general representation. It may be a concept defined in the SNOMED CT ontology. Outside the medical domain, an Information Unit can come from any external resource (ontology or controlled vocabulary). It can be an entity derived as a result of an Information Extraction process (for example, a Person or Place). Finally, an Information Unit can also be an n-gram or term phrase and in its most basic form an Information Unit could be a single term.

Information Units may belong to one or more Information Types.

Definition 2 *Let* $\mathbb{T}$ *denote a set of Information Types.*

A Type ($t \in \mathbb{T}$) may simply be a part-of-speech type or more complex entities such as Person, Place, etc. In the medical domain, $\mathbb{T}$ is the set of Semantic Types explicitly defined in UMLS or SNOMED CT, for example Disease, Treatment or Symptom. Each Information Unit may belong to one or more Type according to a Type relationship.

Definition 3 *Let* T *be a total function which maps Information Units to Information Types.*

$$T : \mathbb{U} \rightarrow \mathbb{T}$$

In medical terminologies such as UMLS and SNOMED CT, types are explicitly defined and each concept is associated with a corresponding type.

In addition to a Type relationship, Information Units are also related to each other in a many-to-many relationship:

Definition 4 *Let* $\mathbb{R} \subseteq \mathbb{U} \times \mathbb{U}$ *define a non-empty set of Information Relationships.*

If the Information Unit comes from an ontology or thesaurus, the relationship may be explicitly pre-defined. This is the case for UMLS or SNOMED CT, which includes explicit relationships between concepts. For other types of Information Unit, such as terms or n-grams, Information Relationships may be determined by term co-occurrences relationships. Other implementations may link Information Units that are semantically similar to each other. The particular implementation will most likely impose further restriction on $\mathbb{R}$; for example, if the relationships are taken from SNOMED CT, which can be represented as a directed acyclic graph, then $\mathbb{R}$ would be irreflexive and antisymmetric.

Bringing together the above definitions, a graph can be constructed where *Information Units* represent vertices and *Information Relationships* represent the edges between *Information Units*. If Information Units are SNOMED CT concepts and Information Relationships are SNOMED CT relationships, then the resulting graph is simply the SNOMED CT ontology represented as a graph.

Definition 5 *Let* $G = \langle \mathbb{U}, \mathbb{T}, T, \mathbb{R} \rangle$ *denote an Information Graph.*

It is the incorporation of queries and documents into this graph representation that provides a representation that facilitates retrieval by inference. We first provide a formal definition of queries and documents within our framework and then describe how they are integrated into the graph representation.

6.2.2 Queries and documents

A query expresses a user's information need.

Definition 6 *A query* q *is a sequence of Information Units.*

$$q = \langle u_0, \ldots, u_m \rangle$$

Definition 7 *A document* d *is a sequence of Information Units:*

$$d = \langle u_0, \ldots, u_n \rangle$$

The sequence captures the order in which Information Units appear within the document. This differs from a bag-of-words (or Bag-of-concepts) set which does not capture word order.

6.2.3 Corpus and Document Representation

An Information Graph can be used to model an entire corpus of documents. This is achieved by first constructing a graph with Information Units as nodes and Information Relationships as edges and then attaching to each node the list of documents or the query in which that Information Unit appears. An example graph created using of this approach is provided in Figure 6.2. There are three Information Units u_0, u_1 and u_2 and two document d_1 and d_2. The Information Unit u_0 is found in document d_1 so d_1 is attached to the u_0 node whereas u_2 is found in both d_1 and d_2 so these documents are attached to u_2.

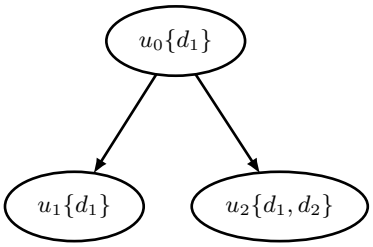


Figure 6.2: Example graph-based corpus representation — basic node-document representation.

Using this method, the graph of Information Units and Information Relationships is the underlying skeleton to which documents and queries are assigned. Rather than just attaching documents and queries to a node, a weight or initial probability can be assigned. We call this an initial probability because it is assigned prior to retrieval and is independent of the query. After estimating the initial probabilities, the node u_0 in Figure 6.2 would no longer contain $\{d_1\}$ and instead contain $\{P(u_0|d_1)\}$, the initial probability of the Information Unit u_0 within the document d_1. Assigning probabilities to each node results in the modified representation shown in Figure 6.3. Note that although the figure shows only the initial probability for the document attached to the node, in reality the initial probability can be estimated for all documents in the collection. How these probabilities are estimated is not constrained by the model and is an implementation-specific decision. They can be implemented using the Maximum Likelihood Estimate (i.e., the normalised term frequency of u_0 in d_1). In this case, if the Information Unit does not appear in the document then its initial probability will be zero. Instead, a Dirichlet smoothing (Equation 3.5)

can be used to refine the probabilities and thus avoid zero probability estimates. In this case, every document will have an initial probability with respect to the Information Unit.

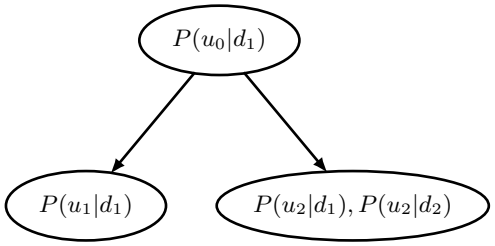


Figure 6.3: Example graph-based corpus representation — node-document representation with initial probabilities assigned to each node.

Alternatively, the initial weights might not be probabilities at all and instead others measures such as a BM25 or tf.idf weight could be assigned. The only requirement is that the weight represent a measure of importance for that Information Unit in the context of the specific document or query.

If SNOMED CT is used as the source of Information Units and Relationships, then SNOMED CT provides the underlying graph structure — the underlying *skeleton*. In this way, the external domain knowledge explicit in SNOMED CT — and the medical domain in general — is encoded within the graph-based representation of the corpus. The representation integrates background formal domain knowledge with data from the particular corpus.

6.2.4 Diffusion Factor

An important requirement for bridging the semantic gap is modelling the strength of association, or measure of uncertainty, between concepts. (This is part of the Inference of Similarity semantic gap problem of Section 2.4.) To account for this, we introduce the *diffusion factor*: a measure of the strength of association, or spread of information, between two Information Units in the corpus graph. The diffusion factor is akin to the similarity measure from the Logical Uncertainty Principle; however, there are some important distinctions. In the Logical Uncertainty Principle, the similarity measure estimates the amount of uncertainty to transition from document d to d', such that $d' \rightarrow q$ is true. Instead, the diffusion factor measures the amount of uncertainty to transition from an Information Unit u to u'. In addition, the diffusion factor can capture more than just a similarity measure: it can also capture a strength of association based on *how* the two Information Units are connected. In our model, this is represented by the Information Unit Relationship (Definition 4). The diffusion

factor is defined as:

Definition 8 *Let δ be a recursive function $\delta : \mathbb{U} \times \mathbb{U} \rightarrow \mathbb{R}^+$ that denotes the maximal diffusion between two information units, $u, u' \in \mathbb{U}$ such that:*

$$\delta(u, u') = \begin{cases} 1, & \text{if } u = u' \\ \delta_0(u, u'), & \text{if } uRu' \\ \arg\max_{u_i \in \mathbb{U}: uRu_i} \delta(u, u_i) \otimes \delta(u_i, u'), & \text{otherwise} \end{cases} \tag{6.3}$$

$\mathbb{R}^+$ represents the set of positive real numbers. The maximal operator accounts for the case of multiple paths to transition between u and u'. In this case, the path with the greatest diffusion factor (least effort) is favoured. As with the Logical Uncertainty Principle, the definition of $\bigotimes$ operator is implementation-dependent. However, if the diffusion factor is implemented using a probability, then the probabilities can be multiplied to combine diffusion factors:

$$\delta(u, u') = \begin{cases} 1, & \text{if } u = u' \\ \delta_0(u, u'), & \text{if } uRu' \\ \arg\max_{u_i \in \mathbb{U}: uRu_i} \delta(u, u_i)\delta(u_i, u'), & \text{otherwise} \end{cases} \tag{6.4}$$

Other alternative implementations for the $\bigotimes$ operator could take into account the actual number of transitions for estimating the diffusion or could implement the overall diffusion factor as the maximum or minimum value of the individual diffusion factors.

Although not imposed above by the general definition, the diffusion factor can be calculated in a number of different ways, both using corpus-based techniques and from domain knowledge. For corpus-based techniques, a semantic similarity measure, such as those mentioned earlier, would capture the strength of association between Information Units; we denote this strength $\text{sim}(u_{i-1}, u_i)$. For domain knowledge-based techniques, the Information Unit Relationship would capture some measure of association; we denote this strength $\text{rel}(u_{i-1}, u_i)$. As an example from SNOMED CT, the *ISA* relationship would have a greater strength of association than the *Procedure site* relationship. The base case of the recursive diffusion factor (δ_0) between u and u' with uRu' can be estimated as a linear interpolation of the two functions:

$$\delta_0(u, u') = \alpha \ \text{sim}(u, u') + (1 - \alpha) \ \text{rel}(u, u') \quad 0 \leq \alpha \leq 1 \tag{6.5}$$

where the parameter α is the *diffusion mix* of the similarity and relationship type measure.

6.2.5 Retrieval Function

Having defined a diffusion factor function, we can now use it as a measure of strength of implication. Ultimately, we wish to estimate the probability of the implication between document and query: $P(d \to q)$. However, before providing this, we first consider the probability of implication between a single Information Unit in the document and a single Information Unit in the query: $P(u_d \to u_q)$, where $u_d \in d$ and $u_q \in q$. The event space is all the concepts in the document and all the concepts in the query. The strength of implication is assumed to be proportional to the diffusion factor required to transition from u_d to u_q:

$$P(u_d \to u_q) \propto \delta(u_d, u_q).$$

This assumption is further refined by recalling that the graph representation of the corpus from Section 6.2.3 also contains an initial probability $P(u_d|d)$ for each Information Unit. Therefore,

$$P(u_d \to u_q) \propto P(u_d|d)\, \delta(u_d, u_q).$$

The initial probability $P(u_d|d)$ represents the strength of the Information Unit u_d in document d. As previously stated, this can be estimated in a number of different ways (for example, as the Maximum Likelihood Estimate or Dirichlet smoothed estimate). In addition, it could be determined by other features such as the Type Relationship (Definition 3) of the Information Unit.

Having provided a means of evaluating $P(u_d \to u_q)$ we can now return to the original problem of inferring the query from the document, i.e. $P(d \to q)$. The single Information Unit inference definition can be extended to that of query and document by evaluating each combination of query Information Unit $u_q \in q$ and document Information Unit $u_d \in d$:

$$\begin{aligned} P(d \to q) &= \bigodot_{u_q \in q} \bigsqcap_{u_d \in d} P(u_d \to u_q) \\ &\propto \bigodot_{u_q \in q} \bigsqcap_{u_d \in d} P(u_d|d)\, \delta(u_d, u_q). \end{aligned} \tag{6.6}$$

This is the general retrieval function of the Graph Inference model. It has two placeholders for operators: $\bigodot$, for Information Units in the query and $\bigsqcap$, for Information Units in the document. Their definitions are left to the specific implementation but we consider two possible alternatives here. First, if the query Information Units are assumed independent (as is the case for many retrieval models) and the document Information Units are also considered

independent, then the probabilities are multiplied; therefore $\bigodot = \prod$ and $\square = \prod$ to derive the retrieval status value function:

$$\text{RSV}(d, q) = \prod_{u_q \in q} \prod_{u_d \in d} P(u_d|d)\, \delta(u_d, u_q). \tag{6.7}$$

In this implementation, the Information Units u_i, related to u_q, are considered as additional information regarding the query, with the diffusion factor controlling the strength of association between the two. This is akin to the query expansion process where additional query terms are derived. The implementation shown above in Equation 6.7 is similar to the approach used in probabilistic language modelling.

An alternative implementation is still to consider query Information Unit as independent but to consider the document Information Units as dependent. In this case, the query placeholder $\bigodot$ is a product ($\bigodot = \prod$), thus multiplying the independent query Information Units, but the related Information Units in the document are summed ($\square = \sum$). This gives the retrieval status value function:

$$\text{RSV}(d, q) = \prod_{u_q \in q} \sum_{u_d \in d} P(u_d|d)\, \delta(u_d, u_q). \tag{6.8}$$

In this case, the Information Units related to u_q via the graph represent an alternative representation of the query Information Unit u_q and provide an additional source of supporting evidence (albeit a weaker source according to the discounting applied by the diffusion factor).

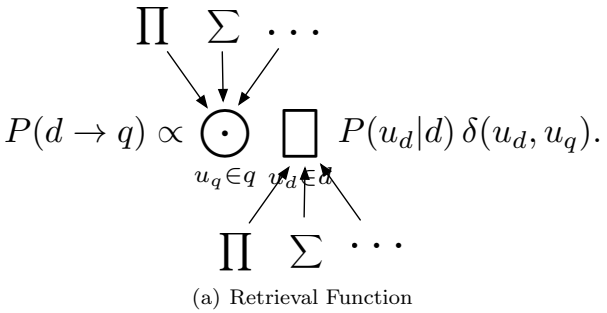
The general retrieval function from Equation 6.6 can be applied in a number of different ways; two are presented above but others are possible. Figure 6.4 shows a number of different possible implementations. The Graph Inference model intentionally generalises these operators so a particular implementation is not imposed by the model. This means that the model can be applied to a number of different scenarios, making it a general model from which particular inference-based retrieval models can be instantiated.

6.2.6 Worked Retrieval Example

This section provides a simple example of evaluating a query using the Graph Inference model. It is provided to highlight a number of characteristics of the model and how they might benefit retrieval.

Consider a query q and three documents d_1, d_2 and d_3:

$$q = \langle u_q \rangle \quad d_1 = \langle u_1, u_2, u_q \rangle \quad d_2 = \langle u_3, u_q \rangle \quad d_3 = \langle u_4 \rangle \tag{6.9}$$



(a) Retrieval Function

$$\delta(u, u') = \begin{cases} 1, & \text{if } u = u' \\ \delta_0(u, u'), & \text{if } uRu' \\ \arg\max_{u_i \in \mathbb{U}: uRu_i} \delta(u, u_i) \otimes \delta(u_i, u'), & \text{otherwise} \end{cases}$$

$\prod$ $\cdots$

(b) Diffusion Factor

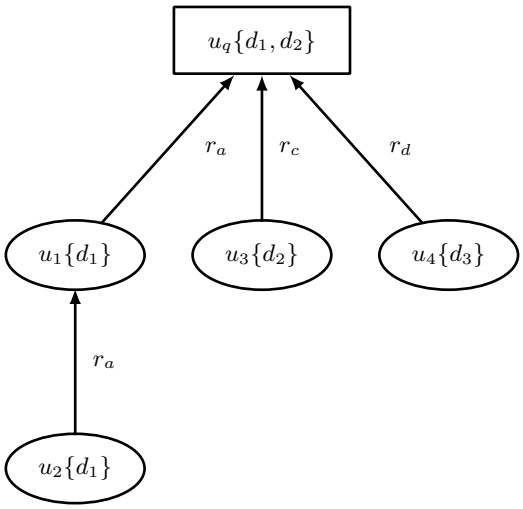
Figure 6.4: Possible implementation options for the Graph Inference model retrieval function and diffusion factor.

The posting list for the documents and query is:

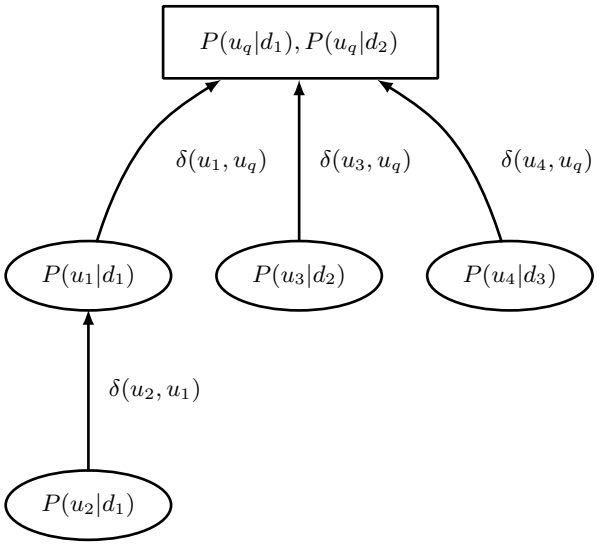
$$\begin{aligned} u_1 &: d_1 \\ u_2 &: d_1 \\ u_3 &: d_2 \\ u_4 &: d_3 \\ u_q &: d_1, d_2, q \end{aligned} \tag{6.10}$$

From the above query and documents, the graph shown in Figure 6.5(a) is created. The query node u_q is indicated as a square node; other document nodes are elliptical. Documents are attached to the Information Unit nodes they encompass. Recall that instead of just attaching the document to a node, an initial probability can be assigned to represent the likelihood of that Information Unit in the context of that document. Once these initial probabilities are estimated, the resulting graph is shown in Figure 6.5(b). Also included in the figure are the diffusion factors representing the strength of association between Information Units.

Now we show how diffusion factors combine to come up with a probability of implication, $P(d \to q)$. We consider the scoring of each document separately and use the retrieval function from Equation 6.7 (i.e., where $\bigodot = \prod$ and $\square = \prod$). Starting with document d_1, Figure 6.6(a) shows the graph traversal used to score d_1. Black nodes and edges relate to the current documents (d_1) and grey nodes and edges relate to other documents. The score for d_1 comes from three sources of evidence. Firstly, d_1 contains the query Information Unit u_q



(a) Basic node-document representation.



(b) Node-document representation with initial probabilities assigned to each node.

Figure 6.5: Corpus and document representation for retrieval example. Square nodes indicate a query node; documents are attached to the node that they encompass.

so d_1 first receives $P(u_q|d_1)$. Secondly, d_1 also contains the Information Unit u_1, which is related to the query u_q; so d_1 receives $P(u_1|d_1)$ but discounted by the effort to move this probability as determined by the diffusion factor $\delta(u_1, u_q)$. Finally, d_1 also contains u_2, which is related to u_q via u_1; so d_1 receives $P(u_2|d_1) * \delta(u_2, u_1) * \delta(u_1, u_q)$. These three different estimates determine the score of d_1 under the GIN. Note that most information retrieval models would consider only the first estimate, that is $P(u_q|d_1)$.

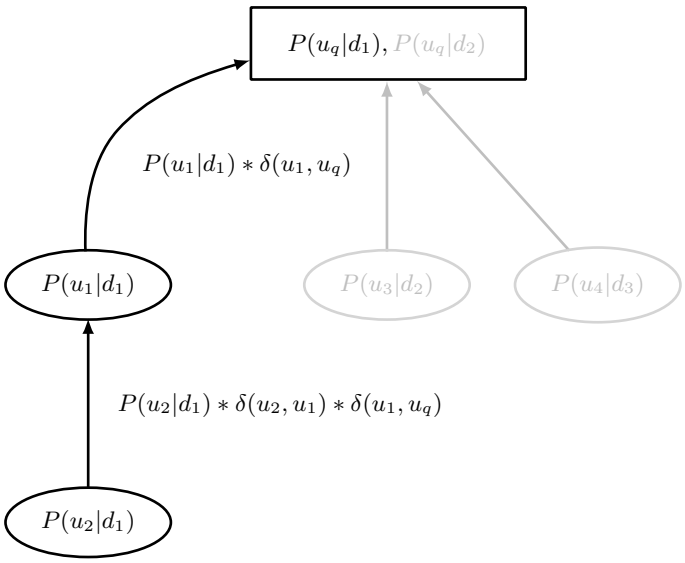
Figure 6.6(b) illustrates the process for d_2. The score for d_2 comes from only two sources: $P(u_q|d_2)$, because the document contains the query; and $P(u_3|d_2) * \delta(u_3, u_q)$, because d_2 contains one other Information Unit related to the query. Both documents d_1 and d_2 contain the query and both contain Information Units related to the query. However, d_1 contains additional evidence in the form of u_2 (which is related to u_q via u_1). This additional evidence may result in d_1 being ranked higher than d_2 (depending on the actual strength of the initial probabilities and diffusion factors).

Figure 6.6(c) illustrates the process for d_3. This example illustrates the situation of scoring a document that does not contain any *query* Information Units. For most information retrieval models, such a document would be ignored.[1] Although d_3 does not contain the query Information Unit, it does contain u_4, which is related to the query. Therefore, even though it does not contain the query, d_3 is still retrieved by the Graph Inference model; its score is determined by $P(u_4|d_3)$ but discounted by the association between u_4 and u_q.

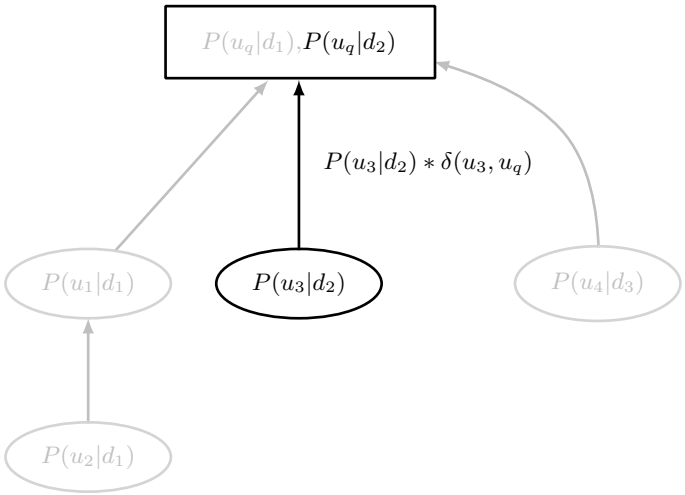
6.3 Graph Inference Model Implementation

The previous section on Graph Inference model theory intentionally omitted a number of implementation aspects to ensure the general applicability of the model. In this section, an efficient implementation of the Graph Inference model is provided. The implementation is divided into two parts: indexing and retrieval. Since the basis of the model is a graph-based representation, the indexing process is responsible for constructing a graph and the retrieval process is responsible for traversing it according to a query.

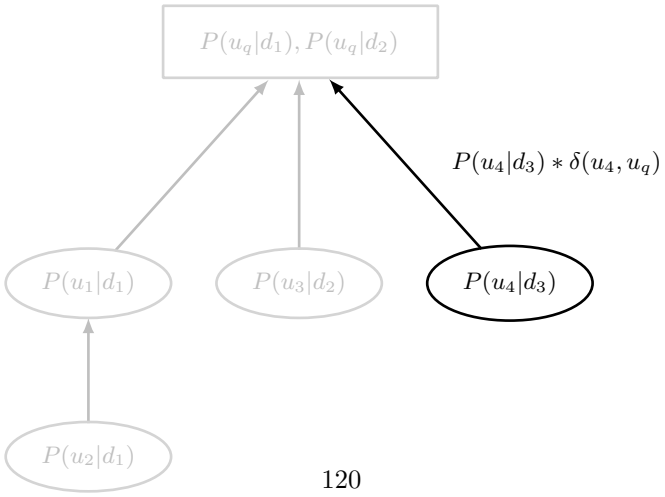
[1] Theoretically, most IR models do not impose the restriction that only documents that contain a query term should be returned; in practice, however, they typically score only documents that contain at least one query term.



(a) Retrieval process for document d_1.



(b) Retrieval process for document d_2.



(c) Retrieval process for document d_3.

Figure 6.6: Retrieval process for three example documents using Graph Inference model.

6.3.1 Indexing

Rather than construct the graph directly from the corpus, the documents are first indexed using a standard IR indexer (in this case, Lemur) to create an inverted file index. The Graph Inference model then uses this efficient data structure to build the graph. Doing so means that the graph can be rebuilt quickly with different options without having to process the corpus again. Additionally, it means that the Graph Inference model can be applied to existing indices without requiring access to the original corpus.

Recall that the nodes in the graph constitute Information Units and the edges constitute Information Unit Relationships. In our implementation, the Information Unit is a term or concept (depending on the representation) in the inverted file index. The relationships are based on the explicit associations taken from some domain knowledge source, such as a medical ontology. Therefore, in addition to the inverted file index, the other input to the Graph Inference model is the set of relationships connecting Information Units.

The Graph Inference model indexing process is detailed in Algorithm 1, which takes as input the inverted file index (denoted Idx) and the set of relationships connecting Information Units (denoted Ont since this is often simply supplied as the ontology itself).

Algorithm 1 Pseudo code for efficient graph indexing.

Require: Idx, Ont ▷ Index, Ontology
Ensure: $G = \langle V, E \rangle$ ▷ Graph (vertices and edges)

```
1:
2: function CREATE_VERTEX(u)
3:     v = vertex(u)
4:     if v ∉ V then
5:         V = V + v                          ▷ Add node to graph
6:         return v
7:
8: function CREATE_EDGE(v1, v2, diffusion)
9:     if (v1, v2, diffusion) ∉ E then
10:        e = edge(v1, v2, diffusion)
11:        E = E + e                          ▷ Add edge to graph
12:        return e
13:
14: for u_i ∈ Idx do
15:     v_i = CREATE_VERTEX(u_i)
16:     for u' ∈ related_concepts(Ont, u_i) do
17:         v' = CREATE_VERTEX(u')
18:         diffusion = δ(u_i, u', α)          ▷ Calculate diffusion factor
19:         e_i = CREATE_EDGE(v_i, v', diffusion)
20: serialize_graph(path(Idx), G)
```

Using this method, each Information Unit (i.e., term or concept) in the collection becomes a node in the graph. The graph also contains many additional nodes representing Information Units not in the corpus but related (via the ontology) to Information Units that are in the corpus. These can provide additional domain knowledge at retrieval time and could link two Information Units that appear in the collection but have no direct edge between them. In the method described here, the initial probabilities on the nodes are not calculated at indexing time; this is left to retrieval time to allow for different weighting models to be selected. Depending on the use case, a more efficient implementation could calculate these at indexing time.

For a large corpus, the indexing process can be run in parallel, provided thread-safe, concurrent access to the graph is managed. After indexing, the resulting graph is serialised to reside with the original inverted file index.

