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
Problems cannot be solved by the same level of awareness that created them.

Albert Einstein

Conclusions

As a starting point for the thesis we observed that common IR approaches have typically used either full-text indexing or indexing using concepts and, moreover, that few methods exist where the two are combined in a principled manner. 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, has enabled the use of the natural language associated with concepts to be included in the retrieval model in a principled and transparent manner. We have investigated how we can employ the actual use of concepts as measured by the natural language that people use when they discuss them. Furthermore, 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.

The main motivation for this thesis has been to verify whether knowledge captured in concept languages and the associations between concepts and natural language texts can be successfully used to inform IR algorithms and improve information access. Such algorithms are able to match queries and documents not only on a textual, but also on a semantic level. We present and evaluate several models and methods and perform and report on extensive experiments. In sum, we have shown that employing the (natural) language use associated with concepts can successfully and significantly improve information access.

In the remainder of this chapter we conclude the thesis. We first answer the research questions governing the preceding chapters (Section 8.1) and then conclude the thesis by discussing several directions for future work.

8.1 Main Findings

The general question governing this thesis has been: “How can we leverage conceptual knowledge in the language modeling framework to improve information access?” We have approached this question as a query modeling problem. That
8. Conclusions

is, we have looked at methods and algorithms to improve textual queries or their representations using concept languages in the context of generative language models. This main question has lead us to formulate five main research questions listed in Section 1.4 which have been answered in the previous chapters. In this section we recall the answers.

We have started the thesis with an overview of current approaches to information retrieval, concept languages, and their combination (Chapters 2 and 3). We have zoomed in on a technique called query modeling, with which the information need of a user can be captured more thoroughly than solely using the initial query.

In Chapter 4, we have employed pseudo and explicit user feedback in the form of relevance ratings at the document level to improve the estimation of the query model. The first research question thus dealt with relevance feedback methods for query modeling. We asked:

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

We have presented two query modeling methods for relevance feedback that are based on leveraging the similarity between feedback documents and the set thereof. By providing 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, we have shown that all the models we have evaluated are able to improve upon a baseline without relevance feedback in the case of explicit relevance feedback. One of our proposed models (NLLR) is particularly suited when explicit relevance assessments are available. In the case of pseudo relevance feedback, we have observed that RM-1 is the most robust model. Parsimonious relevance models, on the other hand, perform very well on large, noisy collections. We have further found that, in the case of pseudo relevance feedback, there exists a large variance in the resulting retrieval performance for different amounts of pseudo relevant documents, most notably on large, noisy collections. We have also concluded that the test collection itself is of influence on the relative performance of the models; there is no single model that outperforms the others on all test collections. As to the observations made when using explicit relevance feedback, here we found that the variance with respect to the number of feedback documents is much less pronounced. Furthermore, one of the two novel methods we introduce consequently outperforms the other models.

Inspired by relevance feedback methods, we then developed a two-step method in Chapter 5 that uses concepts to estimate a conceptual query model. Here, we moved beyond the lexical level by introducing an automatic method for generating a conceptual representation of a query and subsequently using this representation to improve end-to-end retrieval. We asked
RQ 2. What are effective ways of using conceptual information for query modeling to improve retrieval performance?

We have introduced a novel way of using document-level annotations in the form of concepts to improve end-to-end retrieval performance. We have found that our proposed method obtained the highest performance of all evaluated models. We have concluded that, although each step in our method of applying conceptual language models is not significantly different from the other, the full model is able to significantly outperform both a standard language modeling and a pseudo relevance feedback approach.

After that, in Chapter 6, we have considered DBpedia as a concept language, in which case each Wikipedia article constitutes a concept. Here, we have turned to a different way of obtaining concepts pertinent to a user's query based on supervised machine learning. The research question was:

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?

We have developed a novel way of associating concepts with queries that can effectively handle arbitrary features and answered this question in the affirmative. We have concluded that our proposed approach significantly outperforms other methods, including commonly used methods based on a lexical matching between query terms and concept labels.

In the next chapter (Chapter 7), we have moved to the open domain and brought together the ideas presented in all preceding chapters. We have taken the conceptual mapping approach from Chapter 6 to obtain DBpedia concepts. Next, we have used the natural language text associated with each concept (in the form of the accompanying Wikipedia article) to estimate a query model, similar to the conceptual language models presented in Chapter 5. The associated research question was:

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?

We have found that the conceptual mapping method presented in Chapter 6 transfers well to the open domain; the linked concepts seem reasonable and the estimated query models are to the point. When evaluated, we have concluded that our novel method is able to improve recall, precision, and diversity metrics on two large web collections.
8.2 Implications for Future Work

There exist several unexplored avenues of research that are either opened or have not yet been fully addressed by the work presented in this thesis. Here we list these, in no particular order.

In Chapters 4, 5, and 7, we have measured the “quality” of query models by their resulting retrieval performance. While one could argue that improving end-to-end retrieval performance should be the ultimate goal of improving query model estimations, other ways by which to explicitly measure the quality of the query models themselves should be investigated. Examples of such measures would include perplexity measures or scores related to query clarity [84]. What existing measures cannot account for, however, is an intrinsic measure of diversity with which different topical aspects of a query model could be quantified.

