cerning the above, it should be remembered that in a real system there are cases in which the same search request is specified by different users and the same output is evaluated differently by these users. Discrepancies between user evaluations and expert evaluations are not uncommon. Nevertheless, the approach selected by the experiment organizers makes it possible to avoid many technical, methodical, and organizational difficulties in pursuing experiments.

There were 25 systems participating in TREC-1, using a wide range of search techniques. They all received in advance the collection of documents and topics. (We have learned from the experiment organizers that this collection, which was recorded on two disks, cost $2500.) Upon conducting the searches, results were directed to the experiment organizers for comparative analysis. A second workshop (TREC-2) was held in September 1993; 31 systems participated. On the whole, the experiment scheme for conducting TREC-2 was not changed.

In spite of the fact that one of the major objectives of the experiment was comparison of the macroevaluated objects (systems, algorithms, methods, approaches, etc.), the organizers pointed out the following:

> There are so many variables in running the experiment and so many caveats about the evaluation methodologies used that it is very difficult to compare even two systems directly and impossible to come up with a "ranking" of approaches . . . comparisons across systems are very difficult to interpret. (Report on TREC-2, 1993)

Regarding evaluation the experimenters wrote:

> There is a real problem with using the standard measures for evaluation of search in something like TREC; looking at averages of averages is very superficial and hides most of what is actually going on with respect to performance. (Report on TREC-2, 1993)

These statements illustrate the importance of having the ability to evaluate macroevaluated objects, that is, the importance of those approaches and methods discussed in this chapter.

As for the experiment itself, its organizers wrote the following:

> There are many different approaches to information retrieval represented among the TREC-II participants grouped roughly into probabilistic models and variants thereof, vector space approaches, NLP-based, bayesian networks, query expansion and dimensionality reduction, Boolean query construction, combination of results of different retrieval strategies, explorations into document structuring, . . . as well as some outliers like retrieval using n-grams, word pairs, hardware approaches, and some work on efficiency issues. Generalizing results across systems and across approaches is difficult but some trends have already emerged. Simple systems with simple things are still doing really well and the more complex ones are catching up and in some cases surpassing the simple approaches. (Report on TREC-2, 1993)

It should be particularly emphasized that this result was obtained in both TREC-1 and TREC-2 experiments. In this connection note the following: as early as in late 1960s through the early 1970s (see Chapter 5) similar results were obtained. In Chapter 5 we said that as a rule a more thorough syntactical analysis of texts as well as attempts in using semantic analysis in retrieval only led to using more sophisticated programs, increased retrieval duration, and made it more expensive. Some researchers explained all this as insufficient progress in computer science (low speed, inadequate memory, etc.). Later, with progress in computer science, negative effects inherent in complex methods were attributed to the fact that the experiments were run on comparably small collections. Now the TREC experiments have produced results dispelling this reason as well. The existing algorithms of complex methods are seemingly so imperfect that it is too early to consider their positive effects.

The important result of the experiment is the fact that failed attempts of comparing macroevaluated objects are explained not by some deficiencies in the benchmark but by using imperfect methods of evaluating these objects. Thus, we have more support for the importance of selecting evaluation methods, and this should be done before running the experiment. In our opinion, any experiment, especially a comparative one, should be run on the basis of evaluation methods specified in advance and common to all experiment participants.

11.7 Conclusion

This chapter dealt with the problem of evaluating macroevaluated objects. We were interested in an evaluation that would allow for a given search request to predict the results of the search when the search was performed with or within the scope of evaluated object. There is no clear picture in information science as to what macroevaluated objects are expedient to evaluate and which are not. Therefore, we analyzed a number of different objects in an attempt to answer this question for a given object. The approaches used in evaluating these objects are based primarily on the averaging values of functional effectiveness attained in a specially organized series of searches. We showed that in practice, within the scope of the evaluation, it is expedient to apply only one of them, namely, calculation of the arithmetic means of existing values.

An important problem connected with evaluating macroevaluated objects is to determine which search characteristics should be recommended for this evaluation (and which should not be recommended) and in which cases this evaluation could be trusted (and in which cases it cannot be trusted). We showed that a useful instrument in solving this problem for complex search characteristics is the domain of applicability of these characteristics. In addition, for solving this problem we had to introduce the notion of a "confidence threshold," and the analysis for achieving this threshold allows us to solve the problem of "trust"