Diffusion Factor

In our implementation of the GIN, the diffusion factor (line 18 of Algorithm 1) is calculated by mixing two measures, semantic similarity and relationship type, as previously shown in Equation 6.5. Semantic similarity can be implemented as the cosine angle between two term or concept document vectors. (This was described in Section 6.1.2.) The relationship types are the Information Units Relationships explicitly defined in the the input ontology. In SNOMED CT, for example, the Information Units Relationships are the explicit relationships between concepts, for example *ISA*, *causative agent* or *finding site*. These different relationship types can indicate a strength of association: an *ISA* relationship might indicates a strong relationship between two concepts, whereas relationships such as *severity* indicate a much weaker association. The semantic similarity and relationship type measures are mixed according to the diffusion mix parameter α.

6.3.2 Retrieval

The previous section on the theory underlying the GIN concluded with the general retrieval function shown in Equation 6.6. We now expand on this to realise an efficient implementation. The retrieval function evaluates the relevance of a particular document d to a query q, but it does not consider which documents are chosen for scoring. Evaluating all documents in the collection against a query is obviously infeasible, so a subset of possibly relevant documents is therefore required for evaluation. In other retrieval models, this is often simply

determined by those documents that contain at least one query term. However, the GIN has the ability to score potentially relevant documents that do not contain the query but may contain information related to the query (see document d_3 in the worked retrieval example of Section 6.2.6). For feasibility reasons, an alternative method is therefore required to limit which documents should be scored using the GIN. This can be determined by the diffusion factor, which increases exponentially the further the node is from the query. At some point, the effort becomes so large that a document at that node is not worth consideration (its probability being insignificant once weighted by the diffusion factor). As a result, we need consider only the documents attached to Information Unit nodes k edges away from the query node. Retrieval can therefore be modelled as a depth-first-search, originating from the query node, visiting only nodes k edges away. This process is detailed in Algorithm 2. The inputs are: the query, comprising a sequence of Information Units; the graph, created by the previous indexing process; and the depth k, determining the maximum depth of traversal.

Algorithm 2 Pseudo code for efficient depth-first-search graph retrieval.

Input: Idx, Q, G, k ▷ Index, Query, Graph, Max depth
Output: $scores \leftarrow \{d_0, \ldots, d_n\}$ ▷ Document scores

```
1:
2: for u_q ∈ Q do
3:     DFS(u_q, 0)                          ▷ Start traverse from query node, depth 0
4:
5: function DFS(u, depth)
6:     if depth ≤ k then
7:         for d_i ∈ Idx.docs(u) do          ▷ Documents containing this Info. Unit
8:             scores[d_i] = scores[d_i] + P(u|d_i) * δ(u, u_q)   ▷ Score each doc at
9:                                                                ▷ this node
10:        for u' ∈ children(u) do
11:            DFS(u', depth + 1)            ▷ Recursively traverse child nodes
```

When the maximum depth parameter k is set to zero, then the algorithm processes only the query nodes and does not traverse any edges. In this case, if the initial probabilities are Dirichlet smoothed estimates, then $k = 0$ represents a standard probabilistic language model with Dirichlet smoothing. Similarly, if BM25 weights are assigned to nodes, then $k = 0$ is a standard BM25 model. Thus, the GIN incorporates these standard IR models by setting the depth parameter. This is particularly useful for evaluation: the retrieval effectiveness can be measured for different settings of k with $k = 0$ constituting a standard benchmark for comparison.

Computational Complexity Analysis

The computational complexity of Algorithm 2 is based on the number of documents scored each time a node is visited (score function on line 8). At each depth level $l = [0, .., k]$, there are e^l nodes, where e is the average number of edges (degree) for nodes in the graph G. Assuming an average of d documents are attached to each node, then $e^l d$ documents are processed at each depth level. When traversing multiple levels for a *singl*e query concept, the number of documents processed is:

$$\sum_{l=0}^{k} e^l d.$$

For a query of size $|Q|$ concepts, the number of documents processed is:

$$|Q| \sum_{l=0}^{k} e^l d.$$

As stated previously, at a certain depth the diffusion factor becomes so small that documents scored at this level will not change the overall ranking; thus, we need consider only the documents k edges away from the query node.[2] The size of d is determined by the average inverse document frequency of the collection. The size of e (average number of edges per node) is the average degree of G (for SNOMED CT the average degree is 4.4). The size of the query, $|Q|$, is typically small for a retrieval scenario. With e, l and $|Q|$ all small, the retrieval method is computationally efficient.

Reranking

The Graph Inference model can also be used in 'reranking mode'. This is performed by scoring an initial set of documents using the GIN at depth level $k = 0$ and then, at subsequent depth levels, only considering those documents already seen at level 0. If the initial probabilities assigned to each node are Dirichlet smoothed estimates, then the result is Graph Inference model reranking of a standard language model with Dirichlet smoothing. Reranking may be desirable in some cases, although one of the motivating characteristics of the Graph Inference model is its ability to retrieve new documents that do not contain the query — and therefore would not be retrieved at depth level 0 — but are relevant because they contain information related to the query.

[2] The empirical evaluation revealed $k = [0 - 3]$ was preferred.

6.4 Empirical Evaluation

This section contains the evaluation of the Graph Inference model and includes our experimental setup, evaluation methodology and retrieval results.

6.4.1 Experimental Setup

Concept-based Collection and Index

As with previous chapters, the test collection used here was the TREC Medical Records Track. Both documents and queries were converted to SNOMED CT concepts using the method already outlined in Chapter 4, Section 4.1.1. Following this, the concept-based collection was indexed using the Lemur IR library.[3] Each unique SNOMED CT concept in the index represented an Information Unit and the index was the first input to the Graph Inference model indexing process.

Graph Inference Model Indexing

The other input to the Graph Inference model is a set of relationships connecting Information Units. In our implementation, relationships were taken directly from the SNOMED CT ontology. SNOMED CT was chosen over other medical domain knowledge resources for a number of reasons. SNOMED CT covers a wide range of medical knowledge in a single, self contained resource. Other resources are more specific to certain situations; for example, the ICD coding scheme is used for diagnostic coding or the the Medical Subject Headings (MeSH) controlled vocabulary is used for indexing medical journal articles. Although UMLS is general purpose, it was constructed by amalgamating a number of individual medical domain knowledge resources, each with varying coverage and quality. In contrast, SNOMED CT has a quality control process overseen by the International Health Terminology Standards Development Organisation. Finally, SNOMED CT is now mandated as the standard medical terminology in Australia and in many other countries.

With SNOMED CT as the underlying domain ontology, we applied the indexing process described in Section 6.3.1. The construction of the graph was done using the LEMON graph library.[4] The graph was serialised using LEMON and stored inside the Lemur index directory. For the MedTrack corpus, which

[3] Lemur version 4.12; http://www.lemurproject.org/

[4] LEMON (Library for Efficient Modelling and Optimisation in Networks) is a C++ template library providing efficient implementations of common data structures and algorithms with a focus on graphs and networks; see http://lemon.cs.elte.hu/.

has a vocabulary size of 36,467 SNOMED CT concepts, the resulting graph was 4.4MB.

Graph Inference Model Retrieval

The retrieval process requires a number of inputs: (1) the document index, in our case the Lemur index; (2) the graph, which was the LEMON graph, previously created at indexing time, read into memory prior to retrieval; (3) the set of (concept-based) query topics; and (4) the depth parameter k.

Depth setting (k): The depth parameter k controls how many edges are traversed from the query node and reflects how much additional information the model will draw on to score documents. We focus on three different depth settings, 0, 1 and 2, which we denote as lvl0, lvl1 and lvl2, reflecting different levels from the query node.[5] Lvl0 reflects the situation when only the query nodes are processed, which equates to the Bag-of-concepts model from Chapter 4. Lvl0 is therefore the main baseline used to compare the GIN to the Bag-of-concepts. This comparison is provided to understand the effect of the inference mechanism provided as part of the GIN. To further understand how the traversal depth affects retrieval effectiveness, we also examined the retrieval effectiveness for setting of $k = [1, .., 10]$ on a per-query basis. This was to uncover how effective an adaptive method that varies the depth based on the query would be.

Diffusion Factor: The diffusion factor between two concepts is a linear interpolation of two measures: semantic similarity and relationship type (described in Section 6.2.4). Semantic similarity is implemented as the cosine angle between the document vectors of the two concepts; relationship type is based on the SNOMED CT relationship connecting the two concepts. A weight, $[0 - 1]$, was manually assigned to each SNOMED CT relationship type. This was done by the author based on their intuition regarding the strength of association for that relationship. (These weights are provided in Appendix C and more analysis on this weighting scheme is provided in the discussion.) The two measures — semantic similarity and relationship type — were linearly interpolated, with the parameter α controlling the mix (Equation 6.5). To understand the effect of semantic similarity and relationship type, the model was run with different values of α (from 0.0 to 1.0 in 0.1 increments).

[5] Retrieval effectiveness degraded on average for depth values greater than 2 and so we focus on levels 0, 1 and 2.

Weighting schema: To estimate the initial probabilities $P(u|d)$ we used a Dirichlet smoothed language model estimate (Equation 3.5). This estimate has a single parameter μ used to control the effect of document length. The value of μ was set to 22,000 according to the findings of Chapter 4 (the setting that maximised bpref for the Bag-of-concepts model).

6.4.2 Results

Table 6.1 shows the retrieval results for each of the three depth settings. The term baseline from Chapter 4 is also included for comparison. Both bpref and precision @ 10 were lower for the GIN (lvl1 and lvl2) compared against the Bag-of-concepts model (lvl0). To further understand the differences between the three levels, the retrieval effectiveness of individual queries was required. The plots in Figure 6.7 provide this by showing the bpref performance (y-axis) of each of the 81 queries (x-axis). Queries were ordered by decreasing bpref of the lvl0 baseline. The left figure presents the comparison between lvl0 and lvl1 and the right between lvl0 and lvl2. The plots show that both lvl1 and lvl2 made gains on some queries and losses on others. The gains and losses tended to be greater for lvl2 than for lvl1.

Hard queries

Chapter 4 demonstrated that the Bag-of-concepts model generally made greater improvements in hard queries (those that perform poorly on the term baseline). In that chapter, we conjectured that performance improvements on hard queries, but not on easy queries, were a characteristic of semantic search systems in general. To understand if this applied to the Graph Inference model, we provide some analysis of performance on hard queries. In order to determine what constitutes a hard and easy query, we used the results of other teams

Depth (k)	**Bpref**	**Prec@10**
terms	0.3917	0.4975
lvl0	0.4290	0.5123
lvl1	0.4229	0.4481†
lvl2	0.4138	0.4259†

Table 6.1: Graph Inference model retrieval results using TREC MedTrack. $\alpha = 1.0$. The term baseline from Chapter 4 is also included for comparison. † indicates statistical significant differences with lvl0 (paired t-test, $p < 0.05$).

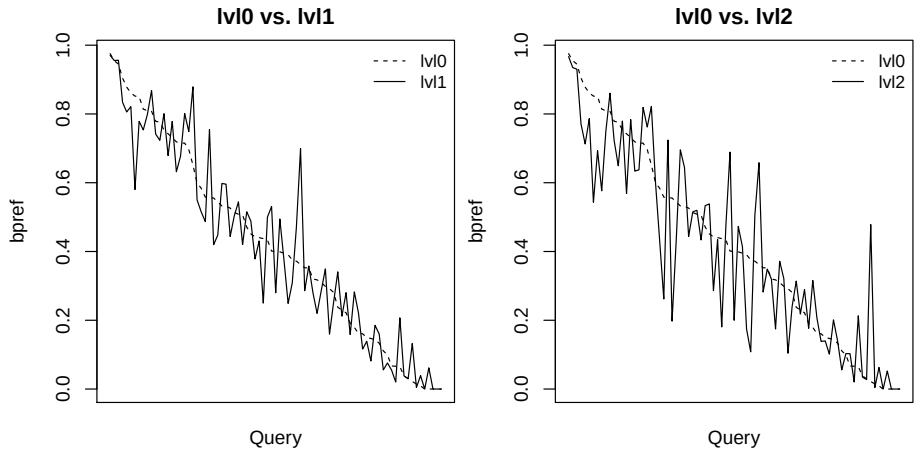


Figure 6.7: Per-query performance comparing the Graph Inference model with Bag-of-concepts baseline (lvl0). Queries are ordered by decreasing bpref of the lvl0 baseline. The left figure presents the comparison between lvl0 and lvl1 and the right between lvl0 and lvl2. The plots show that lvl2 varies more than lvl1 (both greater gains and greater losses). $\alpha = 1.0$.

participating in TREC MedTrack. Specifically, we obtained each team's run and for each query calculated the median bpref for that query. Easy queries represented those with a high median value; hard queries were those with a low median value.

Figure 6.8 shows how the Graph Inference model compared with the TREC median performance. The plot is ordered by decreasing performance according to the TREC median value, representing easy to hard queries. The plot indicates that more gains were observed in those queries that had poor performance in TREC MedTrack. To quantify this, we considered the performance of half the query set with the lowest TREC median bpref value (i.e., out of 81 queries, we selected the 40 queries with lowest TREC median bpref value). The results for the hard query set is shown in Table 6.2. The table confirms that the GIN made greater improvements on hard queries and that these improvements were greater when more of the inference mechanism is applied (i.e., for the GIN at lvl2).

Diffusion Factor Mix

The diffusion mix parameter α controls the mix of semantic similarity and relationship type strength. The effect of retrieval effectiveness for different values of α is shown in Figure 6.9. The best retrieval performance for both bpref and precision @ 10 was observed for $\alpha = 1$. This represents a diffusion factor that

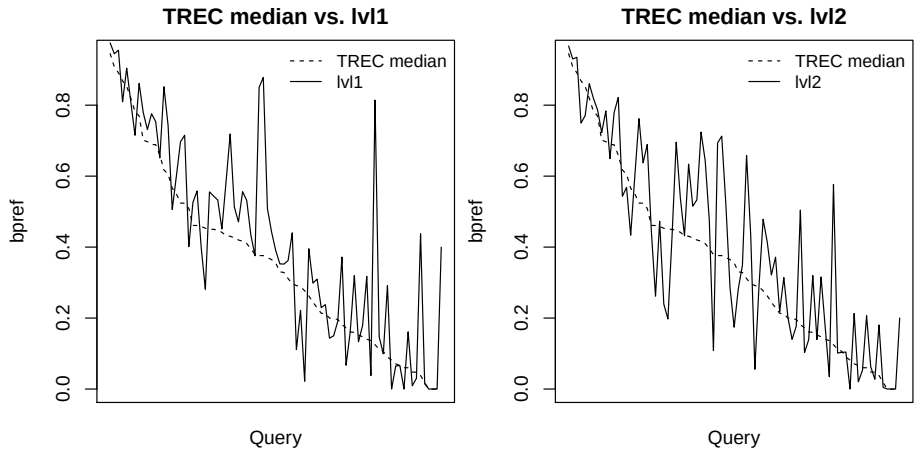


Figure 6.8: Retrieval results for the Graph Inference model compared with the TREC teams. The plot is ordered by decreasing performance according to the TREC median value, representing easy to hard queries. $\alpha = 1.0$.

System	Bpref
TREC Median	0.1514
lvl0	0.1985 (+31%)
lvl1	0.2024† (+34%)
lvl2	0.2072† (+37%)

Table 6.2: Retrieval results for hard queries; GIN compared to the TREC median performance. † indicates statistical significant differences with TREC Median (paired t-test, $p < 0.05$).

made use of only semantic similarity and did not consider relationship type. The relationship was manually assigned by the author and is not likely to be optimal. Further investigation would be needed to determine optimal relationship types.

Per-query Depth Setting

To understand the effect of the depth parameter, retrieval effectiveness using different settings of $k = [1, .., 10]$ were examined on a per-query basis. The heatmap in Figure 6.10 shows the change in bpref compared to the lvl0 baseline for different settings of k. Blue areas indicate that the performance of a query improved for that setting of k when compared to lvl0 ($k = 0$), while red areas indicate that the performance of the query degraded when compared to lvl0. There is considerable variation between different queries. Some queries had a

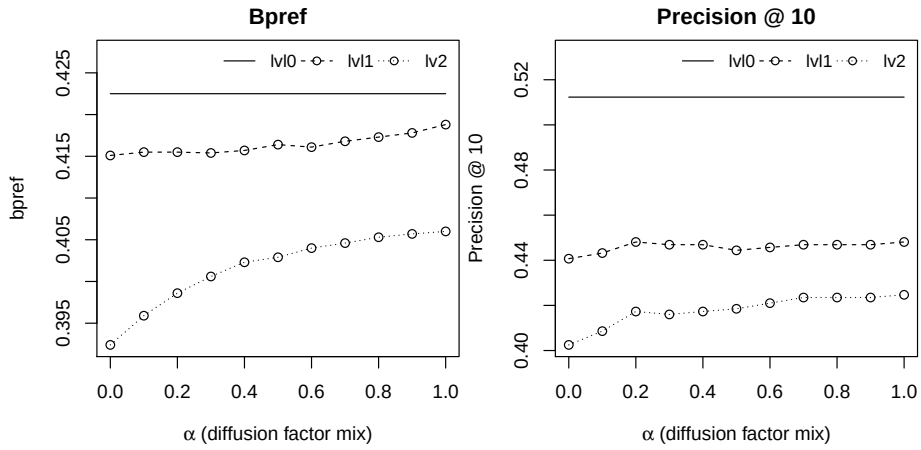


Figure 6.9: Retrieval results for different settings of the diffusion mix parameter α, which controls the mix of semantic similarity and relationship type measures in the diffusion factor. $\alpha = 1$ equates to only semantic similarity.

constant improvement over lvl0 for different depth settings, for example query 108, 140 and 171. Other queries degraded as the depth increased, for example 104, 109 and 161. Some queries improved over lvl0 in the first few levels but then degraded at greater levels, for example 113, 119 and 135. Generally, the best improvements were observed for $k = 1$–3. Finally, the optimal value of k varied considerably based on the query.

6.5 Analysis

This section presents an analysis of a number of queries to understand how the GIN works and under which conditions. The heatmap previously shown in Figure 6.10 was used to group queries according to the performance results that they exhibit at different depth settings. To aid understanding, we provide a graph-based visualisation of the traversal for the query. An example of this visualisation and an explanation of the information provided is shown in Figure 6.11. Each node has a number of statistics in the form $(x, y)\,\#z$, where z is the number of documents that the Information Unit appears in (i.e., the document frequency), y is the portion of z that are relevant documents and x is the portion of y relevant documents that do not contain the query concept (indicated in red).

This query visualisation format is used to explain a number of characteristics of the Graph Inference model.

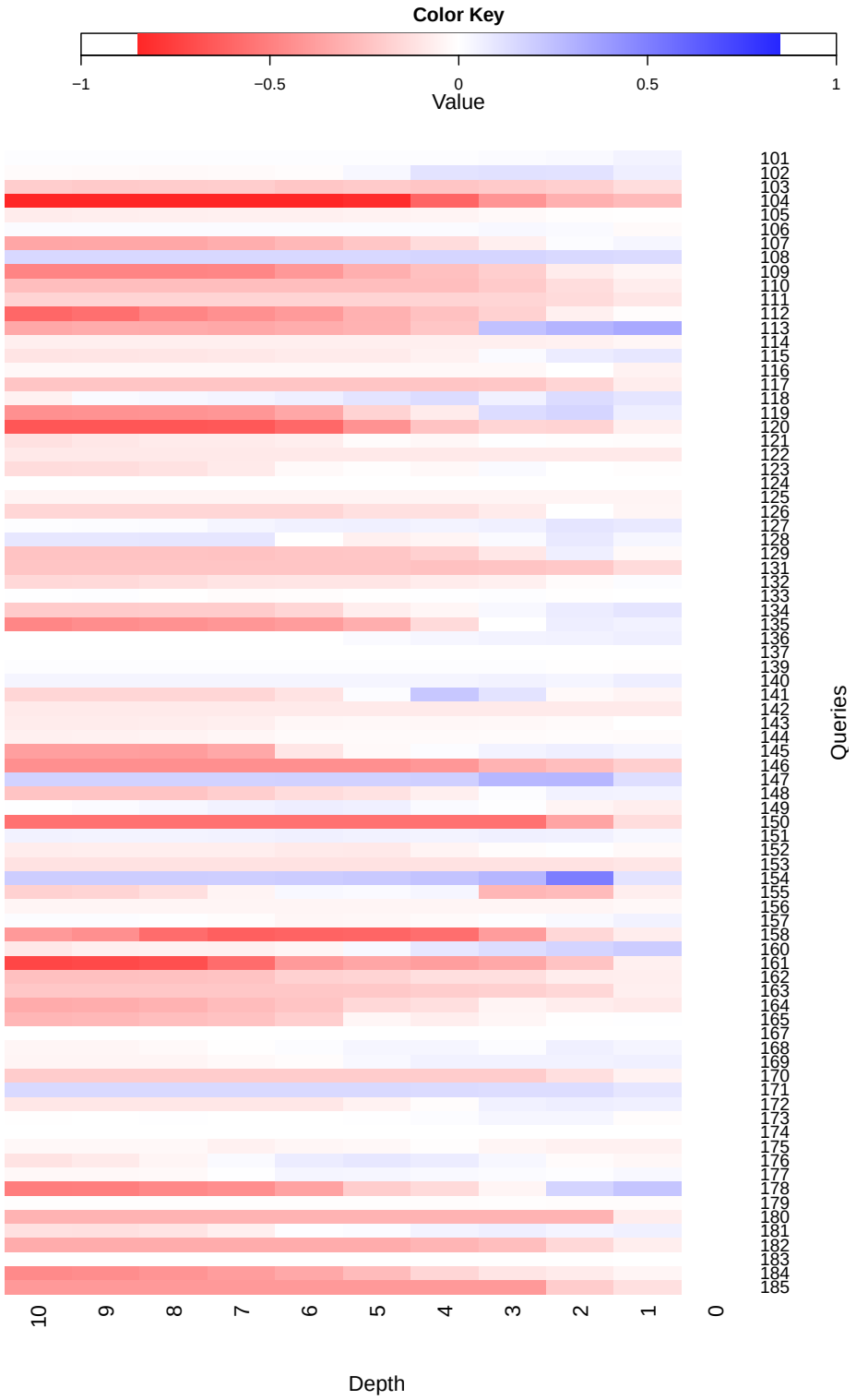


Figure 6.10: Heatmap showing the change in bpref compared to the lvl0 baseline for different depth settings of k.

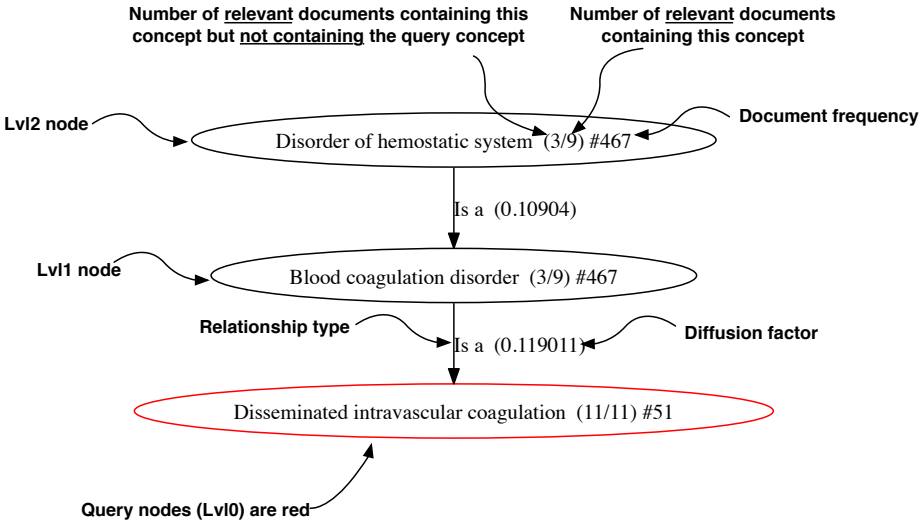


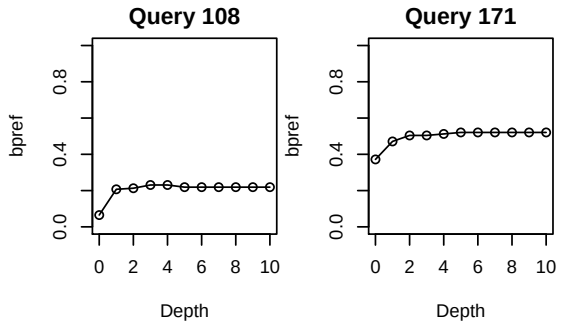
Figure 6.11: Explanation of traversal visualisation graph for a single query.

6.5.1 Consistent Improvements

A number of queries exhibited a consistent improvement over the baseline for different depth settings. Two examples are query 108 and 171, which exhibited the performance shown in Figure 6.12 at different depth settings. (The query keywords are included below the plots.) For query 108, the Graph Inference model returned the same number of relevant documents as the Bag-of-concepts baseline (lvl0) but these were better ranked by the Graph Inference model. Both the query concepts "vascular" and "claudication" had a large number of related concepts in the graph. These concepts often occurred with the query concepts in relevant documents. Thus, the same document was scored multiple times, for both query concepts and related concepts, and therefore these relevant documents were moved higher in the ranking.

Query 171 was an example where SNOMED CT provided valuable domain knowledge to bridge the semantic gap. A partial traversal graph for this query is shown in Figure 6.13. The query specified patients with a specific disease (Thyrotoxicosis). The Graph Inference model was able to infer other relevant documents that contained the cause of Thyrotoxicosis (Hyperthyroidism) and the part of the body affected (Thyroid structure).

These types of queries tended to have valuable related concepts traversed by the Graph Inference model at levels greater than 0 (for example, the Hyperthyroidism concepts in Figure 6.13). Including these valuable concepts always improved performance over the lvl0 baseline. In addition, the diffusion factors were effective at limiting the introduction of noise for greater levels and as a



108: `Patients treated for vascular claudication surgically`
171: `Patients with thyrotoxicosis treated with beta blockers`

Figure 6.12: Queries with consistent improvements (bpref) over the baseline for different depth setting. The query keywords are included below the plots.

