Abstract

The World Wide Web is growing at a rapid rate. A web crawler is a computer program which independently browses the World Wide Web. The size of web as on February 2007 was 29 billion pages. One of the most important uses of web page is in indexing purpose and keeping web pages up to date which can be used by search engine to serve the end user queries. Web is dynamic in nature; hence we need to update the web pages constantly. In this paper, we put forward a technique to update a page stored in web repository. This paper put forward an efficient method to refresh a page. We are proposing two methods for refreshing the page by comparing the page structure. First method compares the page structure with the help of tags used in it. And second method creates a document tree compare structures of pages.

References

- Sergey Brin and Lawrence Page, "The Anatomy of a Large-Scale Hypertextual
Crawling the Web Surface Databases


- Anshika pal, Deepak Singh tomar, S. C srivastava, "effective focused crawling based on content and link structure analysis", international journal of computer science and information security, vol 2, no. 2, June 2009
- Nitdi Tyagi, Deepti Gupta, "A novel architecture for domain specific parallel crawler", Indian journal of computer science and engineering, vol 1, no 1, 44 – 53.
- Bergman, Michael K, "White paper: the deep web : surfacing hidden value", Vol 7, Issue 1, August 2001
- Vipul Sharma, Mukesh Kumar, Renu Vig, A Hybrid Revisit Policy For Web Search, Vol 3, No 1, Feb 2012, Page(s): 36 - 47

Index Terms

Computer Science

Information Sciences
Keywords

Web Crawler  WWW  Spidering  Search Engine  Surface Web  Deep Web

Document Tree Structure