Web Search - Challenges and Opportunities

Berthier Ribeiro-Neto

CS Department, UFMG & Google Engineering
Belo Horizonte, Brazil
berthier@dcc.ufmg.br

Abstract. As Udi Manber, currently a VP of Engineering at Google, likes to say: "Search is hard!" Yet, it is possible to build a rather powerful search engine that provides rather accurate answers most of the time. In the talk, I briefly discuss some of the fundamental concepts and technologies behind search engines, starting with a quick review of Information Retrieval and then moving through 15 years of search engines evolution. Following, I discuss some of the current challenges related to the search task, as well as open opportunities for research.

1 From IR to Search Engines

An Information Retrieval (IR) system deals with the storage, indexing, organization, and access to information items, where an item is a reference to a full document or book, a book chapter, a paper, or even a simple Web page. For simplicity of notation and convenience, we will refer to an information item simply as a document, to the set of all documents stored as the collection, and to the set of all distinct words in the collection as the vocabulary [1]. Once the documents have been stored and indexed, the task of the system is to answer user queries written as a query, i.e., as a sequence of a few words. In its most common form, the answer set constitutes of a list of documents of the collection.

We say that the system is operating well if it returns relevant documents to the user, while returning as few non-relevant documents as possible. The impreciseness in our definition is due to the inherent nature of the problem, i.e., finding *relevant* answers involves interpreting the query and inferring user intention, which is difficult because it is tied to the semantics of the world.

To determine the answers to a user query, the IR system implements a *retrieval model*, i.e., a similarity function that takes a query and assigns a numeric score do each pair query-document in the collection. The most fundamental retrieval model is the *vector space model*, which associates with the queries and the documents weighted vectors in an orthonormal space, in which the basis of the space is composed of unit vectors associated with the terms of the vocabulary [3]. Given that queries and documents are represented as weighted vectors, natural similarity functions are vector comparison functions. Thus, the vector space model adopts as similarity function the cosine of the angle between the vector representing the user query and the vector representing a document in the answer set. This simple retrieval model works with small collections, but breaks down with larger collections. To deal with large collections, such as the ones dealt with by modern search engines, the retrieval model needed to evolve.

In the early days of the Web, search engines were implemented as classic IR systems. Their first big evolutionary step was to recognize that the links among Web pages provide additional information that is of value for ranking. Indeed, a Web page that is pointed to by many links reflects a *degree of authority* on a given topic of interest, a degree of authority that can be computed as a normalized count of the number of inlinks. Thus, when a user poses a query, the engine can return first the pages that match the query and have the highest degrees of authority. This is the idea behind the now famous Page Rank, the ranking function used to build Google [2].

Since then, search engines have evolved to include into the ranking function a variety of signals originary from the document collection, from the query stream, and from the user actions. As a result, modern search engines now combine hundreds of signals into a single ranking function in an attempt to better answer the user queries. During the talk, I will discuss some of these signals and how they are used for ranking purposes.

2 Challenges and Opportunities

While search engines have evolved continuously over the last 15 years, they still face many challenges, particularly with infrequent and detailed queries. A user that is seeking a phone number of their doctor will frequently be frustrated with the answers produced by the search engine. To cope with queries of this nature, search engines need to evolve further. They need to evolve to incorporate knowledge encoded in some form that it can be useful for ranking purposes.

Also important, it is frequently the case that determining the most relevant document for a query requires interpreting the content of the document and determining its central topics. That is, document understanding is an area little understood and which needs much more research, if search is to be further improved. During the talk, I will discuss additional examples of challenges and opportunities in the space of search.

References

1. R. Baeza-Yates, B. Ribeiro-Neto: Modern Information Retrieval—The Technology and Concepts Behind Search Engines. Pearson, 917 pages (2011)
2. S. Brin, L. Page: The anatomy of a large-scale hypertextual Web search engine. World Wide Web Conference (1998)
3. G. Salton, A. Wong, C. S. Yang: A Vector Space Model for Automatic Indexing. Communications of the ACM, 613–620, vol 18, num 11 (1975)