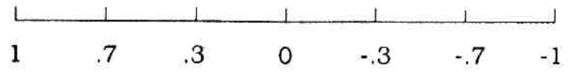

Figure 7.4
Example of partitioning into zones.

descriptors from the positive zones form one part of the final query formulation, and those from the negative zones form another part; furthermore, the descriptors from the negative zones may also form subrequests consisting of one or more descriptors. Then the second (negative) part is connected to the first one by means of the AND NOT operators. It is interesting to note that since the early 1960s, many researchers have been questioning this use of descriptors in a subject search. Having carried out a series of experiments, Dillon and colleagues formed this conclusion on this matter:

> Results using negative feedback in addition to positive feedback are disappointing and somewhat surprising. What was expected (or naively hoped for) was a means of increasing recall through lower thresholds in the positive scale and the achievement of reasonable precision through a judicious application of negative thresholds. Though our experimentation was by no means exhaustive, no negative thresholds were discovered that seemed to hold any promise of improving precision by NOTing documents that did not also diminish recall proportionally. (Dillon et al., 1983)

In general, however, if descriptors with the negative values of Ψ_i are not included in query formulation, that is, if the NOT operator is dismissed, the results of the experiments should be treated as positive.

Between 1983 and 1985, a number of publications appeared (Salton, Buckley, & Fox, 1983; Salton, Fox, & Voorhees, 1985; Salton, Fox, & Wu, 1983) that also dealt with the creation of a method for the automatic construction of query formulations. Their authors noted that within the framework of this method "user's initial natural language statement of information need could be taken as input to an automatic query formulation process. Alternatively, the texts of previously retrieved items judged to be relevant to a given query could be used." (Salton, Fox, & Voorhees, 1985). They especially emphasized that it is precisely the disjunctive normal form (the NOT operator was not used) that would be constructed automatically. The authors proposed that the search weight be calculated not only for individual descriptors in the marked set but also for pairs of descriptors, triples of descriptors, and so on. Moreover, the authors computed a joint occurrence frequency of descriptors (more precisely, the probability of their joint occurrence) for pairs, triples, and so forth based on the occurrence frequency of individual descriptors making up a pair, a triple and so on, by using commonly known formulas of multiplying probabilities for independent events. Thus, the probability of the occurrence of *i* and *j* descriptors jointly in the document profiles of some collection of documents, denoted by

P_{ij}, if one assumes that the occurrence of these descriptors in the document profile is independent (as is assumed by Salton et al.), can be calculated as

$$\frac{R_i \cdot R_j}{N},$$

where R_i and R_j are occurrence frequencies of *i*-th and *j*-th descriptors in the document profile of the collection of *N* documents. In subsequent calculations, the authors use P_{ij} as R_{ij}, the frequency of joint occurrence of descriptors *i* and *j*. Therefore, in calculating the probability of the joint occurrence of three descriptors (three independent events), calculation is reduced to the previous one, namely, to the calculation of two independent events as before:

$$P_{ijt} = \frac{R_t \cdot R_{ij}}{N}.$$

As mentioned earlier, in 1974 (see Voiskunskii & Frants, 1974) we did not follow this path because we found the results of the experiment unsatisfactory, and we argued (in agreement with linguistics) that the occurrence of terms in a document should not be assumed independent. What is more, it is often assumed that in expressing certain meanings, words are used connected with these meanings. Yet Salton et al., making use of the given method of calculation, succeeded in creating an algorithm which, as is clear from the results of the experiment (we discuss this at length later on), leads in a number of cases to the construction of query formulations with better search performance as compared to that conducted by means of query formulations constructed manually.

In calculating the probability of the occurrence of pairs and triples in the document profiles of the collection, Salton et al. actually calculated the number of documents that can be retrieved by every pair or triple of descriptors. In the framework of the proposed method, of most importance for the search (i.e., having the highest weight) are those descriptors that help retrieve the least number of documents. The same holds true for determining the importance of descriptor pairs, triples, and so on. A certain threshold *T*, meaning a number of documents (wanted number) in the output, is used in constructing the disjunctive normal form. This is followed by the selection of *M* more significant (with higher weights) descriptors, and, continuing on the assumption that they do not occur jointly in the documents, calculation is made (by adding together the frequencies of joint occurrence of descriptors) of the number of documents in the output, which can be obtained by means of these *M* descriptors. If this number (Salton et al. call it estret) is far smaller than *T*, then the next (by weight) descriptor is added to the set *M*, and if it is far bigger, then the descriptor (from the set *M*) having the least weight is subtracted from the given set. Then another calculation is performed (both in the first and second cases) and the result is compared with *T*. Whenever in the next calculation estret is higher than *T* and the