Combining concepts and language models for information access
Meij, E.J.

Citation for published version (APA):
Meij, E. J. (2010). Combining concepts and language models for information access

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
If a man will begin with certainties, he shall end in doubts; but if he will be content to begin with doubts, he shall end in certainties.
Sir Francis Bacon

Introduction

The definition of information retrieval (IR), in its broadest sense, is finding relevant information in response to queries issued by users [237, 296]. It is a highly dynamic discipline with a relatively short but rich history in which many techniques and methods have been proposed towards improving effectiveness, i.e., finding more relevant information given a query. In essence, IR subsumes two highly related activities: indexing (which deals with how information is represented) and searching (which deals with matching an expression of an information need—the query—with indexed information). A schematic overview of an IR system is provided in Figure 1.1.

1.1 Indexing

In the earliest days, textual documents were the sole unit of retrieval and most of the initial IR systems were used to search bibliographic databases. Contrary to many modern-day IR systems, they used information from references to documents instead of the contents of the documents themselves for searching. This had two main reasons. First, there was only very limited information storage capacity available and documents could only be represented by punched data cards. Information retrieval in that day and age constituted finding cards that had holes in the right places [157]. As such, documents could only be represented by the presence of a very limited number of “terms.” Second, small-sized controlled vocabularies which unambiguously and precisely represented the content of documents had been in use for a long time in libraries [105]. Obvious vocabularies of choice were the indexing systems commonly used by libraries, although other domain-specific thesauri were also used [157]. Documents were treated in the same fashion as library books; trained annotators would assign to them the terms by which the documents were to be indexed in the retrieval system.

Later, as computing power and storage space increased, the assigned indexing terms were gradually replaced by terms that can be found in the actual content of the documents, i.e., their vocabulary. This development was furtheracceler-
ated by a rapidly increasing number of documents and document types that made manual annotations prohibitively expensive. In the Cranfield experiments, a controlled study was performed to measure the effect of various factors on retrieval effectiveness [75]. In Cranfield II, the indexing languages constituted the performance variable under investigation and the aim was to investigate the retrieval effectiveness of systems using different indexing languages and methods [74]. Here, it was found that retrieval based on vocabulary terms (or: full-text indexing) performed better than retrieval based on assigned indexing terms. This finding was later corroborated by Salton [275] who lead the development of the SMART system in the 1960s [186].

The effectiveness and popularity of indexing using assigned terms and controlled vocabularies further waned, as the size of the documents and collections grew larger (the Cranfield experiments used only 1,400 documents). Today, most of the early retrieval systems have been replaced by full-text search systems, with well-known web search engines including Google, Bing, and Yahoo! as prime examples. As to search engines using assigned indexing terms, MEDLINE is a prime example of such an IR system from the 1960s that still exists today [177].

Unlike assigned terms from controlled vocabularies, the terms occurring in a document are only constrained by the grammar of the language and the imagination of the author. They are, as such, noisier and more prone to ambiguity. Despite the popularity of using full-text indexing, the clear semantics and manual labor involved with assigning indexing terms to documents has many merits, ranging from enabling browsing facilities of a collection to enabling result list segmenta-

Figure 1.1: Schematic representation of an IR system.
1.1. Indexing

Figure 1.2: Excerpt of the MeSH thesaurus, partially showing the concepts below the concept “Central Nervous System Diseases.”

