"Babeş-Bolyai" University, Cluj-Napoca, România
Faculty of Mathematics and Computer Science

Optimizing Information Retrieval in World Wide Web Space

PhD Thesis

Phd Student:
Ioan BĂDĂRÎNZĂ
PhD Supervisor:
Prof. Dr. Florian Mircea BOIAN

2018

Summary

In this thesis, entitled "Optimizing Information Retrieval in World Wide Space", we approached a very interesting subject that is popular among researchers, which is the field of information retrieval.

Nowadays, in the World Wide Web space, the amount of information and web documents grows in a very fast manner. The growing speed of the number of web-pages available in the Internet is caused by the fact that users can add content very fast and easy. They can easily upload all types of data files, from text files to audio and video files and can edit pages and change their content each and every day. The problem that this growth arrises is that it has become very hard for a human person to go and try to find some answers for an information need in all these data that is available in the Internet. In order to make this information available for users, there has been done a lot of research in the field of information retrieval. The problem of information retrieval was first mentioned by Luhn in 1958 when he first proposed a way of determining the usefulness of a word for the document that it belongs to, and starting with this, it evolved a lot in the last decades. Information retrieval is a process through which a software system takes as input a real information need that is expressed through a query, and returns a list of relevant documents that were extracted from a very large collection of documents. In order to be able to perform such extraction in a textual Information Retrieval system, each word in a document must have a weight assigned, a weight that outlines how representative that word is for the document. We can later use these weights in different algorithms that return a ranked list of documents which are the most representative to a user query.

The thesis focuses on two subdomains of information retrieval systems and these are: textual retrieval based in bag of words concept and personalizing query suggestions. The bag of words concept, that is heavily used in information retrieval models, has a major drawback: the meaning of the sentences that are present in a document, is lost because the order of the words in a sentence is ignored. We did a lot of research in this particular area, we implemented different tools that we have used to demonstrate the effectiveness of our proposals of improving these kind of information

retrieval systems. The other part of an information retrieval system that got our attention and focus, is the query suggestion mechanism. Lately, there has been done a lot of research in this area, mostly on personalizing the query suggestions that are presented to a user when he is trying to satisfy an information need. We identified that none of the proposed systems are making use of user's personal browsing history data when offering suggestions. Regarding this, we have build a tool through which we started to collect such data about users, and demonstrated that this personal browsing history can be used to improve the quality of the suggested queries offered to users.

Relevant keywords for thesis subject are: information retrieval, textual search, ranking models, boolean model, probabilistic model, index, inverted index, tokenizing, bm25, clustered index, ranking function, query, query suggestion, subquery, personalized suggestion, syntactic analysis, syntactic index.

The thesis has the following structure:

In the first part of the thesis we are going to describe some background information that is already present in the world of text retrieval. We will start with a brief history about information retrieval, when was it first mentioned and how it became so popular. We'll continue with presenting some fundamental concepts and workflows of an Information Retrieval (IR) system. We will first discuss about the components of an IR system and why the preprocessing stage of retrieval is important, then we will see how the index of the IR is constructed, after which we'll go into details about different ranking models, like the vector space model, probabilistic model and boolean model. We also describe a more recent feature that IR systems have, which is query suggestion, and how this feature helps users in better formulating their queries. We will finish with describing how we can evaluate an IR system and the results returned by it and also how we can evaluate a list of suggestions, returned by a suggest system, for a particular query.

Next, we are going to outline the main contributions of the thesis:

Chapter 3.1 - Clustering and Tired Indexes: In this chapter we present a textual retrieval system based on clustering and tiered indexes, that has as base model, the vector space model. This system can be used for exact phrase matching and also for

improved keyword search by employing term proximity weighting in the similarity measure. The document retrieval process is constructed in an efficient way, based on clusters of documents, so that not all the documents in the database need to be compared against the searched query.

Chapter 3.2 - Syntactic Indexes: This chapter presents three techniques for incorporating syntactic metadata in a textual retrieval system. The first technique involves just a syntactic analysis of the query and it generates a different weight for each term of the query, depending on its grammar category in the query phrase. These weights will be used for each term in the retrieval process. The second technique involves a storage optimization of the system's inverted index that is the inverse index will store only terms that are subjects or predicates in the document they appear in. Finally, the third technique builds a full syntactic index, meaning that for each term in the term collection, the inverse index stores besides the term-frequency and the inverse-document-frequency, also the grammar category of the term for each of its occurrences in a document.

Chapter 3.3 - Custom Term Weights: After performing an analysis of the IDF values for terms in Reuters collection, we have observed that queries that contain pairs of terms, one with very low IDF values and another with very high IDF values, are not returning relevant documents. In this chapter we present a technique which is based on the probabilistic model and which allows the user to alter the weights of query terms in a textual retrieval system so that it returns more relevant results. This technique is not meant to increase the relevancy of results returned for general search queries, but is meant to increase the relevancy of the returned results for some specific queries in which the query terms have disproportionate IDF values.

Chapter 4.1 - Analysing User's Browsing History: In this chapter we are going to present the research that we have done in query suggestion systems with major focus on personalizing suggestions. Query suggestions is a mechanism through which the user is presented with a list of possible queries from where he can choose and be able to perform a search easier and faster. We tried to assess the usefulness of a user's recent web browsing history for generating new query suggestions. We performed a one month experiment in which we collected browsing history logs of several users and searched query terms submitted by those users to Google (using a Chrome plugin) and found that approximately 32% of the queries submitted can be predicted from the user's browsing history.

Chapter 4.2 - Personalized Query Suggestions: After observing, that the personal browsing history can be used to predict what users will try to search for next, in this chapter we proposed a new method for personalizing the order of the query suggestions listed for a subquery, so that suggestions more relevant to the user's

information need are placed first (e.g. have a higher rank). The way we achieved this is by defining a *Personal Temporal Query Suggestion* score for each suggestion in the list, score that is taking into consideration a very recent and short personal browsing history, which is further used to compute a $HybridPageScore$. This chapter also contains the evaluation and experiments that we have performed for the proposed method.

In the conclusions chapter we outline the main contributions of this thesis related to the optimization of the information retrieval systems and the future research directions in this field.