sumption that the joint occurrence of words in documents is an evidence in favor of term similarity or close interrelation. This seems to be the reason why we were not able to find references to the practical use of this idea. However, we have presented it here because it is virtually the only known idea concerning conditional equivalence classes containing more than one word.

The other ideas considered the word occurrence frequency and turned out to be useful both for the creation of descriptor dictionaries of the "uniterm" type (with conditional equivalence classes containing one word each) and for the automatic selection of terms for Mooers-type dictionaries. In the latter case, selected words are grouped into conditional equivalence classes manually.

We will now describe one of the most well-known methods. First, the relative word occurrence frequencies are calculated for the collection of documents (or for its sufficiently representative part). Then the words are selected with relative occurrence frequencies exceeding their relative occurrence in texts of the given language by n% (some predetermined value). Obviously, to perform such a selection one needs to input into the computer both the collection of documents in which the word occurrence frequency is calculated and the dictionary of the relative word occurrence frequencies (a fragment of which is given in Figure 5.2). When one wants to select m words for the dictionary, it is not difficult to find the value of n%. For instance, assume we want our dictionary to contain 1000 words, that is, we want to select about 1000 words that are most typical for the collection of documents entered into the computer. Then let the first approximation of n% be, say, 30%. If we obtain considerably more than 1000 words then we can increase n and repeat the selection. Otherwise n should be reduced and the selection should also be repeated. These procedures can be repeated until the quantity of selected words satisfies the developer. Obviously, link-words and general-use words would not be selected in this case. For example, in the document collection devoted to computing hardware, the occurrence frequency of such a commonly used word as "the" will not increase by, say, 30%, whereas the occurrence frequency of special words used in this subject area will definitely increase. As practice shows, such an increase sometimes exceeds 100%.

The graphical illustration of the method just described is given in Figure 5.4. Line A in this figure represents the average occurrence frequency of words in the language. Points mark the frequencies of the words used in the collection of documents. The line h represents the given value of the excess over the average word occurrence frequency. This value is chosen to fit the minimal required quantitative composition of the dictionary. All words (points) reaching or exceeding the h level are included in the dictionary being created.

Another powerful method to select the most important (representative) terms from the collection of documents is the weighting method. According to this method, a term is considered as important for the given collection of documents if it appears in the collection with maximum frequency *and* appears in common texts rather seldom. Thus weights are calculated as follows:

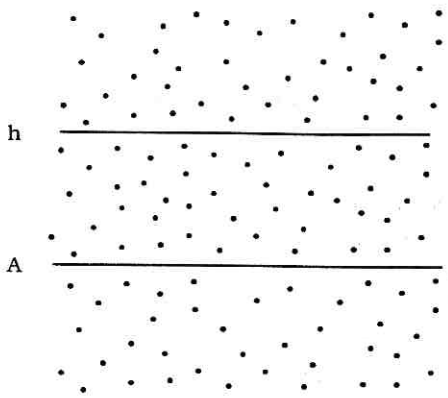

Figure 5.4
The graphic representation of the method used to select words whose relative occurrence frequencies exceed their average values by h%.

$$w = \frac{f_i Q}{q F_i},$$

where

f_i = occurrence frequency of the i-th word in the collection of documents;
q = number of words in the collection of documents;
F_i = occurrence frequency of the i-th word in texts in the language in which the documents are written; and
Q = number of words in the texts that are used for the computation of the occurrence frequency of each word in the language.

Obviously, to select the most weighted words and include them in the dictionary, one should input both the collection of documents (or a sufficiently representative part) and the dictionary of relative word occurrence frequencies into the computer.

Let us look at the following example. Assume we have entered into the computer 500 abstracts of articles about various aspects of information science. Also assume that the average length of an abstract is 100 words. Suppose that the frequency dictionary containing 1,000,000 words (its fragment is presented in Figure 5.2) is also stored in the computer. Now assume that a certain term has appeared 120 times in the collection of documents, whereas its occurrence fre-