Recent semantic web initiatives have sparked a renewed interest in the discussion, development, semantics, and interoperability of concept languages [15, 33, 292]. Berners-Lee et al. [33] define an ontology as a structure of well-defined, i.e., unambiguous, concepts. Ontologies define objects as well as their relations and properties, with an accompanying logic allowing inference. The semantic web, then, is envisaged to be a layer over the current World Wide Web defined in terms of such concepts. To further this goal, the semantic annotation of web pages, their contents, or any other kind of resource using concepts that can not be directly derived from their content is gaining in popularity [6, 110, 251, 306].
In a way, this is a step “back” towards the controlled vocabularies that were in use in the early days of IR [292]. However, recent advances in information extraction have ameliorated the need for manual, labor-intensive mappings. Later in this thesis, in Chapter 6, we will look at several methods for automatically mapping queries to concepts. Furthermore, we will also introduce a method that leverages the manual annotations of documents as well as their full-text representations to improve end-to-end retrieval performance. As we will see later, manual annotations and controlled vocabularies can effectively be used in conjunction with full-text indexing to improve information access. We will present, implement, and evaluate various intuitions about leveraging controlled vocabularies and manual annotations to improve end-to-end retrieval performance by introducing ways of combining information from documents, concept languages, and relevance assessments.

1.2 Searching

Originating from the binary assignment of controlled vocabulary terms to documents, all initial IR systems adopted the Boolean model of searching. Here, a user’s search terms are linked by the Boolean logical operators OR, AND, and NOT; OR is used to link synonyms or alternatives, AND to link conjunctively, and NOT to indicate irrelevant terms, i.e., those terms that should not be assigned to the required documents. Such systems typically return an unordered set of results, although in 1958 Joyce and Needham [157] already proposed the use of a notion of term frequency to sort the list of matching documents. They also suggested the use of aggregated terms (where the set of documents containing the phrase information retrieval is different from the union of the set of documents containing information and retrieval). The imprecise nature of language (as well as “relevance”) have led to a number of developments moving away from the inherently restrictive Boolean model and towards a coordinate-level, ranked output.

A first step was the move towards thesauri that were automatically generated from the documents’ content [95, 291]. Luhn [194] first addressed automatic keyword indexing, in which the terms in the documents were directly searchable. Maron and Kuhns [203] were the first to take a probabilistic view on IR, centered on the notion of relevance. This introduced a principled notion of term weighting (although Maron and Kuhns [203] assumed that human indexers would assign the initial weights). Via advances in automatic speech recognition and the probability ranking principle [263], term weighting obtained a principal role in retrieval models. Current state-of-the-art retrieval approaches employ models of language to compare queries with documents. In this thesis, we will make extensive use of a process called query modeling, where the query is represented as
1.3 Motivation

Previous IR approaches have typically used either full-text indexing or indexing using concepts and few methods exist where the two are combined in a principled manner. We hypothesize that the knowledge captured in concept languages and the associations between concepts and texts (for example, in the form of document-level annotations) can be successfully used to inform IR algorithms. Such algorithms would be able to match queries and documents not only on a textual level, but also on a semantic level. Recent advances in the language modeling for IR framework have enabled the use of rich query representations in the form of query language models. This, in turn, enables the use of the language associated with concepts to be included in the retrieval model in a principled and transparent manner.

Note that we do not pursue a research direction that uses concepts in a language modeling framework. Instead, we investigate how we can employ the actual use of concepts as measured by the language that people use when they discuss them.

Recent developments in the semantic web community, such as DBpedia and the inception of the Linked Open Data cloud, have enabled the association of texts with concepts on a large scale. These developments enable us to move beyond manually assigned concepts in domain-specific contexts and into the general domain. In sum, we will show in the remaining chapters of the thesis how we can successfully apply language modeling techniques in tandem with concepts to improve information access performance.

1.4 Research Questions

The central question governing this thesis is: “How can we leverage concept languages to improve information access?” In particular, we will be looking at methods and algorithms to improve the query or its representation using concept languages in the context of generative language models. Instead of creating, defining, or using such languages directly, however, we will leverage the natural language use associated with the concepts to improve information access. Our central research question leads to a set of more specific research questions that will be answered in the following chapters.
After we have provided a theoretical and methodological foundation of IR, we look at the case of using relevance information to improve a user’s query. A typical method for improving queries is updating the estimate of the language model of the query, a process known as query modeling. Relevance feedback is a commonly used mechanism to improve queries and, hence, end-to-end retrieval performance. It uses relevance assessments (either explicit, implicit, or assumed) on documents retrieved in response to a query to update the query. Core relevance feedback models for language modeling include the relevance modeling and the model-based feedback approach. They both operate under different assumptions with respect to how to treat the set of feedback documents as well as each individual feedback document. Therefore, we propose two models that take the middle ground between these two approaches. Furthermore, an extensive comparison between these models is lacking, both in experimental terms, i.e., under the same experimental conditions, and in theoretical terms. We ask:

**RQ 1.** What are effective ways of using relevance feedback information for query modeling to improve retrieval performance?

