system (Salton, 1971). The authors of this system themselves have characterized this system as follows: "The SMART system is perhaps the best known of the experimental systems" (Salton & McGill, 1983). This seems to be true, at least according to the number of publications devoted to this system, which exceeds the number of publications on any other experimental or practically implemented system. The correspondence criterion used in this system has the distinction of being very original and also of being used exclusively in a computerized environment. In essence, each search request to the SMART system results in the sorting of documents; that is, the number of times the collection of documents is sorted is equal to the number of search requests. Documents sorted according to a concrete search request are placed in order of decreasing values (calculated using the weights) of the correlation between a search request and the documents in the collection. It is assumed that a document with the highest correlation value is the most appropriate for the corresponding search request and should be read first by the user. Let us look at its implementation in more detail.

All descriptors of the system dictionary form the space of terms. Document profiles are built in this space. Each descriptor has the corresponding coordinate in this space. Document profiles are represented by vectors in the term space. If a certain document deals with the i-th descriptor, then the i-th coordinate of its vector is nonzero; otherwise it is zero. Concrete values of nonzero coordinates are identified with their weights in the document profiles considered. Manual indexing implies that weights are specified by a human being and in this case the quality of indexing depends on the skills and experience of the person performing the indexing. However, various methods have been developed for the automatic weight specification. For example, the descriptor weight can be defined as equal to the appearance frequency of this descriptor in the document.

Query formulations are also built as vectors in the term space. The extent of proximity of a document profile (let us denote it as **A**) to the query formulation (let us denote it as **B**) is commonly defined as a cosine of the angle between the corresponding vectors:

$$r(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{AB}}{|\mathbf{A}||\mathbf{B}|},$$

where **AB** is the scalar product; and $|\mathbf{A}|$ and $|\mathbf{B}|$ are lengths of vectors **A** and **B**. The value $r(\mathbf{A}, \mathbf{B})$ is sometimes called the coefficient of the correlation between the query formulation and the document profile. Vectors **A** and **B** are considered close if their correlation coefficient is close to 1. Because all vector coordinates are non-negative, the minimum possible value of the correlation coefficient is zero. This value corresponds to a case when all terms in the query and the document are different. Vectors of both document profile and query formulation consist of the descriptors and the values of nonzero coordinates. Let

vectors **A** and **B** be represented as **A**[16(1), 27(3), 195(4), 327(1), 592(3)] and **B**[16(2), 82(3), 195(2), 327(2), 984(2)]. Then their correlation coefficient is

$$r(\mathbf{A}, \mathbf{B}) = \frac{1 \times 2 + 3 \times 0 + 0 \times 3 + 4 \times 2 + 1 \times 2 + 3 \times 0 + 0 \times 2}{\sqrt{1^2 + 3^2 + 4^2 + 1^2 + 3^2} \times \sqrt{2^2 + 3^2 + 2^2 + 2^2 + 2^2}}$$

$$= \frac{2 + 8 + 2}{\sqrt{36} \times \sqrt{25}} = \frac{12}{30} = 0.4.$$

Thus, the collection of documents is ordered (sorted) in the order of decreasing cosines of angles between query and document vectors. But what does the user obtain? Obviously, it is not practical to select all documents in the collection, even if it is sorted. Therefore, various approaches to the selection criterion have been proposed. For example, it has been suggested that the user should obtain all documents with correlation coefficients above a certain specified value. However, this approach has several shortcomings. First, in situations when all correlation coefficients are below the threshold, nothing will be selected. On the other hand, too many documents may be selected for a given request; that is, too many documents may have correlation coefficients above the specified threshold.

Another proposal has been to select exactly N (say, 10) best documents (with the highest correlation coefficients). In this case, however, specificity and "broadness" of search requests are not considered. Furthermore, this approach implies that all requests have the same number of corresponding documents in the system. The authors of the system consider it important in generating the output to take into account the user's determination of the number of documents in the output. Before the search is started, users are asked to specify the number of documents in the output. For example, when specifying the search request, one user would like to obtain 6 documents, another would like to have 17 documents, and some other user, say, 38 documents. Clearly, each will receive the exact number of documents requested regardless of the number of appropriate documents. In each case the documents obtained will have a higher correlation coefficient to the query than the documents that were not selected.

Note that with the development of more advanced methods for on-line services, this problem has lost its validity. For example, in the on-line search mode, the user can view as many documents as he or she wishes. To a certain extent, a problem still exists in the *selective dissemination information* activities, when a system sends information to its users according to their long-term search requests on a regular basis (for example, once a month). However, the number of such systems is getting smaller and smaller.

Another point is worth noting in connection with the SMART system. In the mid 1960s—by the time the main principles used in the SMART system had been developed—developers were rather pessimistic about the problem of sort-