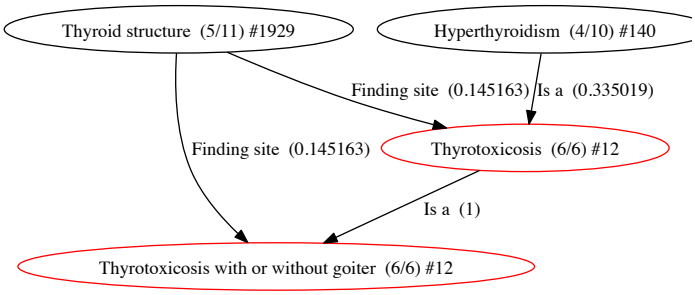
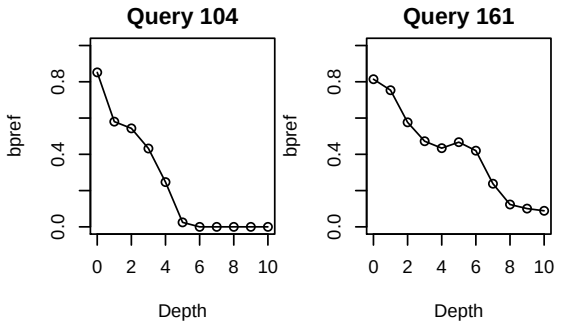


Figure 6.13: Partial traversal graph for query 171.

results no degradation was seen for levels up to 10.

6.5.2 Inference Not Required

A number of queries exhibited decreasing performance at greater depth levels. We focus on the performance of query 104 and 161 shown in Figure 6.14. Query 104 contained only 8 relevant documents. Key to this query was the concept "Robot, device", which was found in all 8 of the relevant documents. All these relevant documents were retrieved by the Bag-of-concepts model, as highlighted by Figure 6.15, which shows the "Robot" portion of the traversal graph. (8/8 of the relevant documents are located at the "Robot" node and no new relevant documents are located at the "Biomedical device" node at level 1.) This constituted an easy query and as such both the Bag-of-concepts and a term baseline achieved good results on this query. No additional valuable information was available to the Graph Inference model at levels greater than 0. This is an example of a query where inference was not required.



104: `Patients diagnosed with localized prostate cancer and treated with robotic surgery`
161: `Patients with adult respiratory distress syndrome`

Figure 6.14: Queries that exhibited decreasing performance at greater depth levels. Typically, such queries were those for which inference was not required.

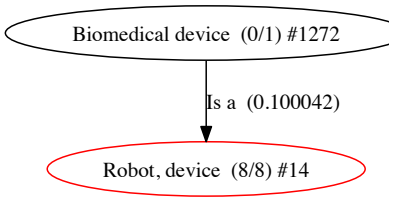


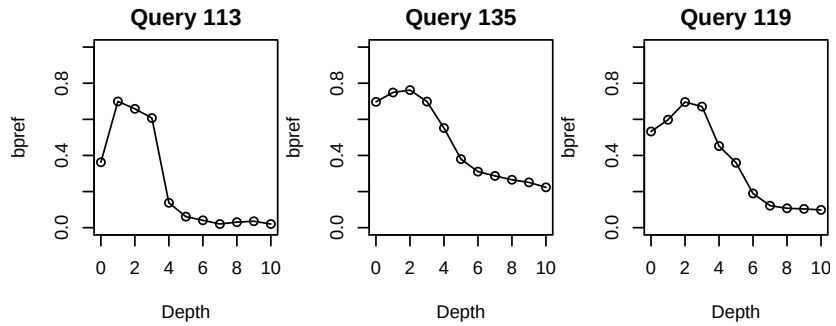
Figure 6.15: Partial traversal graph for query 104.

A similar situation was observed for query 161. This query was previously presented in Chapter 4 as an example of where the Bag-of-concepts model was particularly effective. The query was effectively mapped to the concepts *Adult respiratory distress syndrome* (`67782005`) and *Non-cardiogenic pulmonary edema* (`95437004`). Using these, most relevant documents were ranked effectively for level 0. At greater levels, there were a large number of very general concepts that did not provide any valuable information — again, a query where inference was not needed.

Queries that did not require inference tended to have a small number of relevant documents and an unambiguous query definition: the "Robot" concept (query 104) and the *Adult respiratory distress syndrome* concept (query 161) provided all that was required to retrieve and rank relevant documents.

6.5.3 Reranking

The Graph Inference model was also effective at reranking documents already retrieved for level 0. Queries exhibiting this trend were 113, 119 and 135, shown in Figure 6.16.



113: `Adult patients who received colonoscopies during admission which revealed adenocarcinoma`
119: `Adult patients who presented to the emergency room with with anion gap acidosis secondary to insulin dependent diabetes`
135: `Cancer patients with liver metastasis treated in the hospital who underwent a procedure`

Figure 6.16: Queries with effective reranking using the Graph Inference model.

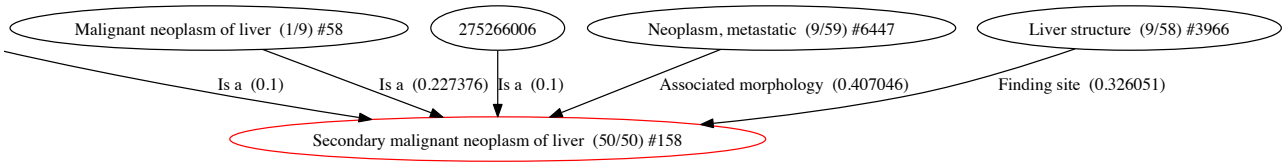


Figure 6.17: Partial traversal graph for query 135.

Query 113 contained only 14 relevant documents, all retrieved at lvl0. For levels 1–3, these documents were reranked based on the presence of other concepts in the document that were related to the query (for example, the presence of general cancer concepts, which were related to the specific "Adenocarcinoma" cancer in the query). Beyond level 3, the concepts were too general and thus performance dropped.

Query 135 is another example of reranking; a portion of the traversal graph for query 135 is shown in Figure 6.17. The query contained a very specific concept (shown in red), while documents were effectively reranked when they contained the more general related concepts from level 1.

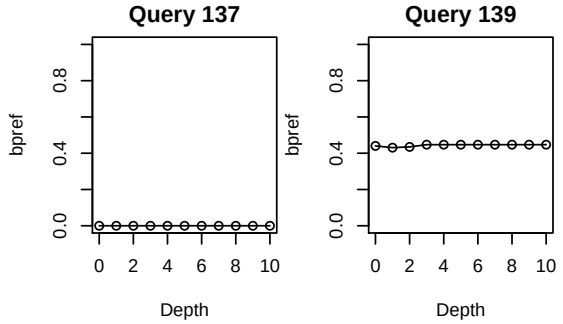
Query 119 (another example of reranking) was a verbose query containing a large number of query concepts. Therefore, the number of nodes visited increased exponentially at greater levels. The consequence was a scoring of a large number of related concepts with only a weak association to the query concepts. This introduced noise at greater depth levels and degraded performance.

The queries that benefitted from reranking tended to have two dependent

aspects to the query, for example query 113 had a procedure ("colonoscopy") and diagnosis ("Adenocarcinoma") and query 119 had a symptom ("anion gap acidosis") and a disease ("insulin dependent diabetes").

6.5.4 Unaffected Queries

Queries 137 and 139 exhibited a near constant performance for different depth settings, as shown in Figure 6.18.



137: `Patients with inflammatory disorders receiving TNF inhibitor treatments`
139: `Patients who presented to the emergency room with an actual or suspected miscarriage`

Figure 6.18: Queries that exhibited constant performance for different depth settings.

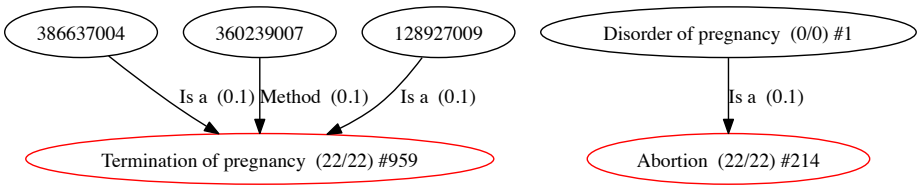


Figure 6.19: Partial traversal graph for query 139.

For query 137, no relevant documents were returned for both the Bag-of-concepts model, Graph Inference model and a term baseline. MetaMap was unable to map the TNF abbreviation to a SNOMED CT concept and for the term baseline TNF was never mentioned in relevant documents. This query highlights the challenge in searching medical data and bridging the semantic gap.

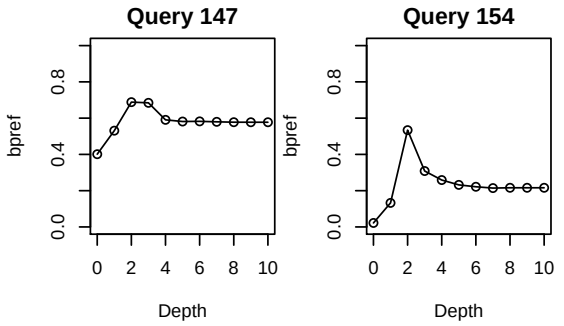
For query 139, there were two key concepts in the query: "Termination of pregnancy" and "Abortion". The portion of the traversal graph with these concepts is shown in Figure 6.19. The graph shows that there were no valuable

related concepts. (The concepts in the graph with numeric labels are concepts related to the query in SNOMED CT but do not ever occur in the document corpus). Very few additional documents were processed at levels greater than 0, therefore the ranking of documents changed little compared to level 0 and consequently performance did not differ.

Unaffected queries were either those that were particularly challenging, such as query 137, which had very poor performance for term, concept and Graph Inference models; or those where no valuable information attached to the query concepts in SNOMED CT.

6.5.5 Inferring New Relevant Documents

Some queries improved by retrieving new relevant documents not retrieved by the lvl0 baseline; these are shown in Figure 6.20.



147: `Patients with left lower quadrant abdominal pain`
154: `Patients with Primary Open Angle Glaucoma`

Figure 6.20: Queries where the Graph Inference model retrieved new relevant document not retrieved by lvl0 baseline.

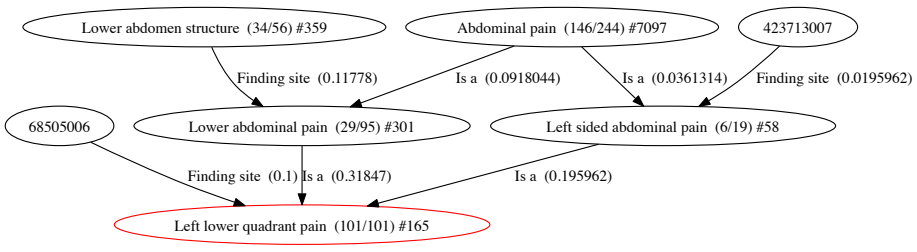


Figure 6.21: Partial traversal graph for query 147.

For query 147, the Bag-of-concepts model retrieved only 101 relevant documents, whereas the Graph Inference model retrieved 136 at level 1, 153 at level

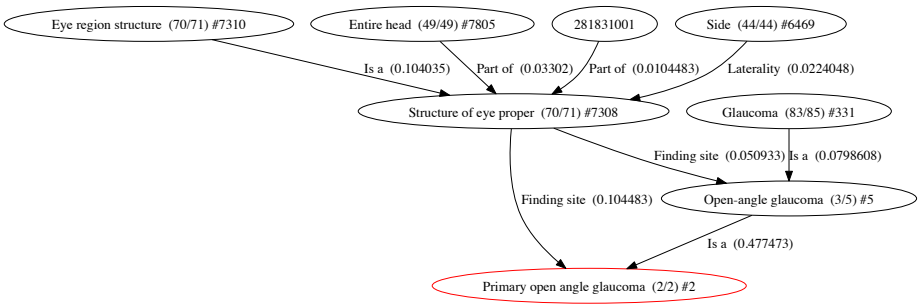


Figure 6.22: Partial traversal graph for query 154.

2 and 189 at level 3. The traversal graph for this query is shown in Figure 6.21. The concepts at level 1 and 2 provided an alternative way of expressing the query concepts.

For query 154, the traversal graph is shown in Figure 6.22. Only 2 relevant document were return at level 0, mainly because the "Primary open angle glaucoma" query concept is too specific. At level 1, the more general concept "Open angle glaucoma" is included, resulting in 3 relevant documents included at this level. Finally at level 2, the "Glaucoma" concept is included and 83 relevant documents are retrieved for this level.

These queries exhibit both granularity and vocabulary mismatch. The related concepts in SNOMED CT, traversed by the Graph Inference model, provided the additional information required to retrieve a large number of relevant documents not retrieved with just the query concepts. For both these queries, the Graph Inference model was more effective than the Bag-of-concepts baseline, no matter the depth setting (although the best performance was found for depth settings 1–3).

6.5.6 Relationships Traversed

The Graph Inference model traversed SNOMED CT relationships, and the relationship type was used to calculate the diffusion factor, so it is important to understand which relationships were being traversed by the model. The traversal graphs from the example queries presented in this section showed a large number of ISA relationships. This was confirmed in general by Figure 6.23, which shows the relationships traversed by the Graph Inference model (lvl1), ordered by frequency of occurrence. ISA relationships dominate those seen by the Graph Inference model. The effect this had on the retrieval performance of the model is considered in the discussion.

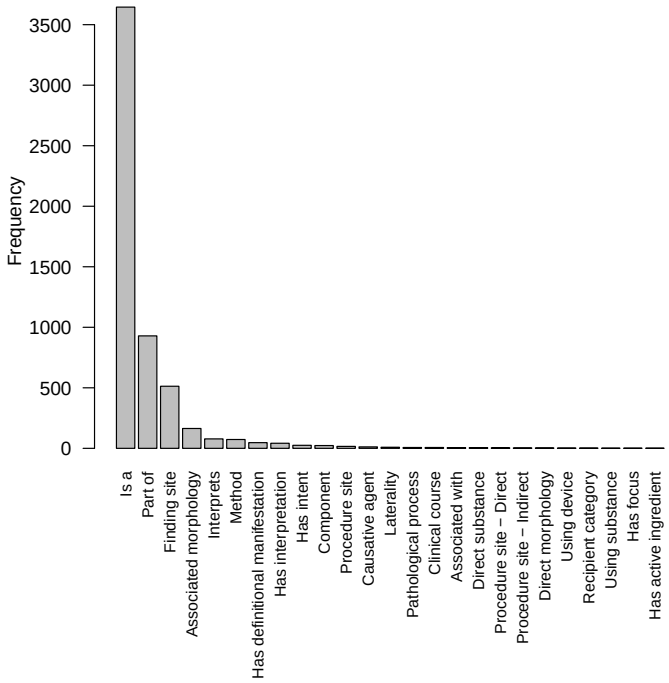


Figure 6.23: Relationships traversed by the Graph Inference model (lvl1), ordered by frequency of occurrence. The ISA relationship is significantly more frequent.

6.6 Discussion

The Graph Inference model specifically addresses a number of semantic gap problems. Regarding vocabulary mismatch, the Graph Inference model utilises the same concept-based representation as the Bag-of-concepts model and thus inherits its benefits for overcoming vocabulary mismatch and to a lesser extent some of the granularity mismatch benefits from the concept expansion process — although the Graph Inference model specifically addresses granularity mismatch by traversing parent-child (i.e. ISA) relationships. The semantic gap problem of Conceptual Implication is where the presence of certain terms in the document infer the query terms, for example where an organism implies the presence of a certain disease. These associations are encoded in SNOMED CT and thus the Graph Inference model specifically addresses Conceptual Implication by traversing these types of relationships. Finally, the semantic gap problem of Inference of Similarity, where the strength of association between two entities are critical, is specifically addressed by the diffusion factor, which assigns a corpus-based measure of similarity to the domain knowledge-based relationship. By integrating domain knowledge and corpus statistics, the Graph Inference model

addresses each of the major semantic gap problems.

A recent model proposed by Herskovic et al. [2012] for the classification of certain medical conditions (e.g., breast cancer), has a number of similarities with the GIN. The model aims to infer the presence of certain concepts (breast cancer was the concept chosen in their evaluation) by analysing free-text patient records. The similarity with the GIN is the graph based representation: nodes represent concepts identified by MetaMap and edges represent either UMLS relationships or are taken from a separate statistically derived relations database; the weights of edges are estimated from a corpus-based measure of similarity, akin to that used in the GIN. While the GIN uses a single graph for the whole corpus, in this model the graph is built for each document. Nodes are assigned an initial weight and a spreading activation process applied to adjust the node weights. The final weights of particular nodes are used to classify the document ("Breast cancer" or "No breast cancer" in their evaluation). The similarities with the GIN are: the goal of performing inference from implicit evidence; and the graph-based representation, combining structured domain knowledge and corpus statistics. The task to which the two model are applied differs (retrieval vs. classification). As such, the dynamics of the model — the retrieval mechanism of the GIN and spreading activation of Herskovic et al. [2012] — sets the two models apart. The model of Herskovic et al. [2012] aims to identify the strength of a single, pre-determined concept within a document; spreading activation is used to estimate this and the documents are treated independently of each other. In contrast, the GIN combines evidence from many concepts in the graph, including their occurrence within certain documents, to produce a ranked list of documents. This is used to determine a ranking of documents given some input set of concepts representing a query.

6.6.1 Understanding when Inference Works

This section characterises when inference using the Graph Inference model works. This is important for both understanding the model itself and the broader theme of search as inference. As part of this analysis, we consider the two different components of the Graph Inference model: the *representation*, which uses a graph constructed from domain knowledge, and the *traversal*, which utilises the graph representation for retrieval.

The Representation

A number of issues arose from the underlying representation, that is, SNOMED CT. The analysis of the relationship types traversed by the Graph

Inference model showed that the ISA relationship far outnumber any other (as shown in Figure 6.23). The ISA relationship captures parent-child associations between concepts. These relationships are valuable for overcoming only granularity mismatch (as shown by Zuccon et al. [2012]) but do not help address the other semantic gap problems. For these, different types of relationships are required, such as *treatment* → *disease* and *organism* → *disease* relationships. The former relationships are not modelled in SNOMED CT as they are not definitional (because opinions may differ on the best treatment for a disease and may change over time). For the latter, although it is valid to model organism → disease relationships in SNOMED CT, the coverage is lacking [Spackman, 2008]. In addition, coverage may also vary considerably for ISA relationships. Some concepts may inherit from very specific parent concepts (for example, "Right ventricle" ISA "Cardiac ventricle"), while others may inherit from very general parent concepts (for example, "Vertebral Unit" ISA "Body Structure"). This affects the Graph Inference model as some ISA relationships may provide valuable information, while others are too general and add noise. In fact, this was the finding for a number of queries, where performance degraded when very general concepts were traversed. (For these cases, work by Boudin et al. [2012], which attempts to identify the granularity of concepts in a medical query, might be applied.) More generally, poor performance in the Graph Inference model was found in queries where there was little valuable information in the representation for levels greater than 0. These issues highlight a limiting factor for the Graph Inference model as the underlying representation, rather than the traversal mechanism that acts on this representation.

Also related to the underlying representation, the wider issue of using an ontology designed for knowledge representation but applied to information retrieval is worth discussing. The purpose of SNOMED CT (or many other such domain knowledge resources) is to represent the concepts belonging to that domain; the information regarding these concepts is *definitional*. The conclusions possible using this definitional information are valid from a conceptual point of view; however, these conclusions may not be valuable from an information retrieval perspective. For example, it is logically true that "Vertebral Unit" is indeed a "Body Structure" but such information is unlikely to be of any value when encountering "Vertebral Unit" in a retrieval scenario. Two types of inference are at play here: *definitional inference*, used in knowledge representation to understand the concepts belonging to that domain, and *retrieval inference*, used to determine whether some information (typically, a document) is relevant given some context-specific situation (typically, a query representing an information need). Differing requirements between these two types of inference mean that many relationships that are definitional are not useful for retrieval. The

strain between definitional and retrieval inference has been highlighted by other researchers as one of challenges in utilising conceptual representations [Frixione and Lieto, 2012]. New conceptual representations that can account for these two differing requirements are under investigation [Frixione and Lieto, 2012].

The Traversal

The traversal component of the GIN is the inference mechanism that acts on the representation. This includes the diffusion factor, the initial probabilities assigned to each node and the general retrieval function that combines the two.

In our instantiation, the diffusion factor between two concepts is determined by semantic similarity and relationship type; the diffusion mix parameter α interpolates these two estimates. The best retrieval results were observed for a diffusion factor that made use of only semantic similarity and did not consider relationship type (as shown in Figure 6.9). One interpretation might be that the relationship type did not provide any valuable information in determining the diffusion factor. However, the analysis from the previous section has already highlighted that most relationships are ISA relationships. Therefore, relationship type might not be discriminating enough and more training data is required to be able to learn a good weighting to assign to each relationship type.[6] One solution might be to include additional implicit non-ISA relationships. For example, consider the SNOMED CT graph shown in Figure 6.24.

[6]Recall that we considered only a fixed, manually assigned weighting for relationship types. This was based on the author's intuition and the assigned weights were most likely not optimal.

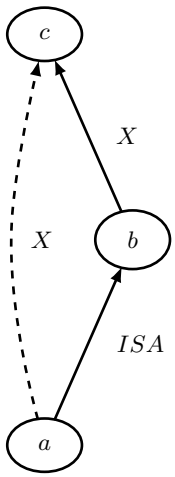


Figure 6.24: Example of deriving implicit relationships in SNOMED CT. The solid edges indicate explicit relationships and the dashed edge indicates an implicit relationship.

The concept a is a child of b and b is related to c via some relationship X. These are the relationships recorded in SNOMED CT and are indicated with solid edges. However, there is an implicit X relationship (dashed edge) between a and c that is not recorded in SNOMED CT but is typically computed by formal reasoning engines that use such ontologies [Lawley and Bousquet, 2010]. These implicit relationships could be derived to provide additional non-ISA relationships in order to improve the graph representation, although the risk is that additional noise may be introduced as a result. The investigation of this is left to future work.

In contrast to relationship type, the semantic similarity measure was effective. A manual review of the diffusion factor values, as determined by the semantic similarity measure, showed reasonable values. (This was also seen in the traversal graphs for the example queries in Section 6.5.) Semantic similarity between concepts was determined as the cosine angles between the two document vectors. Although this method has been shown to be effective [Koopman et al., 2012b], more sophisticated measures are available, for example the Tensor Encoding model [Symonds et al., 2012], and may improve the similarity measure and, hence, the retrieval results.

The depth parameter k controls how many edges are traversed from the query node and reflects how much additional information the model draws on to score documents. For the main empirical evaluation, three different depth settings were evaluated: 0, 1 and 2. In addition, the analysis considered how retrieval effectiveness was affected for depth settings 0–10. Generally, the best performance was achieved for depth 1–4. (See the heatmap of Figure 6.10.) Beyond this, the related concepts were too peripheral to the query concepts and often introduced noise. For some cases, this was mitigated by the diffusion factor, which decreases exponentially the further the concept is from the query concept.

The analysis of different depth settings also revealed a number of insights about how the GIN was working empirically:

- Queries that had consistent improvements over the baseline for different depth settings tended to have valuable related concepts at levels greater than 0. Including these valuable concepts always improved performance and the diffusion factor was effective at limiting the introduction of noise.

- Some queries did not require inference. These tended to be easy queries with a small number of relevant documents and an unambiguous query. Here, the Bag-of-concepts baseline was already performing well. If these easy queries could be identified, then the Bag-of-concepts model, or GIN at lvl0, would be preferred over the GIN for these queries. Previous work

on query performance prediction [Hauff et al., 2008; Boudin et al., 2012] could be investigated for this situation.

- The GIN was effective at reranking those documents already retrieved by the Bag-of-concepts baseline. This was observed in queries that contained two dependent aspects, for example a procedure and a diagnosis. These also tended to be more verbose queries where the key query concepts (for example the procedure and the diagnosis concepts) contained more related concepts than the less important query concepts. The larger number of related concepts attached to key query concepts meant that documents related to these key concepts received greater scores and were ranked higher.

- The GIN was effective at retrieving new relevant documents not retrieved by the baseline; this is where the inference mechanism was particularly effective. In these situations, there was valuable domain knowledge available to the GIN recorded in SNOMED CT. These queries also tended to suffer from multiple semantic gap problems.

- Some queries had very poor performance on term, Bag-of-concept or GIN, highlighting the challenge of search in the medical domain and that additional work is still required to bridge the semantic gap.

The above insights about the working of the Graph Inference model also highlight that inference is required for some queries but not for others (or varying degrees are required). Practically, this equates to adaptively controlling the depth of traversal on a per-query basis. To understand the potential gains that this might provide, we selected the bpref value for the best depth setting for each query and averaged this across all queries; this represents an oracle upper bound for an adaptive depth method. The results are shown in Table 6.3, along with the fixed depth approaches for comparison. As suspected, the adaptive method demonstrates the best performance. More important though is what characteristics or conditions might indicate the optimal depth setting. We have already commented that hard queries required inference and that the Graph Inference model was more effective for these. In contrast, easy queries do not require inference. Therefore, a query performance predictor might inform whether it is worth traversing beyond level 0.

Inference can be risky. For hard queries, there is nothing to lose and adding domain knowledge can bring substantial benefits. For easy queries, adding domain knowledge is not required and can introduce noise. The analysis provided here points to an adaptive approach, where inference is applied on a per-query

Depth Approach	Bpref	Prec@10
Fixed — lvl0	0.4290	0.5123
Fixed — lvl1	0.4229	0.4481
Fixed — lvl2	0.4138	0.4259
Adaptive Depth, 0–10 (Oracle)	0.4731 (+10%)†	0.5741 (+12%)†

Table 6.3: Graph Inference model retrieval results using the best depth setting per-query. This represents an oracle upper bound for an adaptive depth method. The percentages show the improvements of this method against the lvl0 baseline. † indicates statistical significant differences with fixed approaches (paired t-test, $p < 0.05$).

basis, as more appropriate. Future work aimed at the development of an adaptive depth method is considered in Section 8.6.1.

6.6.2 Bias in the Evaluation

Empirically, the Graph Inference model did not demonstrate statistically significant improvements over the Bag-of-concepts baseline (lvl0), but this does not represent the whole story. Manual analysis of the results revealed that the evaluation was underestimating the performance of the GIN. Specifically, a large number of unjudged documents — those never assessed by TREC judges — were retrieved by the GIN. Considering the top 20 documents returned for a query, the number of unjudged documents was 12% for a term baseline, 15% for the Bag-of-concepts baseline (lvl0), 27% for the Graph Inference model at lvl1 and 36% for lvl2. Such a large number of unjudged documents can significantly affect the evaluation measures. For precision @ 10, an unjudged document is considered not relevant; thus greater numbers of unjudged documents will lower precision @ 10. Our results showed that precision @ 10 was significantly lower for the GIN than the Bag-of-concept baseline. In contrast, the bpref measure ignores unjudged documents; this was reflected in our results where bpref differed only slightly between models. Large numbers of unjudged documents would have a significant impact on retrieval measures and could mean the effectiveness of the GIN is underestimated.

