A Brand Scoring System for Cryptocurrencies Based on Social Media Data

Santomauro, G., Alderuccio, D., Ambrosino, F., Fronzetti Colladon, A., & Migliori, S.

Santomauro, G., Alderuccio, D., Ambrosino, F., Fronzetti Colladon, A., & Migliori, S. (2020). **A Brand Scoring System for Cryptocurrencies Based on Social Media Data**. In V. Bitetta, I. Bordino, A. Ferretti, F. Gullo, S. Pascolutti, & G. Ponti (Eds.), Mining Data for Financial Applications. MIDAS 2019. (pp. 127–132). https://doi.org/10.1007/978-3-030-37720-5_11

A brand scoring system for cryptocurrencies based on social media data

– Short regular paper –

Giuseppe Santomauro[1], Daniela Alderuccio[2], Fiorenzo Ambrosino[1]
Andrea Fronzetti Colladon[4], Silvio Migliori[3]

[1] ENEA - C.R. Portici, DTE-ICT-HPC, P.le E. Fermi, 1 - 80055 Portici (NA), Italy
{giuseppe.santomauro, fiorenzo.ambrosino}@enea.it
[2] ENEA - Sede Legale, DTE-ICT-HPC, L. Thaon di Revel, 76 - 00196 Roma - Italy
daniela.alderuccio@enea.it
[3] ENEA - Sede Legale, DTE-ICT, L. Thaon di Revel, 76 - 00196 Roma - Italy
silvio.migliori@enea.it
[4] Department of Engineering, University of Perugia, Via G. Duranti 93 - 06125 Perugia - Italy
andrea.fronzetticolladon@unipg.it

Abstract. In this work, we present an overview on the development and integration in ENEAGRID of some tools to evaluate brand importance of homogeneous financial instruments, such as cryptocurrencies. Our system is based on the analysis of textual data, such as tweets or online news. A collaborative environment called *Web Crawling* Virtual Laboratory allows data retrieval from the web. Below we describe this virtual lab and the ongoing activity aimed at adding a new feature, to allow news and social media crawling. We also provide some details about the integration in ENEAGRID of a new measure of brand importance and its Virtual Laboratory, namely the *Semantic Brand Score*. We aim to test the first version of this new virtual environment on *Twitter* data, to rank digital currencies.

Keywords: social network crawling, semantic brand scoring, cryptocurrency, financial trends

1 Introduction

Cryptocurrencies are digital assets that are designed to work as the economic component of a distributed ledger technology system such as a *blockchain*. They use cryptography to manage several functionalities, as the secure exchange of value between users or the creation of economic supply, over a distributed network, without having to trust central authorities.
The first and most popular cryptocurrency is *Bitcoin* [6]. It was released in 2009; since then more than 3000 new digital coins have been created each of them coming from a different project and having functionalities, from the simple exchange of value to, for example, smart contracts.

Units of these digital assets, sometimes also called *tokens*, are used for the economic incentive mechanism that motivate the different players of the distributed system. This tokens therefore have economic values and are exchangeable on the network itself. This has contributed to the arise of several marketplaces where users interact in trading activities with prices typically depending on demand and supply.
The demand of a token, hence its price, also depends on the quality and uniqueness of the underling service, or at least the way it is perceived. Characteristics include: the number of developers of the platform, who are the long-term investors and the size of the userbase.
The numerosity of users of a platform and their sentiment about the project can generate positive network effects facilitating the involvement of new users. To study these effects we propose to explore the Web (news, forums, tweets, etc.) and analyse messages users share about some cryptocurrencies. In order to compute a rank of cryptocurrencies based on their relevance, we consider a method that extracts important information from the Web, concerning the topic of digital coins and to applies a semantic algorithm that elaborates a score.
The task of downloading a large amount of data from the Internet, that is the World's largest data source, is commonly known as *Web Crawling*. In this context, the task of performing a brand ranking from a large set of text data (news, tweets, etc.) is named *Brand Scoring*. Both operations are critical points in terms of computational costs. For this reason, the advanced computing center of ENEA Portici, hosting the ENEAGRID/CRESCO infrastructure [8] is used to perform these activities.
In the following, we introduce the *Web Crawling* and the *Semantic Brand Score* virtual labs integrated in ENEAGRID, used to retrieve and analyze data from the Web. Finally, we provide details on a work-in-progress activity describing how to obtain financial news from social networks and how to compute a rank of brand awareness for digital coins.

