ABSTRACT

METHOD OF LANGUAGE CODING IDENTIFICATION AND DATA FORMAT THEREFOR

A method of language coding identification for identifying different language coding is composed of the steps of: checking whether or not a text of a language to be identified is ASCII coded, applying the double-byte coding identification method; applying the Latin-based coding identification method; and applying the single-byte coding identification method sequentially to the text in case the previous step is not successful, the single-byte coding identification method is further composed of the steps of: generating a language profile of each supported language coding by analyzing a sample text and extracting occurrence possibility of all two character sequence combinations in the sample text; calculating a confidence value of each language coding by checking the occurrence possibility of all the two character sequence combinations in a text string of the sample text by the respective language profile; and selecting language coding with the highest confidence value.

Figure 1