that this approach cannot be used in systems incorporating weight coefficients in their document profiles. Intended mainly for the Boolean search, the free text search technique is increasingly popular among researchers and is becoming an important alternative in the creation and development of new systems, making them more attractive to the system workers. It essentially simplifies the formation of document profiles providing, at the same time, rather high levels of retrieval efficiency.

It is worth noting that for decades many researchers were convinced that free text search was unacceptable for practical applications. Their confidence was based on two factors. On one hand, until recently the level of technology (first, the level of hardware) did not allow any appropriate practical realization of such a search. In this sense the words of C. J. van Rijsbergen, a well-known researcher in the area of information retrieval, were rather characteristic. He was basing the necessity for indexing on the fact that "the computer is not likely to have stored the complete text of each document in natural language" (van Rijsbergen, 1979). On the other hand, it was believed that the quality of the free text search would be unsatisfactory. This point of view was well described by Salton and McGill, who indicated, "Many experts feel that an uncontrolled indexing vocabulary, which in principle can include the whole variety of natural language, introduces too many opportunities for ambiguity and error" (Salton & McGill, 1983). Indeed, numerous examples of language situations can be given to show the unsatisfactory quality of the free text search (although it is intuitively clear that the original document text has far more content possibilities than the set of its descriptors). However, today this line of development looks promising due, again, to progress in computer science.

In devoting this chapter to the automatic indexing of documents, we use the following considerations. First, this process (indexing) is already implemented in the majority of existing systems. Second, one can regard the free text search method as still being in the experimental phase (although the experiments are rather promising). Still this chapter is not only a nod to the tradition; it will be useful because even in free text search systems it is advantageous to use some traditional indexing procedures. We will consider this aspect after having examined the document indexing process.

To begin, we present the document indexing block (see Figure 4.8) in Figure 6.1 as a black box.

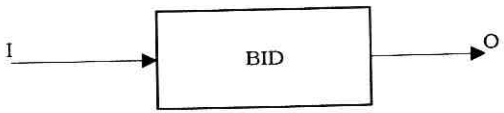

Figure 6.1
Block of indexing of documents.

The input of BID is made up of documents in natural language, whereas the output contains representations of these documents in IRL, that is, the document profiles. Recall that in the framework of the Boolean search criterion, document profiles consist of unordered sets of descriptors. Thus, the function of this subsystem is to translate the input documents from the natural language into IRL. It is clear that the translation result (the document profile) is not intended to be read by a human being (to understand its meaning) but is used for algorithmic (computer-based) information retrieval. Yet if the translation result is not intended for reading (reading gives one an opportunity to evaluate the text), how should one judge the translation quality? In other words, what objectives should the translation meet? We will try to answer this question.

The quality requirement of the input to output translation immediately follows from its function. One can easily see that the BID function is formulated according to the "minimal" quality; that is, in the framework of this function, the indexing subsystem is required only to translate the text from natural language to IRL. In this sense, one can consider the translation speed, stability, and cost as the main characteristics of its quality. However, one would hardly expect that any unordered set of descriptors (document profile), obtained after indexing of a concrete document, will lead to an equally successful (or unsuccessful) search when searching for this document based on the queries of users interested in this document. Clearly, the quality of the translation with respect to the search should be considered. But what does this mean if the translation result is just an unordered set of descriptors? This means that the quality requirements should be taken into account when including (or not including) a certain descriptor into the set. Consequently, when evaluating quality, one should consider the characteristics of individual descriptors (and not the whole set). But what are these characteristics and how can they be determined? Let us consider this problem in more detail.

It seems clear that when indexing documents, one should try to include the most important (in terms of meaning) descriptors in the document profiles. However, some researchers disagree and argue that document profiles should contain descriptors that allow the document to be retrieved successfully, rather than those that well represent a certain meaning. In other words, the document profile should allow the document to be found only by the users that are interested in it. This opinion is based on the assumption that a descriptor that is most appropriate for the purposes of retrieval and one that provides the best representation of meaning are different.

Our own experience, based on analysis of documents from dozens of queries (during an investigation of the efficiency of three different functioning systems, including an experiment that in essence repeated the famous Cranfield project), shows that such situations are not rare. This is partially due to the fact, well known to researchers, that sometimes the success of the document search by different queries is determined by different descriptors of the document pro-