in Chapter 11. We note here that objects that require averaging of the initial set of values for its evaluation will be called *macroevaluated objects*. Thus, the discussed subprocesses of document search can be regarded as macroevaluated objects in certain situations. Some objects, such as information retrieval systems and other objects to be discussed in Chapter 11, can also be regarded as macroevaluated.

Finally, in defining the framework for our consideration of the evaluation problem, we must discuss the following: a human's evaluation of the functional efficiency of a document search is determined by the concept-based views (i.e., views formed in one's mind) of the comparative merits of various outputs (Cherniavsky & Lakhuti, 1970). These concept-based views give a more accurate understanding of functional efficiency and are formed proceeding from a particular task for which the document search is conducted. In other words, it is this task that shapes concept-based views, which determine in the aggregate the position from which functional efficiency of document retrieval is evaluated. We will call these points of view content criteria of functional efficiency evaluation. It should be emphasized that the formation of concrete content criteria is affected to some extent by the scale of pertinence (relevance) that is employed by the user (expert) in analyzing a document. In this and subsequent chapters, we will assume that document analysis will be performed by the user; therefore, in the context of the analysis of a specific document, we will consider only pertinence evaluation. From our point of view, such an analysis is more correct, although it only affects the terminology: the contents of the discussion are not changed by substitution of the term "pertinent" for the term "relevant."

At present, almost always the binary scale is used, which involves two values of pertinence: pertinent and nonpertinent (they may also have numerical equivalents). In our opinion, this situation is due to the fact that it is essential to obtain, as a result of document search, all pertinent documents and nothing but pertinent documents available in the search collection. Generally speaking, the nonpertinence degree of nonpertinent documents (either included or not included in the output) is inessential, because such documents should not be included in the output. This point of view seems to conform to established practice. This is why, in practice, the binary scale of pertinence is sufficient for analyzing document search results. In these cases the documents are considered pertinent if the user judges that they should be included in the output; otherwise they are nonpertinent. It can be presumed that any user is able, in the end, to judge whether the analyzed document must be included in the output. Bearing the previous discussion in mind, we will assume that, unless specified otherwise, a binary scale of pertinence will be used by document users for analysis.

At the same time, during pertinence analysis of a document by a user, there may be situations when the user finds it so difficult to judge whether the analyzed document is pertinent that he or she will deem it helpful to take into account the nonpertinence degree of nonpertinent documents for the evaluation of the document search results. In this event, one of the judgments from a number of alternatives may be necessary: almost pertinent, possibly pertinent, almost nonpertinent, and so forth. Naturally, the binary scale of pertinence is out of the question here. Rather, a fuzzy scale is used, which includes, in addition to the pertinent and nonpertinent values, other values such as almost pertinent, almost nonpertinent, and so on. (All these values can also have numerical equivalents.) We will also consider the possibility of using a fuzzy scale of pertinence in the analysis of documents by the user.

10.3 Problems of Evaluating the Functional Effectiveness of a Document Search

An evaluation of the functional effectiveness of a document search includes the solution to different problems, for example, determining the achieved functional effectiveness level or finding the case of the highest effectiveness. It follows that in the course of the functional effectiveness evaluation one has to assess the quality of the produced output, to compare the outputs in search of the best, and to consider some other questions of a similar nature. As a rule, there are no simple ways to solve these problems. Moreover, in some cases such problems seem impossible to resolve. Suppose, for instance, that two searches (based on a single search request) in a collection of 10,000 documents result in two outputs, one with 9800 nonpertinent documents and 10 pertinent ones and the other with 9900 nonpertinent documents and 15 pertinent ones. It would be impossible to decide which of the two outputs is the best because neither of the two outputs can be considered useful and, in fact, should be rejected by the user irrespective of the number of pertinent documents in the search collection. In other words, there is no sensible way to say which of the two is better. Such outputs are, naturally, considered unacceptable. The outputs that can provide the basis for evaluation of the functional effectiveness we will call *admissible*. For example, the outputs containing only 15 pertinent documents or 20 pertinent and 10 nonpertinent materials can be considered admissible because such outputs allow for the solution of the questions considered in the evaluation process of the functional effectiveness. In the following discussion, unless it is explicitly stated, the outputs in question will be assumed admissible. Of course, our comment on the difficulties of document search evaluation is only relevant to admissible outputs, because with inadmissible outputs a functional effectiveness evaluation would generally seem unnecessary and impossible.

We should stress here that two methods of evaluation are available: formal and by content. With the by-content method, a person evaluates functional effectiveness on the basis of all information needed for this purpose (such as recall or precision level or information obtained in the course of analyzing the pro-