Effective focused retrieval by exploiting query context and document structure

Kaptein, A.M.

Publication date
2011

Chapter 1

Introduction

1.1 Research Objective

Information retrieval (IR) deals with the representation, storage, organisation of, and access to information items such as documents, Web pages, online catalogs, structured and semi-structured records, and multimedia objects (Baeza-Yates and Ribeiro-Neto, 2011). Many universities and public libraries use IR systems to provide access to books, journals and other documents, but Web search engines are by far the most popular and heavily used IR applications.

Let's try to find a particular piece of information using a Web search engine. The search process, depicted in Figure 1.1, starts with a user looking to fulfil an information need, which can vary in complexity. In the simplest case the user wants to go to a particular site that he has in mind, either because he visited it in the past or because he assumes that such a site exists (Broder, 2002). An example of such a navigational information need is:

> I want to find the homepage of the Simpsons.

In more complex cases the user will be looking for some information assumed to be present on one or more Web pages, for example:

> A friend of mine told me that there are a lot of cultural references in the 'Simpsons' cartoon, whereas I was thinking that it was 'just' a cartoon like every other cartoon. I'd thus like to know what kind of references can be found in Simpsons episodes (references to movies, tv shows, literature, music, etc.)[1]

The next step in the search process is to translate the information need into a query, which can be easily processed by the search engine. In its most common form, this translation yields a set of keywords which summarises the information

[1] This is INEX ad hoc topic 464 (Fuhr et al., 2008), see Section 1.3.1.

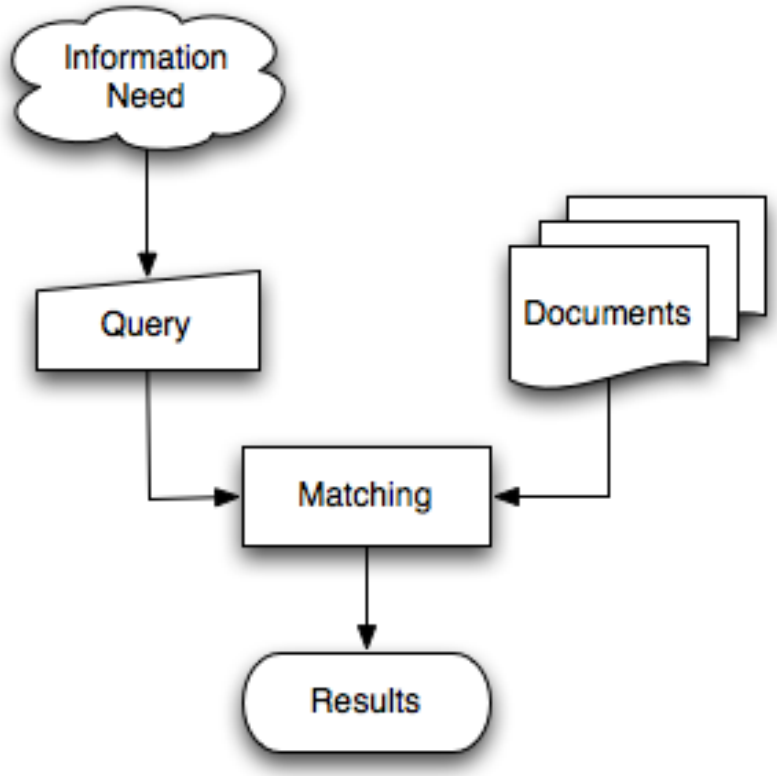


Figure 1.1: Main components of the search process, adaptation of the classic IR model of Broder (2002).

need. For our first simple information need formulating a query is also simple, i.e., the keyword query 'the simpsons' is a good translation of the information need. For our second, more complex information need also formulating the keyword query becomes a more complex task for the user. A possible keyword query is 'simpsons references'.

Given the user query, the key goal of an IR system is to retrieve information which might be useful or relevant to the information need of the user. For our first simple information need, there is only one relevant result: the homepage of the Simpsons, that is `http://www.thesimpsons.com`. When the keyword query 'the simpsons' is entered into Web search engines Google[2] and Bing[3], both these search engines will return the homepage of the Simpsons as their first result, thereby satisfying the user information need.