2 Web Crawling in ENEAGRID

Generally, a crawling technique searches for documents to download by systematically and automatically analyzing the content of a network. A web crawler starts from a list of URLs to visit. When it downloads a web page then it updates this list by new URLs retrieved by parsing the explored document. This process can be infinitely repeated and it can be stopped either when it reaches a target number of pages or after a fixed amount of time. In the next, we provide a description of our web crawling environment installed on ENEAGRID.

2.1 The software solution

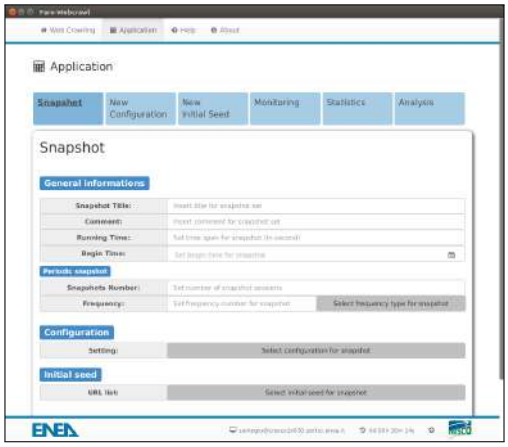
As software solution, we decided to integrate the *BUbiNG* [2] program into ENEAGRID. It is an open source product that allows the parallel execution of multiple crawling agents. Each agent communicates with the others to ensure not

repeated visits the same webpages and to balance computational load. *BUbiNG* also is able to save space up to around 80% by storing contents in compressed *warc.gz* files.
We did some tests to evaluate the performance of our software solution. We checked its efficiency, robustness and reliability by performing long-time and periodic crawling sessions: we obtained good results [9].

2.2 The Virtual Laboratory

We created a collaborative *Web Crawling Project* integrated in ENEAGRID. Here, the main issue consisted in harmonizing 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 combined in an ENEAGRID virtual laboratory, named *Web Crawling*. The virtual lab has a public web site [5] (Fig. 1(a)) where information about the research activity is collected, and a web application (Fig. 1(b)) with analytical tools and the display and clustering of web data.

(a) The Web Crawling Virtual Lab site.



(b) The Web Crawling Virtual Lab GUI.

3 Semantic Brand Scoring in ENEAGRID

The *Semantic Brand Score* (SBS) is a novel metric designed to assess the importance of one or more brands, in different contexts and whenever it is possible to analyze textual data, even big data [3]. The advantage with respect to some traditional measures is that the SBS do not relies on surveys administered to small samples of consumers.

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

3.1 The metric

The measure can be calculated on any source of text documents, such as newspaper articles, emails, tweets, posts on online forums, blogs and social media. The idea is to capture insights coming from honest signals [7], through the analysis of big textual data and combining methods of text mining and social network analysis. Spontaneous expressions of consumers, or other brand stakeholders, can be collected from the places where they normally appear for example a travel forum, if studying the importance of museum brands. This has the advantage of reducing the biases induced by the use of questionnaires, where interviewees know that they are being observed. The SBS can also be adapted to different languages and to study the importance of specific words, or set of words, not necessarily *brands* [3]. The SBS measures brand importance, which is at the basis of brand equity [3]. Indeed the metric was partially inspired by well-known conceptualizations of brand equity and by the constructs of brand image and brand awareness [1, 5]. Brand importance is measured along the three dimensions of prevalence, diversity and connectivity. Prevalence measures the frequency of use of the brand name, i.e. the number of times a brand is directly mentioned. Diversity measures the diversity of the words associated with the brand. Connectivity represents the brand ability to bridge connections between other words or groups of words (sometimes seen as discourse topics). The SBS has been used in different fields, for example to evaluate the transition dynamics that occur when a new brand replaces an old one [3], or for political forecasting [4].