- a. Can we develop a relevance feedback model that uses evidence from both the individual feedback documents and the set of feedback documents as a whole? How does this model relate to other query modeling approaches using relevance feedback? Is there any difference when using explicit relevance feedback instead of pseudo relevance feedback?

- b. How do the models perform on different test collections? How robust are our two novel models on the various parameters query modeling offers and what behavior can we observe for the related models?

Inspired by relevance feedback methods, we then develop a two-step method that uses concepts (in the form of document-level annotations) to estimate a conceptual language model. In the first step, the query is translated into a conceptual representation. In a process we call *conceptual query modeling*, feedback documents from an initial retrieval run are used to obtain a conceptual query model. This model represents the user’s information need at the level of concepts rather than that of the terms entered by the user. In the second step, we translate the conceptual query model back into a contribution to the textual query model. We investigate the effectiveness of our conceptual language models by placing them in the broader context of common retrieval models, including those using relevance feedback information. We organize the following research question around a number of subquestions.

**RQ 2.** What are effective ways of using conceptual information for query modeling to improve retrieval performance?
1.4. Research Questions

a. What is the relative retrieval effectiveness of our method with respect to the standard language modeling and conventional pseudo relevance feedback approach?

b. How portable is our conceptual language model? That is, what are the results of the model across multiple concept languages and test collections?

c. Can we say anything about which evaluation measures are helped most using our model? Is it mainly a recall or precision-enhancing device?

We then move beyond annotated documents and take a closer look at directly identifying concepts with respect to a user’s query. The research questions we address are the following.

RQ 3. Can we successfully address the task of mapping search engine queries to concepts using a combination of information retrieval and machine learning techniques?

a. What is the best way of handling a query? That is, what is the performance when we map individual n-grams in a query instead of the query as a whole?

b. As input to the machine learning algorithms we extract and compute a wide variety of features, pertaining to the query terms, concepts, and search history. Which type of feature helps most? Which individual feature is most informative?

c. Machine learning generally comes with a number of parameter settings. We ask: what are the effects of varying these parameters?

After we have looked at mapping queries to concepts, we apply relevance feedback techniques to the natural language texts associated with each concept and obtain query models based on this information. The guiding intuition is that, similar to our conceptual query models, concepts are best described by the language use associated with them. In other words, once our algorithm has determined which concepts are meant by a query, we employ the language use associated with those concepts to update the query model. We ask:

RQ 4. What are the effects on retrieval performance of applying pseudo relevance feedback methods to texts associated with concepts that are automatically mapped from ad hoc queries?

a. What are the differences with respect to pseudo relevance estimations on the collection? And when the query models are estimated using pseudo relevance estimations on the concepts’ texts?

b. Is the approach mainly a recall- or precision-enhancing device? Or does it help other aspects, such as promoting diversity?
1.5 Main Contributions

The following summarizes the main contributions of this thesis, which adds both theoretical insights and practical contributions to the body of existing work in the field.

1. **Novel relevance feedback methods** — We develop two query modeling methods for relevance feedback that are based on leveraging the similarity between feedback documents and the set thereof.

2. **Comparison of relevance feedback methods** — We provide a comprehensive analysis, evaluation, comparison, and discussion (in both theoretical and practical terms) of our novel and various other core models for query modeling using relevance feedback.