Continuing with our more complex information need, entering the keyword query 'simpsons references' into Google and Bing, leads to the results as shown

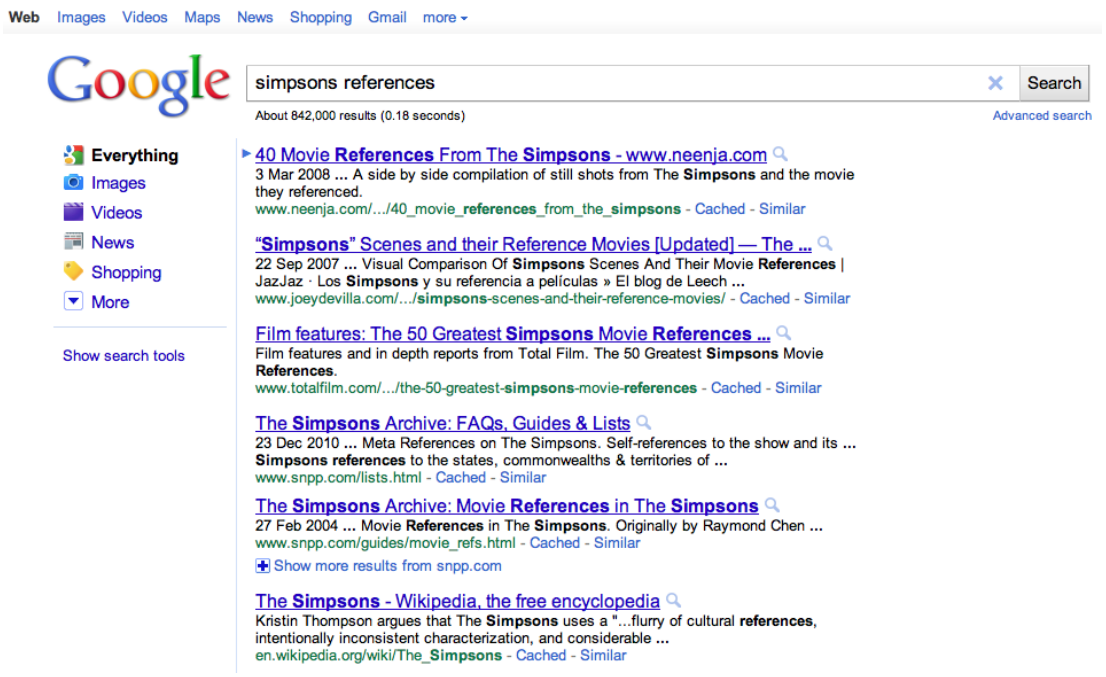
[2] `http://www.google.com/`
[3] `http://www.bing.com/`

in Figure 1.2. The results of the two searches look similar. The search engines return ranked list of results. Each result consists of the title of the Web page, a short snippet of text extracted from the page, and the URL. Clicking on a result will take you to the Web page and hopefully the desired information. Indeed, clicking on the first Google result takes you to a page[4] with references to movies like 'Apocalypse Now', 'Batman' and 'Ben Hur' with side by side images from various episodes of the Simpsons besides the image from the movie scene they refer to. While this document is relevant to the information need, it does not lead to a complete fulfilment of the information need. It does for example not contain information on references to literature or music. Actually, most of the results are about references to movies, and the user has to inspect quite some documents, including documents containing redundant information and non-relevant documents, to find all the types of references he is looking for.

