6.7

Some Issues in Automatic Indexing

We have already mentioned that only automatic indexing with a dictionary has practical implementation in functioning IR systems. Such systems have been used successfully since the late 1960s. Why then have the statistical methods remained experimental for more than 30 years now (note that most scientific literature is devoted to these particular methods)? In answering this question we give several reasons that are, in our opinion, most essential.

First, note that indexing with a dictionary is a rather simple procedure (see, for example, the algorithm described earlier). This cannot be said about statistical methods. For instance, a large number of such methods are based either on term occurrence frequency calculated for the entire collection of documents or on the total number of documents that contain the analyzed term. Obviously, to obtain these data one needs to browse the whole collection of documents. This means that one must input all documents of the collection into the computer in order to begin indexing the first one. Furthermore, each document is indexed by analyzing all documents of the collection. We are not even discussing the technical aspects of such an indexing (such as the volume of computer memory and the time required for indexing one document). In addition, because functioning systems are, as a rule, Boolean ones, selection of descriptors for the document profile requires choosing appropriate thresholds in order to distinguish a significant word (descriptor) from an insignificant one, which is not a simple task. Another problem arises due to the necessity of recognizing word combinations (a necessity that is frequently indicated by researchers). The methods previously described cannot be used for this purpose. Problems of another type (e.g., ideological problems) arise due to insufficient knowledge about statistical laws of natural-language texts. Therefore, assumptions made in various approaches are often very rough approximations of reality. Nevertheless, the argument given by Salton and McGill (1983) seems the strongest.

Salton developed the experimental IR system called SMART, which included several automatic indexing subsystems. Not only were subsystems implementing various approaches to the problem of word significance measurement involved, but subsystems performing indexing with a dictionary also played a part. This situation (in essence, this has been a unique situation) allowed the researchers to conduct the comparative experiment to determine the most effective (clearly, for this particular system) method of automatic indexing. This experiment was unique in providing comparison performed by using the same queries, the same group of researchers, and the same (structural) system. The results obtained by the SMART system were compared to those obtained by the MEDLARS system, which was the most well-known functioning system. The results of this experiment are presented in Figure 6.6.

System Analysis method	Recall	Percent difference from MEDLARS	Precision	Percent difference from MEDLARS
MEDLARS (manual indexing)	0.3117		0.6110	
SMART (word stems with frequency weights)	0.2622	-16	0.4901	-19
SMART (word stems with discrimination value weghting)	0.2872	-8	0.5879	-4
SMART (thesaurus)	0.3223	+4	0.6106	0

Figure 6.6
A comparison of indexing methods. Source: Adapted from G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval* (New York: McGraw Hill, 1983), 104.

The analysis of the results presented in Figure 6.6 shows that the best results are achieved with the method of indexing using a descriptor dictionary (in this case Salton and McGill use the term *thesaurus* for what we call *descriptor dictionary*). Other authors who have conducted other experiments have reported the same result. Because now the majority of functioning systems now involve indexing (remember that this process is not necessary), on the basis of the existing literature one would guess that in many cases indexing is performed automatically. But that is not so. At the same time one would hardly find works supporting the advantages of manual indexing over the automatic indexing. What is the problem then? Why are those engaged in the operation and development of existing IR systems unwilling to automate the indexing process? In the majority of cases, the main reason they give is that the size of the document profile is sharply increased. In other words, the number of descriptors included in the document profile in the course of automatic indexing considerably exceeds the number of descriptors selected during manual indexing. This, they reason, reduces the retrieval quality. Let us consider the main argument starting from the increased size of the document profile.

In discussing the indexing of documents we have not mentioned what kind of documents are considered. This point was omitted on purpose, because only general principles of indexing implementation are important for comprehending the essence of this process. However, now we can admit that not all documents are indexed automatically, for the following reason. During automatic indexing all terms of the text that are contained in the descriptor dictio-