Abstract

Web crawlers are programs designed to fetch web pages for information retrieval systems. Crawlers facilitate this process by following hyperlinks in web pages to automatically download new or update existing web pages in the repository. A web crawler interacts with millions of hosts, fetches millions of pages per second and updates these pages into a database, creating a need for maintaining I/O performance, network resources within OS limits, which are essential in order to achieve high performance at a reasonable cost. This paper aims to showcase efficient techniques to develop a scalable web crawling system, addressing challenges which deal with issues related to the structure of the web, distributed computing, job scheduling, spider traps, canonicalizing URLs and inconsistent data formats on the web. A brief discussion on new web crawler architecture is done in this paper.

References

distrbuted web crawler. In Proceedings of the 18th International Conference on Data
Engineering (ICDE), pages 357-368, San Jose, California. IEEE CS Press.


Int. World Wide Web Conference, 2001

4. Web Crawling, By Christopher Olston and Marc Najork Foundations and Trends R in
10.1561/1500000017.

http://commoncrawl.org/2014/02/common-crawl-move-to-nutch/


the ACM SIGMOD Int. Conf. on Management of Data, pages 117–128, May 2000.

incremental crawler. In Proc. of 26th Int. Conf. on Very Large Data Bases, pages 117–128,
September 2000

9. George Adam, Christos Bouras, Professor Vassilis Pouloupolos, Utilizing RSS feeds for
crawling the Web Conference: Fourth International Conference on Internet and Web
Applications and Services, ICIW 2009, 24-28 May 2009, Venice/Mestre, Italy.

1623-1640.

Internet mathematics, 1(4), pp.485-509.

12. High Scalability, “10 Things You Should Know About Running MongoDB At Scale”
http://highscalability.com/blog/2014/3/5/10-things-you-should-know-about-running-mongodb-at-
scale.html

https://docs.mongodb.com/manual/core/gridfs/

14. Compose, “Better Bulking for MongoDB 2.6 & Beyond –Compose an IBM company”.
https://www.compose.com/articles/better-bulking-for-mongodb-2-6-and-beyond/

15. Castillo, Carlos, and Ricardo Baeza-Yates. Practical Issues of Crawling Large Web

Index Terms

Computer Science

Information Sciences

Keywords

Web Crawler, Distributed Computing, Bloom Filter, Batch Crawling, Selection Policy, Politeness
Policy.