Technologies for Information Access and Knowledge Management

Thomas Mandl
University of Hildesheim, Germany

INTRODUCTION

Internet search engines established themselves as an everyday technology for many users. Search engines provide the main access to information on the Internet. According to estimates, some 500 million queries are sent to search engines every day (http://searchenginewatch.com/reports/article.php/2156461) in order to find the most relevant pages among the billions of pages on the Internet. The underlying technology for search engines is provided by information retrieval (IR), which is a key technology in the knowledge society.

IR deals with the search for information and the representation, storage, and organisation of knowledge. Information retrieval is concerned with search processes in which a user needs to identify a subset of information that is relevant for his or her information need within a large amount of knowledge.

For many years, information retrieval systems were mainly used by professional database hosts that provide online access to bibliographic references. Most knowledge objects were manually (or intellectually) indexed. This means that the representation in the form of index terms is selected by human information specialists usually out of a controlled vocabulary. The dominant retrieval model was the Boolean model that advocates an exact match between query and documents, which is inspired by set theory.

In the 1960s, automatic indexing methods for texts were developed. They had already implemented the "bag-of-words" approach, which still prevails. Although automatic indexing is widely used today, many information providers and even Internet services still rely on human information work.

In the 1970s, research shifted its interest to partial-match retrieval models and proved their superiority over Boolean retrieval models. Vector-space and later probabilistic retrieval models were developed. However, it took until the 1990s for partial-match models to succeed in the market. The Internet played a great role in this success. All Web search engines were based on partial-match models and provided ranked lists as results rather than unordered sets of documents. Consumers got used to this kind of search systems, and all big search engines included partial-match functionality. However, there are many niches in which Boolean methods still dominate, for example, patent retrieval.

The basis for information retrieval systems may be pictures, graphics, videos, music objects, structured documents, or combinations thereof. This article is mainly concerned with information retrieval for text documents.

BACKGROUND

The user is in the center of the information retrieval process. Nevertheless, most research tends either to be more user oriented or more algorithm and system oriented. User-oriented research tries to pursue a holistic view of the process whereas system-oriented research is concerned with measuring the effect of system components and tries to resolve efficiency issues.

The information retrieval process is inherently vague. In most systems, documents and queries traditionally contain natural language. The content of these documents needs to be analyzed, which is a hard task for computers. Robust semantic analysis for large text collections or even multimedia objects has yet to be developed. Therefore, text documents are represented by natural-language terms mostly without syntactic or semantic context. This is often referred to as the bag-of-words approach. These keywords or terms can only imperfectly represent an object because their context and relations to other terms are lost.

As information retrieval needs to deal with vague knowledge, exact processing methods are not appropriate. Vague retrieval models like the probabilistic model are more suitable. As a consequence, the performance of a retrieval system cannot be predicted but must be determined in evaluations. Evaluation plays a key role in information retrieval. Evaluation needs to investigate how well a system supports the user in solving his or her knowledge problem (Baeza-Yates & Ribeiro-Neto, 1999).

Web search engines take the information retrieval process to the Internet. They need to contain the following modules (Arasu, Cho, Garcia-Molina, Paepcke, & Raghavan, 2001).

- A crawler collects pages on the Web by starting from known pages, following the links encountered in these seed pages, and iteratively following all links found in further pages (Baeza-Yates & Castillo, 2002).

- An indexer builds a representation of the pages passed on by the indexer. Well-known information retrieval technology is used for this process including linguistic preprocessing and weighting schemes dealing with several occurrences of the same term.
- The user interface (usually a Web client) allows the user to enter queries, presents the results, and should support user strategies like iterative retrieval.
- The query processor analyzes the queries and compares them to the pages represented in the index. Based on the similarity between page and query, a ranking is produced.

Except for the crawler, the other modules are necessary for any information retrieval system and will be introduced in the following sections.

REPRESENTATION AND RETRIEVAL OF TEXT DOCUMENTS

Information retrieval deals with the storage and representation of knowledge and the retrieval of information relevant for a specific user information need. The information seeker formulates a query trying to describe the information need. The query is compared to document representations that were extracted during the indexing phase. The representations of documents and queries are typically matched by a similarity function such as the cosine. The most similar documents are presented to the users, who can evaluate the relevance with respect to their problem.

Representation

Indexing is a process during which words describing the content of a document are chosen as content representation of this document. During automatic indexing, algorithms assign keywords to documents. The indexing process for natural-language documents typically consists of the following steps.

- Word segmentation
- Elimination of stop words
- Stemming
- Compound analysis (for some languages)

Linguistic preprocessing is limited to the level lexemes and morphology. Syntax and semantics are not analyzed because current technology is not able to achieve satisfying quality for mass data. Consequently, the word provides the core for the content representation. The meaning of a text is seen as a set of basic word forms of words that occur in the text. Each word contains its meaning; however, the specific meaning within the text or a sentence can be reconstructed based on the information in the index.

Segmentation is defining the boundaries between the individual words. In European languages, most boundaries can be found by considering blanks. However, other characters need to be considered additionally. In Asian languages, where words are not segmented by blanks, segmentation is a difficult task.

Subsequently, many words that occur frequently are eliminated. These are called stop words and comprise usually articles, prepositions, pronouns, and conjunctions. These words obviously do have meaning; however, because the content is represented following the bag-of-words method, the words are isolated and taken out of their context. Stop words cannot be used for representation in such an approach. In addition, few users would post queries containing stop words. The elimination of stop words also reduces the corpus size typically by 30% and thus leads to higher efficiency (Savoy, 1999).

The most important operation during linguistic preprocessing is stemming. It maps conjugated word forms to their basic form or their stem (e.g., runs -> run, walking -> walk). Morphological variations of words fulfill their function only within their grammatical context. In a bag-of-words approach, all variations can be reduced to their basic form. Stemming improves efficiency also. Three main methods are used for stemming.

- Rule based
- Lexicon based
- Similarity based

The most important algorithms are rule based. The rules describe which steps are necessary in order to obtain the stem from a word form. The number of rules necessary is still under debate (Savoy, 2006).

All index terms are stored in an inverted list from which the documents belonging to a term can be easily accessed.

Weighting

Weighting determines the importance of a term for a document. A term weight measures how well the term represents a document. These weights mirror different levels of relevance. First, the frequency of each term is counted. Weighting assumes that words occurring more often are better representatives for a document. The relationship is not modeled as linearly increasing but is governed by a logarithmic function. The second important parameter of weighting systems is the frequency of a term in the whole collection. Very frequent words contain little discriminative power. Rare words better