SAC 2012 Tutorial Proposal

Tutorial title: Web crawling
Duration: half-day
Presenter: Denis Shestakov, PhD

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 four topics: architecture and implementation of high-performance web crawler, collaborative web crawling, crawling the deep Web and future directions in web crawling research.

Motivation, target audience, and interest for the SAC community: To generate more interest in web crawlers and specifically to current challenges in the web crawling area. Particularly, participants of SAC technical tracks 3, 7-12, 17, 19, 20, 32, 34 will see the importance of their own fields to web crawling. Others will learn the state-of-the-art techniques and methods for one of the three main underlying technologies (other two are indexing and ranking) required for operation of major web search engines.

Outline: 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. 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, 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.) 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.

In this tutorial, we will address the following topics:

- 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 [2].
- Collaborative web crawling. A collaborative web crawler [1] 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 and techniques used in modern web crawlers [3,4].
- Future directions in web crawling research. Here we discuss open questions in web crawling research and their relevance to other computer science fields.

Literature:

Specific goals and objectives: attract attention to open questions in web crawling research

Expected background of the audience: Master-level (and higher) computer science researchers and IT professionals

A biographical sketch of the presenter: Denis Shestakov\(^1\) is a postdoctoral researcher in the Department of Media Technology at the Aalto University (Kone miehen tieto 2, Espoo, Finland). His research interests span Web data management and massive data processing, with a focus on scalable Web agents. Shestakov received his PhD in Computer Science (topic: improving coverage of web crawlers) from the University of Turku, Finland. He published over

\(^1\) Homepage is available at [http://www.tml.tkk.fi/~denis/index.html](http://www.tml.tkk.fi/~denis/index.html)
15 peer-reviewed publications, most of which on different aspects of web crawling. Contact him at denis.shestakov@aalto.fi

**Equipment:** projector