ENEA
Italian National Agency for New Technologies,
Energy and Sustainable Economic Development

A brand scoring system for cryptocurrencies based on social media data

ECML-PKDD 2019 - European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases.
The Fourth Workshop on MIning DAta for financial applicationS (MIDAS)

Wurzburg, 16-20 September, 2019

Santomauro Giuseppe / *DTE-ICT-HPC, ENEA – Italy*

Alderuccio D., Ambrosino F., Fronzetti Colladon A., Migliori S.

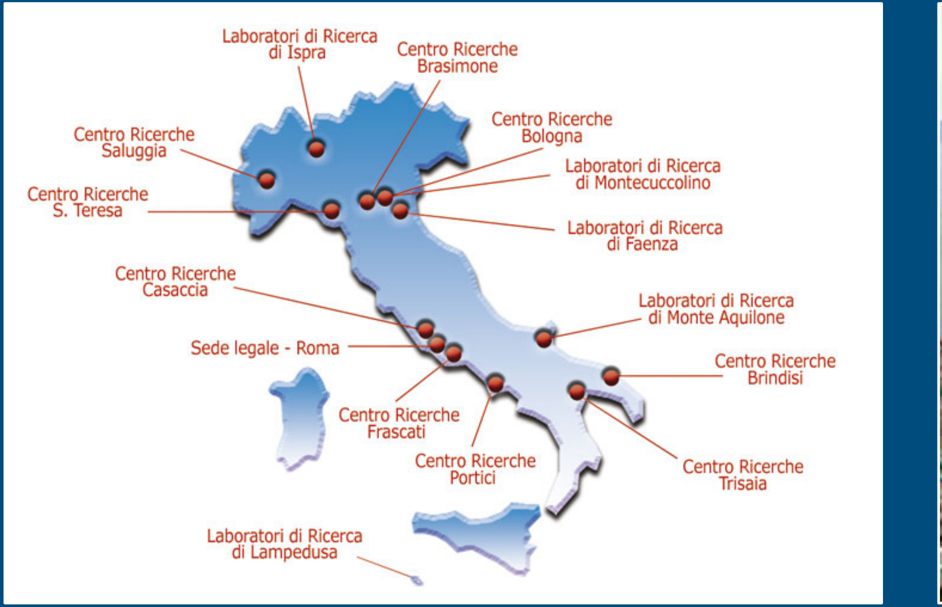
Outline

- **The ENEA context**
  - ENEAGRID environment and CRESCO HPC clusters
  - Web Crawling in ENEAGRID
  - Semantic Brand Scoring in ENEAGRID

- **Proposal of current development**
  - Social Networks Crawling
  - Semantic Brand Score for Cryptocurrencies

- **Conclusions**

1 Headquarters;
9 Research Centers;
5 Laboratories.


Laboratori di Ricerca di Ispra
Centro Ricerche Brasimone
Centro Ricerche Saluggia
Centro Ricerche Bologna
Laboratori di Ricerca di Montecuccolino
Centro Ricerche S. Teresa
Laboratori di Ricerca di Faenza
Centro Ricerche Casaccia
Laboratori di Ricerca di Monte Aquilone
Sede legale - Roma
Centro Ricerche Brindisi
Centro Ricerche Frascati
Centro Ricerche Portici
Centro Ricerche Trisaia
Laboratori di Ricerca di Lampedusa