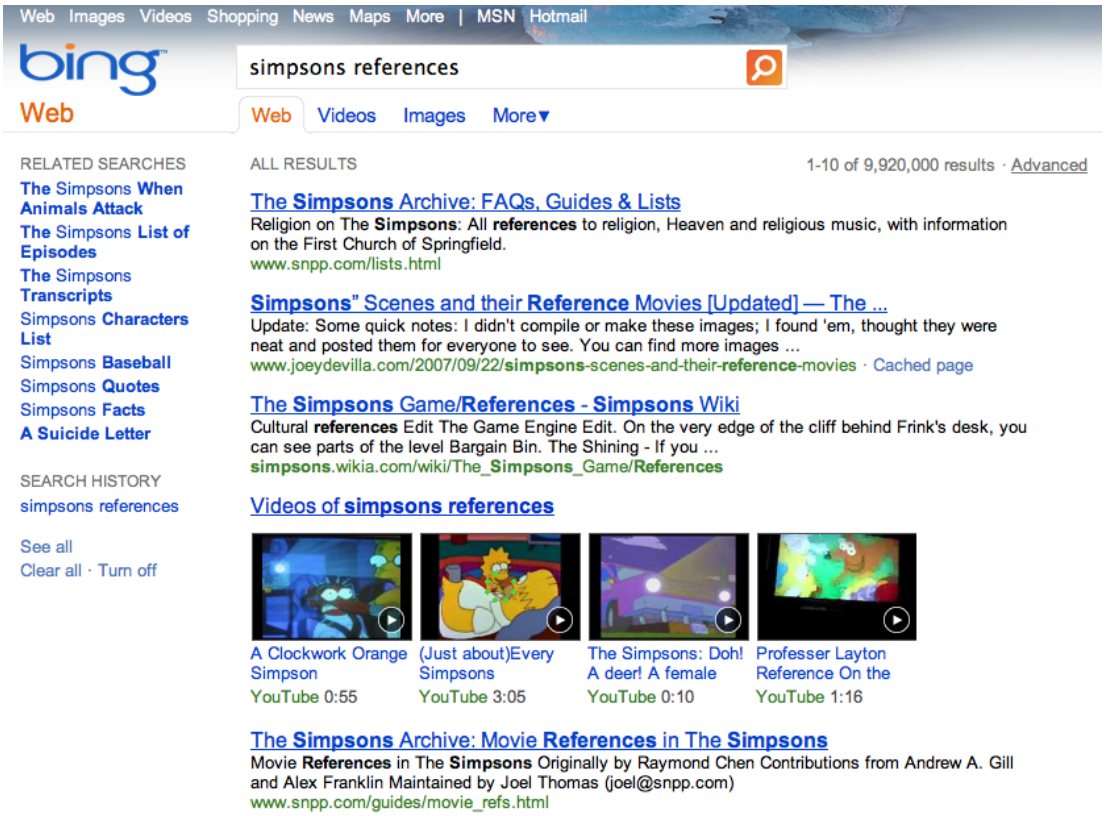
The primary goal of an IR system is to retrieve all the documents which are relevant to a user query while retrieving as few non-relevant documents as possible. To achieve this goal IR systems must somehow 'interpret' the contents of the documents in a collection, and rank them according to a degree of relevance to the user query. The 'interpretation' of a document involves extracting syntactic and semantic information from the document and using this information to match the user information need. The difficulty lies not only in the extraction of this information but also how to use it to decide relevance. The notion of *relevance* is at the center of information retrieval. An issue when evaluating the relevancy of search results for a query, is that relevance is a personal assessment that depends on the task being solved and its context. For example, relevance can change with time when new information becomes available, or it can depend on the location of the user, e.g., the most relevant answer is the closest one (Baeza-Yates and Ribeiro-Neto, 2011).

The search process we just described and is depicted in Figure 1.1 consists of three main elements: query, documents, and results. While for simple navigational information needs the search process is straightforward, for more complex information needs we need focused retrieval methods. The notion of 'focused retrieval' can be defined as providing more direct access to relevant information by locating the relevant information inside the retrieved documents (Trotman et al., 2007). In this thesis we consider the following, broader notion of focused retrieval. There is a loss of focus throughout the search process, because keyword queries entered by users often do not suitably summarise their complex information needs, and IR systems do not sufficiently interpret the contents of documents, leading to result lists containing irrelevant and redundant information. Focused retrieval methods aim to solve these problems.

[4] `http://www.neenja.com/articles/4/40_movie_references_from_the_simpsons`



(a) Google results, retrieved on 9-3-2011.



(b) Bing results, retrieved on 9-3-2011.

Figure 1.2: Web search results for the query 'simpsons references'

Our main research objective is:

Research Objective Exploit query context and document structure to provide for more focused retrieval

In the remainder of this section we examine opportunities that can help to achieve our research objective by looking at each of the three main elements of the search process (query, documents, and results) in more detail.

Query

The first element of the search process is the query. Shallowness on the user side is a major bottleneck for delivering more accurate retrieval results. Users provide only 2 to 3 keywords on average to search in the complete Web (Jansen et al., 2000; Lau and Horvitz, 1999; Jansen et al., 2007). In an ideal situation this short keyword query is a suitable summarisation of the information need, and the user will only have to inspect the first few search results to fulfil his information need. To overcome the shallowness of the query, i.e., users entering only a few keywords poorly summarising the information need, we add context to the query to focus the search results on the relevant context. We define context as: all available information about the user's information need, besides the query itself. The first opportunity we explore is:

> **Queries are posed in a search context**
> Different forms of context can be considered to implicitly or explicitly gather more information on the user's search request. Potential forms of query context are document relevance, and category information.

