A Web Crawling Environment to Support Financial Strategies and Trend Correlation

(Extended Abstract)

Giovanni Ponti[1(✉)], Giuseppe Santomauro[1], Fiorenzo Ambrosino[1], Giovanni Bracco[1], Antonio Colavincenzo[2], Matteo De Rosa[1], Agostino Funel[1], Dante Giammattei[1], Guido Guarnieri[1], and Silvio Migliori[1]

[1] ENEA - ICT Division - Portici Research Center (NA), Portici, Italy
{giovanni.ponti,giuseppe.santomauro,fiorenzo.ambrosino,giovanni.bracco,
matteo.derosa,agostino.funel,dante.giammattei,guido.guarnieri,
silvio.migliori}@enea.it
[2] Accenture Technology Solutions s.r.l. - Assago (MI), Milan, Italy
antonio.colavincenzo@accenture.com

Abstract. We provide an overview on the development and the integration in ENEAGRID of a web crawling tool to retrieve data from the Web, manage and display it, and extract relevant information. We collected all these instruments in a collaborative environment called *Web Crawling Virtual Laboratory*, offering a GUI to operate remotely. Finally, we describe an ongoing activity on semantic crawling and data analysis to discover trends and correlations in finance.

Keywords: Web crawling · Big data · Machine learning · Market trends

1 Introduction

Internet is certainly the World's largest data source. Web data has characteristics that involve a considerable effort of analysis and organization. The ability of extracting strategical information in big data from the Web is becoming a crucial task that involves several contexts, such as cyber security, business intelligence, and finance. All the applications in these fields have to face with computational and storing issues. For this reason, the advanced computing center of ENEA Portici, hosting the ENEAGRID/CRESCO infrastructure [2] offers the possibility to perform this activity. In the following, we introduce the web crawling environment integrated in ENEAGRID to retrieve and analyze data from the Web, and we provide some details on a work-in-progress activity in finance describing how to obtain financial information and correlation with market trends.

C. Alzate et al. (Eds.): MIDAS 2018/PAP 2018, LNAI 11054, pp. 103–107, 2019.
https://doi.org/10.1007/978-3-030-13463-1_8

2 Web Crawling and Web Data Analysis in ENEAGRID

A crawling technique analyzes systematically and automatically the content of a network to search for documents to download. Web crawlers are based on a list of URLs to visit that is continuously updated by new records retrieved by parsing the explored web pages. In the next, we provide a description of our web crawling environment installed and configured in ENEAGRID.

2.1 Web Crawling Tool: *BUbiNG*

We resorted to *BUbiNG* [1] as the web crawling product to integrate in ENEAGRID. This software allows the parallel execution of multiple crawling agents. Each agent communicates with each other one to ensure not repeated visits of same pages and to balance the computational load. *BUbiNG* also allows to set up at runtime all configuration options in a single parameter setting file, such as thread number and initial seeds. *BUbiNG* saves contents in compressed *warc.gz* files. This data compression is very important because it allows to save space up to around 80%.

2.2 Virtual Laboratory and Web Application

We created a collaborative *Web Crawling Project* integrated in ENEAGRID. Here, the main issue consisted in harmonize the tool in a typical HPC environment to exploit infrastructure resources, that are computational nodes, networking, storage systems, and job scheduler. All the web crawling instruments are collected in a ENEAGRID virtual laboratory, named *Web Crawling*[1]. The virtual lab has a public web site (Fig. 1(a)) where information about the research

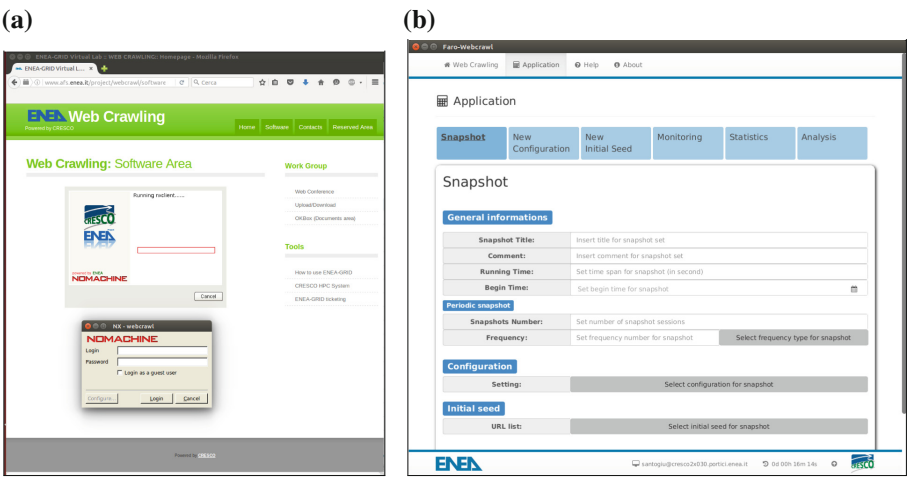

Fig. 1. (a) The virtual lab site. (b) The virtual lab GUI.