Portici Research Center

ENEAGRID & CRESCO HPC Clusters

ENEAGRID

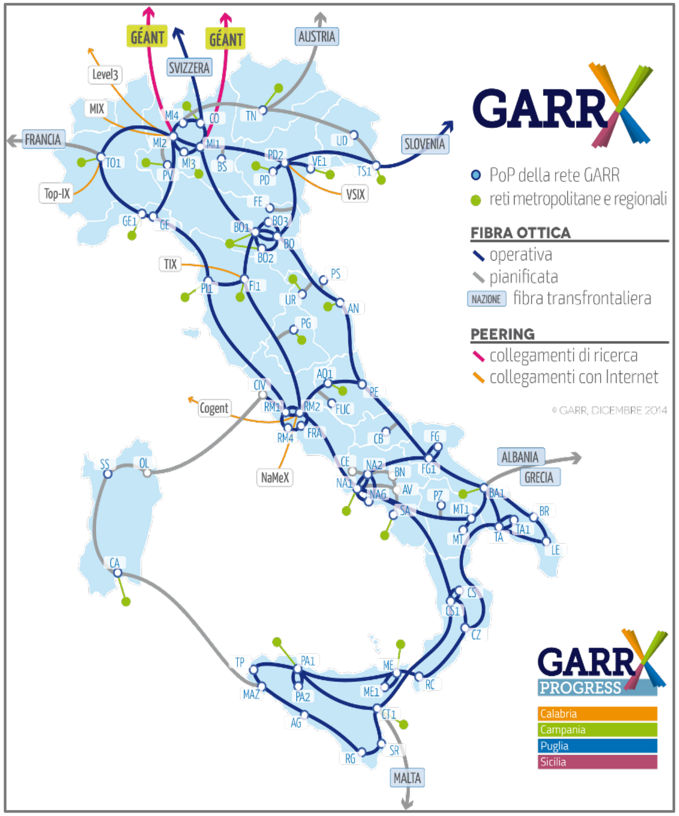
Computation & Storage ***ENEA*** distributed resources interconnected via ***GARR*** network.

CRESCO HPC Clusters:

- 6 *Data Centers* in ENEA (**Portici** is the main site);
- More than 20000 cores;
- More than 400 computing nodes:
  - ➢ *Linux x86_64 + Special systems (GPU, PHI);*
- Storage resources:
  - ➢ *AFS* (distributed);
  - ➢ *GPFS* (parallel high-speed) ~2.2PB;
- Cloud computing facilities *(Openstack, VMWare);*
- More than 1.4Pflops
  - ➢ *in Top500 rank at Nov. 2018.*

CRESCO Project

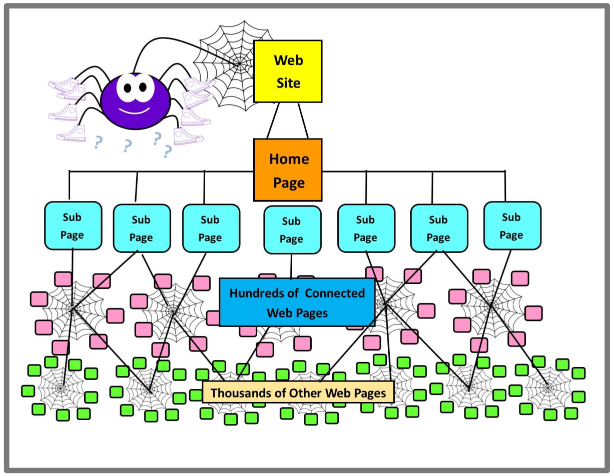
http://www.cresco.enea.it

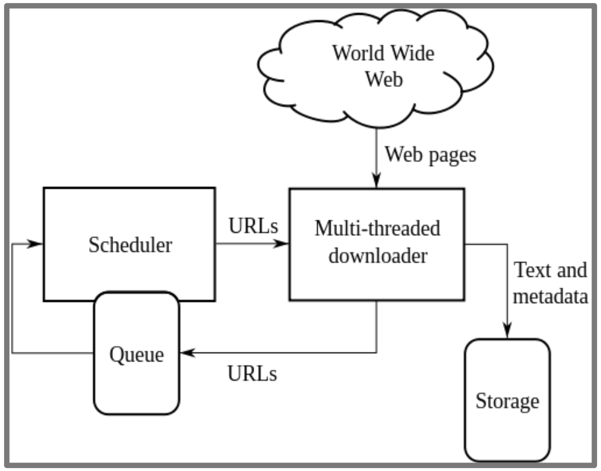


https://www.garr.it

Web Crawling in ENEAGRID

- Activity to browse *www* systematically and download web content;
- *Google*, *Bing* and *Yahoo* periodically download the content of a wide web space;
- Data are stored locally and processed to build indexes, statistics and to structure them;





Application contexts:

