2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology

Babouk: Focused Web Crawling for Corpus Compilation and Automatic Terminology Extraction

Clément de Groc
Syllabs
Univ. Paris Sud & LIMSI-CNRS
France
cdegroc@limsi.fr

***Abstract*— The use of the World Wide Web as a free source for large linguistic resources is a well-established idea. Such resources are keystones to domains such as lexicon-based categorization, information retrieval, machine translation and information extraction. In this paper, we present an industrial focused web crawler for the automatic compilation of specialized corpora from the web. This application, created within the framework of the TTC project[1], is used daily by several linguists to bootstrap large thematic corpora which are then used to automatically generate bilingual terminologies.**

Keywords-focused crawling; web-as-corpus; resources bootstrapping

I. Introduction

Babouk is a focused crawler [4] (also called thematic or topical crawler): its goal is to gather as many relevant webpages as possible on a specialized domain of interest to the user. Compared to traditional crawling, focused crawling allows rapid access to specialized data and avoids computational and financial bottlenecks induced by full web crawling.

Leveraging the web as a source for natural language processing data is not a novel idea [6] and, researchers, either via search engine queries [2] or web crawling [3] have already derived many general- or special-purpose corpora or language models from it. We propose a resource-light user-centered focused crawler where the crawler relies on a small set of user-provided seed terms or URLs. From this input, a weighted lexicon is automatically built using the Web. The lexicon is then used to classify webpages as in or out of domain. This approach provides interpretable results and allows users to easily override the lexicon bootstrapping process at any time.

II. Focused Web Crawling Platform

A. *Focused Crawler*

Babouk starts from a potentially small set of seed terms or URLs and tries to maximize the ratio of relevant documents over the total number of documents fetched (also called “Harvest rate” in the focused crawling literature). To achieve this goal, the crawler selects relevant documents and follows links that are most relevant to the user-defined domain.

The categorization step is achieved using a weighted-lexicon-based thematic filter to compute webpages relevance. This lexicon is automatically built at runtime, during the first iteration of the crawling process: user given input is expanded to a large lexicon using the BootCaT [2] procedure. Then, the new lexicon is weighted automatically using a web-based measure called “Representativity”. Finally, a calibration phase, during which a categorization threshold is computed automatically, is applied. A comparative evaluation against machine learning based models is on-going.

In order to guide the crawler to process the most relevant pages first (a process called “crawl frontier ordering” in the crawling literature [5]), the crawler uses the thematic score of the pages to rank URLs in a way similar to the OPIC criterion [1].

B. *Crawl options*

Crawling the web is a recursive process that will solely stop when no more relevant documents are found. While this strategy is theoretically founded, the crawl duration might still be very long. This is why several stopping criteria were added: users can specify a maximum amount of tokens or documents to retrieve, a maximum crawl depth or even an upper bound time limit.

In order to get high quality webpages, users can also force pages size to be in a certain range (either as number of tokens or as HTML Kilobytes). Moreover, a live content-based web-spam filtering is applied. Finally, users can limit the crawl to specific domains or file formats (such as Microsoft Office, Open Office, Adobe PDF, or HTML) and apply a blacklist of unwanted URLs/domains.

C. *Under the hood*

From a technical point of view, Babouk is based on the latest version of Apache Nutch and is distributed across a cluster of computers (optionally in the cloud). The « scale out » approach ensures scalability in terms of computational resources required for many simultaneous large crawls. Crawl results and documents meta-information are stored in a distributed database.

[1] http://www.ttc-project.eu

978-0-7695-4513-4/11 $26.00 © 2011 IEEE
DOI 10.1109/WI-IAT.2011.253

The web crawler is run and monitored through a dynamic web-based user interface (Fig. 1) that interacts with the distributed crawler through messaging. Users can run and stop crawl jobs as well as get live logs from the distributed crawlers.

Acknowledgment

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement n°248005.

I'd like to thank Helena Blancafort (and the linguistic team of Syllabs) and Javier Couto for their support and fruitful discussions.

