ing (or, as some scientists say, the problem of *ranging*) the whole collection of a document. The main problem was that the calculation of scalar products required considerably more computer operations than a simple comparison did. Hence, information retrieval in the SMART system took much more time than retrieval in Boolean systems (which additionally utilized special file organization allowing the reduction of the number of comparison operations). The increase in CPU time made the search more expensive. Therefore, the idea of splitting the retrieval collection into groups (subsets) of similar (in meaning) documents soon appeared. The creation of the scheme for dividing the collection of documents is usually called the *classification,* while distributing documents into groups (classes) according to this scheme is called the *clusterization.* With the aid of document classification, Salton reasoned, the retrieval can be made more efficient by using sorting only for certain parts (classes) of the collection. Although ignoring the larger part of the collection can cause large losses in the retrieval, the improvement in retrieval time has been considered a more essential factor. Therefore, since the late 1960s a considerable number of publications on classification and clusterization have appeared describing elaborate methods and algorithms mostly for the improvement of the SMART system. However, in this case, new developments in the field of computer science (and the sharp improvement in the internal performance and memory of computers) has been making this problem less and less important. Moreover, in the near future, with the appearance of computers with massive parallel processing, sorting collections containing tens of thousands of documents will not be a problem. This will significantly improve the efficiency of the SMART system.

Thus, generally speaking, changing the semantical component of IRL by assigning weights to lexical units of the language is of certain interest and, in our opinion, is rather promising. At least, one can hardly deny that the potential of this direction is far from being exhausted.

It would seem that changes in the semantical component of IRL should have included the introduction of certain grammatical rules of the natural language for the representation of texts. Such attempts have been made; however, we are not familiar with any of the functioning systems or interesting experimental systems where this approach is used, and therefore we will not consider it in this book.

Having considered the semantical component of IRL, let us discuss its grammar. As we have mentioned, IRL grammar (the grammatical component) is first of all intended for writing (fixing) meaning representations of documents and search requests on a certain physical medium. As seen from the previous consideration, the number of different meaning representations in IR systems is quite small. Therefore, the previous examples are sufficient illustrations of IRL grammar used to represent meanings of documents and queries. However, we have not accentuated the rules for writing meaning representations. We will provide a more detailed explanation of these rules next.

Let us start with Taube's search criterion. In this case, the meanings of documents and search requests is represented in the same form, that is, as sets of descriptors. For example, denoting the document profile as D and the query formulation as Q, we might have the following representation:

$$D = \{a, m, k, x, y, z\} \qquad Q = \{b, f, k, z\},$$

where $a, b, f,$ and so forth are descriptors from the dictionary used in the system. The same representation is also sufficient for a search criterion that takes partial matching into consideration.

When the meaning of the search request is represented by a set of descriptor sets (the Boolean criterion) and the meaning of the document is given by a single set, the expression of the document profile is the same as in the preceding example, whereas the query formulation is given (i.e., is written as) in the disjunctive normal form. In a special case when indexing provides only one set of descriptors, the expression of the query formulation coincides with that of the document profile. For example, the set $S = \{d, n, o, p, r\}$ with descriptors $d, n, o, p,$ and r can represent the contents of both a document and a query given to the system. However, the expression

$$a \vee b \vee c \vee (d \wedge u) \vee (m \wedge n) \vee (m \wedge k \wedge w)$$

can, in the framework of the approaches to meaning representation described above, only be a query formulation (Frants & Shapiro, 1991).

When the semantical component allows us to use descriptor weights, the descriptors of the query formulation—and sometimes those of the document profile—are supplied with weights assigned to them either with or without weight values. For example, the use of weights in the system developed at the U.S. Department of Interior (mentioned earlier) is confined to only marking weighted descriptors by a star. Thus, grammatical rules to represent both document profiles and query formulations in this system have the following form:

$$M = \{a, b^*, c, d, e^*, f^*\}.$$

Here M may represent either a document profile or a query formulation; $a, b, c, d, e,$ and f are descriptors; and (*) marks the descriptor's importance (in our example $b, e,$ and f are important descriptors).

When descriptors are characterized by different values of weights, one can write corresponding sets as follows:

$$N = \{a(3), b(1), c(-2)\},$$

where N may be the representation both of a document and a query; $a, b,$ and c are descriptors; and numbers in brackets give the importance of corresponding descriptors.

All of these examples, in our opinion, are quite exhaustive in describing the existing grammatical means (grammatical components) used to represent documents and queries in IRL.