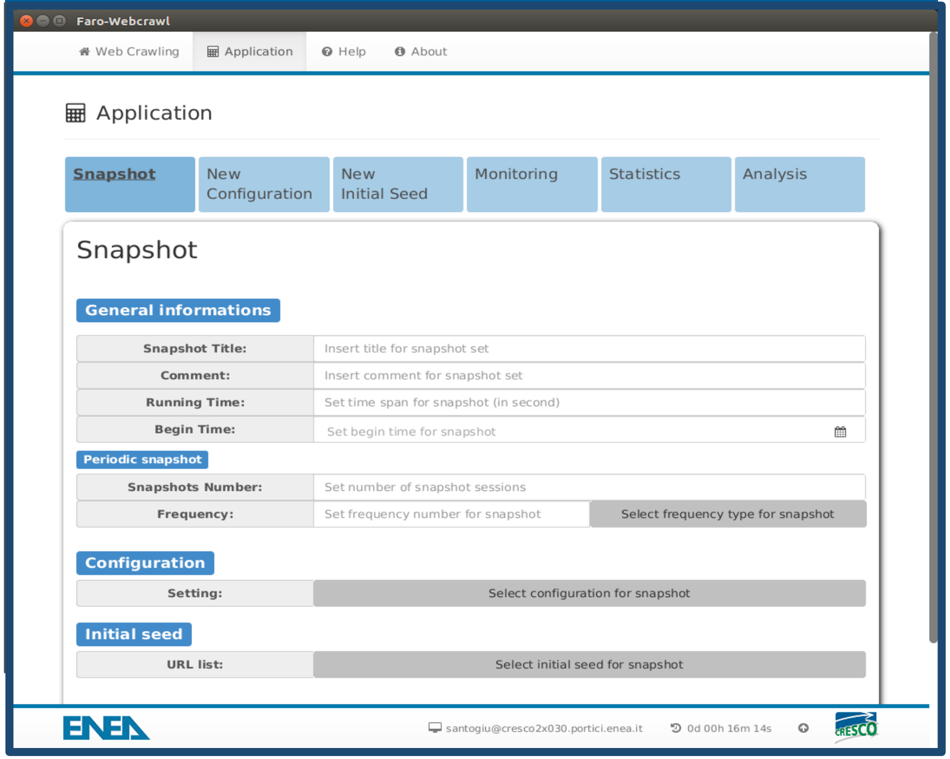
- Web Searching;
- Intelligence & security;
- Blog analysis;
- User behaviors;
- Marketing.

Web Crawling: Software and Virtual Lab

Problems, rules and requirements

- Best practices to avoid improperly using the network (e.g. DDOS);
- Laws to comply with privacy and/or copyright (e.g. GDPR).
- Open source solution (BUbiNG*)

Virtual Laboratory



Snapshot

General informations

Snapshot Title:	Insert title for snapshot set
Comment:	Insert comment for snapshot set
Running Time:	Set time span for snapshot (in second)
Begin Time:	Set begin time for snapshot

Periodic snapshot

Snapshots Number:	Set number of snapshot sessions	
Frequency:	Set frequency number for snapshot	Select frequency type for snapshot

Configuration

Setting:	Select configuration for snapshot

Initial seed

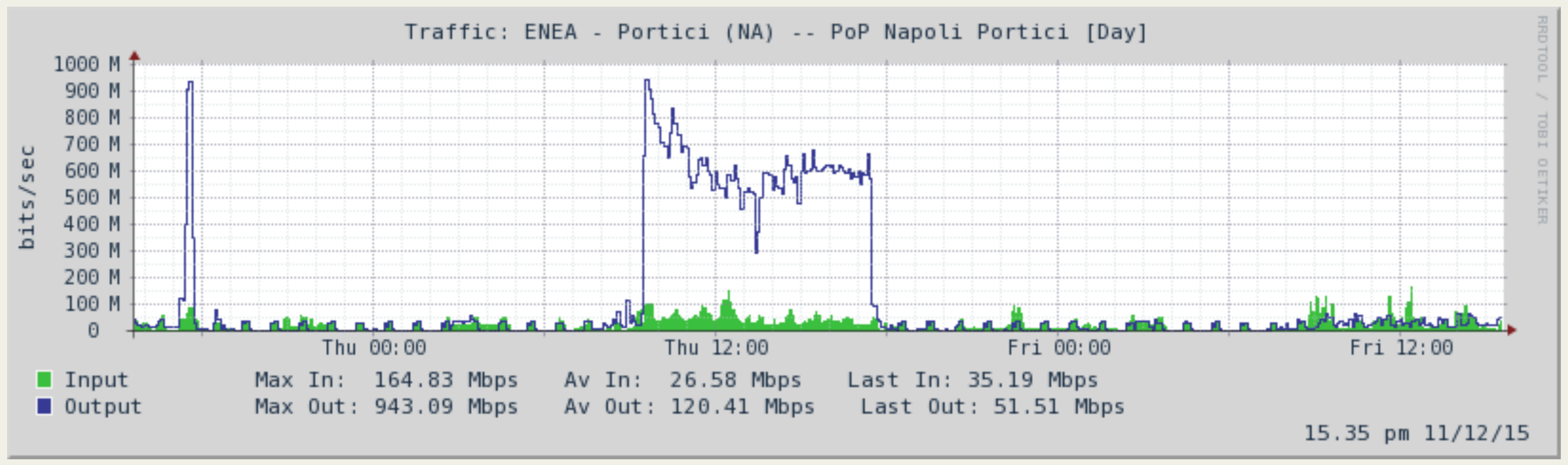
URL list:	Select initial seed for snapshot