Documents

The second element of search we examine are the documents. Documents on the Web are rich in structure. Documents can contain HTML structure, link structure, different types of classification schemes, etc. Most of the structural elements however are not used consistently throughout the Web. A key question is how to deal with all this (semi-)structured information, that is how IR systems can 'interpret' these documents to reduce the shallowness in the document representation.

Structured information on the Web exists in various forms. The semantic Web tries to give meaning to everything on the Web to create a web of data that can be processed directly or indirectly by machines. While they may not have succeeded for the whole Web, a large enough semantic Web has indeed emerged, capturing millions of facts into data triples (Bizer et al., 2009). A structured information resource on the Web is Wikipedia[5]. Wikipedia is a free encyclopedia

[5] `http://www.wikipedia.org/`

that anyone can edit, consisting of millions of articles that adhere to a certain structure. Another structured resource on the Web is the DMOZ directory[6]. This Web directory contains a large collection of links to Web pages organised into categories.

These structured resources provide the following opportunities:

Documents categorised into a category structure
We can use the category structure of Web resources to retrieve documents belonging to certain categories.

Absence of redundant information in structured Web resources
A problem in Web search is the large amount of redundant and duplicate information on the Web. Web pages can have many duplicates or near-duplicates. Web pages containing redundant information can be hard to recognise for a search engine, but users easily recognise redundant information and this will usually not help them in their search. Most structured Web resources have organised their information in such a way that they do not contain, or significantly reduce redundant information.

Results

The third and final element of search we examine are the results. While a query can have thousands or millions of results, e.g., our example query 'simpsons references' has 848,000 results on Google, and 9,920,000 results on Bing, most users only look at the first result page (Jansen and Spink, 2006). Looking at the results of our search for 'simpsons references', we see that 4 out of the 6 Google search results in Figure 1.2(a) are Web pages containing movie references. Also, 2 out of 4 of Bing Web search results (excluding the video results) in Figure 1.2(b) are pages containing movie references. While these are all relevant pages, we are also interested in other types of references, such as references to tv shows, literature, and music. Again we face the problem of redundant and duplicate information. Search results are often dominated by the single most popular aspect of a query. Instead of showing single documents in the result list, documents relevant to the same aspects of a query can be grouped and summarised to provide more focused results. The shallowness on the result side lies in the combination of users only inspecting the first result page, and search engines returning redundant information on this first results page. The last opportunity we explore is:

Multiple documents on the same topic
Result lists often contain redundant information. We study how we

[6] `http://www.dmoz.org/`

can summarise multiple (parts of) documents on the same topic into a single summarised result to create a topically more diverse result list.

Summary

To summarise this section, the main research objective of this thesis is to exploit query context and document structure to provide for more focused retrieval. To tackle this problem we examine each of the three main elements of the search process: query, documents and results. The challenges to face are:

- Shallowness on the query side, i.e., users provide only a short keyword query to search in a huge amount information.
- Shallowness in the document representation, i.e., documents contain structure which is hard to extract and exploit for computers.
- Shallowness on the results side, i.e., users only pay attention to the first 10 or 20 results that often contain redundant information, while a Web search can return millions of documents.

The opportunities described provide ample possibilities to face the challenges and explore our main research objective. The next section will describe the key points that we will focus on in this thesis. Section 1.3 gives information on the methodology, the test collections and evaluation measures, we use. To conclude this chapter in Section 1.4 we give an outline of the contents of the remaining chapters in this thesis.

1.2 Research plan

This section describes the separate components of this thesis and highlights the areas we will focus on. First of all, we study how to add and exploit query context. Secondly, we examine how we can exploit structured resources. Finally, we explore methods to summarise documents in search results.

1.2.1 Adding Query Context

In the first part of this research, we examine how we can use query context to improve retrieval results. Query context is obtained by feedback. In this thesis we consider context obtained together with the query also as feedback, that is if a user for example provides a topical category at the same time as the input of the query, we still consider this feedback on the query. We distinguish between two types of feedback:

- **Implicit** feedback techniques unobtrusively obtain information about queries and users by watching the natural interactions of the users with the system. Sources of implicit feedback include clicks, reading time, saving, printing and selecting documents (Kelly and Teevan, 2003).
- **Explicit** feedback techniques require users to explicitly give feedback through user interaction, such as marking documents or topic categories relevant, or clicking on a spelling suggestion.

