{tag}

{/tag}
International Journal of Computer Applications

Number 7 - Article 2

Year of Publication: 2011

Authors:

M. Sunil Kumar

P.Neelima

10.5120/1963-2629

{bibtex}pxc3872629.bib{/bibtex}

Abstract

The Web is a context in which traditional Information Retrieval methods are challenged. Given the volume of the Web and its speed of change, the coverage of modern web search engines is relatively small. Search engines attempt to crawl the web exhaustively with crawler for new pages, and to keep track of changes made to pages visited earlier. The centralized design of

crawlers introduces limitations in the design of search engines. It has been recognized that as the size of the web grows, it is imperative to parallelize the crawling process. Contents other then standard documents (Multimedia content and Databases etc) also makes searching harder since these contents are not visible to the traditional crawlers. Most of the sites stores and retrieves data from backend databases which are not accessible to the crawlers. This results in the problem of hidden web. This paper proposes and implements DCrawler, a scalable, fully distributed web crawler. The main features of this crawler are platform independence, decentralization of tasks, a very effective assignment function for partitioning the domain to crawl, and the ability to cooperate with web servers. By improving the cooperation between web server and crawler, the most recent and updates results can be obtained from the search engine. A new model and architecture for a Web crawler that tightly integrates the crawler with the rest of the search engine is designed first. The development and implementation are discussed in detail. Simple tests with distributed web crawlers successfully show that the Dcrawler performs better then traditional centralized crawlers. The mutual performance gain increases as more crawlers are added.

Reference

- David Karger, Eric Lehman, Tom Leighton, Matthew Levine, Daniel Lewin, and Rina Panigrahy. (1997) Consistent hashing and random trees: Distributed caching protocols for relieving hot spots on the World Wide Web. In Proc. of the 29th Annual ACM Symposium on Theory of Computing, pages 654-663, El Paso, Texas,
- David Karger, Tom Leighton, Danny Lewin, and Alex Sherman. (1999.) Web caching with consistent hashing. In Proc. of 8th International World-Wide Web Conference, Toronto, Canada.

- Demetrios Zeinalipour-Yazti and Marios Dikaiakos. (2002) Design and implementation of a distributed crawler and filtering processor. In Proc. ofNGITS 2002, volume 2382 of Lecture Notes in Computer Science, pages 58-74.
- Hongfei Yan, Jianyong Wang, Xiaoming Li, and Lin Guo. (2002) Architectural design and evaluation of an efficient Web-crawling system. The Journal of Systems and Software, 60(3): 185-193,.
- Java™ remote method invocation (RMI). http: //Java . sun.com/products/jdk/rmi/.
- Marc Najork and Janet L. Wiener. (2001) Breadth-first search crawling yields high-quality pages. In Proc. of 10th International World Wide Web Conference, Hong Kong, China,.
- Paolo Boldi, Bruno Codenotti, Massimo Santini, and Sebastiano Vigna. (2001) Trova-tore: Towards a highly scalable distributed web crawler. In Poster Proc. of Tenth International World Wide Web Conference, pages 140-141, Hong Kong, China,
- Robert Devine. (1993) Design and implementation of DDH: A distributed dynamic hashing algorithm. In David B. Lomet, editor, Proc. Foundations of Data Organization and Algorithms, 4th International Conference, FODO'93, volume 730 of Lecture Notes in Computer Science, pages 101-114, Chicago, Illinois, USA,. Springer-Verlag.
- Tushar Deepak Chandra and Sam Toueg. (1996) Unreliable failure detectors for reliable distributed systems. Journal of the ACM, 43(2):225-267,.
- Vladislav Shkapenyuk and Torsten Suel. (2002) Design and implementation of a high-performance distributed web crawler. In IEEE International Conference on Data Engineering (ICDE),.

- L. Lamport, “Paxos made simple,” ACM SIGACT News, vol. 32, no. 4, pp. 51–58, December 2001.
- V. Paxson, “End-to-end routing behavior in the Internet,”ACM SICOMM Computer Communication Re-view, vol. 35, no. 5, pp. 43–56, October 2006.
- A. Crespo and H. Garcia-Molina, “Semantic overlaynetworks for p2p systems,” Stanford University, Tech. Rep. 2003-75, 2003.
- M. Bender, S. Michel, P. Triantafillou, G. Weikum, and C. Zimmer, “P2p content search: Give the Web back to the people,” February 2006, international Workshop on Peer-to-Peer Systems (IPTPS).
- M. Shokouhi, J. Zobel, F. Scholer, and S. Tahaghoghi,“Capturing collection size for distributed noncooperative retrieval,” in Proceedings of the Annual ACM SIGIR Conference. Seattle, WA, USA: ACM Press, August 2006.
- L. A. Barroso, J. Dean, and U. H¨olzle, “Web search for a planet: The Google Cluster Architecture,” IEEE Micro, vol. 23, no. 2, pp. 22–28,Mar./Apr. 2003.
- A.-J. Su, D. Choffnes, A. Kuzmanovic, and F. Bustamante, “Drafting behind Akamai travelocity-based detouring),” in Proceedings of the ACM SIGCOMM Conference, Pisa, Italy, September 2006, pp. 435–446.
- S. M. Beitzel, E. C. Jensen, A. Chowdhury, D. Grossman, and O. Frieder, “Hourly analysis of a very large topically categorized web query log,” in SIGIR ’04: Proceedings of the 27th annual international ACM SI- GIR conference on Research and development in in- formation retrieval. New York, NY, USA: ACM Press, 2004, pp. 321–328.
- K. Risvik and R. Michelsen, “Search engines and web dynamics,” Computer Networks, pp. 289–302, 2002.
- F. Cacheda, V. Carneiro, V. Plachouras, and I. Ounis, “Performance analysis of distributed information retrieval architectures using an improved network simulation model,” Information Processing and Management, vol. 43, no. 1, pp. 204–224, 2007.
- W. B. Cavnar and J. M. Trenkle, “N-gram-based text categorization,” in Proceedings of SDAIR-94, 3rd An- nual Symposium on Document Analysis and Informa- tion Retrieval, Las Vegas, US, 1994, pp. 161–175.
- G. Grefenstette, “Comparing two language identification schemes,” in Proceedings of the 3rd interna- tional conference on Statistical Analysis of Textual Data (JADT 1995), 1995.

Index Terms

Computer Science Information Retrieval

Key words

web search engines crawling software architecture