EXPLORING THE GERMAN HEALTH WEB
UTILIZING VARIOUS ASF PROJECTS FOR CRAWLING THE WEB

Richard Zowalla[1]

[1] ASF Member, PMC Member of Apache TomEE, PMC Member of Apache OpenNLP, PMC Chair of Apache Storm, PPMC Member of Apache StormCrawler (Incubating)

BACKGROUND

The Internet has become an increasingly important source of health information in recent years. Lay people, in particular, search online for information about diseases, diagnoses and treatment options. However, with the growing amount of information available on the Internet, it is impossible for people to manually keep track of the evolving and constantly changing health content.

PROBLEM STATEMENT

Current research focuses mainly on the quality and evidence of health information provided on the Internet. However, there is a lack of studies that analyze the structure and semantics of the health information available in the predominantly German-speaking countries of Germany, Austria and Switzerland, hereafter referred to as the German Health Web (GHW), in its entirety. There is also a lack of approaches for carrying out such an analysis with little human effort, i.e. fully automatically.

With the help of scalable software components, a basis is created for answering the following research questions:

- What is the size and scope of the GHW?
- Which information providers have a particularly high "reputation" and thus represent aggregation points for health information?
- Which actors or institutions are behind these important information providers?
- How high is the readability of health information in the GHW? What kind of vocabulary is used?
- Which topic or themes of the GHW have a high relative importance?

METHODS

Due to the enormous size of the web, it is necessary to distinguish whether the content of a website is health-related or not in order to reduce the total number of websites to be analyzed.

This can be done using a focused web crawler. Such a crawler uses machine learning techniques to assess the relevance of a website to a given topic of interest. Figure 1 shows the software architecture of the focused web crawler used. It relies on the Apache StormCrawler (Incubating) SDK for the basic crawler components, uses Apache Storm as a scalable runtime environment, Apache Tika for content parsing and Apache OpenNLP for the necessary natural language processing.

The captured content and link structures can then be analysed in more detail using graph analysis and unsupervised machine learning methods (e.g. topic modelling).