Feedback or the context of a search can entail a number of things related to the user, the search session, and the query itself. We will focus on the individual query context, and do not consider the user context, e.g., his search history, a personal profile or location, or session context, e.g., previously issued queries and clicks in the same search session. Although general Web search engines store and maintain more and more information about the user and session context, this type of information is not publicly available.

The most common and well studied form of query context is relevance feedback, consisting of documents marked by users as relevant to their information needs, or pseudo-relevant documents from the top of the ranking. Pseudo-relevance feedback techniques, also known as blind feedback techniques, generate an initial ranking of documents using the query from the user, and then assume the top ranked documents to be relevant. Relevance feedback can be used for query expansion. From the (pseudo-)relevant documents the most frequent and discriminating words are extracted and added to the initial query and a new document ranking is generated for presentation to the user (Ruthven and Lalmas, 2003).

We found the standard relevance feedback approach works quite well (Kaptein et al., 2008), and think that there is not a lot of room for improvement. Relevance feedback techniques have also been studied extensively (see e.g. (Rocchio, 1971; Salton and Buckley, 1990; Zhai and Lafferty, 2001a; Ruthven and Lalmas, 2003; Buckley and Robertson, 2008)), so in this thesis we will focus on a less common form of feedback: topical feedback. Instead of using (pseudo-)relevant documents as feedback, we use topical categories, i.e., groups of topically related relevant documents as feedback. Topically related documents can be extracted from knowledge sources on the Web such as the Web directory DMOZ or the Web encyclopedia Wikipedia, where documents are organised in category structure. DMOZ topic categories containing sets of documents can be used as topical feedback for queries. This feedback can then be used for query expansion in a similar way as is done for relevance feedback.

Providing topical feedback explicitly might also be more appealing to users than providing relevance feedback. Marking documents as relevant can become a tedious task. Other types of explicit feedback are less static, i.e., the required input from the user depends on the query and the system supports the user by providing intelligent suggestions. For example, Googles spelling suggestions

detect possible spelling mistakes; when your query is 'relevence', on top of the result list Google asks: 'Did you mean: relevance'. Or, when we want to use topical feedback, questions like 'Do you want to focus on sports?' or 'Are you looking for a person's home page?' can be asked. When these follow-up questions are relevant to the query and easy to answer these kinds of interaction might be more appealing to users than simply marking relevant documents.

1.2.2 Exploiting Structured Resources

In the second part of the thesis we study how we can exploit the information that is available on the Web as structured resources. One of the main structured information resources on the Web is Wikipedia, the internet encyclopedia created and maintained by its users. Wikipedia is a highly structured resource: the XML document structure, link structure and category information can all be used as document representations. INEX (Initiative for the Evaluation of XML retrieval) provides a test collection for search in Wikipedia (described in more detail in Section 1.3.1), and in this framework the value of the different sources of information can be explored. Continuing the work in the previous part, adding query context, we focus on the use of category information as query context. We obtain category information through explicit and pseudo feedback.

Structured resources provide two interesting opportunities: 'Documents categorised into a category structure' and 'Absence of redundant information'. Category information is of vital importance to a special type of search, namely entity ranking. Entity ranking is the task of finding documents representing entities of an appropriate entity type that are relevant to a query. Entities can be almost anything, from broad categories such as persons, locations and organisations to more specific types such as churches, science-fiction writers or CDs. Searchers looking for entities are arguably better served by presenting a ranked list of entities directly, rather than a list of Web pages with relevant but also potentially redundant information about these entities. Category information can be used to favour pages belonging to appropriate entity types. Similarly, we can use category information to improve ad hoc retrieval, by using Wikipedia categories relevant to the query as context.

Furthermore, the absence of redundant information is of great importance for the entity ranking task. Since each entity is represented by only one page in Wikipedia, searching Wikipedia will lead to a diverse result list without duplicate entities. When searching for entities on the Web, the most popular entities can dominate the search results, leading to redundant information in the result list. By using Wikipedia as a pivot to search entities, we can profit from the encyclopedic structure of Wikipedia and avoid redundant information.

