Letter	Frequency	Letter	Frequency	Letter	Frequency
О	0,09	М	0,026	Й	0,010
Е	0,072	Д	0,025	Х	0.009
А	0,062	П	0,023	Ж	0,007
И	0,062	У	0,021	Ю	0,006
Т	0,053	Я	0,018	Ш	0.006
Н	0,053	Ы	0,016	Ц	0,004
С	0,045	З	0,016	Щ	0,003
Р	0,040	Ь,Ъ	0,014	Э	0,003
В	0,038	Б	0,014	Ф	0.002
Л	0,035	Г	0,013		
К	0,028	Ч	0,012		

Figure 5.1
Character occurrence frequencies in texts written in Russian.

Term	Frequency
the	0,069,971
of	0,036,411
and	0,028,852
to	0,026,149
a	0,023,237
in	0,021,341
that	0,010,595
is	0,010,099
was	0,009,816
he	0,009,543

Figure 5.2
Relative word occurrence frequencies in texts written in English (the total number of words is 1,000,000). Source: Adapted from H. N. Kucera & W. N. Francis, *Computational Analysis of Present-Day American English* (Providence, RI: Brown University Press, 1968).

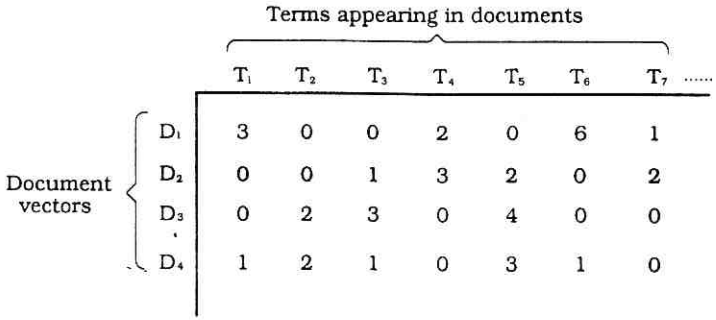

Figure 5.3
Term-document matrix.

constructed for many languages. Figure 5.2 is compiled from the selected texts with the number of words totaling one million (Kucera & Francis, 1968).

As we will show later, several ideas for the automatic compilation of descriptor dictionaries include the use of such word lists. Nevertheless, first approaches did not make use of them. Next we describe one of the first methods, which is rather representative for the mid-1960s. This approach suggests using the terms contained in a set of documents retrieved from a collection of documents considered typical for a given subject area. Then frequencies of word occurrences in the chosen documents are calculated. Each document is identified by its words with high occurrence frequencies. Then a term-document matrix is constructed. An example of such a matrix is presented in Figure 5.3.

The matrix element lying on the intersection of the i-th row and j-th column represents the weight of the j-th term in the i-th document. Given this matrix, one can use well-known methods of statistical processing to calculate coefficients of similarity between terms on the basis of the joint occurrence characteristics of words in the chosen documents. The similarity coefficient is calculated for each pair of terms depending on the frequency of their joint occurrence in the collection of documents. According to Figure 5.3, terms T_1 and T_6 are attributed to both documents D_1 and D_4, though with different weights. At the same time they do not appear in documents D_2 and D_3. This method assumes that these two terms may belong to the same conditional equivalence class of the created descriptor dictionary (Salton & McGill, 1993).

It should be noted that the idea of this method is rather schematic. For instance, nothing is said about the calculation of term weights and why documents should be identified by words with high occurrence frequency (note that, according to Figure 5.2, the most frequently used words are "the," "of," and "and"). Furthermore, the approach discussed is based on a not-so-obvious as-