Information Retrieval

I

Thomas Mandl
Universität Hildesheim, Germany

Christa Womser-Hacker
Universität Hildesheim, Germany

INTRODUCTION

Information retrieval (IR) technology has become a part of everyday life for many people. Search engines process billions of queries every day. Many web sites include site search systems. Overall, there are millions of such systems installed which all rely on information retrieval algorithms. The user experience of knowledge management for unstructured, textual data is to a large extent shaped by information retrieval technology.

By definition, information retrieval deals with the search for information and the representation, storage and organisation of knowledge (Bates, 2011). Information retrieval is concerned with search processes in which users needs to identify a subset of information which is relevant for their information needs within a large amount of knowledge. The underlying information is unstructured and as such information retrieval is different from databases where access is guaranteed by structure.

Information retrieval is often seen as a sub discipline of computer science. Much research also comes from library science which historically dealt with information retrieval without computer system support. Information science as a discipline, with a focus on the user and usage situation, also contributes much research to information retrieval.

In order to access texts, which form the most important data form for information retrieval, indexing is necessary. The words in the text are collected and organized in a data structure for fast access once a term is used within a query. The index represents the texts and allows search and consequently access to the texts. Nowadays, indexing is most often carried out automatically. Nevertheless, manual indexing has not only prevailed in previous decades, manual indexing is still pursued extensively in some areas.

BACKGROUND

In the 1960s, automatic indexing methods for texts were developed. They already implemented the main approach to extract words without their context, a method which still prevails. Although automatic indexing is widely used today, many information providers and even some 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 which consider search as a set operation. Other models which allow a more flexible match were developed, e.g. the vector space model. One influential system was the SMART system (System for the Mechanical Analysis and Retrieval of Text), which was developed at Cornell University by the group of Gerald Salton (Salton, 1971). SMART already incorporated some of the basic term frequency operations which are still used today (Robertson, 2008).

However, it took until the 1990s for partial match models to succeed on 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, e.g. patent retrieval.

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

DOI: 10.4018/978-1-4666-5888-2.ch386

OVERVIEW

The user is in the center of the information retrieval process. Most research tends to be either more user-oriented or more algorithm and system-oriented. User-oriented research tries to pursue a holistic view of the process, observes information behavior and develops measures for user satisfaction. System-oriented research is concerned with developing new algorithms, measuring the effect of system components and tries to resolve efficiency issues.

The information retrieval process of finding documents related to the search intent of a user is associated with various levels of vagueness. The intent of users may be difficult to interpret for the system and even users themselves may be uncertain about the topic they are searching for. In most systems, documents and queries traditionally contain natural language. Natural language has its inherent vagueness. Not all aspects of language can be well analyzed. Robust semantic analysis for very large text collections or even multimedia objects has yet to be developed. As a result, 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 since their context and relations to other terms are lost.

As information retrieval deals 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 users in solving their knowledge problems.

Web search engines are composed of the following modules (Arasu et al., 2001):

- A *Crawler* collects pages on the web by starting from known pages and following the links encountered in these seed pages and iteratively 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 pre-processing and weighting schemes dealing with several occurrences of the same term.
- The *user interface* allows the user to enter queries, presents the results and should support user strategies like iterative retrieval.
- The *query processor* analyses 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 describing an information need. The query is compared to document representations which were extracted during the indexing phase. The representations of documents and queries are typically matched by a similarity function such as the Cosine coefficient. The most similar documents are presented to the users who can evaluate the relevance with respect to their problem.

Creating a Representation of the Document Content

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 key words to documents. The indexing process for natural language documents typically consists of the following steps:

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