The number of unjudged documents retrieved might be an artefact of the semantic search approach we advocate. The motivation for a semantic search approach is that it may retrieve additional relevant documents that are not retrieved by keyword-based approaches. We conjecture that the Graph Inference model does retrieve these new relevant documents but these were never judged by TREC assessors. The set of documents provided to assessors is taken from

the pool of documents obtained from systems participating in TREC — systems that were largely keyword-based [Voorhees and Tong, 2011; Voorhees and Hersh, 2012]. Thus, documents that are not retrieved by keyword-based systems (for example, those that do not contain the query terms) would never make it into the pool and would never be assessed for relevance. This situation highlights the broader issue of how to evaluate semantic search systems and the bias to keyword-based systems of past TREC evaluation campaigns in the medical domain. More specific to the evaluation of the Graph Inference model, the large number of unjudged documents made the estimation of retrieval effectiveness unreliable. To address this, we obtained additional relevance assessments from medical professionals to understand to what extent the Graph Inference model was retrieving new relevant documents; this is the focus of the next chapter.

6.7 Summary

The Graph Inference model integrates external domain knowledge within a corpus of documents. It does this using a graph-based representation: nodes represent Information Units in the corpus but their definition comes from the domain knowledge resource; edges represent the associations between Information Units, also derived from the domain knowledge resource, but the diffusion factor is responsible for incorporating both domain knowledge and corpus statistics for weighting associations. The inference mechanism is realised as the traversal over the graph structure and it is this inference mechanism that is designed to bridge the semantic gap. Theoretically, the traversal mechanism is akin to the process of altering the document from the Logical Uncertainty Principle within logic-based IR; the diffusion factor models the uncertainty of this process.

The Graph Inference model is defined generally (Section 6.2), with implementation decisions left to the particular application. The underlying representation, implementation of the diffusion factor, weight assigned to each node and way concepts are combined in the retrieval function (e.g., multiplied, summed, etc.) are all independent of the model. This was done intentionally to make the model more generally applicable.

Although the model is defined generally, we present an efficient implementation in Section 6.3. The indexing component uses a standard inverted file index to create the graph, while the retrieval component performs a depth-first-search, originating from the query nodes and scoring documents attached to each node visited.

The Graph Inference model addresses the semantic gap in a number of ways. Vocabulary mismatch is addressed by the concept-based representation; granu-

larity mismatch by traversal over ISA relationships; Conceptual Implication by traversal over other relationships; and Inferences of Similarity by using the diffusion factor, which assigns a corpus-based measure of similarity to the domain knowledge-based relationship.

The empirical evaluation highlighted how the underlying representation (that is, SNOMED CT) affected the model. ISA relationships occurred far more frequently. Although traversing ISA relationships alleviated granularity mismatch, other relationships are required to bridge the semantic gap. Specifically, *treatment* $\rightarrow$ *disease* and *organism* $\rightarrow$ *disease* relationships are required. In SNOMED CT, the former is not modelled, while coverage in the latter is poor. More generally, poor performance in the Graph Inference model was found for queries where there was little valuable information in the representation for levels greater than 0. These issues highlight the underlying representation as a limiting factor for the Graph Inference model, rather than the traversal mechanism that acts upon this representation.

The issues with representation also raise the broader topic of the differing requirements of definitional inference versus retrieval inference. The former is concerned with knowledge representation to understand the concepts belonging to that domain, while the latter is used to determine whether some information (typically, a document) is relevant given some context-specific situation (typically, a query). Frixione and Lieto [2012] also raised this issue, describing the strain between compositionality, which is definitional, on the one hand and the need to represent other information important for retrieval, on the other.

Detailed analysis about how the Graph Inference model was working revealed a number of insights. First, that hard queries require inference and easy queries do not. Hard queries tended to be verbose and often contained multiple dependent aspects to the query (for example, a procedure and a diagnosis concept). Reranking using the Graph Inference model was effective here. Easy queries tended to have a small number of relevant documents and an unambiguous query concept. For these queries, inference was not required and the Bag-of-concepts model was most effective. Overall, when valuable domain knowledge was provided by SNOMED CT, then the Graph Inference model was effective — either by returning new relevant documents or by effective reranking. This again highlights the dependence on the underlying domain knowledge.

The limitations of the Graph Inference model can be addressed in a number of ways. The underlying representation can be improved, either by including other domain knowledge resources or by improving the current one (for example, by taking into account implicit relationships in SNOMED CT). The traversal can be improved by selecting the depth in an adaptive per-query manner. In the first instance, this method could use query performance predictors to identify

hard queries requiring inference from easy queries that do not.

Empirically, the Graph Inference model did not show statistically significant improvements. However, the use of the TREC MedTrack test collection might be underestimating the performance of the Graph Inference model. Specifically, the model retrieved a large number of unjudged documents that, we conjecture, may be relevant but were never included in the pool to TREC assessors. Further analysis of this aspect, and the collection of additional relevance assessments, is the focus of the next chapter.

Chapter 7

Relevance Assessment and Evaluating Semantic Search

The most exciting phrase to hear in science, the one that heralds new discoveries, is not "Eureka!" but "That's funny..."

— Isaac Asimov*

This chapter focuses on evaluating semantic search systems. From the evaluation of the Graph Inference model of the previous chapter, we observed that the model was retrieving a large number of unjudged documents — those never judged by TREC assessors. In this chapter, we analyse the effect that these unjudged documents have on the *underestimation* of retrieval effectiveness. This motivated the need to obtain additional relevance judgements. To this end, graduate medical students were recruited to judge those previously unjudged documents. Equipped with additional relevance judgements, we re-evaluate the Graph Inference model and the Bag-of-concepts baseline. The results show that effectiveness improves for both models but greater improvements are observed for the GIN. Finally, we present an alternative to the TREC-style evaluation, aimed at evaluating semantic search systems. This novel evaluation method uses manually coded medical records to generate queries and relevance judgements, thus mitigating the need to recruit human assessors.

*Isaac Asimov (1920 – 1992) was an American author and professor of biochemistry at Boston University and best known for his works of science fiction and for his popular science books.

7.1 Motivation

Systems contributing to the pool for TREC MedTrack were made up of largely keyword-based systems [Voorhees and Tong, 2011; Voorhees and Hersh, 2012].[1] The top ranked documents from these systems were those where the query terms were prominent in the document; therefore, these were the documents pooled for assessment.[2] Semantic search is aimed at making the retrieval model less dependent on the individual terms, retrieving relevant documents where the query terms may not be prominent but may contain other relevant terms. These relevant documents were unlikely to have been retrieved at top rank positions by keyword-based systems and therefore would not have been included in the pool of documents assessed by human judges. We conjecture that the effectiveness of the GIN was underestimated when evaluated using the TREC MedTrack test collection and that the same problem would affect other semantic search systems.

When examining the documents returned by the GIN we observed many were never judged by TREC assessors. These unjudged documents negatively affected the retrieval effectiveness as most evaluation measures assume an unjudged document as not relevant. To understand better the effect of unjudged documents, consider the comparison of different retrieval systems in Table 7.1. For simplicity, we focus on the top 20 documents returned for each query by each model and therefore precision @ 20 as the evaluation measure.[3] The term baseline returned a total of 210 unjudged documents in the top 20 results across the 85 queries. In contrast, the concept baseline returned a total of 257 unjudged documents, an increase of 22%. However, the precision @ 20 for the concept baseline was actually 3.4% higher than the term baseline. This shows that the concept baseline was actually retrieving more relevant documents; specifically, it was returning more *judged* relevant documents than *judged* not relevant documents when compared to the term baseline. However, it was also returning more *unjudged* documents than the term baseline. The concept baseline (lvl0) can be considered a shallow semantic search system that differs from the term baseline but not radically so. However, the GIN is fundamentally different and is designed to rely even less on term occurrences, making it radically different from the term baseline. This is reflected in the fact that the GIN returned far

[1] Note, the GIN was developed after TREC 2012 and as such never contributed to the pool.

[2] The document pool for a single query in TREC MedTrack was constructed by selecting the following documents from each team: all 10 documents from rank positions 1 to 10, a random 10 documents from rank position 10 to 100 and a random 10 documents from rank positions 100 to 1000. Therefore, a maximum of 30 documents per team per query could be added to the pool.

[3] Precision @ 20 is chosen here because this is the evaluation measure used later in this chapter for reporting the results after additional relevance assessments were obtained.

Model	Unjudged documents in top 20 results	P@20
Terms	210 (2.5 docs / query)	0.4244
Bag-of-concepts (lvl0)	257 (3.0 docs / query)	0.4389
Graph-model (lvl1)	468 (5.5 docs / query)	0.4086
Graph-model (lvl2)	616 (7.2 docs / query)	0.3630

Table 7.1: Number of unjudged documents in top 20 rank position and precision @ 20 for different retrieval models.

more unjudged documents in the top 20 results across the 85 queries. As a consequence, it also had a lower precision @ 20.

Additional insights into the effect of unjudged documents can be gained from looking at a specific example query. Consider TREC MedTrack query 119: `Adult patients who presented to the emergency room with anion gap acidosis secondary to insulin dependent diabetes`. Table 7.2 shows the evaluation results for this query; included are the bpref and precision @ 20 results and the number of judged documents (total number judged and number judged relevant) returned in the top 20 results for each model. For lvl0, all the documents returned in the top 20 rank positions were judged — 12 were relevant and 8 not relevant. In contrast, lvl1 had only 12 out of 20 documents judged — 9 relevant and 3 not relevant. For lvl2, 9 out of 20 documents were judged — 8 relevant and 1 not relevant. The table also reports the percentage of judged documents that were relevant (i.e, $\frac{|\text{relevant}|}{|\text{judged}|}$). These results show that the GIN was returning fewer judged documents but that the judged documents it did return tended to be relevant (as shown by the percentage of judged documents that were relevant).

Model	#Judged Docs in Top 20 results			Bpref	P@20
	Total	#Relevant	% of Judged, Relevant		
lvl0	20	12	60%	0.5326	0.6000
lvl1	12	9	75%	0.5978	0.5500
lvl2	9	8	89%	0.6957	0.5000

Table 7.2: The effect of unjudged documents on TREC MedTrack query 119. The GIN (lvl1 and lvl2) returns significantly fewer judged documents but those that it does return are largely relevant.

7.2 Quantifying the Effect of Unjudged Documents

The previous section provided some initial insights into the effect of unjudged documents. In this section, we analyse the effect of unjudged documents on precision @ 20 across all 85 TREC MedTrack queries. The plot in Figure 7.1 shows, for each query (x-axis), the number of unjudged documents (left y-axis) in the top 20 results — for both Bag-of-concepts (lvl0) and the GIN (lvl1). The plot also shows the overlapping documents between lvl0 and lvl1, i.e. the number of unjudged documents that appear in both lv0 and lvl1 top 20 results. Finally, the plot shows the change in precision @ 20 (red line, right y-axis) between lvl0 and lvl1 (i.e., lvl1 minus lvl0). The queries on the x-axis are ordered according to the number of unjudged documents retrieved by lvl1.

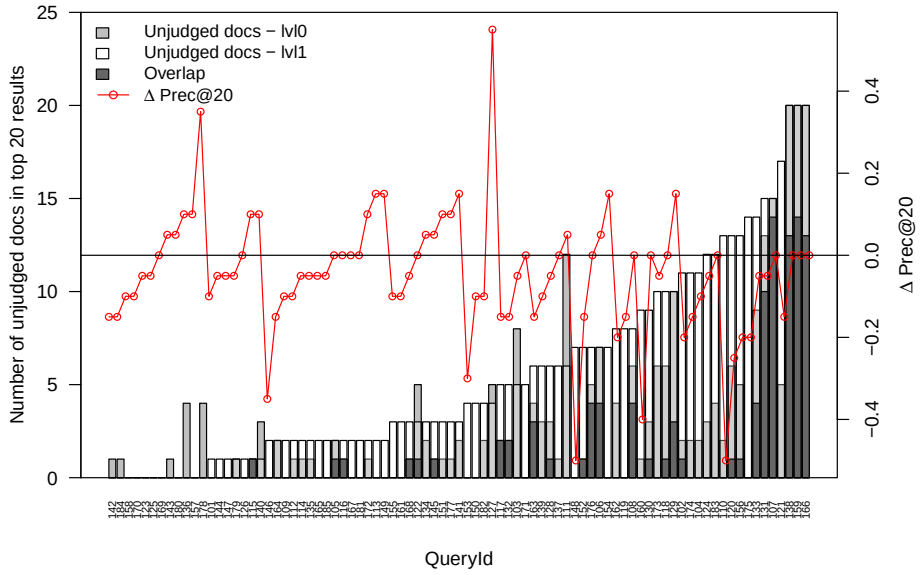


Figure 7.1: The number of unjudged documents in top 20 results (left y-axis) for each query (x-axis), and the corresponding change in precision @ 20 (right x-axis). Queries ordered according to the number of unjudged documents retrieved by lvl1.

The figure provides a number of insights. Clearly, there were far more unjudged documents for lvl1 than lvl0. Therefore, the evaluation was more likely to have underestimated the performance for lvl1. This was highlighted previously and was the initial motivation for obtaining more relevance assessments. In addition, the overlap between the unjudged documents returned by lvl0 and lvl1 was relatively small. This shows that the rankings were quite different. The GIN relies on different information and returns a different set of documents.

Finally, the righthand side of the plot shows a number of queries where lvl1 was returning a significant number of unjudged documents but without a significant degradation in precision @ 20. These queries exhibit the same characteristics as the example query 119 presented in the previous section: many more unjudged documents without a significant loss in precision. We conjecture that these were the queries where the GIN is returning new relevant documents never judged by the TREC assessors. The question, therefore, is: what portion of the unjudged documents returned may have actually been relevant but were never seen by TREC assessors? It is this question that motivates the need for additional relevance assessments.

7.2.1 Simulated Precision

If the GIN was returning many relevant but unjudged documents, then judging these documents would lead to improvements in the measure of retrieval effectiveness. To understand better the potential gains, we provide an analysis in the form of a "simulated" precision measure if all the unjudged documents were assessed. This is done both to understand the potential gains and to contrast how accurate a simulated measure might be compared to the actual measure once complete judgements were obtained through a new assessment exercise. The simulated precision is derived as follows:

- For each query q_i a set of unjudged documents U_i was returned by our system.
- Some portion of U_i may be relevant. The probability of being relevant is $P(r|U_i)$.
- We could assume a uniform probability of relevance, for example, by considering the ratio of the number of judged relevant to total number of judged documents in the TREC qrels (i.e., uniform across all TREC queries). Instead, a better estimate could use other indicators of relevance that are more informative of the potential performance for a given query. One indicator would be the portion of judged relevant to total judged documents in the top 20 results for a given query, i.e., $P(r|U_i) = \frac{|\text{judged\_relevant}|}{|\text{judged}|}$. The intuition here is that if a query contained only relevant and unjudged documents, then the unjudged documents were more likely to be relevant than a query that contained only not relevant and unjudged documents.
- Using the above method of estimating $P(r|U_i)$, we can assign a certain number of documents in U_i as relevant according to $P(r|U_i)$. (This is done

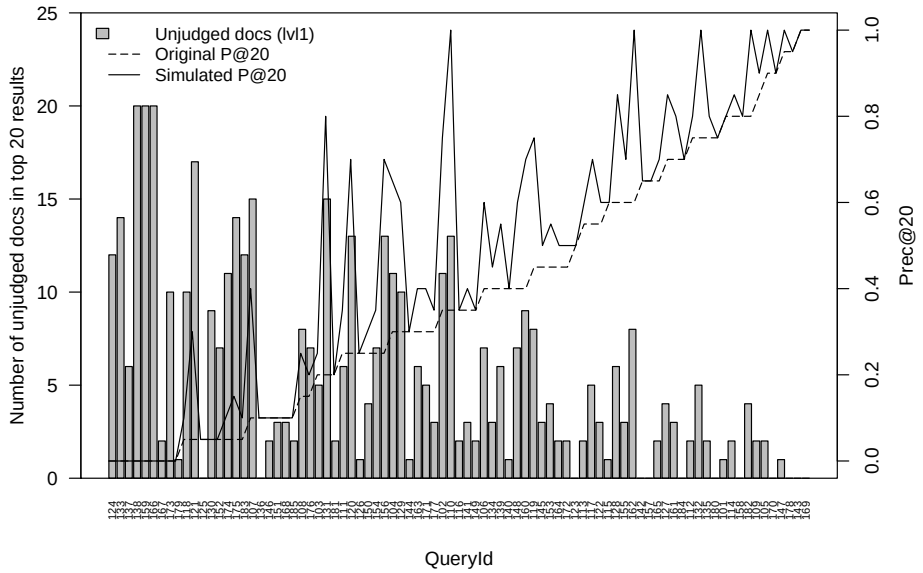


Figure 7.2: Simulated precision for each query, if a portion of unjudged documents are judged relevant.

for each query.) Precision @ 20 is then recalculated using the additional relevant documents, providing a simulated precision measure.

The results of the simulated precision are provided in Figure 7.2. For each query, we show the number of unjudged documents returned by the GIN in the top 20 results. The dashed line is the original precision @ 20 for lvl1 using TREC qrels.[4] The solid line is the simulated precision @ 20. The plot is ordered by increasing original precision @ 20. We observe that the worst performing queries tend to have a higher number of unjudged documents; unsurprising, as these are treated as not relevant. However, there are a number of queries that contain nearly only relevant and unjudged documents — few or no irrelevant documents. These are the queries with the largest gains in simulated precision @ 20 (e.g., the peaks at query 131, 102, 110). Overall, we see increases in simulated precision @ 20 across a large portion of queries.

Although artificially created, these results aim to provide an indication of the improvement we may find from new relevance assessments. These simulated results are revisited after obtaining new assessments to determine how accurate they have been. Further research could investigate other (more reliable) indicators of $P(r|U)$.

[4]We use the term 'original' to denote the evaluation results using the TREC MedTrack. This is used later to contrast against the evaluation results obtained with addition relevance assessments.

There has been previous research into evaluating systems with limited relevance assessments. This includes the development of inferred measures [Yilmaz et al., 2008], which are proposed as a means of obtaining more accurate estimates of retrieval effectiveness when judging a relatively small number of documents (this being the case for TREC MedTrack). These measures are used as part of an approach aimed at evaluating many more queries but with fewer assessed documents per query (as opposed to the more common practice of assessing a small number of queries, each judged to near-completeness) [Carterette et al., 2008]. The reason such methods are not used as part of our evaluation is that the problem is not just that a limited number of documents from each system can be judged. Instead, the problem is that no semantic search systems contributed any documents to the pool. Irrespective of how the pool was formed, if some semantic search system never contributed documents, then potentially relevant documents retrieved by such a system would never be assessed (unless those documents were returned by one of the other keyword-based systems contributing to the pool). The problem is not the limited number of relevance assessments but the type of documents that were available for assessment in the first place.

7.3 Additional Relevance Assessments

This section describes the acquisition of additional relevance assessments by medical professionals. These assessors were recruited to determine the relevance of unjudged documents.

7.3.1 User Experimental Design

Four medical graduates were recruited from the University of Queensland's Bachelor of Medicine Bachelor of Surgery (MBBS) program.[5] All four subjects were completing their fourth and final year of the MBBS program. As part of their training, they had complete rotations in a number of different medical specialities and were familiar with the content of clinical reports. As such, their expertise was equivalent to medical graduates recruited as assessors for TREC MedTrack [Voorhees and Tong, 2011; Voorhees and Hersh, 2012].

[5]University of Queensland MBBS program: http://www.som.uq.edu.au/future-students/bachelor-of-medicine-bachelor-of-surgery-(mbbs).aspx (last accessed 23rd November, 2013).

Pooling Documents for Assessment

The existing corpus and queries from TREC MedTrack were used. The organisers of TREC MedTrack excluded 4 of the 85 queries as these did not have sufficient relevant documents; however, we intentionally included these 4 queries to determine if additional relevant document might be found using the GIN.[6] For each query q_i we proposed to judge a selection of documents U_i that had not previously been judged in TREC MedTrack. These documents were selected by pooling the unjudged documents from the top 20 results of three retrieval runs:

1. The baseline Bag-of-concepts model (lvl0).
2. The Graph Inference model — lvl1;
3. The Graph Inference model — lvl2;

The Bag-of-concepts baseline was included to ensure fairness by including all unjudged documents, not just those returned by the GIN. Using the above pool, each query q_i had between 1 and 60 (20 from each run above) unjudged documents U_i assigned to it for assessment. The average number of unjudged documents in our pool for each query was 11. Using this method, complete judgements were obtained for the top 20 documents returned by each of the three systems listed above: i.e., no unjudged documents would appear in the top 20 ranked position; precision @ 20 would therefore be an accurate evaluation measure.

Control Queries

To familiarise the assessors with our system, we selected two control queries, denoted q_{c1} and q_{c2}. In contrast to all the other queries, which contained only unjudged documents, the control queries comprised documents already judged in TREC MedTrack. For each control query, we selected 4 documents judged relevant in TREC, 4 documents judged not relevant in TREC and 2 documents not judged in TREC (10 document in total). The judged documents were purposely included as part of the control queries to provide inter-coder comparison with TREC assessors. In addition, including some unjudged documents ensured that the control queries contained some semantic search retrieved documents. (This approach was used to train assessors in evaluating the documents for implicit relevance and avoid having them simply seeking out the query terms as indicators of relevance.) Finally, because all assessors completed the same control queries, these could be used to determine inter-coder agreement between

[6] These queries were: 130 from 2011 and 138, 159 and 166 from 2012.

the assessors in our experiment. For diversity, we selected an easy query for q_{c1} and a hard query for q_{c2}. Query difficulty was determined by the performance of the query in the baseline system (lvl0). We conjecture that queries that perform well are also easy to assess and having the assessors complete the easy query first keeps the effect of training noise to a minimum. For q_{c1} we selected query 101 (`Patients with Hearing Loss`) and for q_{c2} we selected 102 (`Patients with complicated GERD who receive endoscopy`). These were shown in the same order (101 and then 102) to all assessors.

Judging Setup

To collect assessments, we developed Relevation!: an open source, web-based system for performing relevance judgements in Information Retrieval system evaluation. Relevation! allows judges to browse queries and documents and then assign relevance assessments. It also supports the collection of qualitative data in the form of questionnaires and comments on specific queries and documents.

The 85 TREC MedTrack queries and a total of 1030 documents from the pool were loaded in Relevation!. The queries were then divided between the four assessors with each query being fully judged by only one assessor. Queries were divided so that each assessor judged, in total, roughly an equal number of documents. For each document, judges were asked to mark the document as either "highly relevant", "somewhat relevant" or "not relevant" with respect to that query (as per TREC MedTrack guidelines). In addition, assessors could optionally provide a free-text comment regarding their decision. On completion of judging all documents for a query, the assessor was also asked to answer the following questions:

- "How difficult was this query to judge?" Options included: "Very difficult", "Moderately difficult" or "Easy".
- "How would you rate the quality of the assessments you have provided for this query?" Options included: "High quality", "Average in quality" or "Poor quality (not confident in my judgements)".
- "Other comments?" Here judges could provide qualitative comments regarding the particular query.

The task description given to assessors was the same as that of the original TREC MedTrack task: recruitment of patients, matching a certain inclusion criteria, for clinical trials. Assessors worked together in the same room and were free to discuss their interpretation of queries, documents or their choices in relevance assessment. They were also free to consult any external resources

of information in making the decisions, including subscription-based medical reference sources or searching online for free information.

A total of 76 hours (19 hours per assessor) of judging was required to complete the 1030 documents. The average time spend per document was 4.4 minutes.

7.3.2 Judging Results

Inter-coder agreement

Inter-coder agreement between our four assessors was calculated based on the two control queries, which all four assessors completed. Agreement was found to be 0.85. This is in line with an inter-coder agreement of 0.8 found by the TREC MedTrack organisers.[7] Recall that the control queries also contain documents already judged by TREC assessors. Therefore, the TREC assessor can be added as a fifth assessor. Agreement between all five was 0.80. Individual agreement between assessors and the TREC assessors is detailed in Table 7.3.

Assessor	Agreement with TREC
One	0.72
Two	0.78
Three	0.81
Four	0.75
Average	**0.76**

Table 7.3: Inter-coder agreement of assessors with the TREC assessors.

Characteristics of New Relevance Judgements

Assessors rated each query according to how difficult it was to judge and a self-assessment of the quality of their judgements. Results are shown in Figure 7.3. For difficulty, most queries were easy or moderate, with only one query[8] considered very difficult to judge. For quality, assessors were confident in the judgements they provided. (No queries were marked as low quality.)

The frequency of documents according to relevance status — highly relevant, somewhat relevant and not relevant — is shown in Figure 7.4. Of the 1030

[7] Based on personal communication with Bill Hersh, TREC MedTrack organiser, 29 May 2013.

[8] Query 149: `Patients with delirium hypertension and tachycardia.`

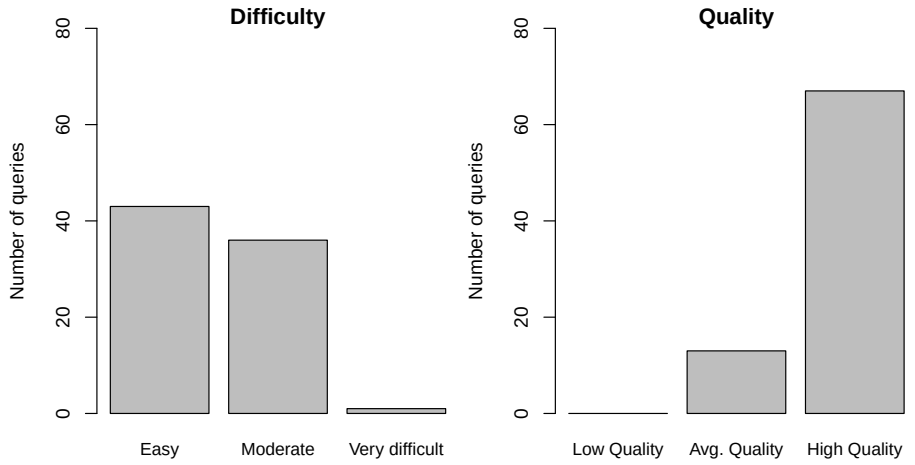


Figure 7.3: Query quality and difficulty

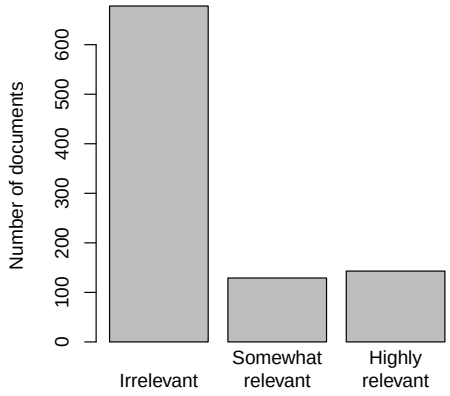


Figure 7.4: Frequency of documents according to relevance status.

documents judged, a large portion were found to be not relevant, while there were almost an equal numbers of "somewhat relevant" and "highly relevant" documents. In total, 29% of documents were judged as relevant. The original relevance assessments provided by TREC contained only 18% relevant documents. Therefore, the pool of documents from our systems (lvl0, lv1 and lv2) contained more relevant documents than the pool of documents provided by systems participating in TREC.

