Abstract

Today, there is a great deal of information available in Web world and the only way to access them is through search relationships. Web crawler is an automated script that independently browses the web. Web crawler starts its task with a "seed URL" and then traces links available in each page. This encountered many available crawlers with essential difficulties. Identification of search intermediate and selection of a proper inquiry, on one hand, and retrieving documentaries returned by the web as the result, on the other hand, are issues that intensify challenges available for web crawlers. The aim of the present paper is to investigate available challenges and guidelines in the field of deep web and intensive crawling.

- Change K. C. C., He B., Li C., Patel M., Zhang Z. 2004 Structured databases on the web: Observations and implications. SIGMOD Record
- Chakrabarti S., Berg M. V. D., Dom B. 1997 Distributed Hypertext Resource Discovery through Example” ; In 25th International Conference on Very Large Data Base, USA
- Cho J., Garcia-Molina H. 2000 the Evolution of the Web and Implications for an Incremental Crawler. In 26th International Conference on Very Large Data Bases, USA, pp. 200-209
- Cho J., Garcia-Molina H. 2000 Synchronizing a Database to Improve Freshness. In ACM SIGMOD International Conference on Management of Data, USA, pp. 117-128
- Diligenti M., Coetzee F., Lawrence S. 2000 Focused Crawling Using Context Graphs. In 26th International Conference on Very Large Databases (VLDB), Cairo, Egypt, pp. 527-534
- Alvarez M., Pan A., Raposo J. and Vina A. 2006 Crawling the client-side hidden web
- Broder A., Carnel D. 2005 Sampling search-engine results. In 14th international Conference on world Wide Web, Chiba, Japan
- Qin J., Chen H. 2005 Using Genetic Algorithm in Building Domain-Specific Collections: An Experiment in the Nanotechnology Domain. In 38th Annual Hawaii International Conference on System Sciences, USA
- Rennie J., McCallum A. 1999 Using Reinforcement Learning to Spider the Web Efficiently. In 16th International Conference on Machine Learning, USA, pp. 335-343
- Shkapenyuk V., Suel T. 2001 Design and Implementation of a High-Performance Distributed Web Crawler. In 18th International Conference on Data Engineering, USA, pp. 357-368
- Gulli A., Signorini A. 2005 The Index able Web is More than 11.5 billion pages. In 14th International World Wide Web Conference, Chiba, Japan
- Gravano L., Garcia-Molina H., Tomasic A. 1999 GI OSS: Text source discovery over the Internet. ACM TODS
- Ipeirotis P. G., Gravano L. 2002 Distributed Search over the Hidden web: Hierarchical Database Sampling and Selection. In 28th VLDB Conference, Hong Kong, China
- Barbosa L., Freire J. 2004 Siphoning Hidden-Web Data through Keyword-Base Interfaces. In SBBD
- Castillo C. 2004 Effective Web Crawling. In ACM SIGIR. Vo. 39, Issue 1

Index Terms

Computer Science
Information Sciences

Keywords

Intensive crawler  search engine  genetic algorithm  deep web