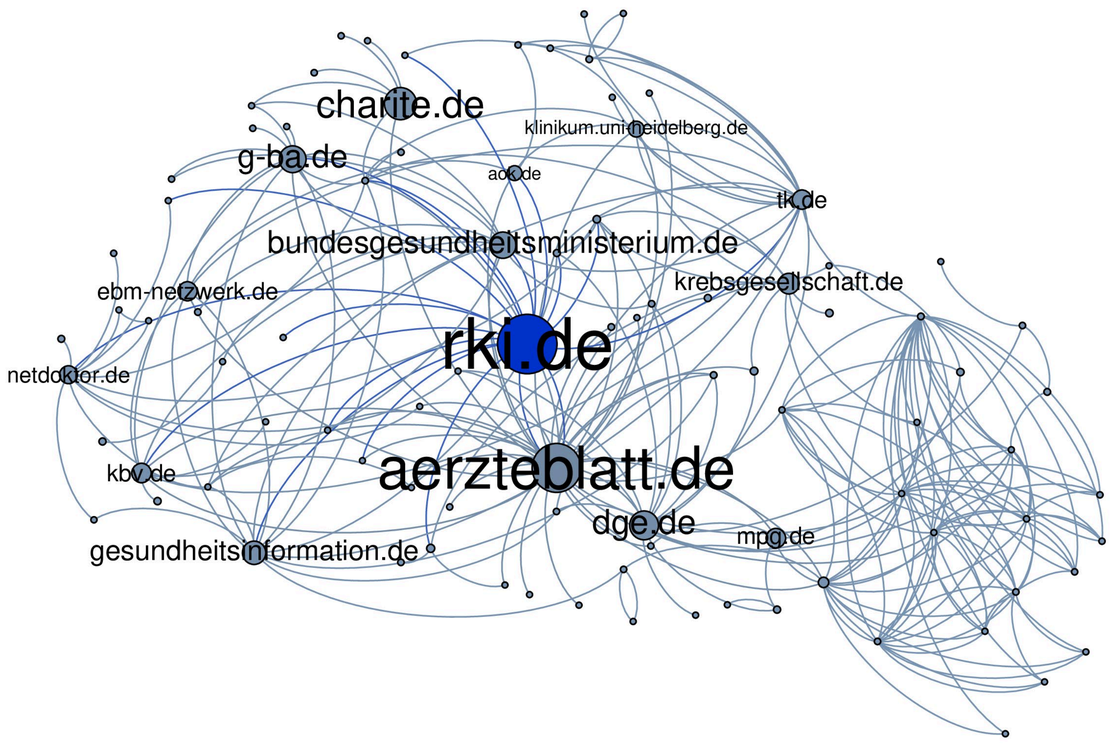

Fig. 2: A small section of the GHW graph focusing on the website www.rki.de. The surrounding nodes represent websites with a maximum link distance of two, starting from www.rki.de. An edge between two nodes means that there is at least one hyperlink in both directions between some of the web pages of the website in question. Only sites with a strong health content are included.