santogiu@cresco2x030.portici.enea.it 0d 00h 16m 14s

*[Boldi, P., Marino, A., Santini, M., Vigna, S.: BUbiNG: Massive Crawling for the Masses. (2016)]

First Test: **Single Snapshot**
(Efficiency and Robustness)

Number of agents: 16;
Running time: 8 h;
Amount of downloaded data: 2.94 TB;
Amount of downloaded resources: 66.806.790 Pages;
Data downloading speed: 850 Mbps;
Resources downloadeding speed: 2305 Pages/Sec.

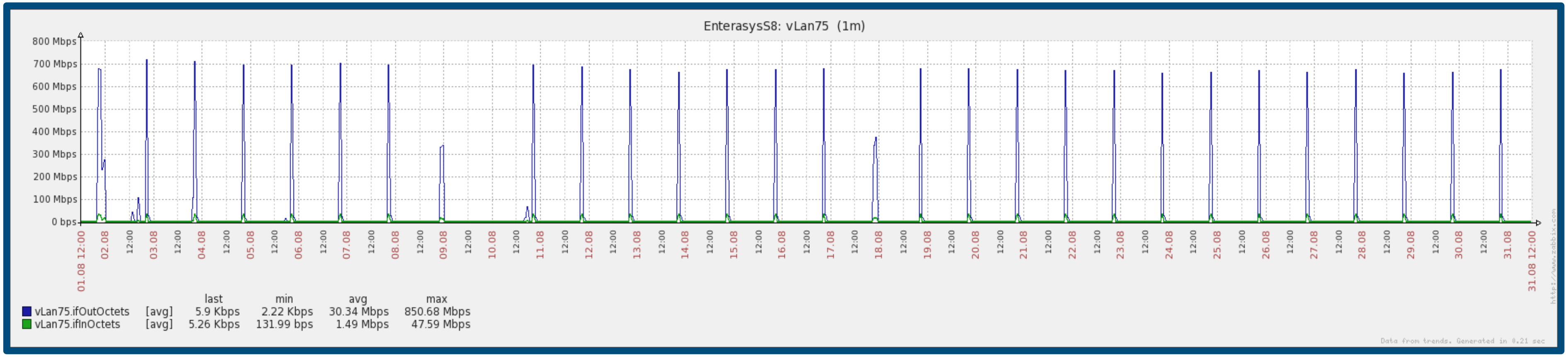


Network traffic measured by GARR on Napoli-Portici PoP.

[Santomauro, G. et Al.: A collaborative environment for web crawling and web data analysis in ENEAGRID. In: DATA 2017, (2017)]

