Abstract

The rapid expansion of internet has made web a popular place for disseminating and collecting information and also it opens up many research topics on various research fields. Since last few years, several attempts have been made on Web-based research, particularly based on HTML web pages because of their huge availability. So that many Research Data Sets have been created and most of them are made available on web. But W3 consortium stated that, HTML does not provide a better description of semantic structure of the web page contents. To overcome this drawback, Web developers started to develop Web page(s) on XML, Flash kind of new technologies [1]. It makes a way for new research methods. This article mainly focuses on Data Set creation on XML Web pages by using Sequential Search, Link Extraction and
String based Classification methods for future research avenues on XML Web pages.

References
- Book: Magdalini Eirinaki, ‘WEB MINING: A ROADMAP’
- Lan Yi, Bing Liu, and Xiaoli Li, 2003, ‘Eliminating noisy information in web pages for data mining’. In KDD ’03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 296{305, New York, NY, USA. ACM.
- http://www.w3c.org/DOM/
- Soumen Chakrabarti, 2000, ‘Data mining for hypertext: A tutorial survey’ Volume 1, Issue 2 - page 1 ACM SIGKDD.

Index Terms

Computer Science
Computational Intelligence
Keywords
URL data set  XML URL's  URL Extraction  URL Classification