1.2.3 Summarising Search Results

In the third and final part of this thesis we study summarisation of sets of search results. The Web contains massive amounts of data and information, and information overload is a problem for people searching for information on the Web. A typical query returns thousands or millions of documents, but searchers hardly ever look beyond the first result page. Furthermore, even single documents in the result list can be sometimes as large as complete books. Here, we explore opportunity 'Multiple documents on the same topic'. In the previous section we introduced the problem of entity ranking where the goal is to find documents representing entities. Very often we will find multiple documents that represent one entity. Since space on the result page is limited, we cannot show each document (summary) in the result list. Therefore we study whether we can summarise these sets of search results into a set of keywords. Similarly, using the context of documents, e.g., category information from DMOZ or Wikipedia, search results can be clustered and summarised. Through user interaction, that is the user selecting the cluster of interest, we can then provide more focused search results.

In this thesis we do not focus on the clustering of the documents, but we focus on how we can reduce (sets of) documents into a set of keywords which can give a first indication of the contents of the complete document(s). The social Web, part of Web 2.0, allows users to do more than just retrieve information and engages users to be active. Users can now for example add tags to categorise Web resources and retrieve your own previously categorised information. By sharing these tags among all users large amounts of resources can be tagged and categorised. These generated user tags can be visualised in so-called tag clouds where the importance of a term is represented by font size or colour. To summarise sets of search results we will use word clouds. Word clouds are similar to tag clouds, but instead of relying on users to assign tags to documents, we extract keywords from the documents and the document collection itself.

1.3 Methodology

We describe the methodology used to study our research objective. The information retrieval community has developed standard test collections that fit our purposes. This section provides information on the test collections and evaluation measures used in this thesis.

1.3.1 Test Collections

To evaluate retrieval methods standard test collections have been developed in the information retrieval field. We use data from two of the main evaluation forums: TREC (Text Retrieval Conference) and INEX (Initiative for the Eval-

uation of XML retrieval). The purpose of TREC[7] is to support research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies. Each year NIST (National Institute of Standards and Technology) provides test collections consisting of search topics for different tasks. Participants run their own retrieval systems on the data, and return to NIST a list of the retrieved top-ranked documents. NIST chooses a set of documents from the submitted result lists for evaluation (a technique known as pooling), judges the retrieved documents for correctness, and evaluates the results.

INEX[8] provides a forum for the evaluation of focused retrieval. The goal of focused retrieval is to not only identify whole documents that are relevant to a user's information need, but also to locate the relevant information within the document. The documents in their test collections contain (XML) structure to allow for focused retrieval. In contrast to TREC where topics are created by NIST, at INEX the participants themselves provide search topics they believe are suitable for experimental purposes. These are collected, verified, and de-duplicated by INEX before being distributed back to the participants. Participants run their own retrieval systems, and return their results to INEX. After pooling the results, the documents are distributed back to the original authors of the topics to make judgments as to which documents are relevant and which are not for each topic. Finally, all participant's results lists are evaluated.

TREC and INEX consist of multiple tracks, in each track certain tasks and/or document collections are explored. We discuss here only the tasks and document collections relevant for this thesis.

Tasks

TREC and INEX run a number of tracks each year in which different tasks related to information retrieval are explored. *Ad hoc retrieval* is the most standard information retrieval task, where a system aims to return all documents from within the collection that are relevant to an user information need.

TREC ad hoc topics consist of three components, i.e., title, description and narrative. The title field contains a keyword query, similar to a query that might be entered into a Web search engine. The description is a complete sentence or question describing the topic. The narrative gives a paragraph information about which documents are considered relevant and/or irrelevant. An example query topic is shown in Figure 1.3. Ad hoc topics at INEX also consist of a title, narrative and description, but in addition also structured queries and phrase queries can be included in the topic (Fuhr et al., 2008; Kamps et al., 2009).

[7] `http://trec.nist.gov/`

[8] `http://www.inex.otago.ac.nz/`

```
<top>
<num> Number:  701

<title>
U.S. oil industry history

<desc> Description:
Describe the history of the U.S. oil industry

<narr> Narrative:
Relevant documents will include those on historical exploration and
drilling as well as history of regulatory bodies.  Relevant are history
of the oil industry in various states, even if drilling began in 1950
or later.

</top>
```

Figure 1.3: TREC ad hoc query topic 701

Document Collections

In this thesis we use the following document collections in our experiments:

.GOV2 This collection is meant to represent a small portion of the general Web and consists of Websites crawled in the ".gov" domain.