3. **Concept-based query modeling** — We develop a way of using document-level annotations to improve end-to-end retrieval performance. Our model naturally generates concept models, which may serve to support, for example, interaction tools for users or which can be used to determine semantic similarity between concepts using the language observed in the documents associated with the concepts.

4. **Novel method for linking queries to concept languages** — We develop and evaluate a novel way of associating concepts with queries that effectively handles arbitrary features. For example, features pertaining to the query, concepts, search history, etc.

5. **Understanding of relevant features for concept identification in queries** — We provide insights why some (groups of features) perform better than others in the context of linking queries to concepts.

6. **Wikipedia-based query modeling** — We show that using the linked concepts can be effectively used to improve diversity and ad hoc retrieval effectiveness on two large test collections.

7. **State of the art retrieval effectiveness** — Through extensive experimental evaluations on various test collections (including those from the biomedical, web, social science, and news domains) we validate and analyze our proposed models. In most cases we show consistent and significant improvements over established and state-of-the-art methods on ad hoc retrieval.

1.6 Overview of the Thesis

- **Chapter 2 - Related Work** — We survey, identify, and describe related work for leveraging concept languages for information access.
• **Chapter 3 - Experimental Methodology** — The basic building blocks pertaining to the evaluation of information retrieval experiments, the test collections we use in the thesis, and the setting of various parameters are presented.

• **Chapter 4 - Query Modeling Using Feedback Information** — We look at and evaluate various query modeling methods for relevance feedback in the context of generative language models. We explicate the relation between two popular models and introduce two novel methods that estimate a query model using information from each feedback document individually and combined. While most previous approaches focus either on features of the entire set or of the individual relevant documents, our models exploit features of both.

• **Chapter 5 - Query Modeling Using Concepts** — We then turn to using concept languages to estimate a query model. In this chapter we propose generative concept models as an extension to query modeling within the language modeling framework, which leverages manual document annotations using controlled vocabularies to improve retrieval. By means of relevance feedback the original query is translated into a conceptual representation, which is subsequently used to update the query model.

• **Chapter 6 - Linking Queries to Concepts** — Next, we take a closer look at identifying relevant concepts with respect to a user’s query. In the previous chapter we used existing document annotations and relevance feedback to obtain concepts for queries. In this chapter we look at how we can apply supervised machine learning models to this task and compare it to several baseline methods including a straightforward lexical match and a purely retrieval based approach.

• **Chapter 7 - Query Modeling Using Linked Concepts** — In this chapter we bring techniques from the previous chapters together. We apply the supervised machine learning method presented in Chapter 6 to queries associated with two web-scale test collections. We link each query to Wikipedia articles and apply the ideas presented in Chapters 4 and 5 to estimate a query model.

• **Chapter 8 - Conclusions and Future Work** — Here we summarize our contributions and describe potential areas for future work.

Chapter 2 and Chapter 3 serve as introductory chapters to the field of information retrieval, language modeling for information retrieval, mapping free text to structured knowledge sources, and experimental evaluation in the context of information retrieval. We recommend that the reader first get familiarized with the
material presented there before reading other chapters. Many of the contributions made in the thesis converge in Chapter 7 and to be able to appreciate the results presented there, we encourage the reader to start with earlier material, in particular with Chapter 4 and Chapter 6. In appendix A (See page 179), we include a nomenclature and list of abbreviations.

1.7 Origins

This thesis is based on the following publications that have arisen as part the thesis work. Full details of these papers below can be found in the bibliography. The models presented in Chapter 4 were introduced in [216, 220] and this chapter is further based on [214]. The concept-based language models in Chapter 5 were introduced in [209] and further built upon in [207, 208, 212, 215, 221]. The work on linking search engine queries to structured knowledge sources in Chapter 6 was published in [219] and expanded in [222]. The work in Chapter 7 is based on material published in [213]. Finally, material from a number of other papers, including [26, 126, 138, 210, 217, 218, 227, 317, 338], have been incorporated at various points in the thesis.