Another special acknowledgment go to Ulrich Heid and his students from the Institute for Natural Language Processing (IMS) at Stuttgart University for their in-depth testing of the tool.

References

[1] S. Abiteboul, M. Preda, and G. Cobena, "Adaptive on-line page importance computation, " Proceedings of the twelfth international conference on World Wide Web - WWW, New York, New York, USA: ACM Press, 2003, pp. 280-290.

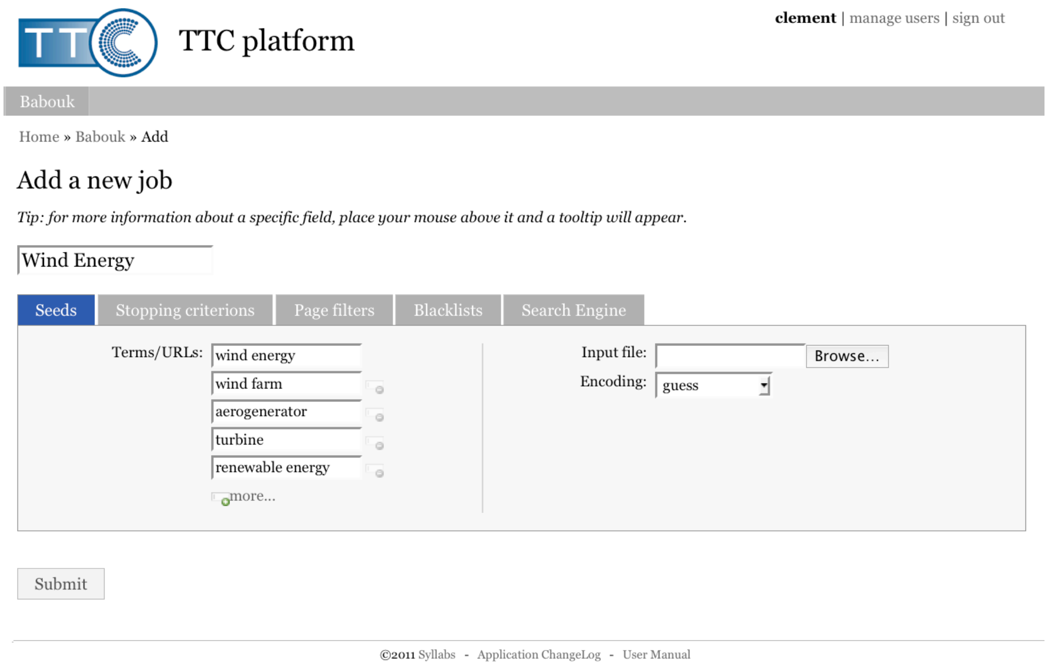
[2] M. Baroni and S. Bernardini, "BootCaT: Bootstrapping Corpora and Terms from the Web," Proceedings of the LREC 2004 conference, 2004, pp. 1313-1316.

[3] M. Baroni and M. Ueyama, "Building general- and special-purpose corpora by Web crawling," NIJL International Symposium, Language Corpora, 2006, pp. 31-40.

[4] S. Chakrabarti, M.V. den Berg, and B. Dom, "Focused crawling: a new approach to topic-specific Web resource discovery," Computer Networks, 1999, pp. 1623-1640.

[5] J. Cho, H. Garcia-Molina, and L. Page, "Efficient crawling through URL ordering," Computer Networks and ISDN Systems, 1998, pp. 1-7.

[6] A. Kilgarriff and G. Grefenstette, "Introduction to the Special Issue on the Web as Corpus," *Computational Linguistics*, vol. 29, Sep. 2003, pp. 333-347.



Add a new job

Tip: for more information about a specific field, place your mouse above it and a tooltip will appear.

Wind Energy

Terms/URLs: wind energy
wind farm
aerogenerator
turbine
renewable energy

Input file:

Encoding: guess

Figure 1. Web-based UI to create a crawl job