World Academy of Science, Engineering and Technology
International Journal of Computer and Systems Engineering
Vol:10, No:06, 2016

Proxisch: An Optimization Approach of Large-Scale Unstable Proxy Servers Scheduling

Authors : Xiaoming Jiang, Jinqiao Shi, Qingfeng Tan, Wentao Zhang, Xuebin Wang, Muqian Chen

Abstract : Nowadays, big companies such as Google, Microsoft, which have adequate proxy servers, have perfectly implemented their web crawlers for a certain website in parallel. But due to lack of expensive proxy servers, it is still a puzzle for researchers to crawl large amounts of information from a single website in parallel. In this case, it is a good choice for researchers to use free public proxy servers which are crawled from the Internet. In order to improve efficiency of web crawler, the following two issues should be considered primarily: (1) Tasks may fail owing to the instability of free proxy servers; (2) A proxy server will be blocked if it visits a single website frequently. In this paper, we propose Proxisch, an optimization approach of large-scale unstable proxy servers scheduling, which allow anyone with extremely low cost to run a web crawler efficiently. Proxisch is designed to work efficiently by making maximum use of reliable proxy servers. To solve second problem, it establishes a frequency control mechanism which can ensure the visiting frequency of any chosen proxy server below the website’s limit. The results show that our approach performs better than the other scheduling algorithms.

Keywords : proxy server, priority queue, optimization algorithm, distributed web crawling
Conference Title : ICICS 2016 : International Conference on Information and Communications Security
Conference Location : San Francisco, United States
Conference Dates : June 09-10, 2016