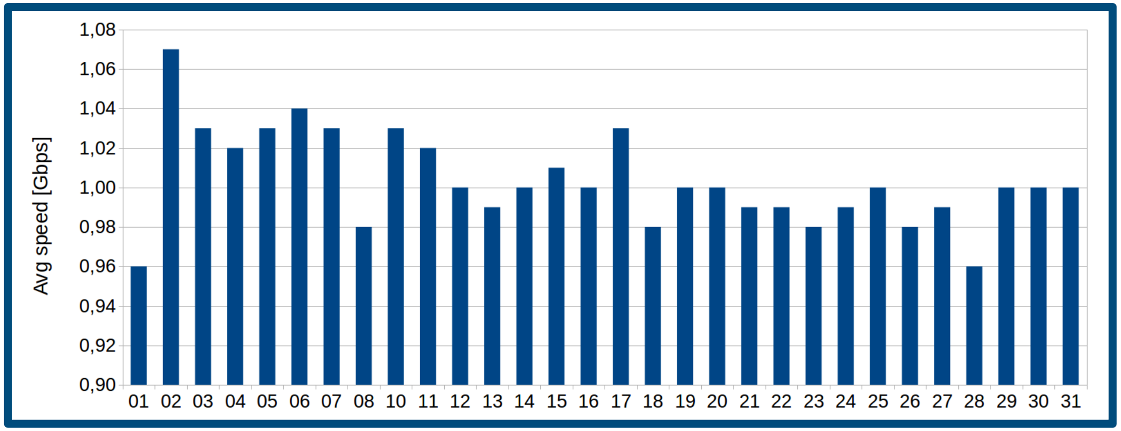
Second Test: **Periodic Snapshots** (Reliability)

Daily web crawling sessions, during **one month**, each of them kept alive for **one hour** (from 21:00 to 22:00), by considering only web pages from *.it* domain.



		last	min	avg	max
🟦 vLan75.ifOutOctets	[avg]	5.9 Kbps	2.22 Kbps	30.34 Mbps	850.68 Mbps
🟩 vLan75.ifInOctets	[avg]	5.26 Kbps	131.99 bps	1.49 Mbps	47.59 Mbps

Network traffic in the snapshot period.



Average of download speed for each snapshot.

- 15 TB of downloaded data (484 GB/snapshot);
- 3,3 TB saved on the storage (111 GB/snapshot);
- **Downloading Speeds**: ***μ***: 1,00 Gbps, ***σ***: 0,0005 Gbps.

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.

- The advantage with respect to some traditional measures is that the SBS do not relies on surveys administered to small samples of consumers.

The metric is measured along the 3 dimensions:

- **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;

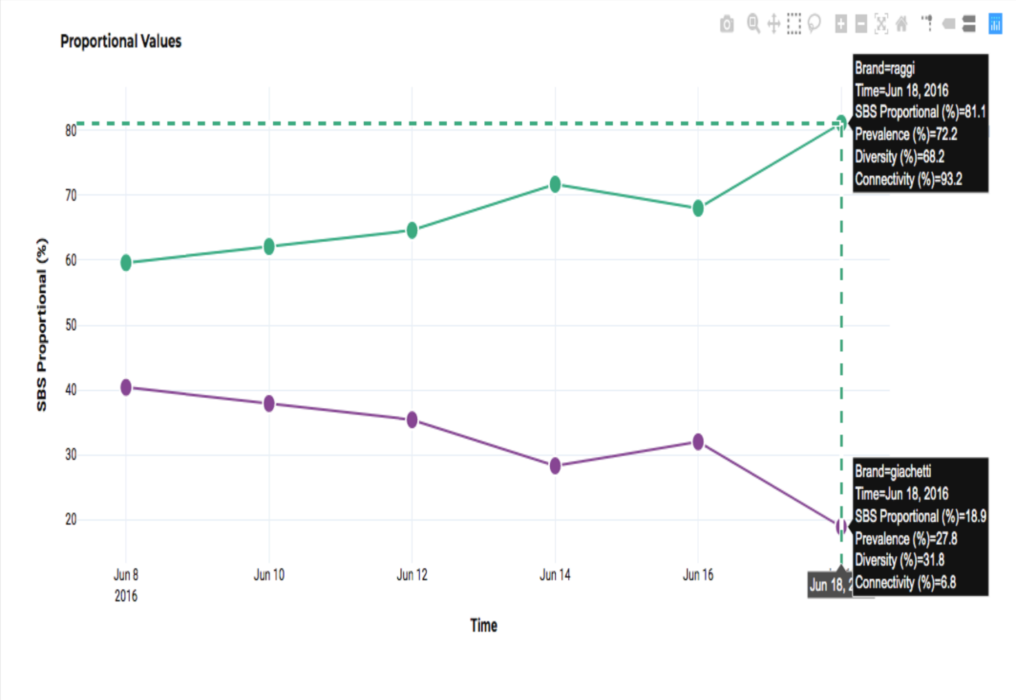
- **Connectivity** represents the brand ability to bridge connections between other words or groups of words (sometimes seen as discourse topics).

