

## **METHODS AND APPARATUS FOR REMOVING A DUPLICATED WEB PAGE**

### **ABSTRACT**

Methods and Apparatuses are disclosed for removing a duplicated web page. An exemplary method may include acquiring a plurality of web pages of a predetermined type extracting a feature code of a current web page and a number of text characters contained in the current web page for each web page. The method may also include looking up a data table to determine whether the feature code is contained in the data table. If the feature code is contained in the data table, the method may further include reading a number of text characters of the web page in the data table corresponding to the feature code, and discarding the current web page when a difference between the read number of text characters and the extracted number of the text characters is within a range.

**Fig. 1**