Information retrieval from the bibliography of coal technology and general organic petrography (CBIB) using automatic key-word indexing on a personal computer

Neely H. Bostick
U.S. Geological Survey
Denver Federal Center, Denver, CO 80225-0046

KWIC (keyword in context) indexing is a form of automatic indexing available on mainframe computers since the mid 1960's. A program of this type was used to assemble and index both the full bibliographic database of coal properties (CBIB) at the Illinois State Geological Survey and the selected part on petrography, catagenesis and geochemistry of dispersed organic matter (MODBIB) published in 1975. Today the power of personal computers matches early mainframes and exceed them in one important way: Information retrieval can be done in the geoscientist's office in just a few seconds. The CBIB database has been imported to MS-DOS systems and has been indexed with commercially available software.

One program to index and retrieve information from the CBIB data file on a personal computer has the following search features: (1) single or alternative words at one time, (2) multiple word logical connectors AND, OR, NOT, (3) exact phrases up to 160 characters, (4) associated words (WITHIN a certain distance in text), (5) logical parentheses [apatite AND (coal OR lignite OR anthracite)], and (6) wild cards for word roots [oxid* => oxidation + oxidative + oxide + oxidised + oxidized + oxidizing]. Successful searches retrieve all files that contain the designated "hit words." These words are automatically highlighted in context when the complete text is displayed.

Location and retrieval of information from text files such as bibliographies, sample records, and laboratory notes can be accomplished in two ways: (1) pre-index by assigning codes from a restricted list, such as certain geologic age names, constituents or topics, (2) automatically index English (or other) language words. The first approach requires building a thesaurus of "defined" categories, manually assigning them, then indexing them accordingly. This requires much discipline and time for input and is difficult when new categories of interest are to be added; the advantages are small storage space and possible simple retrieval. The second "automatic" approach uses indexing of unassigned natural text (except trivial words: FOR, THROUGH, BUT ...). The advantages are low initial preparation and easy expansion interest; the disadvantages are greater storage requirements and complex retrieval.

A seed group of interested petrologists has been formed to update and expand the existing data file (now about 8000 entries) and to provide information and demonstrations of software that can be used to access the data.