***[Fronzetti Colladon, A.: The semantic brand score. Journal of Business Research 88, 150 – 160 (2018)]**

Semantic Brand Score: Test and Virtual Lab

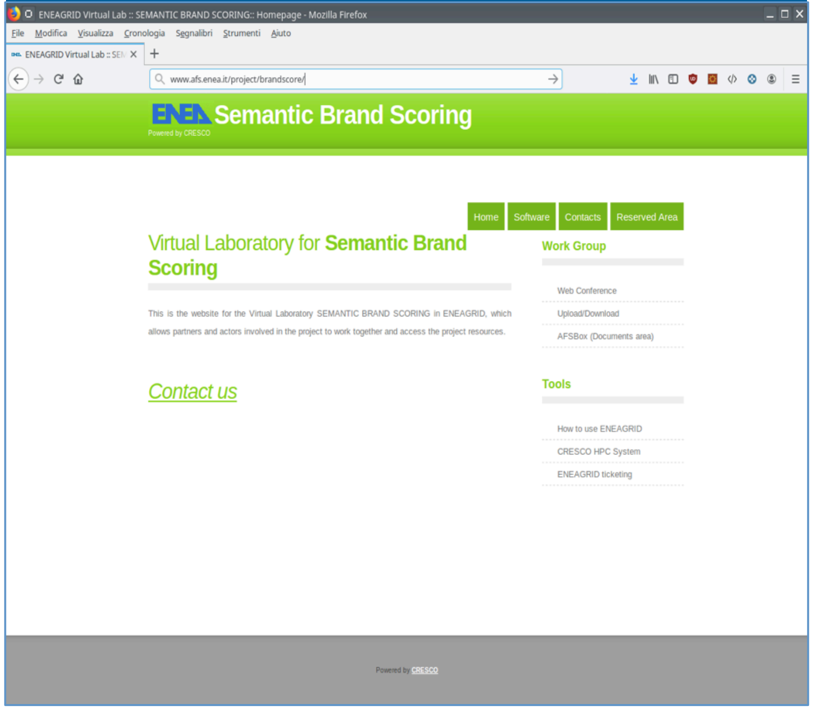
Preliminary tests on the configuration and on the performance demonstrate a correct integration

Test: Rome Mayor election



https://semanticbrandscore.com

Virtual Laboratory


ENEAGRID Virtual Lab :: SEMANTIC BRAND SCORING:: Homepage - Mozilla Firefox

File Modifica Visualizza Cronologia Segnalibri Strumenti Aiuto

ENEAGRID Virtual Lab :: SEM

www.afs.enea.it/project/brandscore/

ENEA Semantic Brand Scoring

Powered by CRESCO

Home | Software | Contacts | Reserved Area

Virtual Laboratory for **Semantic Brand Scoring**

This is the website for the Virtual Laboratory SEMANTIC BRAND SCORING in ENEAGRID, which allows partners and actors involved in the project to work together and access the project resources.

Contact us

Work Group

- Web Conference
- Upload/Download
- AFSBox (Documents area)