Four queries were excluded by the organisers of TREC MedTrack because insufficient relevant documents were found for these queries. However, these queries were included in our judging to determine if additional relevant documents could be found using the GIN. For query 166, no relevant documents were found by TREC assessors, whereas pooling using the GIN provided 6 relevant

documents. This was a sufficient number for this query to be re-introduced into the query set. (TREC organisers set a minimum of 5 relevant documents for a query to be included in the test collection [Voorhees and Tong, 2011; Voorhees and Hersh, 2012].) Details of the number of relevant documents for the four excluded queries, before and after our assessment, are provided in Table 7.4. This also highlights that none of the systems participating at TREC were able to retrieve any of these relevant documents in top ranked positions; instead, the GIN was able to retrieve these relevant documents.

Query	Number of Relevant Documents	
	TREC	Ours
130	1	1
138	0	4
159	0	3
166	0	6

Table 7.4: The four queries excluded by TREC MedTrack organisers for lack of relevant documents. After additional relevance assessment using the GIN, query 166 had a sufficient number of relevant documents to be re-introduced in the query set.

7.4 Graph Inference Model Re-evaluation

In this section, we re-evaluate the Graph Inference model using the new relevance assessments. For clarity, the original relevance assessments pertaining to TREC MedTrack are denoted "TREC" qrels, while the new relevance assessments provided by University of Queensland medical students are denoted "UQ" qrels.[9]

7.4.1 Retrieval Results

Table 7.5 presents the retrieval results of the GIN (lv1, lv2) and the Bag-of-concepts baseline (lvl0) using the old qrels (TREC) and the new qrels (TREC + UQ). The percentages indicate how the measure has changed between the old and new qrels.

Considering bpref, there was little change in overall effectiveness using the new qrels. This is not surprising as bpref considers only judged documents so

[9]In TREC, the term "qrels" is often used to denote relevance assessments; henceforth we adopt this terminology.

Qrel set	System	Bpref	P@10	P@20
TREC	lvl0	0.4309	0.5123	0.4389
	lvl1	0.4294	0.4481	0.4086
	lvl2	0.4208	0.4247	0.3630
TREC + UQ	lvl0	0.4252 (-1%)	0.5415 (+6%)†	0.4732 (+8%)†
	lvl1	0.4264 (0%)	0.5037 (+12%)†	0.4604 (+12%)†
	lvl2	0.4113 (-2%)	0.4878 (+15%)†	0.4220 (+16%)†

Table 7.5: Retrieval results using old (TREC) and combined (TREC + UQ) qrels. The percentages indicate how the measure has changed using the qrels. † indicates statistical significant differences between the TREC and TREC + UQ qrel sets (paired t-test, $p < 0.05$).

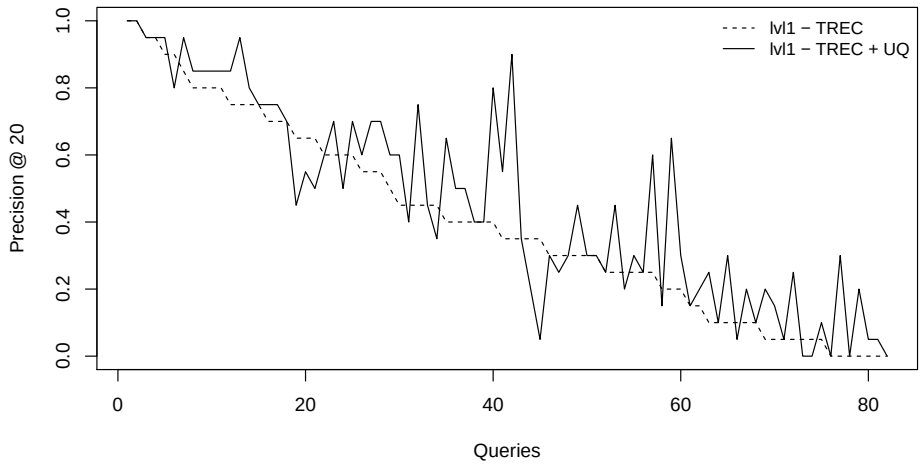


Figure 7.5: Graph Inference model performance of individual queries between the old (TREC) and new qrels (TREC + UQ). Greater number of improvements was observed in hard queries.

the large number of unjudged documents in the TREC qrels did not significantly affect this evaluation measure. However, for precision @ 10 and precision @ 20, all three systems were deemed more effective when evaluated with the new qrels. The percentages indicate by how much the effectiveness of the system was underestimated using only the TREC qrels. The effectiveness was underestimated for all three systems but was significantly more so with the GIN. Furthermore, lvl2, which leverages more of the GIN inference mechanism, was underestimated more than lvl1. This means that lvl2 was returning a larger number of unjudged but relevant documents.

Considering only precision @ 20, Figure 7.5 shows how the performance of individual queries changed between the old and new qrels. A significant

number of queries had improved performance using the new qrels, with only a handful showing degradation. Additionally, a greater number of improvements was observed in hard queries (those with poor performance using the TREC qrels; righthand side of the plot). This highlights that hard queries were the ones where performance was most underestimated.

7.4.2 Analysis and Discussion

Besides the quantitative relevance assessments, assessors also provided substantial qualitative comments regarding their relevance choices. This feedback highlighted how the notion of relevance within medical IR can be complex and subjective.

Assessors worked together in the same room and at times discussed their decisions regarding relevance assessments. Although they were confident in their assessments, they stated that the interpretation of the query was subjective and often required careful consideration regarding different possible interpretations. For the control query 101: `Patient with Hearing Loss`, assessors debated whether a patient born deaf could be considered as exhibiting hearing loss. (Technically, if they never had any hearing, then they never had a loss of hearing.) One assessor marked such a document as relevant, while another assessor marked the document as not relevant. A medical encyclopaedia was consulted and assessors agreed to include patient born deaf as relevant. This disagreement could be identified and resolved for the control queries, where assessors judged the same documents, but not for the actual queries where there was no overlap.

The task description given to assessors (recruitment of patients, matching a certain inclusion criteria, for clinical trials) also affected their decisions regarding relevance. Certain documents described patients who had hearing loss on admission but the hearing loss was treated and resolved by discharge. In this case, assessors decided these patients would not be eligible for the clinical trial and were therefore not relevant to the query. For other tasks, for example finding how hearing loss is treated, these documents may have been highly relevant. These cases highlight the complex and often subjective information needs of clinical information retrieval.

Queries with multiple dependent aspects received more debate by assessors and were also among the hardest queries (in terms of lower performance in the empirical evaluation). The second control query (query 102 `Patients with complicated GERD who receive endoscopy`) was one example. Gastroesophageal reflux disease (GERD) is caused when stomach acid comes up from the stomach into the esophagus. It is a common condition and is therefore found in many patients' records. The difficulty in interpreting this query was

whether the endoscopy was performed because of the GERD or for some other, unrelated condition. There were a number of documents where patients had GERD but received the endoscopy for another reason; these were marked as not relevant. A similar query was 103 `Hospitalized patients treated for methicillin resistant Staphylococcus aureus (MRSA) endocarditis`, where endocarditis and MRSA were mentioned in the same document, but the cause of the endocarditis was not the MRSA. Again, these documents were marked as not relevant. These queries all have multiple dependent aspects to the query; even if both aspects are present in a document, that document may still not be relevant unless the dependence between them can be determined.

Temporality also played a significant role in relevance assessments. The most common situation was when information pertaining to the query was found in the patient's past medical history section. Assessors had to decide whether the information was still valid. Some conditions are ongoing, for example, Gastroesophageal reflux disease (GERD), so the fact that this was stated in past medical history does not affect the relevance of the document; others are temporal and are unlikely to still be valid. In certain cases, assessors consulted the actual dates of the past medical history information to determine how recent the information was and whether it might still apply.

Simulated Precision Revisited

In Section 7.2.1 we provided a simulated precision @ 20 measure if completed judgements were obtained for the top 20 rank positions. We revisit that analysis here in light of the actual results obtained.

The correlation coefficient between the *simulated* performance estimate and the *actual* performance estimate was 0.92, whereas the correlation coefficient between the *original* performance estimate and the *actual* performance estimate was 0.89. This shows that the simulated estimate was more accurate than the original estimate. A plot comparing the three estimates — original, simulated and actual — for individual queries is shown in Figure 7.6. The simulated estimate generally follows the trend of the actual estimate, except for a few cases where the actual estimate was lower than the original estimate. Although the simulated estimate diverges from the actual estimate in these cases, it does provide a more accurate estimate of retrieval effectiveness than the original estimate that used the relevance judgements from TREC MedTrack. It can, therefore, be used as one possible indicator of retrieval effectiveness when large numbers of unjudged documents are retrieved by a system.

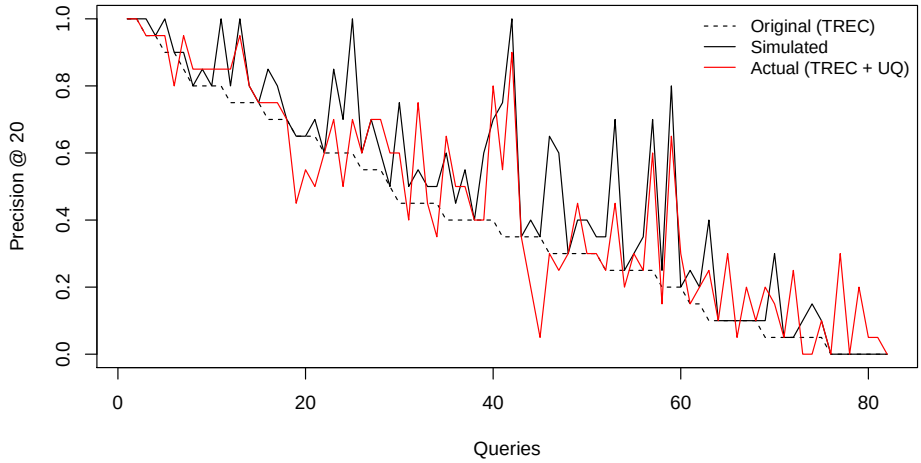


Figure 7.6: Per-query precision @ 20 retrieval effectiveness comparing the original qrels from TREC, simulated performance and actual performance using TREC + UQ qrels.

7.5 ICD Evaluation Method

In this section, we present an alternative evaluation method to TREC-style test collections. This method uses implicit relevance assessments in the form of ICD diagnosis codes, which are manually assigned to clinical reports by clinical terminologists as part of the reporting, billing and administrative requirements of hospitals and governments. The manually assigned ICD codes are used to devise both a set of queries and associated relevance assessments. No manual assessment of documents is required. Finally, an evaluation of the Graph Inference model is provided using this new evaluation method.

7.5.1 Documents and ICD Codes

As the collection of clinical documents, we use the BLULab NLP collection from the University of Pittsburgh[10]. This collection is the same set of documents used as part of the TREC MedTrack. An example medical record is provided in Figure 7.7. The highlighted codes within the `<admit_diagnosis>` and `<discharge_diagnosis>` XML elements are part of the International Statistical Classification of Diseases and Related Health Problems (ICD) coding scheme. ICD is a coding of diseases and signs, symptoms, abnormal findings, complaints, social circumstances and external causes of injury or diseases, as classified by the World Health Organization.

[10]BLULab NLP Repository, University of Pittsburgh, http://nlp.dbmi.pitt.edu/nlprepository.html. Last accessed July, 2001.

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<report>
<checksum>20070901DS-WByC8eeIy9cv-848-182262802</checksum>
<subtype>NEUROSURG DISCHARGE</subtype>
<type>DS</type>
<chief_complaint>POST LAMINECTOMY SYNDROME</chief_complaint>
<admit_diagnosis> 724.5 </admit_diagnosis>
<discharge_diagnosis>
 724.5 , 424.0 , 787.01 , E935.2 , E849.7
</discharge_diagnosis>
<year>2007</year>
<downlaod_time>2009-08-18</downlaod_time>
<update_time/>
<deid>v.6.22.07.0</deid>
<report_text>[Report de-identified (Safe-harbor compliant)
by De-ID v.6.22.07.0]


NEUROSURGERY DISCHARGE SUMMARY
PATIENT NAME:  **NAME[AAA, BBB M]
ACCOUNT #  **ID-NUM

ATTENDING PHYSICIAN:  **NAME[YYY XXX ZZZ]
ADMISSION DATE: **DATE[Aug 29 2007]
DISCHARGE DATE:  **DATE[Sep 01 2007]
PRINCIPAL DIAGNOSES: POST LAMINECTOMY SYNDROME, STATUS POST FAILED
TRIAL OF INTRATHECAL OPIOID PUMP.

REASON FOR ADMISSION:  This is a **AGE[in 40s]-year-old female
with signs  and symptoms of post laminectomy syndrome.  It was
felt that a trial of an  implanted intrathecal opioid pump might
be of benefit to the patient.   She entered the hospital and began
a trial of a morphine pump on  **DATE[Aug 29 07].

