the fact that a writer normally repeats certain words as he advances or varies his arguments and as he elaborates on an aspect of a subject" (Luhn, 1958). In the framework of this idea, Luhn proposed a method for measuring word significance for the purposes of search ("resolving power" of the index word). The essence of this method is the following. First the occurrence frequency of a word in the indexed text is calculated. Because in any "normal" text stop words, such as *the, of,* and *and* occur most frequently, it is suggested that the most frequently appearing words be regarded as insignificant in the given text and that they be excluded from the document profile. Luhn also proposed that the most seldom used words (e.g., words encountered only once) be considered insignificant and that they be excluded from the document profile. The largest resolving power value of the index words extracted from the document text should peak in the middle-frequency range. To illustrate the proposed approach, Luhn used the graph shown in Figure 6.5.

Thus he considered that after successful lower and upper cut-offs have been determined, all the words remaining between them can be regarded as descriptors making up a document profile of rather high quality. Although this particular approach turned out to be practically unacceptable, it was a starting point in the development of many statistical indexing methods. Since that time various researchers have proposed a considerable number of far more successful methods for measuring the significance of descriptors (see, for example, Damerau, 1965). For instance, some researchers have suggested that one should take into account not only the occurrence frequency of a word in the document text but also the frequency of its occurrence in the entire retrieval collection of documents. (Such a frequency has been obtained by totaling all of the occurrence frequencies of the word being calculated for all documents in the collection of documents.) They have suggested that if the occurrence frequency of the

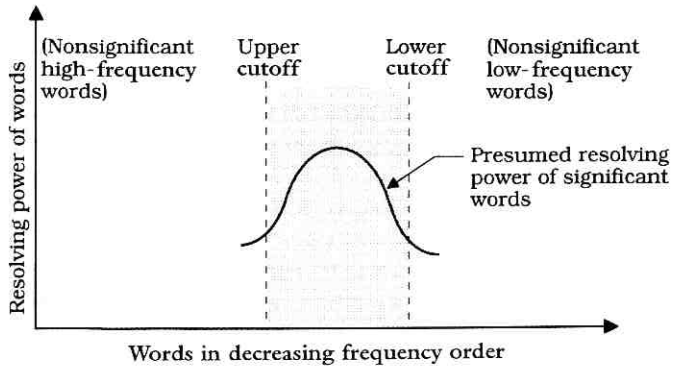

Figure 6.5
Resolving power of significant words.

word is rather high for the given document and relatively low for the entire collection, then the word is important for this particular document (i.e., it can be included in the document profile as a descriptor). Let us explain this idea with the following example.

Let us denote the frequency of occurrence of a term t in the document i as f_{ti}. Let F_t be the total occurrence frequency of the word in n documents. Then

$$F_t = \sum_{i=1}^{n} f_{ti}.$$

Let the significance of the term t for document i be denoted as IMP_{ti}. This value can be calculated as

$$IMP_{ti} = \frac{f_{ti}}{F_t}.$$

If the value of IMP_{ti} exceeds a certain preset threshold, then the term t is included in the document profile of document i as a descriptor.

According to another approach (Edmundson & Wyllys, 1961), one should take into account the frequency of occurrence of a given term in the document (f_{ti}) and the number of documents in the collection that contain this term (D_t). It is thought that if this term is rather frequently encountered in the document being indexed and the number of documents that contain it is rather small, then this term can be used as a descriptor. In this case, the simplest way to measure the significance of the term is

$$IMP_{ti} = \frac{f_{ti}}{D_t}.$$

Again, if the value of IMP_{ti} exceeds a certain preset threshold, then the term t is included in the document profile.

As was noted earlier, many approaches to measuring the importance of a term exist (see, for example, (Bookstein & Kraft, 1977; Bookstein & Swanson, 1974; Robertson & Sparck-Jones, 1976; Salton, 1975; and Sparck-Jones, 1972). We could mention one approach proposed by Sparck-Jones to calculate the so-called inverse document frequency weight (Sparck-Jones, 1972) or the approach proposed by Salton, in which the so-called discrimination values of a term are calculated in order to determine the degree to which the use of the term will help to distinguish the documents from each other (Salton, 1973, 1975). Because all these approaches are aimed mainly at the use of weights and are more of theoretical rather than practical interest, we refer interested readers to the book *Introduction to Modern Information Retrieval* by G. Salton and M. J. McGill (1983), in which many approaches of this kind are considered in detail.

It is worth noting that, because the very approach of measuring the word significance implies an uncontrolled vocabulary, stemming algorithms should be used before the occurrence frequencies are calculated.