[1] http://www.afs.enea.it/project/webcrawl/.

activity is collected, and a web application (Fig. 1(b)) to submit snapshot and to use tools for analysis, displaying and clustering of web data.

2.3 Tests and Experimental Results

We performed experiments to analyze the performance of our solution for web crawling tool integrated in the ENEAGRID infrastructure. For this reason, we designed two types of experiments. In the first one, we performed long-time crawling sessions (of more than 8 h), in order to assess the ability of the tool in crawling and storing web contents at a high network speed, i.e., efficiency and robustness. The second experiment consisted in periodic crawling to test software reliability, a typical scenario to collect periodic snapshots to analyze changes in the network. Both tests provide good results [3].

3 Proposal of Current Development

We are currently working in extending our tool to support *semantic crawling* and apply it in finance, in order to discover how news and discussions in the Web on a specific topic are correlated with market trends and how can influence them.

3.1 Thematic Web Crawling

Working on proper crawling settings and pre-processing strategies, it is possible to have a reduced version of the crawled dataset on a specific topic. In this way we reached two main goals: saving memory space and speeding up the post-crawling indexing time. To obtain this result, we have developed a proper filter that selects web pages according to the topic. Such a filter does not take into account only at the page body, but also title and tags. We integrated it into BUbiNG source code (in *JAVA*) to have thematic snapshot sessions.

3.2 Web Crawling for Financial Strategies

By using the filtered dataset we aim to discover news and discussions in the Web on a specific topic. Information retrieval and deep learning algorithms can be employed to extract strategical information. More specifically, we want to reach two important results: (*i*) searching for a correlation index of web news with market trends and their influence, (*ii*) and developing a tool in order to predict a price behaviour and then to adopt appropriate trading strategy. Below we explain the five steps that we have considered:

1. First of all, for any day d_i we run a web crawling filtered on web news about a financial topic to build a dataset D_i of $g_{i,j}$ web pages:

$$D_i = \{g_{i,1}, g_{i,2}, \ldots, g_{i,N_i}\};$$

2. After, for any web page $g_{i,j}$ we apply a *Sentiment Analysis* algorithm, based on Natural Language Processing, (e.g. the VADER Sentiment Analysis coded in *JAVA*[2]) to compute a weight of positive/negative opinion:

$$w_{i,j} = w(g_{i,j}) \in [-1; +1], \quad \forall j \in [1; N_i];$$

3. Then, we compute a normalized daily opinion index:

$$w_i = \frac{\sum_{j=1}^{N_i} w_{i,j}}{N_i};$$

4. By means of a machine learning approach, we train a neural network (i.e., a Recurrent Neural Network - RNN) to estimate the daily increasing/decreasing rate r_i for the asset:

$$r_i = \frac{p_{i+1} - p_i}{p_i},$$

where p_{i+1} is the estimated price at the day d_{i+1} obtained by the RNN computation.

5. Finally, we compute a correlation between rate R and opinion index W applying the *Pearson correlation coefficient*:

$$cov(R, W) = \frac{E[RW] - E[R]\,E[W]}{\sqrt{E[R^2] - E[R]^2}\sqrt{E[W^2] - E[W]^2}}.$$

For our purpose, in a day d_i, we want to discover a correlation between the expected increasing/decreasing rate r_i and the overall opinion index w_i.

4 Conclusions

To summarize, we provided a parallel implementation of a web crawling product to periodically download contents from web and to analyze them. The tool is fully integrated in our HPC ENEAGRID/CRESCO infrastructure, in order to use computation and storage power. Currently we are equipping our framework with a sentiment analysis tool and training a neural network to correlate opinions and price trend. In the future work we want to perform experiments to tune our framework and refine our semantic filter to obtain a more accurate dataset.

Acknowledgements. The computing resources and the related technical support used for this work have been provided by ENEAGRID/CRESCO High Performance Computing infrastructure and its staff [2]. ENEAGRID/CRESCO High Performance Computing infrastructure is funded by ENEA, the Italian National Agency for New Technologies, Energy and Sustainable Economic Development and by Italian and European research programmes, see http://www.cresco.enea.it/english for information.

[2] https://github.com/apanimesh061/VaderSentimentJava.

References

1. Boldi, P., Marino, A., Santini, M., Vigna, S.: BUbiNG: massive crawling for the masses. CoRR abs/1601.06919 (2016)
2. Ponti, G. et al.: The role of medium size facilities in the HPC ecosystem: the case of the new CRESCO4 cluster integrated in the ENEAGRID infrastructure, pp. 1030–1033 (2014)
3. Santomauro, G., et al.: A collaborative environment for web crawling and web data analysis in ENEAGRID. In: DATA 2017, 24–26 July 2017, Madrid, Spain, pp. 287–295 (2017)