Navigating the government website network using parallel breadth-first crawler

Web Information Extraction and Retrieval 2018/19, Faculty of Computer and Information Science, University of Ljubljana

Matej Klemen, Andraž Povše, Jaka Stavanja

***Abstract*—BOII**

I. Introduction

A web crawler is a program that visits web pages and downloads them for some purpose. It begins with some amount of seed URLs, which are put into a frontier. The crawler then selects a web page from frontier according to some strategy, visits the web page and extracts the web page's content and outgoing links. The outgoing links are put into the frontier and the content is used for some purpose or stored. This process is then repeated until a certain stopping criteria is reached (e.g. the frontier being empty) [1]. The crawling can also be a continuous process ("without end"), but we do not study these in this project.

In this project we implement a web crawler that crawls Slovenian government websites. We are given a set of 4 government websites and choose additional 5 of them and proclaim them as seed websites, from which we start the crawling. While crawling these websites, our crawler only follows links to other Slovenian government websites. The crawler crawls these websites, downloads images and files from them and stores them into a database. The crawler visits new links in a breadth-first fashion. We also implement a method to detect near-duplicate websites based on their content. For this, we use a locality-sensitive hashing algorithm based on min-hashing [2].

The rest of the report is structured as follows. In chapter **TODO** we present **TODO**. In chapter **TODO** we present **TODO** ...

II. Crawler

In this section we present the architecture of the crawler. The full source code is available online [1].

The crawler consists of multiple components: link, sitemap and robots parser, database manager (and the database itself), content deduplicator and a core unit which uses the other components and adds on top of them to produce a parallel crawler that uses breadth first search to navigate the websites.

[**TODO: describe each component briefly (mention its role).**]

PARAMETER THAT OUR CRAWLER TAKES

We used Selenium to make requests, because it also renders the Javascript code on the website. It does, however, not include HTTP status code, which is why we also had to make another request to get that information.

EXPLAIN HOW BREADTH FIRST STRATEGY WORKS Each worker gets an even share of new links. We then wait for each of the workers to finish and collect the new links they have gathered and place them into the frontier. Afterwards, we repeat the first step and share these new links amongst all workers. With this strategy, we are using Breadth-First approach to visit new websites.

[1] https://github.com/matejklemen/govrilovic

We did not use the database as a frontier storage —— We stored our frontier in the program itself.

III. Robots and Sitemaps

Key element of the crawler is also reading the 'robots.txt' file and sitemap. Based on the robots file, we know whether we can visit a certain link or whether it is not allowed for a crawler to visit that link. When we read a sitemap, we add all the URLs mentioned in it, into the frontier.

IV. Database

The database was build inside a Docker container. We used the same schema, with one minor change, which is adding a column 'lsh_hash' into the page table. The column stores the value that LSH hash produced for the content of that website. Access to database is made by each worker when he visits a new page.

V. Link

VI. Image and files extracting

@JAKA — commnet on the issue with absolute path, very interesting:D

We stored images and files into the file system, since it gives a nicer overview of the gathered information. This was made possible by creating a folder for each new site in which we stored the information. We then stored the location of the file/image into the provided data field in the database.

VII. LSH deduplication

We implemented LSH algorithm based on CITE CITE CITE. For the words, we use (top 1000 slovenian words???)

VIII. Results - 2 seed URLs

For the sites that are given in the instructions' seed list and also for the whole crawldb together (for both separately) report general statistics of crawldb (number of sites, number of web pages, number of duplicates, number of binary documents by type, number of images, average number of images per web page, ...). Visualize links and include images into the report. If the network is too big, take only a portion of it. Use visualization libraries such as D3js, visjs, sigmajs or gephi.

IX. Results - 9 seed URLs

References

[1] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[2] A. Gionis, P. Indyk, and R. Motwani, "Similarity search in high dimensions via hashing," in *Proceedings of the 25th International Conference on Very Large Data Bases*, ser. VLDB '99. Morgan Kaufmann Publishers Inc., 1999, pp. 518–529.

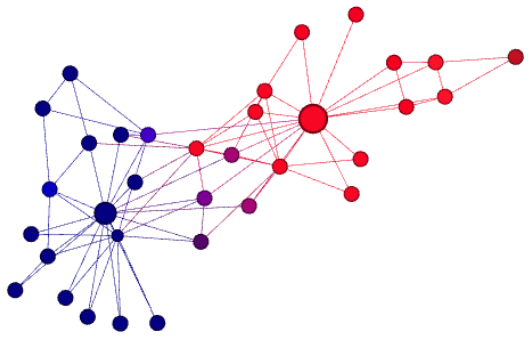

Figure 1. Example of an image.

Table I
EXAMPLE OF A TABLE.

Richard Karstark	0.013661
Jon Arryn	0.012869
Joyeuse Erenford	0.012869
Trystane Martell	0.010761
Willen Lannister	0.010684
Martyn Lannister	0.010684
Robb Stark	0.010304
Joffrey Baratheon	0.009875
Master Caleotte	0.009495
Lysa Arryn	0.009383