Current Challenges in Web Crawling

Denis Shestakov

Department of Media Technology, Aalto University
P.O. Box 15500, FI-00076 Aalto, Finland
denis.shestakov@aalto.fi
https://mediatech.aalto.fi/~denis/

Abstract. Web crawling, a process of collecting web pages in an automated manner, is the primary and ubiquitous operation used by a large number of web systems and agents starting from a simple program for website backup to a major web search engine. Due to an astronomical amount of data already published on the Web and ongoing exponential growth of web content, any party that want to take advantage of massive-scale web data faces a high barrier to entry. In this tutorial, we will introduce the audience to five topics: architecture and implementation of high-performance web crawler, collaborative web crawling, crawling the deep Web, crawling multimedia content and future directions in web crawling research.

Keywords: web crawling, web crawler, web spider, web robot, web structure, web growth, web coverage, web graph, collaborative crawling, web ecosystem, web harvesting, crawler architecture, focused crawling, distributed crawling, web mining, web retrieval, deep Web.

1 Introduction

Web crawling [1], a process of collecting web pages in an automated manner, is the primary and ubiquitous operation used by a large number of web systems and agents starting from a simple program for website backup to a major web search engine. For example, search engines such as Google or Microsoft Bing use web crawlers to routinely visit billions of web pages, which are then indexed and made available for answering user search requests. In this way, the characteristics of obtained web crawls such as coverage or freshness directly affect on the quality of web search results served to users. Besides web search, the web crawling technology is central in such applications as web data mining and extraction, social media analysis, digital preservation (i.e., ensuring continued access to information and all kinds of records, scientific and cultural heritage existing in digital formats), detection of web spam and fraudulent web sites, finding unauthorized use of copyrighted content (music, videos, texts, etc.), identification of illegal and harmful web activities (e.g., terrorist chat rooms), virtual tourism, etc.

Due to an astronomical amount of data already published on the Web and ongoing exponential growth of web content, any party (whether it be an individual, company, government agency, non-profit or educational organization, etc.)

F. Daniel, P. Dolog, and Q. Li (Eds.): ICWE 2013, LNCS 7977, pp. 518–521, 2013.

that want to take advantage of massive-scale web data faces a high barrier to entry. Indeed, only network costs associated with the downloading of web-scale size collection by themselves lead to expenses that are not affordable by the majority of potential players.

For those with flexible budgets, there is a next barrier: operating web-scale crawl (at least, hundreds of millions of pages) is a challenging task that requires skills and expertise in distributed data retrieval and processing, not to mention large operational costs. Finally, for the parties who nevertheless manage to overcome the above obstacles but interested in specific subsets of web information, the results of crawl are often wasteful, as majority of retrieved pages do not match their criteria of interest.

As a result, while there are many parties crawling the Web, the large-scale web crawling is done mostly by commercial companies, specifically by web search engines (e.g., Google). Currently, search engines' crawlers are aware of more than one trillion links and probably of more than one hundred billion pages that are re-visited on a regular basis to keep their indexes fresh.[1] Unlike web crawling under the industrial settings, the scale of non-industrial web crawling is modest and does not usually exceed several hundred million pages. Besides the dramatic difference in scale, the crawl datasets collected by commercial web crawlers are not in a public domain, not to mention that their algorithms and techniques are proprietary and kept in secret. As a result, only crawls of small sizes are available to the research community as well as to the general audience. It is clearly unsuitable since such datasets could facilitate research not only in the area of web information retrieval and more generally in computer science but also in other disciplines such as biology, epidemiology, linguistics, sociology, mathematics, etc. [2]. Furthermore, analysis of web datasets (e.g., investigating how web sites are ready to the next 'wave' of users who browse the Web using mobile devices) is of key importance for business and media companies.

In this tutorial, we will address the following topics: architecture and implementation of high-performance web crawler, collaborative web crawling, crawling the deep Web, crawling multimedia content available on the Web, and future directions in web crawling research. We will also provide some background on the structure of the Web and the role of crawling in the Web ecosystem.

2 Tutorial Synopsis

The material will be presented in the following six modules:

– **Web structure&ecosystem.** We start with some necessary background on the structure&ecosystem of the Web [3,4] and provide some useful estimates for the amount of content on the Web [5,6].
– **Architecture and implementation of high-performance web crawler.** Here we present 'traditional' challenges in building an efficient web-scale crawler system and describe state-of-the-art techniques and approaches [7,8].

