6

Automatic Indexing of Documents

6.1

Introduction

In Chapter 4 we considered the goal, function, and structure of an IR system. However, the construction of a specific IR system—that is, the construction of each element that makes up an IR system's structure—can be quite different. First of all, as noted earlier, system construction depends on the IRL and the document selection criterion. The following discussion mainly concentrates on IR systems utilizing the descriptor IRL and the Boolean search criterion. Therefore, when describing various methods of information processing (i.e., the system construction), we will orient ourselves to the Boolean search. Such an orientation is justified because the Boolean search is used in the majority of existing IR systems, and one of the goals of this book is to help improve existing systems and, as a consequence, to enhance the quality of service to users.

The construction of every IR system element is based on the following global requirement: complete automation of all processes in the IR system. Thus, the construction of each system element should constitute an algorithm. Moreover, the IR system construction, described in this chapter and in a number of the chapters that follow, allows the creation of a Boolean system capable of adapting to the user and of performing the optimal search for each individual user. In other words, such a construction permits the implementation of the IR system function formulated in Chapter 4.

We begin the construction of the IR system with one of its basic elements, namely, the block of indexing of documents (BID) (see Chapter 4, Figure 4.8). In other words, this chapter is devoted to the consideration of various approaches to developing automatic document indexing algorithms, that is, algorithms to translate documents from natural language to IRL. We also give an example of such an algorithm. We would like to stress that we are interested only in the creation of completely automated IR systems. Therefore, we will not consider empirical (manual) methods and recommendations to indexers (people who perform indexing manually).

6.2

On the Problem of Indexing

In previous chapters, in analyzing the process of information retrieval, the IRL, and the structure of IR systems, we showed that the problem of indexing arises because the natural language cannot be used in formal retrieval. Consequently, one needs to represent both retrieval objects (documents) and the search requirements (queries) by means of a certain retrieval language. In the framework of the Boolean criterion, the representations of documents and queries are different: documents are represented by unordered sets of descriptors, whereas queries are represented by disjunctive normal forms. Therefore, the processes of their translation are also different. That is why the indexing of documents and the indexing queries are commonly considered separately.

However, is it always necessary to perform indexing within the system; that is, is the indexing process necessary in all systems? This question may seem strange. Indeed, if queries and documents should be represented in IRL, then it seems reasonable that indexing processes should always be performed. However, this is not necessarily true. For instance, it is a well-known fact that a comparatively large number of existing systems do not receive user queries at all. Remember that by definition (see Chapter 3) a query is an expression of the user's IN (in our case POIN) in a natural language (only). Because these systems do not receive queries, query formulations are not constructed by the system. How then is the retrieval performed? When dealing with such systems, the user should provide query formulations instead of queries. In other words, users are required to express their INs not in natural language, but in IRL. Such systems are not something unusual, and therefore the absence of special query translation processes in them is not considered surprising.

The indexing of documents is a different matter. Until recently it has been assumed that indexing of documents in IR systems is necessary. However, an increasing number of practically implemented systems do not use any document indexing. From the retrieval point of view, document texts are not considered texts in a natural language, but rather document profiles that are directly entered into the system. Furthermore, from the point of view of the system, the author of the document does not write it (in the normal sense) but instead creates the document profile. Such a document profile (being for the system just an unordered set of descriptors) is input into the system as if it has been created by indexers. Attempts to eliminate indexing processes in IR systems started quite a while ago. For instance, some 20 years ago, many journals began to ask authors to send along with their manuscripts a set of key words characterizing the contents of the manuscript as precisely as possible. However, the very use of document texts as document profiles was the major step in this direction. As has been mentioned, such an approach is commonly called free text searching. It is clear