In Chapter 7 we have used a mapping from queries to concepts to automatically estimate query models and shown that this resulted in improved retrieval performance. Besides having the potential of automatically improving the retrieval performance on certain topics we believe that, similar to our observations in Chapter 5, the biggest improvements may be realized when a user selects the most relevant concepts. Future work should indicate if this is a valid assumption and whether such conceptual representations are appreciated by and useful to an end user.

In Chapters 5 and 6 we have obtained an explicit conceptual representation of the query. Several ideas may be employed to improve the performance of this step. For example, a form of query segmentation could be used to identify significant phrases in the queries [118]. Such information could then be used to inform the conceptual mapping process. Additional features can also be added, for example structural ones such as those pertaining to the structure of the ontology. Although we have found that the method presented in Chapter 6 obtained convincing results and improvements over the two baselines, we believe that further improvements may be obtained by considering the graph structure of DBpedia (or the LOD cloud in general). One example of such an improvement could be to use the candidate concepts and the graph structure to “zoom in” on a relevant subgraph of the knowledge structure. This information could subsequently be used for disambiguation purposes, by determining the concepts closest to or contained in this graph. Indeed, in the work presented in this thesis, we have not made any explicit use of any relations (known, discovered, or otherwise) between concepts. In [317] we have introduced a method which uses language modeling estimations to determine the relatedness of two concepts, an approach which is taken further by Trieschnigg et al. [318]. Such estimations effectively “anchor” the perceived meaning of concepts in the language use surrounding each concept and we believe that this avenue of research deserves further investigation when mapping
queries to concepts. Furthermore, related concepts may also be used to perform a kind of “semantic smoothing,” in the context of our proposed conceptual query models.

Our task definition in Chapter 6 required fairly strict matches between the queries and DBpedia concepts, comparable to finding skos:exactMatch or even owl:equivelentClass relations in an ontology matching task. However, our task can also be interpreted in a more liberal sense, where not only exact matches but also semantically related concepts are suggested [26, 218]. For example, when a query contains a book title that is not represented by a concept in DBpedia, we could suggest its author (assuming the book title is mentioned in the author's Wikipedia page). Similarly, instances for which no concept is found can be linked to a more general concept. We believe that our approach can be adapted to incorporate such semantically related general instances of a specific concept could be defined as a correct concept for mapping.

One other aspect that we intend to investigate in the context of Chapter 6 is how to incorporate information from other parts of the LOD cloud. Our current approach has focused on DBpedia, which might be too limited for some queries. We have shown in [26] that, although DBpedia covers the open domain well, it does not exhaustively cover entity-related information. Future work should indicate whether traversing links to other, connected knowledge repositories would yield additional relevant concepts. It would also be interesting to consider more “noisy” types of concept languages that were excluded from the thesis, such as Twitter hashtags or del.icio.us tags [96]. Another interesting angle would be to consider a form of automatic keyphrase extraction in this context [116, 220]. Recent research into supervised topic models and labeled LDA [39, 255], as well as work done for word sense disambiguation [44] and topic identification [80] could also provide an interesting link between observed text and concept languages.

Furthermore, another abstraction layer may be imposed upon concepts in the form of concept types. Examples of such types are sets of Wikipedia articles, grouped together by a common category or by a shared infobox. In previous work we have shown that information pertaining to such concept types can be useful for generating query suggestions for rare or unseen queries [217, 218] and query log analysis [227]. In this thesis we have solely looked at concepts in their own right, discarding any potential type-based information. If and how this kind of information can be used to improve ad hoc retrieval is an interesting continuation of work presented in the thesis.

Recently, several evaluation campaigns have started to investigate methods for retrieving entities, a task known as entity finding [25, 87]. Both of these define entities as Wikipedia articles, much in the same way as we have used Wikipedia articles as concepts. So, phrased in this way, the goal is not to use Wikipedia based information for ad hoc retrieval but rather the other way around: use documents as evidence towards concept retrieval. Some of the models presented in this
thesis (for example those presented in Chapter 6) can be modified or applied to this new task. Another interesting application would be so-called undirected informational queries [270], where the information need is “open” and the user is solely interested in learning more on a certain topic. We are currently only taking the first steps in these new directions and future work should indicate in which ways the methods and models developed in the thesis can be applied to such tasks.

Finally, answering information needs which contain an explicit relationship type between concepts is a current research topic, as witnessed by a dedicated task at the recently launched TREC Entity Track. As we have shown in [26], this particular task is currently one of the best candidates for developing techniques which will bridge the gap between semantic web technologies and information retrieval. In [26] we further show that both semantic web and IR approaches fair well on the task of related entity finding, with each yielding unique sets of relevant results. We have argued then and maintain the position now that semantic web and IR are two sides of the same coin. Especially with the advent of DBpedia, YAGO, and, more generically, the LOD cloud, semantic web requires IR techniques and methodologies for handling the growing volumes of data, whereas IR needs a form of semantic anchoring of the obtained results.