Wikipedia '06 and '09 These document collections consist of dumps of the complete Wikipedia. The '09 collection is annotated with semantic concepts.

ClueWeb Cat. A and Cat. B This collection is meant to represent the general Web. Cat. B is a subset of the pages in Cat. A, i.e., the first 50 million English pages. The complete Wikipedia is also included in the collection.

DMOZ This document collection we created ourselves. It consists of all the Web pages from the top four levels of the DMOZ directory we were able to crawl.

Parliamentary debates This document collection consist of the proceedings of plenary meetings of the Dutch Parliament, on data from 1965 until early 2009. For our experiments we use only an example document that contains the notes of the meeting of the Dutch Parliament of one particular day (September 18, 2008).

Table 1.1: Document Collection Statistics

Name	Forum	Year	Size	# Documents
.GOV2	TREC	2004	42.6GB	25 million
Wikipedia '06	INEX	2006	4.5GB	659 thousand
Wikipedia '09	INEX	2009	50.7GB	2.7 million
ClueWeb (Cat. A)	TREC	2009	5TB (compressed)	1 billion
ClueWeb (Cat. B)	TREC	2009	230GB (compressed)	50 million
DMOZ		2008	1.8GB (compressed)	460 thousand

We only use the English language parts of all the document collections, except for the collection of parliamentary debates that is completely in Dutch. Some basic collection statistics of these collections can be found in Table 1.1.

1.3.2 Evaluation Measures

To evaluate the quality of a ranking we use different performance measures. The two basic measures for information retrieval effectiveness are:

- *Precision*: the fraction of retrieved documents that are relevant.
- *Recall*: the fraction of relevant documents that are retrieved.

For Web search it is important to measure how many good results there are on the first result page, since this is all most users look at (Jansen and Spink, 2006). Precision is therefore measured at fixed low levels of retrieved results, such as 10 or 20 documents, so-called *Precision at k*, e.g precision at 10 (P10).

A standard measure in the TREC community is *Mean Average Precision* (MAP), which provides a measure of the quality of the ranking across all recall levels. For a single information need, average precision is the average of the precision values obtained for the set of top k documents in the ranking after each relevant document is retrieved. MAP is the average of the average precision for a set of information needs. MAP is calculated as follows (Manning et al., 2008):

$$MAP = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} \text{Precision}(R_{jk}) \tag{1.1}$$

where the set of relevant documents for an information need $q_j \in Q$ is $\{d_1, \ldots, d_{m_j}\}$ and R_{jk} is the set of ranked retrieval results from the top result until you get to document d_k.

A relatively novel performance measure that handles graded relevance judgements to give more credit to highly relevant documents is *Discounted Cumulative Gain* (DCG) (Croft et al., 2009). It is based on two assumptions:

1. Highly relevant documents are more useful than marginally relevant documents.

2. The lower the position of a relevant document in the ranking, the less useful it is for the user, since it is less likely to be examined.

The gain or usefulness of examining a document is accumulated starting at the top of the ranking and may be reduced or discounted at lower ranks. The DCG is the total gain accumulated at a particular rank k and is calculated as:

$$DCG_k = rel_1 + \sum_{i=2}^{k} \frac{rel_i}{log_2 i} \tag{1.2}$$

where rel_i is the graded relevance level of the document retrieved at rank i. To facilitate averaging across queries with different numbers of relevant documents, DCG values can be normalised by comparing the DCG at each rank with the DCG value for the perfect or ideal ranking for that query. The *Normalised Discounted Cumulative Gain* (NDCG) is defined as:

$$NDCG_k = \frac{DCG_k}{IDCG_k} \tag{1.3}$$

where IDCG is the ideal DCG value for that query. NDCG can be calculated at fixed cut-off values for k such as $NDCG_5$, or at the total number of R relevant documents for the query ($NDCG_R$).

Finally, the reciprocal rank measure is used for applications where there is typically a single relevant document, such as a homepage finding task. It is designed as the reciprocal of the rank at which the first relevant document is retrieved. The mean reciprocal rank (MRR) is the average of the reciprocal ranks over a set of queries.

