7

Automatic Indexing of Search Requests

7.1 Introduction

This chapter continues the discussion of how to construct fully automated IR systems. Again we are interested in the element of the system's structure that is used for indexing, but this time the element will be a block of indexing of search requests (see the structure of an IR system illustrated in Chapter 4, Figure 4.8). We consider the indexing of search requests for two reasons: first, we will seek to implement this process in the system itself and, second, the search request is the most convenient and common form of expressing the psychological human condition known as the information need (IN) (in our case, this IN is called problem oriented information need, or POIN). Recall that the search request is a text in a natural language, a product of the user's attempt to express his psychological condition (POIN). Thus, the search request can be regarded as the most popular representation of POIN in the IR system, and the process of search request indexing can be considered the process of constructing query formulations.

The necessity for indexing and some of its aspects were discussed at length in Chapter 6. Recall that retrieval is not feasible without a language (IRL in our case) and that this language is used to represent both the objects of retrieval (documents in our case) and search requests for them (queries). It follows from the nature of any retrieval process that indexing presents one of the central problems in the creation of IR systems. The algorithm of the automatic indexing of documents (the process of the automatic construction of document profiles) considered in Chapter 6 is one of the alternatives for constructing a structural element of the IR system as a block of indexing of documents (BID). In this chapter, the emphasis is again on creating alternatives, this time for the BIQ (block of indexing of queries, or search requests). Because we are interested in Boolean information retrieval systems, the result of search request indexing (a query formulation) in such systems should be in a disjunctive normal form (as explained in Chapter 5).

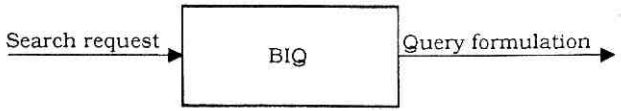

Figure 7.1
BIQ as a black box.

Therefore, we will consider later various approaches to designing algorithms for the creation of such query formulations and will give a specific example of BIQ construction; that is, we will give a detailed description of a successfully functioning algorithm. Moreover, the algorithm described in the following discussion will be designed to work together with an algorithm for the automatic indexing of documents. This means that the query formulation obtained by the algorithm we describe in this chapter will depend on the indexing of documents used in the system. In other words, query formulation obtained by the described algorithm of search request indexing in the system with the automatic indexing of documents and query formulation constructed by the same algorithm and in the same system but with the manual indexing of documents may substantially differ from each other. In addition, results of a search involving query formulations constructed with the automatic indexing of documents are far better as a rule. The importance of this aspect of the algorithm was emphasized in Chapter 6.

Figure 7.1 illustrates the problem under discussion. In the figure, BIQ is presented as a black box.

The input into the BIQ is a flow of information about the user's POIN (search request) and the output is a query formulation in Boolean form. The BIQ itself should be a mechanism (algorithm) that transforms the input information into the output. Because the input is in a natural language and the output is represented in an artificial language, the mechanism should be defined as the algorithm that translates search requests from a natural language into IRL. It is exactly this mechanism that we will now consider.

7.2 Some Aspects of Constructing Query Formulations

Today, in practically all functioning IR systems, query formulations are constructed manually. This is not due to any advantages of the known manual techniques for constructing disjunctive normal forms. On the contrary, the literature contains many complaints from both the IR system designers and those who construct query formulations about the labor-intensiveness and low quality of this process. Why is it that current practice has not changed? Why are the manual methods being used so widely? Clearly, researchers are not opposed in