

CJK UNICODE LANGUAGE IDENTIFICATION METHOD

ABSTRACT

5 The present invention provides a model to identify the unknown CJK
language of a graph of Unicode combined text data by their UTF-16 values. At
first, the model goes through the text data content with its digital character value.
According to the statistical analysis for each character, the model generates the
possibility for each CJK language from the text data. Then the model identifies
10 the language by selecting the largest possibility most favorably.

Figure 1