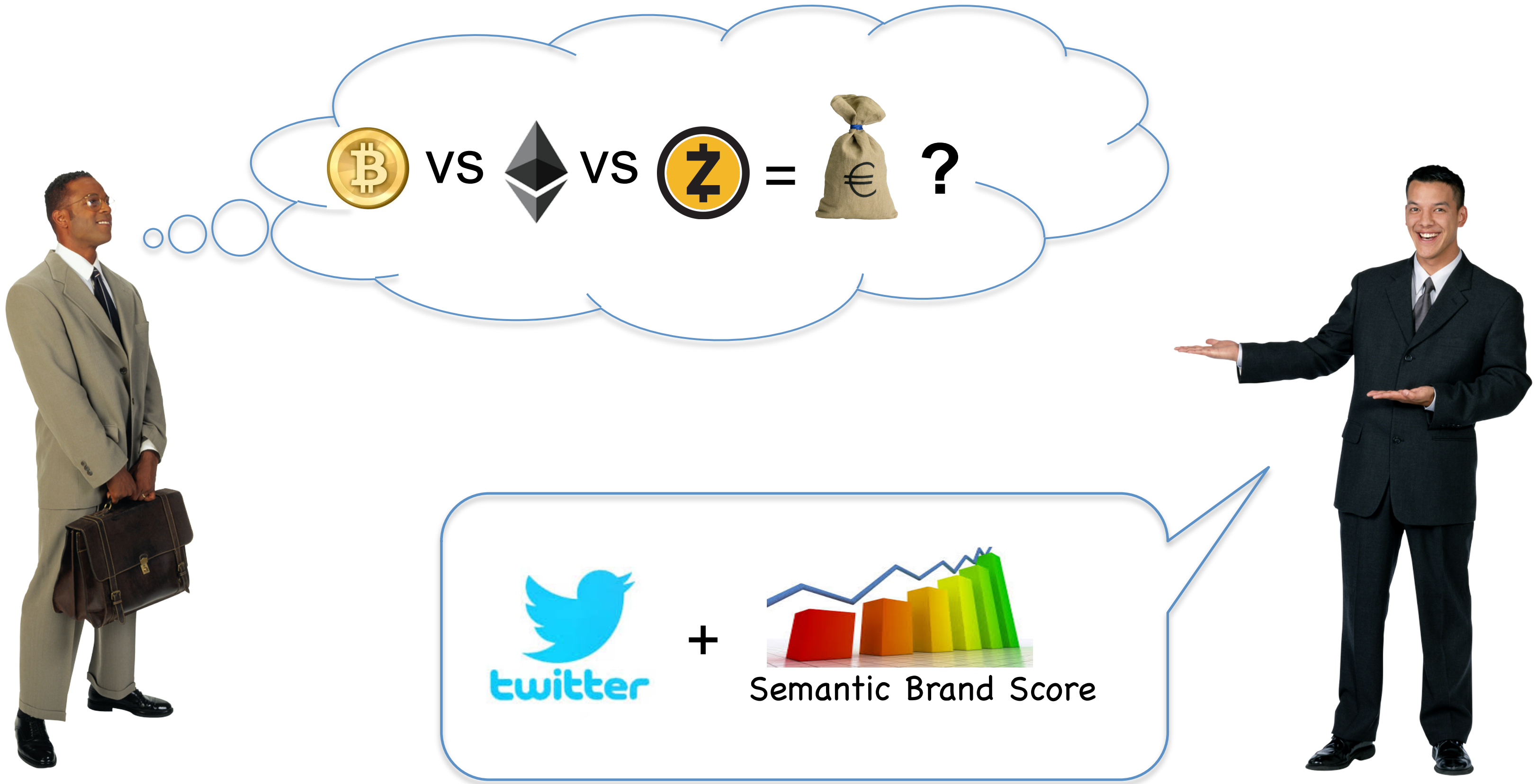
Tools

- How to use ENEAGRID
- CRESCO HPC System
- ENEAGRID ticketing

Powered by CRESCO

http://www.afs.enea.it/project/brandscore/

Proposal



Trader

Our idea
of solution

Current developments

Social Networks Crawling

- **Aim:**
  - Creating dataset on a specific topic;
  - Downloading texts, messages, news from Social Networks.
- **Methodology:**
  - Installing and configuring a social crawler for *Twitter;*
  - Use of parallel developer accounts to avoid the limitation on the number of tweets downloaded per user.

Semantic Brand Score for Cryptocurrencies

- **Aim:**
  - Creating a rank among most popular cryptocurrencies (e.g. *Bitcoin*, *Ethereum*, *Zcash*);
- **Methodology:**
  - Running periodic sessions of crawling in order to create a database of tweets that concern news and discussions about digital coins;
  - Applying the Semantic Brand Score to rank cryptocurrency importances.

Conclusion

✓ We provided an overview of activity about the implementation of web crawler integrated in our HPC ENEAGRID/CRESCO infrastructure;

✓ Currently we are also equipping our framework with a social media crawler that downloads contents from *Twitter* and *a* Semantic Brand Scoring (SBS) tool which uses ENEA computational and storage power;

✓ First tests on the social crawler and on the SBS software demonstrate good results.

Future work…

✍ *Performing experiments to tune our framework;*

✍ *Refining our semantic filter to obtain a more accurate dataset.*

Thanks for the attention

Dr. Giuseppe Santomauro
giuseppe.santomauro@enea.it