COMPUTERIZED METHODS OF DATA COMPRESSION AND ANALYSIS

ABSTRACT

A computerized method and apparatus compresses symbolic information, such as text. Symbolic information is compressed by recursively identifying pairs of symbols (e.g., pairs of words or characters) and replacing each pair with a respective replacement symbol. The number of times each symbol pair appears in the uncompressed text is counted, and pairs are only replaced if they appear more than a threshold number of times. In recursive passes, each replaced pair can include a previously substituted replacement symbol. The method and apparatus can achieve high compression especially for large datasets. Metadata, such as the number of times each pair appears, generated during compression of the documents can be used to analyze the documents and find similarities between two documents.

Figure 1