HOSPITAL COURSE:  The patient remained alert and oriented,
afebrile with stable vitals during her stay.However, she
experienced significant nausea and vomiting with very little
relief in her pre-existing pain during the morphine trial.  As a
result, the  intrathecal medication was changed from morphine to
Dilaudid.  By postop day  number 2, her nausea had cleared and she
was tolerating p.o. intake.   However, despite large increases in
the intrathecal administration rate, she  received essentially no
relief of her pretrial pain.She also complained of a severe
positional headache on postop day number 1.   This was treated
with IV fluids and flat bed rest, and it resolved on its own prior
to discharge.
```

Figure 7.7: Example medical record (`report1.xml`) from the BLULab corpus. ICD codes highlighted.

ICD Code	Description
724.5	Backache; Vertebrogenic pain syndrome
424.0	Mitral valve disorders
787.01	Nausea with vomiting
E935.2	Other opiates and related narcotics: Codeine [methylmorphine], Morphine, Opium (alkaloids), Meperidine [pethidine]
E849.7	Residential institution: Children's home, Dormitory, Hospital, Jail, Old people's home, Orphanage, Prison, Reform School

Table 7.6: ICD code descriptions for the codes listed in Figure 7.7.

The example record in Figure 7.7 has been coded with five unique ICD codes; the descriptions of these codes are shown in Table 7.6. The ICD codes used to classify the medical documents in the BLULab collection form the basis of our evaluation framework. They represent a human gold standard for the key concepts contained within the particular medical record.

It is important to consider the regional differences affecting the manual assignment of ICD codes. In the U.S.A, coding using ICD is conducted for billing purposes; therefore, the codes are far less reliable as indicators of clinical facts. In other countries, the codes are applied primarily to classify the medical diagnoses and conditions pertaining to the record and would be more reliable.

7.5.2 Queries and Relevance Judgements

The process for developing queries and relevance judgements from the BLULab collection is illustrated in Figure 7.8.

The steps required are:

❶ For each medical record (document) the ICD codes assigned to that record were extracted;

❷ Each ICD code was considered an individual query: the query id was the code id and the query text was the ICD code description as defined in the ICD taxonomy;

❸ The ICD code and document id (filename) were then added to the relevance judgement file.

A total of 3500 queries was generated using this method. Every document was assigned at least one ICD code so every document was relevant to at least one query (i.e., there were no unjudged documents).

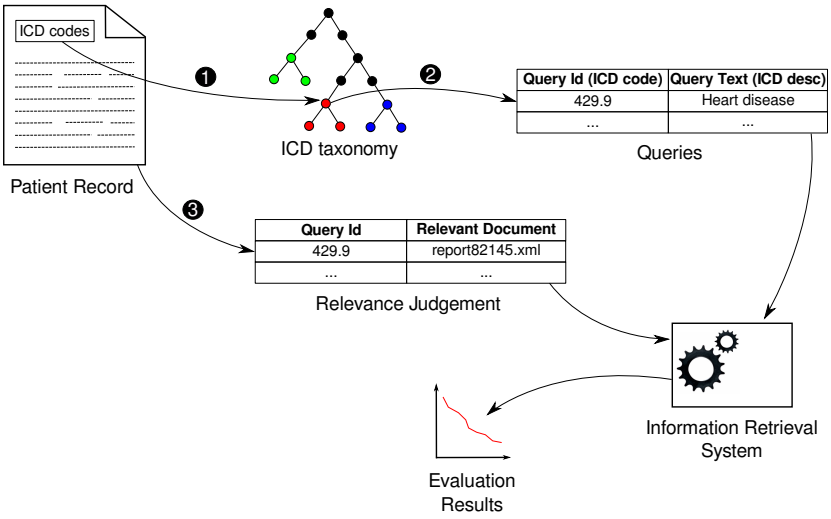


Figure 7.8: Evaluation architecture for creating an IR test collection from the BLULab collection.

A number of retrieval experiments were conducted using the ICD test collection described here. The findings from these experiments and discussion of issues in using ICD codes for relevance assessment are provided in Koopman et al. [2011]. Additionally, the queries and qrels were made available online at `http://aehrc.com/med_eval`.

Manually assigned codes or categories have been used previously for IR evaluation. Lewis [1992] applied such a method to evaluate a phrase-based indexing and retrieval system, while Sanderson and Joho [2004] advocated using manually assigned categories to create test collections without the need for human relevance assessment.

The ICD evaluation proposed here provides an IR test collection within the medical domain without the need to gather relevance assessments from human judges. The clinical terminologists who assign ICD codes to documents do so by reading the documents and determining the specific diagnoses relevant to that patient. Determining these diagnoses sometimes requires interpreting the raw medical data and inferring an implicit diagnosis. For these cases, the implicit diagnosis may not have been mentioned in the text of the document; therefore, a semantic gap exists between queries and documents. Thus, a test collection formed in this way is a realistic resource for evaluating medical IR systems.

7.5.3 Graph Inference Model Evaluation

In this section, we evaluate the GIN using the ICD evaluation method. Two query subsets are devised: The first contained 167 queries with at least 100 associated relevant documents; we denote this "Min100RelDocs". The second contained 114 queries that were mapped to concepts containing only a single SNOMED CT concept; we denote this "'SingleConceptQuery". This second subset was devised to determine the effect that multiple query concepts and, therefore, query dependence had on retrieval effectiveness. In line with the evaluation of the GIN in Chapter 6, we performed three retrieval runs: lvl0, lvl1, lvl2, with lvl0 representing the Bag-of-concepts baseline.

Table 7.7 presents the retrieval results using the two different query sets. For MAP, there is little difference between the three systems; however, recall is significantly higher for the GIN. Precision @ 10 is less for the GIN using the Min100RelDocs queries, but greater for the SingleConceptQuery queries.

Query Set	System	MAP	# Relevant Returned	P@10
Min100RelDocs	lvl0	0.1740	22179	**0.4162**
	lvl1	**0.1754**	**24454**	0.3725
	lvl2	0.1497	23387	0.3144
SingleConceptQuery	lvl0	0.2325	859	0.1053
	lvl1	**0.2370**	1527	**0.1184**
	lvl2	0.2157	**1583**	0.1079

Table 7.7: Evaluation of the GIN using the ICD evaluation method.

The increase in recall demonstrates that the GIN is returning many more relevant documents that were never retrieved by the Bag-of-concepts baseline. This is where the inference mechanism is working — traversing the SNOMED CT relationships to find documents containing concepts related to the query concepts and bridging the semantic gap.

Precision @ 10 degrades for Min100RelDocs but not for SingleConceptQuery. Min100RelDocs contains cases of multiple dependent query aspects, whereas SingleConceptQuery contains only single query concepts and therefore no dependent query aspects. In the GIN, if a single query concept contains a large amount of related concepts, many of which appear together in a document, then that document will receive a higher score and may appear at top ranked positions, even though the document may contain only one of the query aspects. To handle such cases, a query dependence model would be required that ensured that documents containing multiple aspects were preferenced. Further

discussion on query dependence is provided as part of future work in Chapter 8.

Finally, the retrieval results using the ICD evaluation method are in line with those found using the TREC MedTrack test collection in terms of how the three systems compared with each other. This shows that the ICD evaluation method is accurate and that an implicit test collection can be devised without the use of human judges for relevance assessments.

7.6 Summary

How to measure the effectiveness of a semantic search system is critical to the evaluation of the models put forward in this thesis. Semantic search systems are aimed at making the retrieval model less dependent on the individual terms, retrieving relevant documents that may not have been returned by keyword-based systems. It is these keyword-based systems that largely contributed to the judging pool of documents given to human assessors.

In this chapter, we quantify the effect that large numbers of unjudged documents found in retrieval rankings of the GIN have on its retrieval effectiveness estimates. Although the GIN returns many unjudged documents, in some cases this does not lead to a significant degradation in performance for these queries. These are examples of where the inference mechanism of the GIN is working — returning new relevant documents never retrieved by systems in TREC and therefore never assessed for relevance. This analysis into the effect of unjudged documents was the motivation for obtaining additional relevance assessments.

Additional relevance assessments were obtained with the help of four graduate medical students, who judged those documents previously not judged by TREC assessors. Documents were selected by pooling three retrieval runs: Bag-of-concepts (lvl0) and the GIN (lvl1 and lvl2). Using the new relevance assessments, these three systems were re-evaluated. The results showed that the effectiveness of all three systems was underestimated using the TREC qrels and that the underestimation was worse for the GIN (especially for the greater inference provided by lvl2). Furthermore, the underestimation was worse for hard queries — those more suited to the GIN. These results confirm our hypothesis that the inference mechanism in the GIN is returning new relevant documents that were not retrieved by other systems (either TREC or Bag-of-concepts). In fact, one of the queries, previously excluded in TREC for lack of relevant documents, could now be re-introduced as the GIN found sufficient relevant documents for this query.

Qualitative feedback from our assessors highlighted how the notion of relevance within the medical domain can be complex and subjective. A number of

different interpretations of a query are possible and these can have a significant effect on document relevance. Queries with multiple, dependent query aspects were particularly ambiguous. The specific task description of eligible patients for a clinical trial also played an important role in assessors' decisions of relevance. Finally, temporality, which was introduced as one of the semantic gap problems in Chapter 2, proved to be a significant issue requiring future work.

This chapter also provides an alternative evaluation method, one that uses implicit relevance assessments in the form of ICD diagnosis codes, manually assigned by clinical terminologists. In some cases, these codes are assigned based on the terminologist's interpretation of the documents, where the document may not explicitly mention the query terms. Thus, the test collection we provide also contains a number of queries exhibiting semantic gap issues, making it a realistic resource for evaluating medical IR systems. An evaluation of the GIN using the ICD method showed that the GIN returned many more relevant documents (increased recall) but precision was affected by queries with multiple dependent aspects. Overall, the results using the ICD evaluation were in line with those found using the TREC test collection, showing that an implicit test collection can be devised without the use of human judges for relevance assessments.

Additional discussion regarding the issues in evaluating semantic search systems, how the inference mechanism in the GIN works and future work that arose from this chapter, are covered in the next chapter on Discussion and Future Work.

Chapter 8

Discussion and Future Work

The ultimate authority must always rest with the individual's own reason and critical analysis.

— Dalai Lama

The major aim of this thesis was to bridge the semantic gap in searching medical data. In this chapter, we reflect on the ability of the three models we proposed — Bag-of-concept, Graph-based Concept Weighting and Graph Inference Model — to bridge this semantic gap. We revisit the main hypothesis of a unified model of semantic search as inference. We provide an understanding of the different types of inference and when they should be leveraged in semantic search. In addition, we discuss the differences in definitional inference used by ontologies and reflect on the types of inference required for effective retrieval. The challenges in evaluating semantic search systems are discussed; in particular, we consider how these might be addressed. Finally, we present those characteristics that a successful semantic search model would need to have in order to fully bridge the semantic gap. In the future work section, we consider how the application of the GIN can be extended beyond medical IR into other areas, including large-scale web search using structured knowledge resources such as the Google Knowledge Graph.

8.1 Bridging the Semantic Gap

Table 8.1 presents which semantic gap issues are addressed by each of the three models proposed in this thesis.

Semantic Gap	Bag-of-concepts	Graph Weighting	Graph Inference
Vocabulary Mismatch	●	●	●
Granularity Mismatch	○	○	●
Conceptual Implication			●
Inference of Similarity		○	●

Table 8.1: Semantic gap issues addressed by each model presented in this thesis. A ● indicates that the model specifically addressed the issue; ○ indicates that the model partially or indirectly addressed the issue.

Recall that as part of the process of converting terms to concepts, semantically similar variations of term phrases are conflated. Whilst this is not 100% precise, it did address the vocabulary mismatch problem at the level of terms (see Section 4.2.1). As all three models used a concept-based representation, they all addressed vocabulary mismatch in this way.

Granularity mismatch occurs when the same information is expressed with different levels of granularity, for example the general class of drugs "antipsychotics" and the specific drug "Diazepam". Granularity mismatch was only partially addressed by the Bag-of-concepts and Graph Weighting model. This is a result of the concept expansion process, where the expanded concepts were potentially more specialised instantiations of the source terms. However, the concept expansion process of mapping to more specialised concepts occurred only in certain cases. In addition, it did not account for the reverse case of deriving more general concepts from the source terms. Therefore, granularity mismatch was only partially addressed by the concept-based representation of the Bag-of-concepts and Graph Weighting model. In contrast, the GIN specifically tackled granularity mismatch. This was achieved by traversing parent-child (i.e., ISA) relationships to infer more specialised and more general concepts from the query concept.

Conceptual Implication is where the presence of certain terms in the document infer the query terms, for example where an organism implies the presence of a certain disease. Deriving these associations and tackling conceptual implication can be difficult. Even though such associations are usually implicit in the corpus, they are often explicit in domain knowledge resources, for example,

SNOMED CT encodes them as relationships between concepts, such as *Varicella Zoster virus* → *Chicken Pox*. The Bag-of-concepts model did not utilise these relationships and the Graph Weighting model used only the number of relationships (rather than the actual relationship) as an indicator of importance for a concept. Only the GIN specifically addressed Conceptual Implication by traversing these types of relationships, inferring concepts that implied the query concepts and as a consequence scoring documents that contained the implied concept.

In Inferences of Similarity, the presence of a certain concept indicates high likelihood of another, or the two concepts are semantically similar in some way. In these cases, an IR system needs to account for both the dependence between medical concepts and the strength of association between them, in order to be effective. The Graph Weighting model captured the dependence between two concepts within a document as an edge within the document graph. In our implementation of the model, edges were determined by the co-occurrence of concepts with a context window. However, the model did not capture the strength of association between concepts so the model only partially addresses the problem. In the GIN, the associations were taken from SNOMED CT, so the GIN leverages the explicit dependence information provided by the domain knowledge resource. In addition, the GIN also captures the strength of association by means of the diffusion factor, which assigns a corpus-based measure of similarity to the domain knowledge-based relationship. Thus, the GIN captures both the type of association and the strength of the association required for the problem of Inference of Similarity.

8.2 Unified Model of Semantic Search as Inference

The aim of this thesis was to develop a unified theoretical model of semantic search as inference, which is expressive enough to integrate structured domain knowledge (ontologies) and corpus-based, statistical methods. We now revisit this aim in light of the Graph Inference model proposed in this thesis (Chapter 6).

We claim that the GIN is a unified model of semantic search. Structured domain knowledge is integrated using a novel graph-based representation of a corpus: nodes represent Information Units in the corpus but their definition and associations are derived from the domain knowledge resource. We also claim that the GIN is general, as Information Units, associations and the inference

mechanisms can be instantiated in a variety of ways; this makes the model flexible and adaptable in that:

- Any knowledge resource — domain-specific or general — can be used; provided it can be represented as a graph. This includes ontologies or thesauri such as WordNet or other resources such as Freebase or DBpedia.[1] Further comment on this is provided as part of the future work section.
- Different scoring methods can be used, simply by changing the initial probabilities or weighting scheme on the node. This allows the integration of existing, standard IR models such as language models, BM25 or others but also provides an easy means to integrate new models still being developed.
- Any diffusion factor measure can be applied: corpus-based such as semantic similarity, or relationship type-based.
- An efficient implementation of the model makes it attractive to large scale retrieval task; more on this in future work.

Graph-based representations have proved effective as the unifying framework by capturing data, structured ontologies, domain knowledge and associations within a single representation. Beyond using graphs for the representation of information, graph-based algorithms also provide a powerful means of utilising this information. In the Graph Weighting model, the graph-based algorithm, PageRank, is used to identify important concepts in a document. In the GIN, the inference mechanism is realised as the traversal over a graph; it is this inference mechanism that is designed to bridge the semantic gap.

8.3 Understanding Inference

Transacting inference to improve retrieval effectiveness can be risky. Starting with Salton's study of the use of thesauri for query expansion in the 1960's [Salton, 1968], a variety of studies over subsequent decades have confirmed that inference can realise significant improvements in effectiveness for some queries, and massive degradation for others. In this sense, employing inference for information retrieval has been somewhat like an unreliable genie. Despite the upsurge in interest in inference via the logic-based IR drive in the 1990's, most researchers nowadays would probably hold the opinion that the genie be best

[1] DBpedia is a resource of structured information extracted from Wikipedia. Freebase is a graph database of structured general human knowledge.

left in the bottle. We do not subscribe to that view — and this thesis can be seen as an attempt to let the genie out of the bottle, albeit cautiously.

For hard queries, inference is worth it; shown by the fact that all three models generally made more improvements on hard queries. Hard queries often suffer from multiple semantic gaps and, sometimes, there is nothing to lose by applying inference. For easy queries, the inference mechanism is not required and sometimes is detrimental.

An important outcome of the empirical evaluation of the GIN was an understanding of the characteristics of queries that require inference and those that do not. A post-hoc analysis allows queries to be clearly categorised according to the degree of effectiveness the inference achieved; these are presented in Table 8.2. Included are example queries from TREC MedTrack; the keywords for each of the queries are provided in Appendix D.

Queries are divided into five broad categories. For each category, a number of characteristics of the queries comprising that category are provided. For each category, the effect of the inference mechanism on retrieval effectiveness is also stated.

The information presented here provides a greater understanding of how the inference mechanism is working. Such information is valuable because it provides a means both to improve the models proposed here and to provide a foundation for future models of semantic search.

8.3.1 Definitional vs. Retrieval Inference

The discussion of the GIN in Chapter 6 also raised the broader topic of the differing requirements of definitional inference versus retrieval inference. Ontologies such as SNOMED CT are largely definitional, meaning that they are concerned with providing domain specific semantics of concepts. As a consequence, ontologies capture the "what" of concepts. For example, SNOMED CT represents, by way of definition, "what" diabetes is and its relationships with other concepts. As a consequence, valid conceptual inferences amount to inferences that essentially preserve definitional validity. This is perfectly fine if one wants to extract appropriate *implied* conceptual knowledge from the concepts present in a document. However, it begs the question as to what degree such inferences are appropriate for *retrieving* relevant documents. In this thesis, we have come to the conclusion that inferences that preserve definitional validity are not sufficient to guarantee inferences that promote effective retrieval. This conclusion is perhaps unsurprising and should not be construed as an admission that the genie should remain in the bottle. Rather, we advocate that a clearer understanding is necessary regarding the conceptual inferences needed to promote

Category	Characteristic	Effect on Retrieval	Example Queries
Query with consistent improvements using inference:	Valuable related concepts from the ontology.	Inference always improves results and diffusion factor controls noise.	108, 171
Query where no inference is required:	Easy, unambiguous queries, often with small number of relevant documents.	Inference degrades performance.	104, 161
Query requiring reranking:	Verbose queries with multiple dependent query aspects. The key query aspects contained many related concepts.	Small amounts of inference required (depth 1–2).	113, 135, 119
Queries requiring the inference of new relevant documents:	Domain knowledge essential in interpreting the query. Relevant documents that do not contain any query terms. Queries with multiple semantic gap issues.	Inference always improves performance.	147, 154
Queries unaffected by inference:	Very hard queries; semantic gap cannot be bridged. No domain knowledge available for terms/concepts in the query.	Inference has no effect.	137, 139

Table 8.2: Categories and characteristics of queries and the effect that the inference mechanism in the GIN has on them. Included are example queries from TREC MedTrack; the keywords for each of the queries is provided in Appendix D.

retrieval — an understanding that is unburdened by the need to preserve (near) definitional validity. It is certainly true that definitional validity does not necessarily translate into an easily assumed counterpart, namely retrieval precision. In short, useful inferences for retrieval revolve around the "how" rather than the "what" of concepts.

As an example of this, consider the concept of diabetes and two possible related queries: 1) *Patients with insulin dependence* and 2) *Patients likely to be subject to chronic renal failure*. SNOMED CT tell us that diabetes is a "disorder of glucose metabolism" and a "disorder of the endocrine system" and it affects the "structure of the endocrine system". Such information clearly defines diabetes and makes the definition distinct from other concepts; it provides the "what". However, it does not include the "how": "how" diabetes is treated with insulin and "how" diabetes results in chronic renal failure. Such information is not part of the definition of diabetes but is required to handle the example queries effectively: the fact that insulin is used to treat diabetes can be used to infer that patients with diabetes are relevant to the first query. The fact that diabetes can cause chronic renal failure can be used to infer that patients with diabetes are relevant to the second query. These examples again illustrate that inferences of definitional validity are not sufficient to guarantee inferences that promote effective retrieval.

The tension between definitional and retrieval inference mirrors a tension identified in artificial intelligence. Frixione and Lieto [2012] describe the situation as a strain between compositionality, which is definitional on the one hand, and the need to represent *prototypical* information[2] (which includes some of the "how" information is used) on the other.

Given that ontologies are largely definitional, how can the inference mechanism that utilises them be improved? How can we distinguish the concepts and relationships, useful for retrieval, from less useful, definitional concepts and relationships? One solution may be to leverage some measure of quality for a fragment of domain knowledge from the perspective of inference; for example, a hypothetical heuristic used by the GIN that indicated the quality of the portion of SNOMED CT that it was about to traverse. Such a heuristic could take into account the granularity or coverage of a particular part of an ontology; very general concepts could be avoided, whereas specific "leaf" concepts might be favoured. Corpus statistics could be used to augment the measure of quality; for example, the IDF of a concept could aid in identifying general concepts. However, the previous solutions do not yet capture the "how".

[2] Frixione and Lieto [2012] provide the following definition of prototypical information: "According to the prototype view, knowledge about categories is stored in terms of prototypes, i.e. in terms of some representation of the"best" instances of the category. For example, the concept CAT should coincide with a representation of a prototypical cat."

An alternative approach, assuming the need to represent specific information for retrieval, is to devise a domain knowledge resource specifically suited to retrieval inference. What if the resource were specifically constructed to represent information with retrieval inference in mind? Ideally, what would constitute such a resource and what information would it contain? We identify some of the key characteristics of such as resource:

Vocabulary: The resource should cover vocabulary: how things are *described* (synonyms, variants, etc.), not how they are *defined*.

Associations: The resource should capture how things are associated and the strength of that association.

Granularity: It should capture granularity such as specialisation and generalisations but these should be quantified by how specific or how general a parent or child concept is.

Uncertainty: A measure of certainty (such as "known" or "suspected") should be included. Ontologies such as SNOMED CT only represent conceptual implications such as *organism* $\rightarrow$ *disease* and do not capture pragmatic conceptual relationships such as *treatment* $\rightarrow$ *disease*. (This is because opinions may differ on the best treatment for a disease and may change over time.) Instead, these types of relationships should be included in a resource aimed for retrieval but qualified with a measure of certainty.

Table 8.3 presents which of the above requirements are met by the SNOMED CT ontology. The requirement of Vocabulary is met by SNOMED CT, while the requirements of Associations and Granularity are only partially met; Uncertainty is not provided by SNOMED CT.

Some knowledge resources do exhibit some of the characteristics described above as desirable for retrieval. In recent years, there has been an effort to semi-automatically derive large structured knowledge resources. Initiatives such as

Requirement	**SNOMED CT**
Vocabulary	●
Associations	○
Granularity	○
Uncertainty	

Table 8.3: The requirements of a domain knowledge resource specifically suited to retrieval inference and how these are met by the SNOMED CT ontology. ● indicates that the requirement has been fully met, while ○ indicates that the requirement has been partially met.

DBpedia [Auer et al., 2007], Freebase [Bollacker et al., 2008] and the Google Knowledge Graph [Singhal, 2012] are examples of these. Such resources are constructed by analysing web content, from Wikipedia or by combining other knowledge resources together (e.g. LinkedData initiatives [Bizer et al., 2009]). Critics argue that such resources, being semi-automatically generated from data, lack rigour; however, being generated from data, they capture much of the associational information desirable for IR. In the future work section, we consider how such resources might be utilised by the GIN.

8.4 Evaluating Semantic Search

Issues of how to evaluate a semantic search system played a significant role in this thesis. The Bag-of-concepts model was developed prior to the advent of the TREC Medical Records track so an alternative evaluation method was required. This was done using implicit relevance judgements in the form of ICD codes assigned by clinical coders (described in Section 7.5). The advent of the TREC MedTrack provided a standard test collection for evaluation but Chapter 7 showed that evaluation using the relevance judgements associated with this collection underestimated the effectiveness of the GIN. In this section, we consider the challenges for evaluating semantic search systems and how they might be overcome.

8.4.1 Pooling for Semantic Search

One major issue for evaluating semantic search using TREC-style evaluations is how the test collection is constructed: the pooling method. Recall that the driving motivation for a semantic search and inference approach is that it may retrieve documents that share few or even no keywords with the query. Such documents are unlikely to be retrieved by a keyword-based IR system. If the pool was derived from predominately keyword-based systems, then documents that are not retrieved by keyword-based systems would never make it into the pool and would never be assessed for relevance. Ideally, the solution to this problem is to ensure diversity within the pool by having semantic search systems contribute to the pool. This is a well known problem, as TREC collections are extensively utilised for many years after they are constructed and testing is performed using systems that never contributed to the pool [Voorhees and Harman, 2005].

If there were only a few semantic search systems contributing to the pool and a large number of keyword-based systems, then the portion of documents

contributed by these semantic search systems and judged for relevance would still be small. This would be the case for test collections with a large number of contributing systems and a large document collection; for example, the modern TREC WebTrack. In this case, other strategies could be applied when constructing the document pool. Zobel [1998] proposed varying the number of documents to be judged for each query based on its characteristics. The number of relevant documents at the top of the ranking was used as an indicator of how many more would be found further down the ranking. Thus, shallow pooling was performed for queries with poor performance, while deeper assessment was performed for queries with many relevant documents in top-rank positions. This approach could be adapted to dealing with semantic search systems by focusing in on those queries where the number of relevant documents in top-rank positions differed considerably between different contributing systems. This might indicate queries with a diversity of results that require a greater depth of judging. Another approach is to judge more documents from certain systems. Cormack et al. [1998] noted that some systems contributing to the pool are more effective at returning relevant documents; they argue more documents should be assessed from such systems. This approach could be adapted by biasing systems that add diversity to pool (but are still reasonable in terms of the number of relevant documents returned in top-rank positions). Both the approaches of Zobel [1998] and Cormack et al. [1998] were found to produce test collections as good as TREC [Sanderson and Joho, 2004]. Finally, diversity could be improved by considering characteristics of the documents themselves; for example, including documents that contained few or no query terms. Such documents are more likely to have been retrieved by semantic search systems (or by other novel systems, for example, those applying some form of unorthodox query expansion).

A number of techniques for forming the document pool, outlined here, can be used to improve the way semantic search systems are evaluated. Further research is needed to determine the effectiveness of these approaches.

8.4.2 Dealing with Unjudged Documents

If the relevance assessment process cannot be influenced, researchers should at least be aware of the effect that unjudged documents may have on estimating retrieval effectiveness. The analysis presented in Chapter 7 is an example of one method that can be applied to understand this effect. When different systems are compared by means of their retrieval results (typically a test and benchmark system), it is also valuable to report the number of unjudged documents retrieved. This provides an insight into the effect of these unjudged documents

and how the two systems being compared may differ and their performance be underestimated.

The choice of evaluation measure is also an important consideration for evaluating semantic search systems. Measures such as bpref do not consider unjudged documents so may be preferable in some situations. In contrast, MAP assumes that an unjudged document is assumed not relevant; large numbers of unjudged documents would thus significantly impact effectiveness estimates. If the pool contains complete judgements for the top k results of each system, then there are no unjudged documents in the top k results and precision @ k provides a reliable estimate; however, it provides no measure of recall. Although there is no ideal evaluation measure, consideration should be given to the most appropriate measure given the task at hand. Using the two measures above in conjunction can provide an indication as to the number and effect of unjudged documents: if bpref increases but MAP decreases, this could indicate that unjudged documents are having a significant effect.

Recruiting assessors to perform additional judging has been the approach taken in this thesis. Although this process can be costly and time consuming, it can provide a definitive result in terms of the effectiveness of a system. In addition, in the case of the work exposed in this thesis, observations and discussions with the assessors provided valuable insights both in terms of the information need of such users and the workings of the systems they were evaluating. (These were presented in Section 7.4.2.)

Evaluating semantic search systems presents some specific challenges [Uren et al., 2010]. However, it also provides a number of interesting avenues of research that may have implications for evaluating IR systems in general.

8.5 Characteristics of a Successful Semantic Search Model

In concluding our discussion, we consider the characteristics of a successful semantic model, one that combines structure domain knowledge with corpus-based statistical techniques. This conveys both the lessons learnt from this thesis and is a precursor to future work. A successful semantic search model should have:

- A good source of domain knowledge, one that contains not just definitional information about the concept making up the domain, but also associational information capturing how these concepts are used in the data — i.e., both the "what" and the "how". The resource should have

sufficient coverage: the major topics constituting that domain should be modelled, but the resource should also have suitable consistency in coverage, i.e., avoiding the situation where certain topics are modelled in great depth, while for others no detail are provided.

- An effective means of mapping free-text to domain knowledge. In our case, this was provided by the MetaMap system but alternatives are under active development [Suominen et al., 2013]. In the medical domain, natural language remains an important means by which medical professionals communicate. It is unlikely that this will be replaced by structured reporting. In addition, legacy reports (potentially covering an entire lifespan of a patient) will still need to be interpreted. Therefore, an effective means to map free-text to domain knowledge will remain an important requirement for effective semantic search.

- An adaptive mechanism to know when and how much inference to apply (for example, an adaptive depth method in the GIN). A finding of this thesis has been that inference is needed in certain cases and not needed in others. (Some characteristics of these cases were outlined in Table 8.2.) A successful system would use features like those in Table 8.2 to adapt the amount of inference on a per-query basis. Our findings have shown that inference generates consistent improvements on hard queries. If these queries could be determined in advance, then an adaptive mechanism would pay dividends. Another way to determine when to apply inference would be to have the user manually specify this. In a medical IR scenario, where users are medical professionals with complex information needs, this may be desirable.

- An effective evaluation method that is suitable for a semantic search system, either with a suitable test collection or at least with an understanding of the effect of unjudged documents.

The above is not intended as an exhaustive list of features of a semantic search system. Instead, it provides four key characteristics that should be considered when developing such systems.

8.6 Future Work

A number of areas of future research arise from this thesis, primarily related to the Graph Inference model from Chapter 6.

8.6.1 Adaptive Depth

A clear area of future work is the development of an adaptive depth method that controls the amount of inference to apply on a per-query basis. The question is whether such queries might be automatically identified by the retrieval system and dealt with differently. Previous work on query performance prediction [Hauff et al., 2008], including recent work within the medical domain [Boudin et al., 2012], could be applied here. Many additional features are also available in the three models that could support a predictive model of query difficulty — features such as the semantic type of the query: for example, are queries that are of type "Gene" more difficult than those mentioning "Symptoms"? Ambiguity measures can also be used, for example, the number of candidate concepts provided when mapping from free-text. Other potentially useful features include differences between the corpus-wide distributions of terms in, and concepts extracted from the query; the number of concepts in the query; the granularity of the query concepts in the UMLS hierarchy; and the degree of the concept in the SNOMED CT relationship graph.

There is a rich array of information available to train a predictive model of query difficulty. In this discussion, we have also presented five query categories and their characteristics (Table 8.2). Along with query difficulty, these categories could be used to inform an adaptive method.

8.6.2 Web Search using the Graph Inference Model

The GIN, although developed within the medical domain, was advocated as a general model of semantic search as inference. To demonstrate this, we consider how it can be applied to web search.

In 2012, Google announced the release of the Google Knowledge Graph [Singhal, 2012]: a very large knowledge base designed to enhance the results of the Google search engine with semantic-search information. The knowledge base is constructed from a number of resources, including the CIA World Factbook, Freebase and Wikipedia. Unfortunately, the entire resource is not publicly available; however, the Freebase component is available. Freebase is a graph database of structured general human knowledge [Bollacker et al., 2008]. As of 2013, it contains 1.9 billion triples covering a wide variety of concepts. An example of the concept for "Nelson Mandela" is shown in Figure 8.1; the concept is related to four other concepts according to the specified relationships. Freebase provides a structured domain knowledge resource suitable to implement a Graph Inference model tailored to web search. While SNOMED CT was the domain knowledge resource used for the medical domain, Freebase is the general knowledge resource

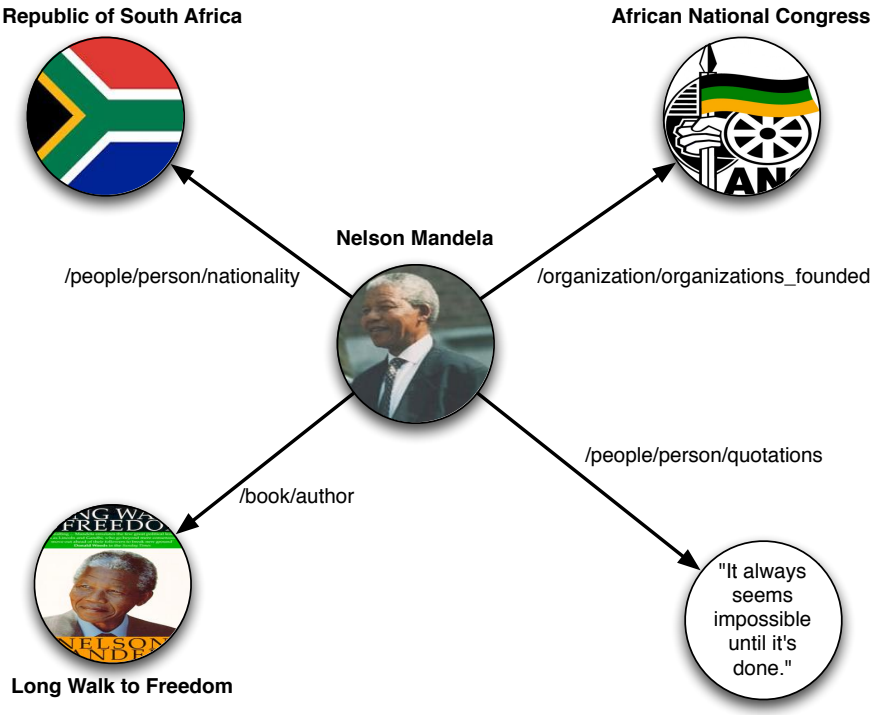


Figure 8.1: Freebase concept for "Nelson Mandela"; the concept is related to four other concepts according to the specified relationships.

suitable for the web domain.

Web-scale evaluations are often done using the TREC Web Track Task [Collins-Thompson et al., 2012], which uses the ClueWeb document collection.[3] As part of the Knowledge Graph project, Google has released a Freebase annotated version of the entire ClueWeb12 collection.[4] Also included are Freebase annotations of the TREC Web Track query topics. These annotated resources are the mapping of the free-text web documents and queries to structured Freebase entities. (They are equivalent to what MetaMap provides in the medical domain.) Using Freebase as the underlying structure, the ClueWeb12 annotated documents can be attached to the relevant nodes in the graph and retrieval can be performed using the GIN.

Compared to SNOMED CT, Freebase also provides a different type of underlying representation, one that is less definitional and more associational. Therefore, applying the GIN to web search also evaluates the model using a potentially more suited knowledge resource.

[3]ClueWeb is a crawl of approximately 1 billion webpages; http://lemurproject.org/clueweb12 (last accessed 13th June 2014).

[4]ClueWeb12 Related Data: Freebase Annotations of the ClueWeb Corpora, v1 (FACC1): http://lemurproject.org/clueweb12/FACC1 (last accessed 20th November, 2013).

8.6.3 Navigation and Visualisation using the Graph Inference Model

The graph-based representation used in the GIN can be used outside retrieval for navigation and visualisation. For navigation, the GIN could be extended to support an exploratory search interface. This may be particularly suited to situations where a user's information need is uncertain or changing; for example, in exploratory search tasks [Campbell and van Rijsbergen, 1996]. Researchers have developed specific approaches that cater for dynamic and developing information needs; these approaches are referred to as ostensive browsing [Joho et al., 2007; Leelanupab and Jose, 2008]. Using the GIN, a corpus of documents can be used to implement an ostensive browsing approach. Users can explore the document corpus, starting from an initial query node and navigating the concepts and relationships of the underlying graph. This provides users with a high-level understanding of the given domain, based on domain knowledge resource, while also providing them with access points into the documents attached to each node in the graph.

To understand better the domain and document collection, users could also be presented with visual interfaces implementing the ostensive browsing method. In this manner, the GIN graph provides the actual interface by which users navigate the system. The path users navigate via the graph can be recorded to capture an entire retrieval session. This allows users to retrace their steps or view paths that other users have taken.

The GIN provides a number of potential applications for interactive information retrieval systems, with the underlying graph structure providing a means to support navigation and visualisation.

8.6.4 Query Dependence

The implementation of the GIN presented in Chapter 6 assumed independence between query terms. However, the semantic gap problems of Inference of Similarity highlighted that a query dependent model may be advantageous. The development of models of query dependence is an active area of research in information retrieval; a common model used within the language modelling framework is the Markov Random Field method [Metzler and Croft, 2005]. It is important to note that the GIN supports a query dependent model. This is achieved by a query dependent instantiation of the $\odot$ operator in the general retrieval function (Equation 6.6). There are a number of resources that might inform query dependence within the GIN (and concept-based systems in general). MetaMap could provide a number of indicators. In fact, when a query

is mapped from free-text to concepts, it is broken first into phrases, then further into a list of candidate concepts and finally into a list of mapped concepts. (See Figure 4.1 for an example of this process.) Dependence exists between the concepts for a given phrase in that they all represent possible concept-based expressions of the phrase. Dependence also exits between different phrases as they could represent different aspects of the query. Finally, an additional source of dependence information are the semantic types (for example, disease, symptom, treatment) of the query concepts.

A simple method to encode dependence information within the GIN is to create edges in the graph between the dependent query nodes at retrieval time. (These edges would be removed when the processing of the query is complete.) This method could be used to capture dependence between the phrases of a query (as identified by MetaMap). Further research is needed to investigate this.

8.6.5 Query Reduction

A finding from the evaluation of the Graph-based Concept Weighting model was that query reduction was an effective method for improving retrieval effectiveness. Clinical queries, such as those from TREC MedTrack, are complex and verbose. Previous studies have shown that verbose queries may benefit from query reduction methods, with an upperbound of approximately 30% improvement in retrieval effectiveness if an ideal query subset is used [Kumaran and Carvalho, 2009; Bendersky and Croft, 2008]. An initial investigation into query reduction on the Bag-of-concepts model showed similar potential improvements. Query reduction may also help improve the effectiveness of the GIN, where very general query concepts provided little valuable information and may have led to the introduction of noise when the GIN traversed these concepts. An effective predictive model for query reduction using the GIN is as of yet undeveloped and remains an area for future study. Where previous query reduction methods used mainly basic corpus statistics [Kumaran and Carvalho, 2009; Bendersky and Croft, 2008], within concept-based representations or the GIN there are instead a rich set of additional features. Features such as the output from MetaMap, the semantic type of a concept or the retrieval path used in the GIN could all be used to train a predictive query reduction model.

8.6.6 Summary

The general applicability of the Graph Inference model provides a number of avenues for its application, both in the medical domain and more generally in information navigation and visualisation. The graph-based and concept-based representation used in the GIN provides more expressive power over corpus-based statistical representations. This information is potentially valuable in developing query dependence models, adaptive depth methods and query reduction models; all extending the current Graph Inference model.

Chapter 9

Conclusion

There is no real ending. It's just the place where you stop the story.

— Frank Herbert

9.1 Overview of the Research

Bridging the semantic gap involves addressing two issues: *semantics* and *inference*. To this end, three semantic search retrieval models were developed as part of this thesis: Bag-of-concepts, Graph-based Concept Weighting and Graph Inference model.

The Bag-of-concepts model (Chapter 4) focused on *semantics*. It utilised a concept-based rather than a term-based representation of queries and documents. We showed that conceptual representations differ both semantically and statistically from terms. This was as a result of three processes: term encapsulation, conflating term-variants and concept expansion. We empirically demonstrated that it was these differences that resulted in superior retrieval effectiveness using concepts. However, the Bag-of-concepts model addressed mainly vocabulary mismatch and did not account for the innate dependencies that exist between (medical) concepts.

The Graph-based Concept Weighting model (Chapter 5) extended the Bag-of-concepts model to a graph-based representation that naturally captured dependencies between concepts. In addition, the model extended previous graph-based approaches by incorporating domain knowledge that estimated the im-

portance of a concept within the global medical domain. The empirical evaluation showed that the Graph-based Concept Weighting model provided superior retrieval effectiveness. Although effective, the model still did not address all four of the major semantic gap problems. However, the evaluation did demonstrate the potential benefits from incorporating domain knowledge into the retrieval model. This motivated the development of a model that made extensive use of domain knowledge: a unified model of semantic search as inference.

In understanding how this has been achieved, we return to the original hypothesis proposed in the introduction, which was to investigate and develop:

> "A unified theoretical model of semantic search as inference, achieved by the integration of structured domain knowledge (ontologies) and statistical, information retrieval methods, provides the necessary mechanism for inference required for effective semantic search of medical data."

The *unified model of semantic search* was the Graph INference (GIN, Chapter 6). The *integration of structured domain knowledge* was achieved by a novel graph-based representation of a corpus: nodes were concepts from the ontology and edges were relationships between concepts from that ontology. The GIN utilised concept-based representations, which were shown to be effective in the Bag-of-concepts model.

The *statistical, information retrieval methods* components were provided by the probabilistic relevance estimation (in our case, using language model estimates) and by the diffusion factor, which measured the strength of association, or spread of information, between concepts in the graph-based representation of the corpus.

The *necessary mechanism for inference* was provided by the GIN as a traversal over the graph, originating from the query concepts and scoring those documents containing concepts related to the query concepts via the domain knowledge relationships. The theoretical foundations for the GIN were intuitively inspired by logic-based IR.

This thesis also provides a greater understanding of how and when inference works. Inference was needed for some queries and can provide significant benefits but was not required for other queries, where it could lead to degradation. Section 8.3 outlined the characteristics of queries that required inference: verbose queries with multiple dependent aspects, where the GIN was effective in reranking, and queries with multiple semantic gap problems and no mention of the query terms in relevant documents, where the GIN leveraged essential domain knowledge to retrieve new, relevant documents. This section also outlined the characteristics of queries that did *not* require inference: easy, unambiguous

queries, often with a small number of relevant documents. This information provides a greater understanding of how the inference mechanism was working and is valuable for both improving the models proposed here and in the development of new models of semantic search.

Determining "effective semantic search" requires empirical evaluation and empirical evaluation has had a central focus in this thesis. The TREC Medical Records Track (MedTrack) was the primary resource used in evaluating all three models. However, this test collection was created by pooling the runs from primarily keyword-based retrieval systems. Semantic search systems can fundamentally differ from keyword-based systems and return a different set of documents — those that may not contain the query terms in high frequency (or at all) but are still highly relevant. The evaluation of the GIN confirmed that it returned many documents never judged by TREC assessors. Additional assessors were recruited to judge these documents. The results showed that many of these documents were relevant and that TREC MedTrack was indeed underestimating the effectiveness of the GIN. The evaluation of the GIN also raised the broader issue of how to evaluate semantic search systems effectively. For this, we revised and proposed adaptations of previous techniques for forming the document pool (Section 8.4.1). In addition, we devised an alternative evaluation method that used manually coded medical records to generate queries and relevance judgements, thus mitigating the need to recruit human assessors (Section 7.5).

9.2 Contributions

The main contributions of this thesis are:

1. The development, analysis and evaluation of concept-based representations for medical IR. Concept-based representations differs from term-based representations and it is these differences that led to superior retrieval effectiveness, mainly by addressing vocabulary mismatch. This is provided by the Bag-of-concepts model from Chapter 4.

2. A Graph-based Concept Weighting model, which accounts for the innate dependencies that exist between medical concepts. Important concepts within a document are identified by a graph-based weighting method and important concepts within the larger medical domain are identified by incorporating a domain knowledge measure. This model is presented in Chapter 5.

3. The core theoretical contribution of this thesis: the Graph Inference model. The GIN integrates structural domain knowledge (via the graph-based representation of the corpus) and uses statistical, IR methods (node weights and diffusion factor). The GIN addresses all four major semantic gap problems. The GIN is presented in Chapter 6.

4. An empirical evaluation of all three different retrieval models: Bag-of-concepts, Graph-based Concept Weighting and Graph Inference. This provides an understanding of inference — when and why semantic search as inference succeeds and when it fails. In addition, a categorisation of the types of queries that benefit from inference and those that do not is provided. This analysis also reveals how the quality of the ontology affects retrieval and how the notion of 'definitional inference' in an ontology differs from 'retrieval inference' in an IR scenario. This is summarised in Section 8.3.

In addition, the thesis provides a number of minor contributions:

1. The identification and categorisation of the semantic gap problems and the types of inference required to overcome it. This is provided in Chapter 2.

2. An analysis and discussion on the challenges and requirements for evaluating a semantic search system, including how IR test collections developed through pooling keyword-based system underestimate the effectiveness of semantic search systems.

3. Evaluation methods specific for semantic search, including the development of a medical IR test collection that uses manually coded medical records, thus mitigating the need to recruit human assessors.

9.3 Final Remarks

This work represents a significant step forward in the integration of structured domain knowledge and data-driven information retrieval methods. This allows IR systems to exploit valuable information often trapped in domain knowledge resources. The Graph Inference model, although developed within the medical domain, is generally defined and has implications in other areas, including web search, where an emerging research trend is to utilise structured knowledge resources for more effective semantic search.

Appendix A

Converting terms to concepts

This section provides an example of the process of converting a textual, medical document into a sequence of SNOMED CT concepts. Figure A.1(a) shows the original textual document. This document is first converted to a sequence of UMLS concepts (b) by the MetaMap system [Aronson and Lang, 2010]. UMLS concepts are then mapped to SNOMED CT concepts (c) using the UMLS to SNOMED CT mapping provided as part of UMLS. The description for each of the SNOMED CT concepts is provided in Table A.1.

(a) Original medical document

