Effective Focused Retrieval by Exploiting Query Context and Document Structure

Rianne Kaptein
University of Amsterdam
amkaptein@hotmail.com

October 6, 2011

Abstract

The classic IR model of the search process consists of three elements: query, documents and search results. A user looking to fulfil an information need formulates a query usually consisting of a small set of keywords summarising the information need. The goal of an IR system is to retrieve documents containing information which might be useful or relevant to the user. Throughout the search process there is a loss of focus, 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. The main research objective of this thesis is to exploit query context and document structure to provide for more focused retrieval.

The short keyword query used as input to the retrieval system can be supplemented with topic categories from structured Web resources such as DMOZ and Wikipedia. Topic categories can be used as query context to retrieve documents that are not only relevant to the query but also belong to a relevant topic category. Category information is especially useful for the task of entity ranking where the user is searching for a certain type of entity such as companies or persons. Category information can help to improve the search results by promoting in the ranking pages belonging to relevant topic categories, or categories similar to the relevant categories. By following external links and searching for the retrieved Wikipedia entities in a general Web collection, we can also exploit the structure of Wikipedia to rank entities on the general Web. Wikipedia, in contrast to the general Web, does not contain much redundant information. This absence of redundant information can be exploited by using Wikipedia as a pivot to search the general Web.

A typical query returns thousands or millions of documents, but searchers hardly ever look beyond the first result page. Since space on the result page is limited, we can show only a few documents in the result list. Word clouds can be used to summarise groups of documents into a set of keywords which allows users to quickly get a grasp on the underlying data. Instead of using user-assigned tags we generate word clouds from the textual contents of documents themselves as well as the anchor text of Web documents. Improvements over word clouds that are created using simple term frequency counting include using a parsimonious term weighting scheme, including bigrams and biasing the word cloud towards the query. We find that word clouds can to a certain degree quickly convey the topic and relevance of a set of search results. Available online at: `http://dare.uva.nl/record/395691`