Abstract

The access to huge amount of information sources on the internet has been limited to browsing and searching due to the heterogeneity and the lack of structure of the web information sources. This has resulted in the need for automated Web Information Extraction (IE) tools that analyze the Web pages and harvest useful information from noisy content for any further analysis. The goal of this survey is to provide a comprehensive review of the major Web IE tools that used for Web text and based on Document Object Model for representing the web pages. This paper compares them in three dimensions: (1) the source of content extraction, (2) the techniques used, and (3) the features of the tools, moreover the advantages and disadvantages for each tool. Based on this survey, we can decide which suitable Web IE tool will be integrated in our future work in Web Text Mining.

References


**Index Terms**

Computer Science Information Sciences

**Keywords**

Knowledge Engineering Document Engineering Information Extraction Document Object
<table>
<thead>
<tr>
<th>Model</th>
<th>Documents</th>
<th>Web</th>
</tr>
</thead>
</table>