3.2 The Virtual Laboratory

We assembled all the instruments for the Semantic Brand Score into a virtual laboratory, named *Brand Score*. This project is integrated into ENEAGRID by respecting the rules of the infrastructure. There is a software area, where the software for SBS is installed; there is a volume that holds all launcher scripts; and there is an area where is published a web portal [6] of the brand scoring virtual lab (Fig. 2). Preliminary tests on the configuration and on the performance demonstrate a correct integration.

4 Proposal of current development

We are currently working to an extension of our web crawling tool to retrieve data from social media, in order to discover news and discussions on a specific financial topic, such as digital coins, and to calculate the Semantic Brand Score.

4.1 Social Networks Crawling

In recent years, news on politics, sport and the economy have grown considerably on social media. For this reason, we decided to extend the features of our tools

[6] http://www.afs.enea.it/project/brandscore/

Fig. 2: The Semanitc Brand Score Virtual Lab site.

with a social media crawler. We are currently equipping our environment with a *Twitter* crawler. The software solution that we adopted is based on the *JAVA* language. By considering the *Twitter* access rules that limit the number of tweets downloaded per user and by exploiting the APIs of this social network, we created manifold developer accounts. In this way, we can launch parallel sessions of the *JAVA* software (agents) on a specific topic or on a set of themes. We can collect the *tweets* in JSONs files indexed for *hashtag* and for downloading timestamp. Preliminary tests confirm a good performance in terms of number of tweets per time unit.

4.2 Semantic Brand Score and Cryptocurrencies

Once the configuration of the *Twitter* crawler will be fully optimized, we are plan running periodic sessions of crawling in order to create a database of tweets that concern news and discussions about digital coins. This data will be analyzed through the Semantic Brand Score, to rank cryptocurrency importances.

5 Conclusions

To summarize, we provided an overview of activity about the implementation of a social media crawler that downloads contents from *Twitter*. The tool is integrated in our HPC ENEAGRID/CRESCO infrastructure. Currently we are also equipping our framework with a semantic brand scoring tool which uses ENEA computational and storage power. First tests on the social crawler and on the SBS software demonstrate good results.

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 [8]. 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.

References

1. A. Aaker, D.: Measuring brand equity across products and markets. California Management Review **38**, 102–120 (04 1996). https://doi.org/10.2307/41165845
2. Boldi, P., Marino, A., Santini, M., Vigna, S.: BUbiNG: Massive Crawling for the Masses. CoRR **abs/1601.06919** (2016)
3. Fronzetti Colladon, A.: The semantic brand score. Journal of Business Research **88**, 150 – 160 (2018). https://doi.org/https://doi.org/10.1016/j.jbusres.2018.03.026, http://www.sciencedirect.com/science/article/pii/S0148296318301541
4. Fronzetti Colladon, A.: Forecasting election results by studying brand importance in online news. International Journal of Forecasting, in press. (2019)
5. Keller, K.L.: Conceptualizing, measuring, and managing customer-based brand equity. Journal of Marketing **57**(1), 1–22 (1993), http://www.jstor.org/stable/1252054
6. Nakamoto, S.: Bitcoin: A peer-to-peer electronic cash system (Dec 2008), https://bitcoin.org/bitcoin.pdf, accessed: 2019-06-01
7. Pentland, A.S.: Honest Signals: How They Shape Our World. The MIT Press (2010)
8. 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)
9. Santomauro, G., Ponti, G., Ambrosino, F., Bracco, G., Colavincenzo, A., Rosa, M.D., Funel, A., Giammattei, D., Guarnieri, G., Migliori, S.: A collaborative environment for web crawling and web data analysis in ENEAGRID. In: DATA 2017, Madrid, Spain, July 24-26, 2017. pp. 287–295 (2017)