For a more extensive treatment of performance measures and a complete introduction to the field of information retrieval, we refer to (Baeza-Yates and Ribeiro-Neto, 2011; Büttcher et al., 2010; Croft et al., 2009; Manning et al., 2008).

1.4 Thesis Outline

In this section we give a short outline of the research problems and questions for each chapter.

Chapter 2: Topical Context

In this chapter we explore how topical context can be used to improve ad hoc retrieval results. In particular, we study the use of the DMOZ Web directory.

Category information from DMOZ is used for topical feedback in a similar fashion as document relevance feedback. We study how to assign topical categories to queries automatically and manually by users. We analyse the performance of topical feedback on individual queries and averaged over a set of queries. We also study the relations between topical feedback and document relevance feedback.

This chapter is based on work published in (Kaptein and Kamps, 2008, 2009c, 2011a). In this chapter we want to answer the following research question:

RQ1 How can we explicitly extract and exploit topical context from the DMOZ directory?

Chapter 3: Exploiting the Structure of Wikipedia

In this chapter we investigate the problem of retrieving documents and entities in a particular structured part of the Web: Wikipedia. First, we examine whether Wikipedia category and link structure can be used to retrieve entities inside Wikipedia as is the goal of the INEX Entity Ranking task. Category information is used by calculating distances between document categories and target categories. Link information is used for relevance propagation and in the form of a document link prior.

Secondly, we study how we can use topical feedback to retrieve documents for ad hoc retrieval topics in Wikipedia. Since we only retrieve documents from Wikipedia, we can use an approach similar to the entity ranking approach. We study the differences between entity ranking and ad hoc retrieval in Wikipedia by analysing the relevance assessments and we examine how we can automatically assign categories to queries.

Finally, we examine whether we can automatically assign target categories to ad hoc and entity ranking queries. Automatically assigning target categories relieves users from the task of selecting a particular category from the large collection of categories.

This chapter is based on work done for the INEX Entity Ranking track and is published in (Kaptein and Kamps, 2009a,b; Koolen et al., 2010; Kaptein and Kamps, 2011b) In this chapter we want to answer the following research question:

RQ2 How can we use the structured resource Wikipedia to retrieve entities and documents inside of Wikipedia?

Chapter 4: Wikipedia as a Pivot for Entity Ranking

In this second entity ranking chapter, we use Wikipedia as a pivot to retrieve entity homepages outside Wikipedia. To rank entities inside Wikipedia we use the techniques described in the previous chapter. Then, as a second step we try to find entity homepages on the Web corresponding to the retrieved Wikipedia

pages. Web pages are retrieved by following external links on the Wikipedia pages, and by searching for Wikipedia page titles in an anchor text index.

This chapter is based on work published in (Kaptein et al., 2010b). In this chapter we want to answer the following research question:

RQ3 How can we use the structured resource Wikipedia to retrieve documents and entities on the Web outside of Wikipedia?

Chapter 5: Language Models and Word Clouds

In this chapter we study how we can create word clouds to summarise (groups or parts of) documents. First, we investigate the similarities between word clouds and language models, and specifically whether effective language modelling techniques also improve word clouds. We then examine how we can use structure in documents, in this case meeting notes of parliamentary debates, to generate more focused word clouds. These meeting notes are long and well structured documents, and are therefore suitable for summarisation in the form of a word cloud. This chapter is based on work published in (Kaptein et al., 2010a; Kaptein and Marx, 2010). In this chapter we want to answer the following research question:

RQ4 How can we use language models to generate word clouds from (parts of) documents?

Chapter 6: Word Clouds of Multiple Search Results

In this chapter we study how well users can identify relevancy and topic of search results by looking only at summaries in the form of word clouds. Word clouds can be used to summarise search results belonging to the same subtopic or interpretation of a query, or to summarise complete search result pages to give an indication of the relevancy of the upcoming search results.

This chapter is based on work published in (Kaptein and Kamps, 2011c). In this chapter we want to answer the following research question:

RQ5 How can we use word clouds to summarise multiple search results to convey the topic and relevance of these search results?

Chapter 7: Conclusions

In the final chapter we draw our overall conclusions. We summarise each chapter by looking at the answers to our research questions, draw overall conclusions on how we exploited the opportunities to solve our main research objective: to exploit query context and document structure to provide for more focused retrieval. Finally, we look forward to how this work can be continued in further research.