lowing formula was suggested:

$$\Psi_i = \frac{x_i \cdot X}{\gamma \cdot Y_i},$$

where x_i is the number of documents on the marked set containing the i-th descriptor; X is the total number of documents in the search file; γ is the total number of marked documents in the request; and Y_i is the number of documents in the search file containing the i-th descriptor.

This formula assigns a high degree of importance to terms (descriptors) occurring in only few documents of a collection. Within the framework of the M-algorithm, it was accepted that if Ψ_i was larger than a certain specified value α, then the i-th descriptor could be included in the query formulation as an independent set (consisting of one descriptor) and combined with other sets by the operator OR. Note that independent sets included in the query formulation and combined by the operator OR are referred to as *subrequests*. This can be illustrated by the following example. Let us assume that in a certain system the query formulation looks like this:

$$D \vee B \vee (E \wedge A) \vee (C \wedge A \wedge F \wedge G),$$

where A, B, C, D, E, F, and G are descriptors. The query formulation in this example consists of four sets connected by the logical operator OR: (1) D, (2) B, (3) (E $\wedge$ A), and (4) (C $\wedge$ A $\wedge$ F $\wedge$ G). These four sets are subrequests in the given query formulation.

After identifying subrequests consisting of one descriptor, the remaining descriptors are partitioned into zones so that the first zone includes descriptors with the occurrence frequency in the entire collection exceeding K_1, the second zone includes descriptors with a frequency higher than K_2 (but less than K_1), and so on. Figure 7.3 illustrates this breakdown.

Figure 7.3 shows that frequency zones were formulated as follows. A fixed step value (in this case 100) was chosen for a transition from one frequency zone to another. The first zone in the given example included descriptors with an occurrence frequency in the entire collection exceeding 500, in the second zone it ranged from 400 to 500, in the third zone it ranged from 300 to 400, and so on. Then subrequests consisting of two and more descriptors were constructed in the framework of the M-algorithm. From the last zone (zone 6 in the example), subrequests of two descriptors were constructed (i.e., the operator AND combined two descriptors); from the next zone (zone 5), subrequests of three descriptors, were constructed; and so forth. Such an approach evidently takes into account the fact that if a subrequest includes descriptors having a comparatively high frequency of occurrence in the collection of documents, then to avoid excessive noise the number of descriptors in the subrequest must be sufficiently large. At the final stage all the obtained subrequests (including those with

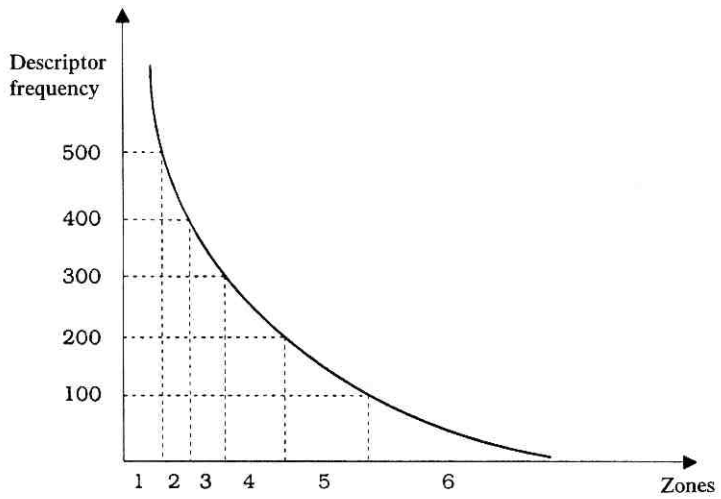

Figure 7.3
An example of the formulation of zones.

one descriptor) were combined by the operator OR into the query formulation finally constructed.

We have described the basic ideas used in the creation of the M-algorithm. Important as they are in the creation of the first automatic query indexing method for the Boolean IR system, they are even more important in the sense that, to a certain extent, they form a basis for almost all the automatic methods available today. Now let us take a closer look at these methods.

We start with the work published by J. Rickman (1972), who, generally speaking, considered some approaches to automating the process rather than focusing on a concrete path (to say nothing of an algorithm) of query indexing. The form of search request that he considered was also marked documents, although in his approach not only were the documents pertinent to the POIN taken into consideration, but so were any other documents assessed (positively or negatively) by the user (and included in the marked set). This choice was due primarily to the fact that Rickman did not seek the automatic construction of query formulations in response to a search request. He considered automation only as a feedback process, that is, to be used only when the user had already appraised the output—regardless of how it was retrieved—by either the manual or the automatic method of query indexing.

Rickman suggested a few methods of constructing query formulations. In one case for "refining a previous search," he suggested that all the descriptors from the document profile of the documents marked by the user as pertinent be