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

Approaches to Bringing Internet Services to WAP Devices. In Proceeding of 9th International
World-Wide Web Conference.

data extraction tools. In Proceeding of SIGMOD Conference.
- C. Chang, M. Kayed, M. Girgis, and K. Shaalan, "A Survey of Web Information
Extraction Systems"; Journal of IEEE TRANSACTIONS ON KNOWLEDGE AND DATA
ENGINEERING, 2006.
- Fiumara, G. 2007. Automated Information Extraction from Web Sources: a Survey. In
Proceeding of the 3rd International Conference on Communities and Technology.

- Soderland, S.  1997. Learning to extract text-based information from the World Wide
Web. In Proceeding of 3rd International Conference on Knowledge Discovery and Data Mining
(KDD).
at Computer Science.
mining. In Proceeding of the 9th ACM SIGKDD International Conference.
- S. Debnath, P. Mitra, N. Pal, and C. Giles, "Automatic Identification of
Informative Sections of Web-pages"; In Journal IEEE Transactions on Knowledge and
Data Engineering, 2005.
In Proceeding of the 9th International PhD Workshop on Systems and Control: Young
Generation Viewpoint. Slovenia.
- M. Asfia, M. Pedram and A. Rahmani, "Main Content Extraction from Detailed

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>