```
LEFT ANKLE:
**DATE[Jul 3 07] 8:59 PM
FINDINGS: There is
moderate soft tissue
swelling. There is no
fracture or dislocation.
The ankle mortise is
intact. IMPRESSION: NO
ACUTE FRACTURE. J4 END OF
IMPRESSION
```

(b) UMLS concepts

```
C1280015
C0230448
C0011008
C0243095
C0205081
C0037580
C0016658
C0012691
C0003086
C0003087
C1283839
C0039316
C0205266
C0564590
C0205178
C0016658
C0442779
C0444930
C0442779
C1522314
C0564590
```

(c) SNOMED CT concepts

```
241784008
51636004 118573002
246188002 6736007
298349001 72704001
157257005 344001
70258002 361292008
108371006 11163003
286781002 53737009
72704001 260253008
261782000
260253008
422117008
286781002
```

Figure A.1: Example document from the BLULab corpus represented as original text, UMLS concepts and SNOMED CT concepts (Report Id: `20070703RAD-0JXYWK9UldBF-392-867771537`).

SNOMED CT Id	Preferred term
241784008	Entire left ankle (body structure)
51636004	Structure of left ankle (body structure)
118573002	Date (property) (qualifier value)
246188002	Finding (finding)
6736007	Moderate (severity modifier) (qualifier value)
298349001	Soft tissue swelling (finding)
72704001	Fracture (morphologic abnormality)
260253008	J4 (finding)
422117008	Stop (qualifier value)
286781002	Character trait finding of level of suggestibility (finding)
70258002	Ankle joint structure (body structure)
361292008	Entire ankle region (body structure)
108371006	Bone structure of tarsus (body structure)
11163003	Intact (qualifier value)
286781002	Character trait finding of level of suggestibility (finding)
53737009	Acute (qualifier value)
157257005	[Dislocations &/or sprains &/or strains] or subluxations (disorder)
261782000	End (qualifier value)
344001	Ankle region structure (body structure)

Table A.1: Concept descriptions for SNOMED CT concepts taken from Figure A.1(c).

Appendix B

Corpus-driven Measures of Semantic Similarity

This appendix[1] evaluates a number of different corpus-bases measures of semantic similarity between medical concepts. Measures of semantic similarity between medical concepts are central to a number of techniques in medical informatics, including query expansion in medical information retrieval. We evaluate the effectiveness of eight common corpus-driven measures in capturing semantic similarity and compare these against human judged concept pairs assessed by medical professionals. Our results show that certain corpus-driven measures correlate strongly (≈ 0.8) with human judgements. An important finding is that performance is significantly affected by the choice of corpus used in priming the measure, i.e., used as evidence from which corpus-driven similarities are drawn. We conclude with some guidelines for the implementation of semantic similarity measures for medical informatics and implications for medical information retrieval.

B.1 Methods

Evaluation of 8 corpus-driven measures was performed against two separate datasets of human judged medical concept pairs. An example of a concept pair is (*Congestive heart failure*, *Pulmonary edema*). Semantic similarity between concept pairs was computed using the following measures:

1. Random Indexing [Sahlgren, 2005] (RI): a technique that constructs an

[1] Previously published as Koopman et al. [2012b].

approximation of the full term-document matrix by assigning each term a unique *index* vector. The index vector is of fixed length and sparsely consists of randomly assigned -1s, 0 and 1s. Similarity was measured as the cosine angle between the index vectors of two concepts. Random Indexing was evaluated using 50, 150, 300, and 500 dimensions; results were averaged over 10 runs for each dimensional setting.

2. Latent Semantic Analysis (LSA): evaluated on 50, 150, 300, and 500 dimensions. Similarity was computed as the cosine angle between reduced concept vectors.[2]

3. Hyperspace Analogue to Language [Lund and Burgess, 1996] (HAL): constructs a full term-term co-occurrence matrix with context window of size 5[3]. Similarity was calculated as the cosine of the angle between the two HAL-based concept vectors.

4. Document Vector Cosine Similarity (DocCosine): cosine angle between concepts represented by document vectors; weighted with tf-idf.

5. Positive Pointwise Mutual Information [Bullinaria and Levy, 2007] (+PMI): variation of PMI where negative values are substituted by zero-values. Bullinaria and Levy [2007] found that negative PMI values, which correspond to a less-than-expected number of co-occurrences, indicate a poor coverage of the concepts in the corpus. This is often the case in the medical domain due to infrequently appearing concepts referring to specific diseases or rare conditions. In preliminary experiments, +PMI significantly outperformed PMI.

6. Cross Entropy Reduction [Trieschnigg et al., 2008] (CER): distance between the unigram language models of two concepts. A concept language model θ_c is defined as a distribution over concepts based on the concatenation of all documents containing concept c; background smoothing using Jelinek-Mercer.

7. Language Model + Jensen-Shannon divergence (LM JSD): unigram concept language model (constructed in the same manner as CER) but comparison was performed using standard Jensen-Shannon divergence.

8. Latent Dirichlet Allocation (LDA): topic model evaluated using 50, 150, 300 and 500 topics. Similarity between two concepts was determined

[2]Both RI and LSA were implemented using the SemanticVectors software package: http://code.google.com/p/semanticvectors

[3]Lund and Burgess [1996] found HAL was most effective with small context windows in this range.

by comparing their topic distributions $P(\text{topic}|c)$ using Jensen-Shannon divergence.

B.2 Experimental Setup

Two separate datasets of human judged concept pairs were used for evaluation. The first dataset consisted of twenty-nine[4] UMLS medical concept pairs, as developed by Pedersen et al. [2007], involving 3 physician and 9 clinical terminologists; inter-coder correlation was reported to be 0.85. A concept pair example is (*Brain tumor*, *Intracranial hemorrhage*), judged as having a similarity of 2.0 on a scale of 1.0 (unrelated) to 4.0 (synonymous). We refer to this dataset as Ped. The second dataset, from Caviedes and Cimino [2004], contained forty-five MeSH/UMLS concept pairs[5] judged by three physicians on a scale of 1 to 10; Cavides and Cimino reported "consensus" amongst judges, but no precise value was reported. This dataset is referred to as Cav.

Two separate corpora were used as data to prime each corpus-driven method. The first corpus was MedTrack, a collection of 100,866 clinical record documents used in the TREC 2011 Medical Records Track. Documents belonging to a single patient's admission were treated as sub-documents and were concatenated together into a single document called a patient *visit* document. The corpus then contained 17,198 patient visit documents. This was done to encapsulate the closely related content of different reports (e.g. pathology report and surgical report) belonging to the same patient admission[6]. The second corpus used was OHSUMED, a MEDLINE subset consisting of 348,566 medical journal abstracts, as used in TREC 2000 Filtering Track. Statistics for each corpus are provided in Table B.1.

Corpus	#Docs	Avg. doc. len.	#Vocab.
MedTrack	17,198*	932	54,546
OHSUMED	293,856	100	55,390

*100,866 original reports collapsed to 17,198 patient *visit* documents.

Table B.1: Collection statistics of the test corpora: MedTrack, collection of clinical patient records; and OHSUMED, MEDLINE abstracts.

[4]One concept pair (*Lymphoid hyperplasia*) was removed from Pedersen's original 30 as it was not found in our test collections.

[5]10 pairs containing the concept C0030631, not present in the test corpus, were removed.

[6]Collapsing reports to patient visits was a common practise among many TREC MedTrack participants [Voorhees and Tong, 2011].

For both corpora, the original textual documents were translated into UMLS concept identifiers using MetaMap, the biomedical concept identification system [Aronson and Lang, 2010]. After processing, the individual documents contained only UMLS concept ids; for example, the phrase *Congestive heart failure* in the original document will be replaced with `C0018802` in the new document. More details of this approach are provided in [Koopman et al., 2012a]. Both test datasets, Ped and Cav, contained UMLS concept pairs (which may actually represent term phrases rather than single terms); converting the test corpora to concepts thus allows direct comparison of the single concept pairs contained in the two datasets.

Each of the 8 models outlined in the Methods section provides a representation of a concept; for example, in DocCosine a concept is a vector based on the documents that the given concept appears in. Similarity can be determined by comparing the representations of two concepts. For each similarity measure, comparison was made against human judges for each dataset (Ped and Cav) using Pearson's correlation coefficient.

B.3 Results & Discussion

Results showing the correlation coefficient against human judges for each corpus-driven method are reported in Figure B.1. The x-axis is ordered by decreasing correlation averaged across all datasets/corpora[7].

The first observation we make is that similar types of measures demonstrate similar results: the three probabilistic language model measures, +PMI, CER

[7] LDA (avg.) is the average for LDA across 50, 150, 300, 500 topics, all of which exhibit almost equivalent results.

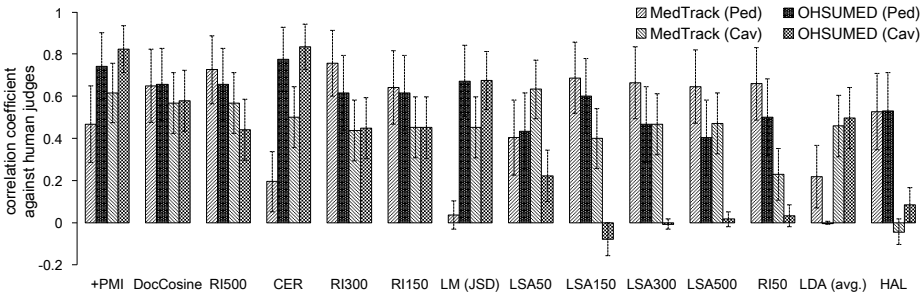


Figure B.1: Correlation coefficient against human judged similarity for each corpus-driven semantic similarity measure. Judgements made against two gold standard datasets (Ped & Cav) using two corpora (MedTrack & OHSUMED). x-axis ordered by decreasing correlation averaged across all datasets/corpora; error bars signify confidence interval at 95%.

and LM (JSD), exhibit comparable performance profiles across datasets / corpora. Similarly, the vector-based measures (RI and LSA and DocCosine) exhibit similar profiles between each other and across different dimensions.

Considering the best performing measures, Table B.2 provides a breakdown of the top 3 semantic similarity measures for each dataset / corpus.

	Dataset	
Corpus	Ped	Cav
MedTrack	RI300, LSA150, DocCosine	LSA50, +PMI, DocCosine
OHSUMED	CER, +PMI, LM/DocCosine	CER, +PMI, LM

Table B.2: Top 3 semantic similarity measures for each corpus and dataset.

Consensus is observed between the two datasets Ped and Cav. However, the best measure differs significantly between the two corpora. In general, vector-based measures perform best when primed with the MedTrack corpus, while probabilistic measures are most effective primed with OHSUMED. This may be explained by the different characteristics of the two corpora: MedTrack contains detailed clinical notes from patient encounters, whereas OHSUMED contains MEDLINE article abstracts. As a result, the *scope* of concepts found in a document differs between the two collections. Clinical notes relating to a patient's admission may cover a wide range of different concepts, especially if they have been admitted with multiple conditions or for a lengthy period. In contrast, journal abstracts are descriptions of a particular topic and are therefore typically narrower in scope. The probabilistic measures use the whole document as the "context window" for determining co-occurrence, OHSUMED's documents of narrower scope therefore offer more precise context windows, whereas the wider scoped MedTrack documents may contain more noise. In addition to the nature of the documents found in each corpus, the average *document length* differs considerably — MedTrack documents are about an order of magnitude larger (Table B.1). Intuitively, longer documents will, in general, cover more topics and be wider in scope. The vector-based measures benefit from the additional context found in the longer documents, which is in contrast to the probabilistic measures.

The nature of the *language* also differs between the two corpora. MEDLINE abstracts contain precise descriptions of a particular topic, whereas clinical records are often terse narratives with considerable jargon and shorthand — and in some cases typographic errors.

Given the differences in *scope*, *document length* and *language* of the two corpora, we could hypothesise that OHSUMED appears a higher quality corpus

for similarity judgements and that measures primed with MedTrack would exhibit degraded performance. However, the results do not affirm this hypothesis. Probabilistic measures primed with OHSUMED display excellent results; however, the longer, less consistent documents found in MedTrack still provide good evidence for similarity judgements when used with vector-based methods.

Table B.2 also highlights the robustness of +PMI and DocCosine, which both occupy three out of four cells. The traditional IR measure of DocCosine, although not producing the best results on a single test, is particularly stable across both corpora and datasets. Both +PMI and DocCosine are simple and computationally efficient, making them more attractive than more computationally intensive measures such as LSA and language model-based measures. Certain measures may perform well on one particular collection / dataset, but have poor performance on others — LM (JSD), LDA and HAL all exhibit this behaviour.

More generally, the results reaffirm the findings of Pedersen et al. that corpus-driven approaches outperform path-based measures, which failed to yield a correlation greater than 0.5[8]. Additionally, our findings using *vector-based* measures are in line with Petersen et al. who reported a 0.69 correlation obtained using their *Context Vector* measure on the Mayo Clinic Corpus of Clinical Notes; our vector-based measure results using MedTrack were ≈ 0.7. MedTrack and the Mayo Clinic Corpus are of similar size and nature (both being clinical records)[9].

An outcome of this study is a set of guidelines for the implementation of corpus-based semantic similarity measures for medical text:

1. The choice of corpus used to prime the similarity measure is an important consideration that may significantly affect the performance of the particular measure.

2. More specifically, the characteristics of individual documents should be considered. If documents cover a range of topics, vector-based measures are preferable whereas if they are smaller in scope, probabilistic methods are then preferred. Average document length can be an indicator of scope — large documents typically cover more topics. Additionally, the type of language (e.g., clinical notes vs. medical literature) should be taken into consideration.

3. +PMI and DocCosine are robust across collections and datasets and have the added advantage of being computationally efficient. As other meas-

[8]Path-based measures are *corpus independent*, based on the UMLS network. As such, Pedersen's results can be used for a direct comparison in our study.

[9]Note that the Mayo Clinic Corpus of Clinical Notes corpus is not publicly available.

ures may perform well on certain collection / datasets, but can perform extremely poorly in certain cases, it may be best to avoid these measures.

4. When implementing a semantic similarity on a particular corpus, the two datasets can be used to find a measure most appropriate to the nature of the corpus documents. Both Ped and Cav are publicly available.

The reported findings may have important impacts for medical information retrieval, specifically for systems making significant use of query expansion and relevance feedback, as was the case with participants of TREC MedTrack. Firstly, the effectiveness of corpus-based query expansion varied significantly between participants of TREC MedTrack — some techniques showed gains, while others degraded performance. Although a number of factors affect query expansion performance, a poor semantic similarity measure could certainly be a major contributor. The most appropriate similarity measure, based on the findings of this study, should be considered when employing corpus-based query expansion.

Finally, having highlighted the choice of corpus as an important consideration, we conjecture that in some cases it may be advantageous to prime the similarity measure with a separate corpus from the one being used for retrieval. For example, when searching medical literature (e.g. OHSUMED), priming with clinical records (e.g. those found in MedTrack) may increase effectiveness. In the literature there is evidence supporting the use of Wikipedia as a background priming corpus [Bendersky et al., 2011]. An in-depth evaluation of this aspect is left to future work.

B.4 Conclusion

In this chapter we evaluated eight different corpus-driven approaches to determining the semantic similarity between medical concepts. Corpus-driven approaches exhibited strong correlations (up to ≈ 0.8) with human judged concept pairs provided by medical professionals. Our findings showed that the choice of corpus used to prime the similarity measure significantly affected performance. We provided a number of guidelines for the use of semantic similarity measures that included consideration of document scope, length and language. Simple measures such as +PMI and DocCosine demonstrated effective and robustness results across evaluations. This work provided an in-depth review of corpus-driven semantic similarity measures, a technique central to medical informatics.

Appendix C

SNOMED CT Relationship Type Weights used in the Diffusion Factor

The weights manually assigned to each SNOMED CT relationship type and used as part of the relationship type component of the diffusion factor. See Section 6.4.1.

Appendix C: SNOMED CT Relationship Type Weights used in the Diffusion Factor

Relationship Id	Description	Weight
116676008	Associated morphology	0.6
116680003	Is a	1.0
116686009	Has specimen	0.6
118168003	Specimen source morphology	0.6
118169006	Specimen source topography	0.6
118170007	Specimen source identity	0.6
118171006	Specimen procedure	0.6
123005000	Part of	0.8
127489000	Has active ingredient	1.0
149016008	MAY BE A	0.6
159083000	WAS A	0.8
168666000	SAME AS	1.0
246075003	Causative agent	1.0
246090004	Associated finding	0.6
246093002	Component	0.8
246112005	Severity	0.2
246454002	Occurrence	0.6
246456000	Episodicity	0.6
246513007	Revision status	0.2
255234002	After	0.4
260507000	Access	0.4
260686004	Method	0.4
260870009	Priority	0.2
263502005	Clinical course	0.8
272741003	Laterality	0.2
363589002	Associated procedure	0.6
363698007	Finding site	0.6
363699004	Direct device	0.8
363700003	Direct morphology	0.6
363701004	Direct substance	0.8
363702006	Has focus	0.4
363703001	Has intent	0.4
363704007	Procedure site	0.4
363705008	Has definitional manifestation	0.8

Appendix C: SNOMED CT Relationship Type Weights used in the Diffusion Factor

Relationship Id	Description	Weight
363709002	Indirect morphology	0.6
363710007	Indirect device	0.6
363713009	Has interpretation	0.6
363714003	Interprets	0.6
370124000	REPLACED BY	1.0
370125004	MOVED TO	1.0
370129005	Measurement method	0.4
370130000	Property	0.2
370131001	Recipient category	0.2
370132008	Scale type	0.2
370133003	Specimen substance	0.4
370135005	Pathological process	0.4
405813007	Procedure site - Direct	0.6
405814001	Procedure site - Indirect	0.4
405815000	Procedure device	0.4
405816004	Procedure morphology	0.4
408729009	Finding context	0.4
408730004	Procedure context	0.4
408731000	Temporal context	0.2
408732007	Subject relationship context	0.6
410675002	Route of administration	0.6
411116001	Has dose form	0.4
418775008	Finding method	0.4
419066007	Finding informer	0.2
424226004	Using device	0.4
424244007	Using energy	0.4
424361007	Using substance	0.4
424876005	Surgical approach	0.6
425391005	Using access device	0.6
42752001	Due to	0.6
47429007	Associated with	0.6

Table C.1: Manually assigned weights for SNOMED CT relationship as used in the diffusion factor.

Appendix D

TREC Medical Records Track Queries

List of query topics and their keywords used in the TREC 2011 and 2012 Medical Records Track [Voorhees and Tong, 2011; Voorhees and Hersh, 2012].

101 `Patients with hearing loss`

102 `Patients with complicated GERD who receive endoscopy`

103 `Hospitalized patients treated for methicillin resistant Staphylococcus aureus MRSA endocarditis`

104 `Patients diagnosed with localized prostate cancer and treated with robotic surgery`

105 `Patients with dementia`

106 `Patients who had positron emission tomography PET magnetic resonance imaging MRI or computed tomography CT for staging or monitoring of cancer`

107 `Patients with ductal carcinoma in situ DCIS`

108 `Patients treated for vascular claudication surgically`

109 `Women with osteopenia`

110 `Patients being discharged from the hospital on hemodialysis`

111 `Patients with chronic back pain who receive an intraspinal pain medicine pump`

112 `Female patients with breast cancer with mastectomies during admission`

113 `Adult patients who received colonoscopies during admission which revealed adenocarcinoma`

114 `Adult patients discharged home with palliative care home hospice`

115 `Adult patients who are admitted with an asthma exacerbation`

116 `Patients who received methotrexate for cancer treatment while in the hospital`

117 `Patients with Post traumatic Stress Disorder`

Appendix D: TREC Medical Records Track Queries

118 Adults who received a coronary stent during an admission

119 Adult patients who presented to the emergency room with with anion gap acidosis secondary to insulin dependent diabetes

120 Patients admitted for treatment of CHF exacerbation

121 Patients with CAD who presented to the Emergency Department with Acute Coronary Syndrome and were given Plavix

122 Patients who received total parenteral nutrition while in the hospital

123 Diabetic patients who received diabetic education in the hospital

124 Patients who present to the hospital with episodes of acute loss of vision secondary to glaucoma

125 Patients co infected with Hepatitis C and HIV

126 Patients admitted with a diagnosis of multiple sclerosis

127 Patients admitted with morbid obesity and secondary diseases of diabetes and or hypertension

128 Patients admitted for hip or knee surgery who were treated with anti coagulant medications post op

129 Patients admitted with chest pain and assessed with CT angiography

131 Patients who underwent minimally invasive abdominal surgery

132 Patients admitted for surgery of the cervical spine for fusion or discectomy

133 Patients admitted for care who take herbal products for osteoarthritis

134 Patients admitted with chronic seizure disorder to control seizure activity

135 Cancer patients with liver metastasis treated in the hospital who underwent a procedure

136 Children with dental caries

137 Patients with inflammatory disorders receiving TNF inhibitor treatments

139 Patients who presented to the emergency room with an actual or suspected miscarriage

140 Patients who developed disseminated intravascular coagulation in the hospital

141 Adult inpatients with Alzheimer s disease admitted from nursing homes with pressure ulcers

142 Patients admitted with Hepatitis C and IV drug use

143 Patients who have had a carotid endarterectomy

144 Patients with diabetes mellitus who also have thrombocytosis

145 Patients with lupus nephritis and thrombotic thrombocytopenic purpura

146 Patients treated for post partum problems including depression hypercoagulability or cardiomyopathy

147 Patients with left lower quadrant abdominal pain

148 Patients acutely treated for migraine in the emergency department

149 Patients with delirium hypertension and tachycardia

