16

Document Information Retrieval

Stefan Klink, Koichi Kise, Andreas Dengel, Markus Junker,
and Stefan Agne

16.1 Introduction

Due to the widespread use of ubiquitous electronic tools like laptops, PDAs and digital cameras for the acquisition, production and archiving of documents, more and more information exists in electronic form. The ease with which documents are produced and shared in the World Wide Web has led to a potentiation of information reachable by each user. This has created a growing demand of adaptive and intelligent access to relevant information.

A study of the International Data Corporation (IDS) shows that the capacity of data in enterprise networks will increase from 3200 petabytes in the year 2002 up to 54,000 petabytes and growing in the year 2004. Cleverdon estimates that the number of new publications in the most important scientific journals will be 400,000 per year [1]. Storing this mass of information is a problem, but searching specific information is a challenge that has become an even greater object of public concern. These tremendous masses of data make it very difficult for a user to find the "needle in the haystack" and nearly impossible to find and flip through all relevant documents and images. Because a human can no longer gain an overview over all information, the risk of missing important data or of getting lost in the haystack is very high.

The task of document information retrieval is to retrieve relevant documents in response to a query that describes a user's information need. Its basic operation is to measure the similarity between a document and a query. Documents with high similarity are presented to the user as the result of retrieval. Although this research field has several decades of history, there still exist some open problems.

The rest of this chapter is organized as follows. In Section 16.2 we give an overview of the vector-space model and basic techniques commonly

utilized in document retrieval. In Section 16.3 three innovative applications are introduced that make use of these techniques.

16.2 Document Retrieval Based on the Vector-Space Model

Throughout the previous few decades of information retrieval science, many models have been invented and a huge amount of variations have been tested. In the field of document retrieval, only a few are established. The most popular retrieval model is the vector-space model (VSM) introduced by Salton [2–4].

16.2.1 Identification of Index Terms

The fundamental unit of an automatic statistical indexing is a word. Each word of a document can be seen as a discriminative feature between the current document and all other documents in the document collection. This feature can be quantified and could also be negative.

During the late 1950s, Luhn showed that the most discriminative words are those that occur with a relatively average frequency within a document collection [5]. If a word – e.g. a pronoun or preposition – occurs very often, then it cannot characterize the content of a document. On the other hand, words occurring very rarely in documents are also rarely used in a user's query. These observations are the foundation of frequency-based techniques for automatic term extraction methods.

Nowadays, it is common not to extract single words according to their frequency but to simply extract all words of the documents and then assign them a frequency-based weight. Words with high frequency will get weights near zero. Due to memory restrictions, they will probably not be included in the index, or they will be skipped over entirely in the case that a *stop word list* is employed. Such a list contains functional words like "with", "also", "can", "the", etc. Furthermore, this list may contain words with no discriminative meaning or words that are given by the semantic of the document collection – e.g. 'computer' or 'program' in a collection of computer science documents.

Another linguistic problem is that the same meaning of a word can be represented by several similar but not identical words – e.g. "sofa" vs. "couch". In the same manner, singular and plural of nouns are often different. This is a significant problem for matching functions that use just the exact occurrence of query terms within the documents.

Several techniques have been proposed to avoid this problem. They are mostly based on a mapping of words to a set of descriptors for word families. Such methods that identify variants of word forms are also known as