evaluation was performed. Assume further that during the search based on one half of the queries, outputs contained 1 pertinent document and only 1, whereas for the second half of the queries each output contained 500 documents with 20 pertinent ones. The collection of documents included a total of 20 pertinent documents. It is not difficult to determine that in this situation arithmetic means of recall and precision values achieved are equal to $R^{ave} = 0.525$ and $P^{ave} = 0.52$, respectively. Thus, having evaluated the search service on the basis of the set of values cited, we will conclude that it is of acceptable quality, whereas a user should hardly expect satisfactory search results from this service. This example shows that there are cases for which the evaluation based on a set of averaged values of general search characteristics cannot be trusted, at least in theory. Thus, it is necessary to have a clear idea of whether such cases occur in practice and, if they occur, to determine the conditions under which such an evaluation can be trusted. In other words, the question of when such an evaluation can be trusted and when it cannot remains open.

11.5

Comparative Evaluation of Macroevaluated Objects

The next issue that we will discuss in this chapter is the possibility of performing a comparative evaluation of macroevaluated objects without averaging the search characteristics values. Suppose that we want to select the best search strategy from two strategies, and to do so we have performed a search for the same search requests using both strategies. Assume further that for every search request the functional effectiveness of the search based on the first strategy was found to be higher than the functional effectiveness of the search based on the second strategy in terms of, say, a complex search characteristic I_2. Also assume that all I_2 values were obtained in suitable situations. It is clear that in this case we can legitimately select the first strategy as the best one (among the two compared). In other words, a comparative evaluation of macroevaluated objects can be done without averaging the values of a corresponding complex search characteristic. Note that a comparative evaluation based on averaging I_2 values will lead to the same conclusion.

Moreover, in our opinion, this can be done under less severe constraints than in the example cited. We have in mind the following. Suppose, as in the preceding, that for the purpose of performing a comparative evaluation of macroevaluated objects within the scope of each of them, a search is performed for the same search requests. Also suppose that all search requests were determined for which the functional effectiveness of the search performed within the scope of the first object was found to be higher (in terms of a certain complex search characteristic) than the functional effectiveness of the search performed within the scope of the second object, and the values of the complex search

characteristic leading to these conclusions were attained in suitable situations. If the obtained search requests formed a "qualified majority" in relation to all search requests for which the search was performed, then, in our opinion, there is a basis to consider the first object superior to the second one. In other words, in this case it is also reasonable to evaluate (comparatively) macroevaluated objects without averaging values of a corresponding complex search characteristic. The use of the considered approach in information practice will require solving a number of theoretical and practical issues, in particular the problem of a "qualified majority." Besides, it should be noted that this approach has a specific constraint—namely, when comparing more than two objects, there is a good probability that not only a qualified majority of search requests, but also a simple majority, will not be "formed" for either of them. At the same time this approach has a useful advantage: in some cases it allows us to considerably reduce time-consuming operations in the comparative evaluation of macroevaluated objects in relation to the situation in which averaging is necessary. At this point we will not discuss when such a reduction can occur because all of these issues were discussed in detail in Chapter 10. We stress only that the given advantage of the considered approach to the comparative evaluation of macroevaluated objects justifies the need for further studies in this direction.

11.6

Some Experiments on the Evaluation of Macroevaluated Objects

We have considered the problems and approaches related to the evaluation of macroevaluated objects. Because many researchers have for some time fully appreciated the importance of such an evaluation, it is not surprising that at different times various experiments have been conducted within the scope of which these objects have been evaluated. The well-known Cranfield tests were perhaps the first of these experiments. Research started with experiments in indexing languages, such as the Cranfield I tests. The Cranfield II studies showed, for example, that the automatic indexing of documents was comparable to manual indexing, and this and the availability of computers created a major interest in the automatic indexing and searching of texts. In information science these experiments are justly considered as classic because they led to a number of fundamental results. Some of them are mentioned in Chapter 10.

Among macroevaluated experiments performed in recent years, the most prominent are those conducted within the scope of the Text Retrieval Conference (TREC). These experiments began in 1992 and continue up to the present time. The idea of these experiments is not to evaluate a specific system, its components, or various realizations of these components, but to allow a comparison of systems and any of their components or their various realizations on the basis of specially organized benchmark data. Conceptually, the benchmark data in-