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 limit, 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 deals 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

Improvised Architecture for Distributed Web Crawling

http://commoncrawl.org/2014/02/common-crawl-move-to-nutch/
12. High Scalability, “10 Things You Should Know About Running MongoDB At Scale”
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/

Index Terms

Computer Science

Information Sciences

Keywords

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