150 Patients who have cerebral palsy and depression

151 Patients with liver disease taking SSRI antidepressants

152 Patients with Diabetes exhibiting good Hemoglobin A1c Control 8 0

Appendix D: TREC Medical Records Track Queries

153 Patients admitted to the hospital with end stage chronic disease who are offered hospice care

154 Patients with Primary Open Angle Glaucoma POAG

155 Heart Failure HF Beta Blocker Therapy for Left Ventricular Systolic Dysfunction LVSD

156 Patients with depression on antidepressant medication

157 Patients admitted to hospital with symptomatic cervical spine lesions

158 Patients with esophageal cancer who develop pericardial effusion

160 Patients with Low Back Pain who had Imaging Studies

161 Patients with adult respiratory distress syndrome

162 Patients with hypertension on antihypertensive medication

163 Patients treated for lower extremity chronic wound

164 Adults under age 60 undergoing alcohol withdrawal

165 Patients who have gluten intolerance or celiac disease

166 Patients who have hypoaldosteronism and hypokalemia

167 Patients with AIDS who develop pancytopenia

168 Patients with Coronary Artery Disease with Prior Myocardial Infarction on Beta Blocker Therapy

169 Elderly patients with subdural hematoma

170 Adult patients who presented to the emergency room with suicide attempts by drug overdose

171 Patients with thyrotoxicosis treated with beta blockers

172 Patients with peripheral neuropathy and edema

173 Patients over 65 who had Pneumonia Vaccination Status presently or previously

174 Elderly patients with ventilator associated pneumonia

175 Elderly patients with endocarditis

176 Patients with Heart Failure HF on Angiotensin Converting Enzyme ACE Inhibitor or Angiotensin Receptor Blocker ARB Therapy for Left Ventricular Systolic Dysfunction LVSD

177 Patients treated for depression after myocardial infarction

178 Patients with metastatic breast cancer

179 Patients taking atypical antipsychotics without a diagnosis schizophrenia or bipolar depression

180 Patients with cancer who developed hypercalcemia

181 Patients being evaluated for secondary hypertension

182 Patients with Ischemic Vascular Disease

183 Patients presenting to the emergency room with acute vision loss

184 Patients with Colon Cancer who had Chemotherapy

185 Patients who develop thrombocytopenia in pregnancy

Bibliography

A. R. Aronson and F.-M. Lang. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3): 229–236, 2010. 3.4.1, 3.5, 4.1.1, A, B.2

S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. Dbpedia: A nucleus for a web of open data. In *The Semantic Web*, pages 722–735. Springer, 2007. 8.3.1

L. Azzopardi and D. E. Losada. An efficient computation of the multiple-bernoulli language model. In *Advances in Information Retrieval*, pages 480–483. Springer, 2006. 3.3.1

M. Bendersky and W. B. Croft. Discovering key concepts in verbose queries. In *Proceedings of the 31st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 491–498, Singapore, 2008. 5.5, 8.6.5

M. Bendersky, D. Metzler, and W. B. Croft. Parameterized concept weighting in verbose queries. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 605–614, Beijing, China, 2011. B.3

A. Biswas, S. Mohan, A. Tripathy, J. Panigrahy, and R. Mahapatra. Semantic Key for Meaning Based Searching. In *IEEE International Conference on Semantic Computing*, Berkeley, CA, USA, 2009. 3.2.2

C. Bizer, T. Heath, and T. Berners-Lee. Linked data-the story so far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22, 2009. 8.3.1

R. Blanco and C. Lioma. Graph-based term weighting for information retrieval. *Information Retrieval*, 15(1):1–39, 2012. 3.3.1, 5.1, 5.2, 5.2, 5.4.1, 5.5

K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, pages 1247–1250, Vancouver, Canada, 2008. 8.3.1, 8.6.2

F. Boudin, J.-Y. Nie, and M. Dawes. Using a medical thesaurus to predict query difficulty. In *Proceedings of the 35th European Conference in Information Retrieval*, pages 480–484, Moscow, Russia, 2012. 6.6.1, 6.6.1, 8.6.1

C. Buckley and E. M. Voorhees. Retrieval evaluation with incomplete information. In *Proceedings of the 27th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 25–32, Sheffield, UK, 2004. 3.3.2

J. A. Bullinaria and J. P. Levy. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39(3): 510, 2007. 5

I. Campbell and C. J. van Rijsbergen. The ostensive model of developing information needs. In *Proceedings of the 3rd International Conference on Conceptions of Library and Information Science*, pages 251–268, 1996. 8.6.3

B. Carterette, V. Pavlu, E. Kanoulas, J. A. Aslam, and J. Allan. Evaluation over thousands of queries. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 651–658, Singapore, 2008. 7.2.1

J. E. Caviedes and J. J. Cimino. Towards the development of a conceptual distance metric for the UMLS. *Journal of Biomedical Informatics*, 37(2):77–85, April 2004. B.2

W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan. Evaluation of negation phrases in narrative clinical reports. In *Proceedings of the AMIA Symposium*, page 105. American Medical Informatics Association, 2001. 2.5.1

C. W. Cleverdon. The significance of the Cranfield tests on index languages. In *Proceedings of the 14th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3–12, Chicago, USA, 1991. 3.3.2, 3.3.2

T. Cohen and D. Widdows. Empirical distributional semantics: Methods and biomedical applications. *Journal of Biomedical Informatics*, 42(2):390–405, 2009. 3.2.2

K. Collins-Thompson, P. N. Bennett, , and F. D. adn Charles L. A. Clarke. Overview of the TREC 2012 Web Track. In *Proceedings of 22nd Text REtrieval Conference (TREC 2013)*, Gaithersburg, U.S.A, November 2012. 8.6.2

I. o. M. Committee on Comparative Effectiveness Research Prioritization. *Initial National Priorities for Comparative Effectiveness Research*. The National Academies Press, 2009. 4.3.1

G. V. Cormack, C. R. Palmer, and C. L. A. Clarke. Efficient construction of large test collections. In *Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrievall*, pages 282–289, Melbourne, Australia, 1998. 8.4.1

F. Crestani. *A Study of Probability Kinematics in Information Retrieval*. PhD thesis, Department of Computer Science, University of Glasgow, 1998. 6.1.1, 6.1.2

F. Diaz and D. Metzler. Improving the estimation of relevance models using large external corpora. In *Proceedings of the 29th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 154–161, 2006. 3.3, 3.3.2

H. Dong, F. K. Hussain, and E. Chang. A survey in semantic search technologies. In *IEEE International Conference on Digital Ecosystems and Technologies (DEST)*, Phitsanulok, Thailand, 2008. IEEE. 3.2.2, 3.2.2

T. E. Doszkocs, J. Reggia, and X. Lin. Connectionist models and information retrieval. *Annual Review of Information Science and Technology*, 25:209–262, 1990. 3.3.1

O. Egozi, S. Markovitch, and E. Gabrilovich. Concept-Based Information Retrieval using Explicit Semantic Analysis. *ACM Transactions on Information Systems*, 29 (2):1–38, 2011. 3.4.1, 3.4.1

W.-D. Fang, L. Zhang, Y.-X. Wang, and S.-B. Dong. Toward a semantic search engine based on ontologies. In *International Conference on Machine Learning and Cybernetics*, volume 3, Guangzhou, China, 2005. 3.2.2

M. Frixione and A. Lieto. Representing concepts in formal ontologies: Compositionality vs. typicality effects. *Logic and Logical Philosophy*, 21(4):391–414, 2012. 3.1, 3.2.1, 3.2.2, 6.6.1, 6.7, 8.3.1, 2

P. Gärdenfors. Symbolic, conceptual and subconceptual representations. In V. Cantoni, V. di Ges, A. Setti, and D. Tegolo, editors, *Human and Machine Perception: Information Fusion*, pages 255–270. Plenum Press, New York, 1997. 3.1

P. Gärdenfors. How to Make the Semantic Web More Semantic. In A. C. Varzi and L. Vieu, editors, *Formal Ontology in Information Systems: proceedings of the third international conference (FOIS-2004)*, volume 114 of *Frontiers in Artificial Intelligence and Applications*, pages 17–34. IOS Press, 2004. 3.2.2

L. Q. Ha, E. I. Sicilia-Garcia, J. Ming, and F. J. Smith. Extension of Zipf's law to words and phrases. In *Proceedings of the 19th International Conference on Computational Linguistics*, pages 1–6, 2002. 4.2.2

C. Hauff, D. Hiemstra, and F. de Jong. A survey of pre-retrieval query performance predictors. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 1419–1420. ACM, 2008. 6.6.1, 8.6.1

W. R. Hersh and D. Hickam. Information retrieval in medicine: the SAPHIRE experience. *Journal of the American Society for Information Science*, 46(10):743–747, January 1995. 3.4.1

W. Hersh. *Information retrieval: a health and biomedical perspective*. Springer Verlag, New York, 3rd edition, 2009. 4.1.1

W. R. Hersh, E. Pattison-Gordon, D. Evans, and R. Greenes. Adaptation of meta-1 for saphire, a general purpose information retrieval system. In *Proceedings of Annual Symposium on Computer Application Medical Care*, pages 156–160, Washington D.C., USA, 1990. 3.4.1

J. R. Herskovic, T. Cohen, D. Subramanian, M. S. Iyengar, J. W. Smith, and E. V. Bernstam. Medrank: Using graph-based concept ranking to index biomedical texts. *International Journal of Medical Informatics*, 80(6):431 – 441, 2011. 5.3.2

J. R. Herskovic, D. Subramanian, T. Cohen, P. A. Bozzo-Silva, C. F. Bearden, and E. V. Bernstam. Graph-based signal integration for high-throughput phenotyping. *BMC bioinformatics*, 13(Suppl 13):S2, 2012. 6.6

D. Hiemstra. A linguistically motivated probabilistic model of information retrieval. In *Research and advanced technology for digital libraries*, pages 569–584. Springer, 1998. 3.3.1

E. Hovy. A Future for IR: Beyond Question Answering and Text Summarization. Keynote address, SIGIR Conference. New Orleans, USA, 2001. 3.1

H. Joho, R. D. Birbeck, and J. M. Jose. An ostensive browsing and searching on the web. In *Proceedings of the 2nd International Workshop on Context-based Information Retrieval*, Copenhagen, Denmark, 2007. 8.6.3

M. Y. Kim, Q. Dou, O. R. Zaiane, and R. Goebel. Unsupervised mapping of sentences to biomedical concepts based on integrated information retrieval model and clustering. In *Proceedings of the 1st ACM International Conference on Bioinformatics and Computational Biology*, Niagara Falls, USA, 2010. 4.1.1

B. King, L. Wang, I. Provalov, and J. Zhou. Cengage learning at TREC 2011 medical track. In *Proceedings of 20th Text REtrieval Conference (TREC 2011)*, Gaithersburg, USA, 2011. 3.3.2

B. Koopman, P. Bruza, L. Sitbon, and M. Lawley. Analysis of the effect of negation on information retrieval of medical data. In *Proceedings of the Fifteenth Australasian Document Computing Symposium (ADCS)*, pages 89–92, Melbourne, Australia, December 2010. 2.5.1

B. Koopman, P. Bruza, L. Sitbon, and M. Lawley. Evaluating medical information retrieval. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1139–1140, Beijing, China, 2011. 7.5.2

B. Koopman, P. Bruza, L. Sitbon, and M. Lawley. Towards Semantic Search and Inference in Electronic Medical Records: an approach using Concept-based Information

Retrieval. *Australasian Medical Journal: Special Issue on Artificial Intelligence in Health*, 5(9):482–488, 2012a. 3.4.1, B.2

B. Koopman, G. Zuccon, P. Bruza, L. Sitbon, and M. Lawley. An Evaluation of Corpus-driven Measures of Medical Concept Similarity for Information Retrieval. In *21st ACM International Conference on Information and Knowledge Management (CIKM)*, Maui, USA, 2012b. 3.2.2, 6.6.1, 1

G. Kumaran and V. R. Carvalho. Reducing long queries using query quality predictors. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 564–571, Boston, USA, 2009. 5.5, 8.6.5

F. W. Lancaster. *Vocabulary Control for Information Retrieval*. Arlington, Virginia, Arlington, Virginia, 2nd edition, 1986. 2.1

M. Lawley and C. Bousquet. Fast classification in Protégé: Snorocket as an OWL 2 EL reasoner. In *Proceedings of the Australasian Ontology Workshop*, pages 45–50, Adelaide, Australia, 2010. 6.6.1

T. Leelanupab and J. M. Jose. An adaptive browsing-based approach for creating a photographic story. In *Proceedings of the 3th International Conference on Semantic and Digital Media Technologies*, Koblenz, Germany, 2008. 8.6.3

D. D. Lewis. An evaluation of phrasal and clustered representations on a text categorization task. In *Proceedings of the 15th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 37–50, Copenhagen, Denmark, 1992. 7.5.2

N. Limsopatham, C. Macdonald, R. McCreadie, and I. Ounis. Exploiting Term Dependence while Handling Negation in Medical Search. In *Proceedings of the 35th International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 1065–1066, Oregon, USA, 2012. ACM. 2.5.1

N. Limsopatham, C. Macdonald, and I. Ounis. Aggregating Evidence from Hospital Departments to Improve Medical Records Search. In *Proceedings of the 35th European Conference on Information Retrieval (ECIR)*, pages 279–291, Moscow, Russia, 2013a. 4.4.1

N. Limsopatham, C. Macdonald, and I. Ounis. A Task-Specific Query and Document Representation for Medical Records Search. In *Proceedings of the 35th European Conference on Information Retrieval (ECIR)*, Moscow, Russia, 2013b. 5

K. Liu, W. R. Hogan, and R. S. Crowley. Natural Language Processing methods and systems for biomedical ontology learning. *Journal of Biomedical Informatics*, 44(1): 163–79, 2011. 3.4.1

Z. Liu and W. W. Chu. Knowledge-based query expansion to support scenario-specific retrieval of medical free text. *Information Retrieval*, 10(2):173–202, 2007. 3.4.1

H. P. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165, 1958. 4.2.2

T. Lukasiewicz and U. Straccia. Managing uncertainty and vagueness in description logics for the semantic web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(4):291–308, 2008. 3.2.2

K. Lund and C. Burgess. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instrumentation, and Computers*, 28 (2):203–208, 1996. 3

C. Mangold. A survey and classification of semantic search approaches. *International Journal of Metadata, Semantics and Ontologies*, 2(1):23–34, 2007. 3.2.2, 3.2.2

C. D. Manning, P. Raghavan, and H. Schtze. *Introduction to Information Retrieval*. Cambridge University Press, 2008. 3.3.2

E. Meij, D. Trieschnigg, M. de Rijke, and W. Kraaij. Conceptual language models for domain-specific retrieval. *Information Processing & Management*, 46(4):448–469, July 2010. 3.4.1

D. Metzler and W. Croft. A Markov random field model for term dependencies. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 472–479, Salvador, Brazil, 2005. ACM. 3.3.1, 3.3.2, 4.4.2, 5.1, 8.6.4

S. Meystre and P. J. Haug. Natural language processing to extract medical problems from electronic clinical documents: Performance evaluation. *Journal of Biomedical Informatics*, 39(6):589–599, 2006. 3.4.1, 4.1.1

P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman. Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, 18(5):544–551, August 2011. 4.1.1

J. Nie. An information retrieval model based on modal logic. *Information Processing & Management*, 25(5):477–491, January 1989. (document), 6.1.1, 6.1, 6.1.1

L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: bringing order to the web. *Technical Report, Stanford Digital Library Technologies*, 1999. 3.3.1, 5.1, 5.2

C. Patel, J. Cimino, J. Dolby, A. Fokoue, A. Kalyanpur, A. Kershenbaum, L. Ma, E. Schonberg, and K. Srinivasclass. Matching patient records to clinical trials using ontologies. *The Semantic Web*, 4825:816–829, 2007. 1, 3.1.1

T. Pedersen, S. V. S. Pakhomov, S. Patwardhan, and C. G. Chute. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics*, 40(3):288–299, 2007. 3.2.2, 6.1.2, B.2

J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 275–281. Melbourne, Australia, 1998. 3.3.1

W. Pratt and M. Yetisgen-Yildiz. A study of biomedical concept identification: MetaMap vs. people. In *Proceedings of American Medical Informatics Association Symposium (AMIA)*, pages 529–533, January 2003. 3.4.1, 4.1.1, 4.1.1, 4.4.3

W. V. O. Quine. Natural kinds. In *Ontological Relativity and Other Essays*, pages 114–138. Columbia University Press, New York, 1969. 3.2.2

D. Ravindran and S. Gauch. Exploiting hierarchical relationships in conceptual search. In *Proceedings of the 13th annual international ACM CIKM conference on information and knowledge management*, pages 238–239, Washington, DC, USA, 2004. 3.4.1, 3.4.1

S. E. Robertson and S. Walker. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 232–241, Dublin, Ireland, 1994. 3.3.1

S. E. Robertson. The probability ranking principle in ir. *Journal of documentation*, 33(4):294–304, 1977. 3.3

M. Sahlgren. An introduction to random indexing. In *Proceedings of the 7th International Conference on Terminology and Knowledge Engineering*, pages 1–9, Leipzig, Germany, 2005. 1

G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, November 1975. 3.3.1

G. Salton. *Automatic Information Organization and Retrieval.* McGraw Hill Text, 1968. 3.3, 8.3

M. Sanderson and H. Joho. Forming test collections with no system pooling. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 33–40, Sheffield, United Kingdom, 2004. 7.5.2, 8.4.1

A. Sheth, C. Ramakrishnan, and C. Thomas. Semantics for the semantic web: the implicit, the formal and the powerful. *International Journal on Semantic Web and Information Systems*, 1:1–18, 2005. 3.2, 3.2.2

A. Singhal. Introducing the knowledge graph: things, not strings. Technical report, Google Inc., May 2012. 8.3.1, 8.6.2

J. F. Sowa et al. *Knowledge representation: logical, philosophical, and computational foundations*, volume 13. MIT Press, 2000. 2.3

K. Spackman. SNOMED Clinical Terms Basics. International Health Terminology Standards Development Organisation Technical Report, August 2008. 3.2.1, 6.6.1

N. Stokes, Y. Li, L. Cavedon, and J. Zobel. Exploring criteria for successful query expansion in the genomic domain. *Information Retrieval*, 12(1):17–50, October 2008. 3.4.1

H. Suominen, S. Salantera, S. Velupillai, W. W. Chapman, G. Savova, N. Elhadad, S. Pradhan, B. R. South, D. L. Mowery, G. J. F. Jones, J. Leveling, L. Kelly, L. Goeuriot, D. Martinez, and G. Zuccon. Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. In *Advances in Multilingual and Multimodal Information Retrieval*. Springer Berlin Heidelberg, 2013. 4.4.3, 8.5

M. Symonds, G. Zuccon, B. Koopman, P. Bruza, and A. N. Nguyen. Semantic Judgement of Medical Concepts: Combining Syntagmatic and Paradigmatic Information using the Tensor Encoding Model. In *Proceedings of the Australasian Language Technology Workshop*, Dunedin, New Zealand, 2012. 6.6.1

D. Trieschnigg. *Proof of concept: concept-based biomedical information retrieval*. PhD thesis, University of Twente, 2010. 3.4.1

D. Trieschnigg, E. Meij, M. de Rijke, and W. Kraaij. Measuring concept relatedness using language models. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 823–824, Singapore, 2008. 6

D. Trieschnigg, D. Hiemstra, F. de Jong, and W. Kraaij. A cross-lingual framework for monolingual biomedical information retrieval. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 169–178, Toronto, Canada, 2010. 3.4.1, 3.4.1

H. Turtle and W. B. Croft. Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9(3):187–222, July 1991. 3.3.1, 5.1

V. Uren, M. Sabou, E. Motta, M. Fernandez, V. Lopez, and Y. Lei. Reflections on five years of evaluating semantic search systems. *International Journal on Metadata Semantics and Ontologies*, 5(2):87–98, 2010. 8.4.2

M. Uschold and M. Gruninger. Ontologies: Principles, methods and applications. *Knowledge engineering review*, 11(2):93–136, 1996. 3.2.2

C. J. Van Rijsbergen. A non-classical logic for information retrieval. *Computer Journal*, 29(6):481–485, 1986. 3.3, 6.1.1

C. J. van Rijsbergen. Another Look at the Logical Uncertainty Principle. *Information Retrieval*, 2(1):17–26, February 2000. 6.1.1

K. van Rijsbergen. *Information Retrieval*. Butterworth & Co, London, 2nd edition, 1979. 4.2.2, 5.4.2

E. Voorhees and D. K. Harman. *TREC : experiment and evaluation in information retrieval*. MIT Press, Cambridge, Mass., 2005. 3.3.2, 8.4.1

E. M. Voorhees. Query expansion using lexical-semantic relations. In *Proceedings of the 17th annual international ACM SIGIR conference on research and development in information retrieval*, pages 61–69, Dublin, Ireland, 1994. 3.4.1

E. M. Voorhees and W. Hersh. Overview of the TREC 2012 Medical Records Track. In *Proceedings of 21st Text REtrieval Conference (TREC 2012)*, 2012. 3.3.2, 4.3.1, 4.3.1, 6.6.2, 7.1, 7.3.1, 7.3.2, D

E. M. Voorhees and R. M. Tong. Overview of the TREC 2011 Medical Records Track. In *Proceedings of the Twentieth Text REtrieval Conference (TREC 2011)*, Gaithersburg, Maryland, USA, November 2011. 3.3.2, 3.3.2, 4.3.1, 6.6.2, 7.1, 7.3.1, 7.3.2, 6, D

W. A. Woods. Meaning and links: A semantic odyssey. In *Proceedings of the 9th Internatonal Conference on the Principles of Knowledge Representation and Reasoning*, pages 740–742, 2004. 3.2.2

S. T. Wu, H. Liu, D. Li, C. Tao, M. A. Musen, C. G. Chute, and N. H. Shah. Unified medical language system term occurrences in clinical notes: A large-scale corpus analysis. *Journal of the American Medical Informatics Association*, 19(1):149–156, 2012. 4.2.2, 6

E. Yilmaz, E. Kanoulas, and J. A. Aslam. A simple and efficient sampling method for estimating ap and ndcg. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 603–610, Singapore, 2008. 7.2.1

C. Zhai. Notes on the Lemur TFIDF model. Technical report, School of Computer Science, Carnegie Mellon University, 2001. 3.3.1, 4.4.1

C. Zhai. Statistical Language Models for Information Retrieval A Critical Review. *Foundations and Trends in Information Retrieval*, 2(3):137–213, 2007. 3.3.1, 4.3.3, 4.4.1

H.-T. Zheng, C. Borchert, and Y. Jiang. A knowledge-driven approach to biomedical document conceptualization. *Artificial Intelligence in Medicine*, 49(2):67–78, 2010. 4.1.1

M. Zhong and X. Huang. Concept-based biomedical text retrieval. In *Proceedings of the 29th International SIGIR Conference on Research and Development in Information Retrieval*, pages 723–724, Seattle, USA, 2006. 3.4.1

W. Zhou, C. Yu, V. Torvik, and N. Smalheiser. A concept-based framework for passage retrieval in genomics. In *The Fifteenth Text REtrieval Conference (TREC 2006) Proceedings*, pages 14–17, 2006. 3.4.1

W. Zhou, C. Yu, N. Smalheiser, V. Torvik, and J. Hong. Knowledge-intensive conceptual retrieval and passage extraction of biomedical literature. In *Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval*, pages 655–662, Amsterdam, The Netherlands, 2007. 3.4.1

D. Zhu and B. Carterette. Combining multi-level evidence for medical record retrieval. In *Proceedings of the 2012 International Workshop on Smart Health and Wellbeing*, pages 49–56, 2012a. 3.3

D. Zhu and B. Carterette. Exploring evidence aggregation methods and external expansion sources for medical record search. In *Proceedings of 21st Text REtrieval Conference (TREC 2012)*, Gaithersburg, USA, 2012b. 3.3.2

D. Zhu and B. Carterette. An adaptive evidence weighting method for medical record search. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1025–1028, Dublin, Ireland, 2013. 4.4.1

J. Zobel. How reliable are the results of large-scale information retrieval experiments? In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 307–314, Melbourne, Australia, 1998. 8.4.1

G. Zuccon, L. Azzopardi, and C. J. van Rijsbergen. Revisiting logical imaging for information retrieval. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 766–767, Boston, USA, 2009. 6.1.2

G. Zuccon, B. Koopman, A. Nguyen, D. Vickers, and L. Butt. Exploiting Medical Hierarchies for Concept-based Information Retrieval. In *Proceedings of the Seventeenth Australasian Document Computing Symposium*, Dunedin, New Zealand, 2012. 6.6.1