


Project Draft (2)



Large Scale Web Crawling and Distributed Search Engines:
Techniques, Challenges, Current Trends, and Future Prospects

Group No. 12

Asadullah Al Galib - 21266015

Md Humaion Kabir Mehedi - RA

Ehsanur Rahman Rhythm - ST



Related Work

- Systematic literature review of over 1488 articles. [4]
- Performance metrics of various web crawlers. [4]
- Efficiently crawling board sites. [5]
- Performance measurement against traditional breadth-first crawling approach. [5]
- Extensive analysis of deep web crawling techniques and challenges. [3]
- New high performance web crawler that is distributed, scalable, and extensible. [6]



Asynchronous Crawler

- **Limitations & Scopes**
- Respect robots.txt to prevent overwhelming sites.
- Handle rate limiting.
- Connection reset error.
- Resume previously failed sites and items.
- Comprehensive logs.
- Integrate distributed crawling with asynchronous process.



Reference

1. Brin, S., \& Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. Computer networks and ISDN systems, 30(1-7), 107-117.
2. <https://www.statista.com/chart/19058/number-of-websites-online/>
3. Hernández, I., Rivero, C. R., \& Ruiz, D. (2019). Deep Web crawling: a survey. World Wide Web, 22, 1577-1610.
4. Kumar, M., Bhatia, R., \& Rattan, D. (2017). A survey of Web crawlers for information retrieval. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 7(6), e1218.
5. Guo, Y., Li, K., Zhang, K., \& Zhang, G. (2006, December). Board forum crawling: a Web crawling method for Web forum. In 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06) (pp. 745-748). IEEE.
6. Najork, M., & Heydon, A. (2002). High-performance web crawling (pp. 25-45). Springer US.