[1] See blog entry at
`http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html`

- **Collaborative web crawling.** A collaborative web crawler [9] is a service that crawls the Web on the behalf of its many client applications that define filters to be evaluated against each crawled page.
- **Crawling the deep Web.** We describe the challenges in accessing information available in myriads of online web databases [10] and techniques used in modern web crawlers [11,12]. We also address here complications for web crawlers caused by new web standards, techniques and practices (e.g., rich internet applications) [13].
- **Crawling multimedia content.** We overview this rather unexplored subarea, which is poorly covered in the literature.
- **Future directions.** Here we discuss some open questions in web crawling research (e.g., crawling utilizing web content structure [14]) and conclude with references to literature, datasets, relevant projects, self-study materials, etc.

3 Biographical Sketch

Denis Shestakov is a postdoctoral researcher at the Department of Media Technology, Aalto University, Finland. He spent one year as a visiting researcher at INRIA Rennes, France. Denis obtained his doctoral degree at University of Turku, Finland in 2008. In his doctoral work [15], Denis addressed the limitations of web crawlers, specifically the poor coverage of information available in online databases (a.k.a. the deep Web). His current research interests lie in the area of distributed algorithms for big data processing, with particular applications in web crawling and large-scale multimedia retrieval. Denis is maintaining an open group on research works in the area of web crawling (see http://www.mendeley.com/groups/531771/web-crawling/). Contact him at denis.shestakov@aalto.fi or visit his homepage at https://mediatech.aalto.fi/~denis/.

References

1. Olston, C., Najork, M.: Web crawling. Foundations and Trends in Information Retrieval 4(3), 175–246 (2010)
2. Barabasi, A.-L.: Scale-Free networks: A decade and beyond. Science 325(5939), 412–413 (2009)
3. Kleinberg, J.M., Kumar, R., Raghavan, P., Rajagopalan, S., Tomkins, A.S.: The Web as a graph: measurements, models, and methods. In: Asano, T., Imai, H., Lee, D.T., Nakano, S.-I., Tokuyama, T. (eds.) COCOON 1999. LNCS, vol. 1627, pp. 1–17. Springer, Heidelberg (1999)
4. Schonfeld, U., Shivakumar, N.: Sitemaps: Above and beyond the crawl of duty. In: Proc. of WWW 2009, pp. 991–1000 (2009)
5. Bar-Yossef, Z., Gurevich, M.: Random sampling from a search engine's index. JACM 55(5) (2008)
6. Shestakov, D.: Sampling the national deep Web. In: Hameurlain, A., Liddle, S.W., Schewe, K.-D., Zhou, X. (eds.) DEXA 2011, Part I. LNCS, vol. 6860, pp. 331–340. Springer, Heidelberg (2011)

7. Shkapenyuk, V., Suel, T.: Design and implementation of a high-performance distributed web crawler. In: Proc. of ICDE 2002, pp. 357–368 (2002)
8. Lee, H.-T., Leonard, D., Wang, X., Loguinov, D.: IRLbot: Scaling to 6 billion pages and beyond. ACM Transactions on the Web 3(3) (2009)
9. Hsieh, J., Gribble, S., Levy, H.: The architecture and implementation of an extensible web crawler. In: Proc. of NSDI 2010 (2010)
10. Shestakov, D.: Deep Web: databases on the Web. Entry: Handbook of Research on Innovations in Database Technologies and Applications, pp. 581–588 (2009)
11. Madhavan, J., Ko, D., Kot, Ł., Ganapathy, V., Rasmussen, A., Halevy, A.: Google's deep-Web crawl. In: Proc. of VLDB 2008, pp. 1241–1252 (2008)
12. Shestakov, D.: On building a search interface discovery system. In: Proc. of VLDB Workshops 2009, pp. 81–93 (2009)
13. Duda, C., Frey, G., Kossmann, D., Matter, R., Zhou, C.: AJAX crawl: Making AJAX applications searchable. In: Proc. of ICDE 2009, pp. 78–89 (2009)
14. Lin, S.-H., Ho, J.-M.: Discovering informative content blocks from web documents. In: Proc. of SIGKDD 2002, pp. 588–593 (2002)
15. Shestakov, D.: Search interfaces on the Web: Querying and characterizing. Doctoral thesis, University of Turku (2008)