ON THE ROLE OF WORDS AND PHRASES IN AUTOMATIC TEXT ANALYSIS

G SALTON
Cornell University

Automatic indexing normally consists in assigning to documents either single terms, or more specific entities such as phrases, or more general entities such as term classes. Discrimination value analysis assigns an appropriate role in the indexing operation to the single terms, term phrases, and thesaurus categories. To enhance precision it is useful to form phrases from high-frequency single term components. To improve recall, low-frequency terms should be grouped into affinity classes, assigned as content identifiers instead of the single terms.

Collections in different subject areas are used in experiments to characterize the type of phrase and word class most effective for content representation.

The following typical conclusions can be reached:

a) the addition of phrases improves performance considerably;

b) use of phrases is better with corresponding deletion of single terms in practically all cases;

c) the use of both high-frequency and medium-frequency phrases is generally more effective than the use of either phrase-type alone;

d) the most effective thesaurus categories are those which include a large number of low-frequency terms;

e) the least effective classes either consist of only one or two terms, or else they include terms with unequal frequency characteristics permitting the high-frequency terms to overcome the others.

The discrimination value theory is developed and appropriate experimental output is supplied.