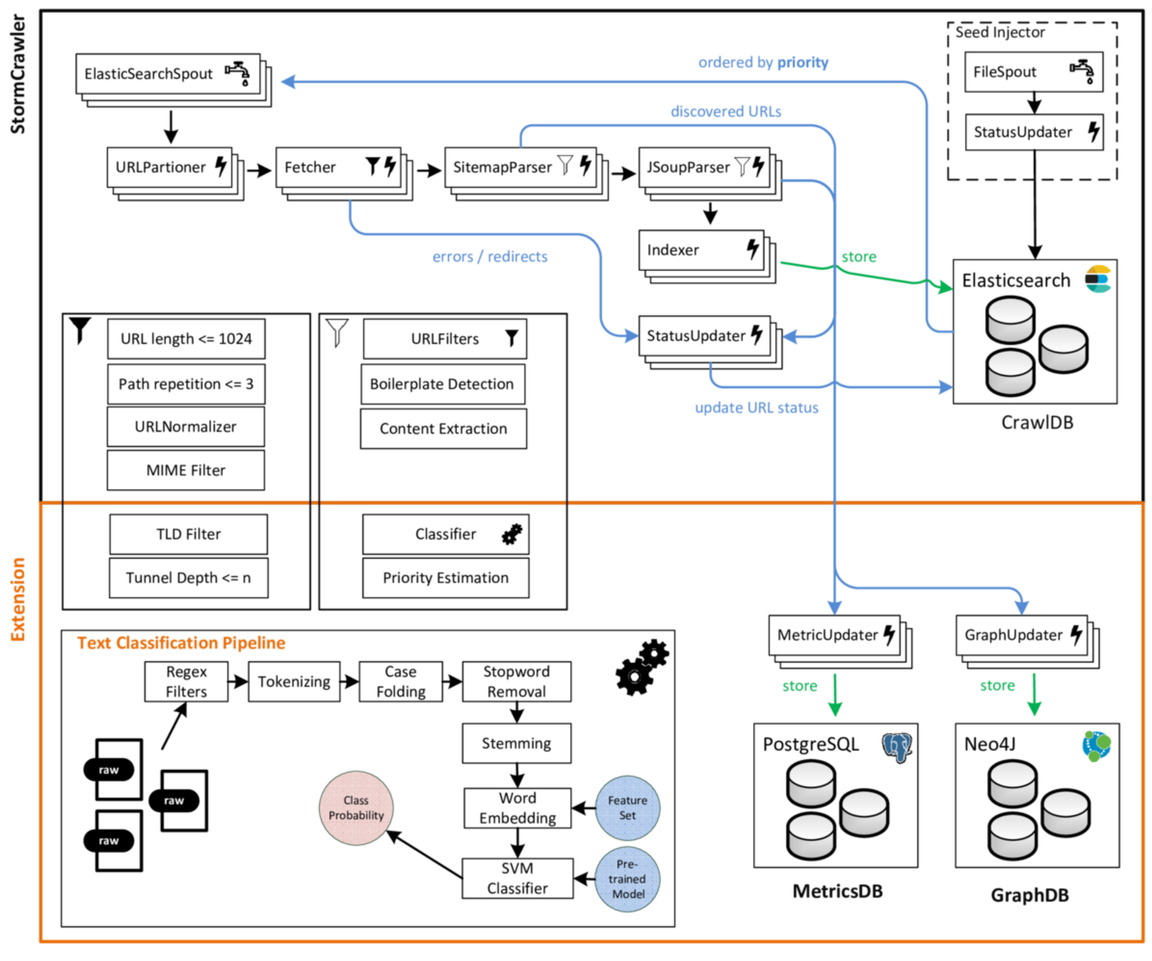

Abb. 1: Architecture of a focussed web crawler based on Apache Storm and Apache StormCrawler. Spouts (water tap symbol) output data (here: URLs), bolts (lightning bolt symbol) process data (i.e. download, parse and save the extracted content). Bolts can be extended with URL filters (white filter symbol) or parse filters (black filter symbol). URL filters are used to remove URLs based on predefined criteria. Parse filters include URL filters, but are primarily used to clean up the parsed content and calculate the health relevance and priority.

RESULTS

The focused web crawling system ran from 27 May 2019 to 31 May 2020 and identified a total of 14,193,743 health-related websites. The focused web crawler achieved a download rate of seven to eleven websites per second. This amounts to 370 days of pure web crawling and the classification of approximately 341 million websites during the study period.

The resulting aggregated GHW graph contains a total of 231,733 nodes (web pages) connected by 429,530 edges (links between web pages). In total, 82.63% (191,479/231,733) of the websites belong to the domain extensions "de", 7.89% (18,272/231,733) to "at" and 9.48% (21,976/231,733) to ".ch". The network diameter of the graph is 25.

Figure 2 shows a small part of the GHW, focusing on the website of the Robert Koch Institute. Figure 3 shows the distribution of subject areas per domain ending.

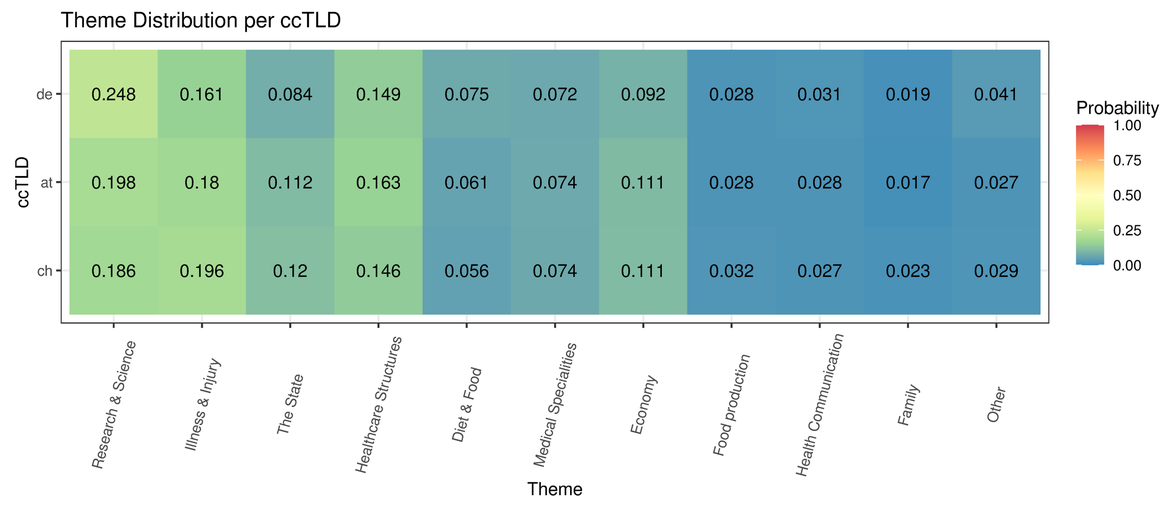

Fig. 3: Topics within the GHW for the country-code top level domains "de", "at" and .ch".

FURTHER INFORMATION

Unfortunately, I will not be able to attend Community Over Code in person. If you have any questions, please feel free to contact me via email or ASF Slack. More information and details can be found on the project websites or in the two scientific publications behind the two QR codes:

- stormcrawler.apache.org
- storm.apache.org
- opennlp.apache.org