



# Analisi e sviluppo di nuove tecniche per l'estrazione di informazioni da grandi moli di dati provenienti dal web

**Giuseppe SANTOMAURO**

*Tutor:* Ing. **Giovanni Ponti**  
(DTE-ICT-HPC, ENEA C.R. Portici)



Agenzia nazionale per le nuove tecnologie,  
l'energia e lo sviluppo economico sostenibile



# Primo Anno: sintesi

## Predisposizione di un ambiente per il web crawling

**Web Data Retrieving:** esplorazione dei contenuti in una rete in maniera sistematica e automatizzata al fine creare archivi di dati web.

### Operazione eseguite:

- Installazione di strumenti di web crawling sul cluster HPC CRESCO presente nel Centro Ricerche di Portici;
- Integrazione in ENEAGRID;
- Definizione di un'infrastruttura hardware ad-hoc per il crawling e confinamento delle risorse;
- Tuning dei parametri del crawler;
- Pianificazione ed esecuzione di sessioni di web crawling in diverse condizioni e con diversa durata.



# Primo Anno: metodologie e strumenti

## Problematiche e Normative

- Ricerca delle best practices al fine di evitare eccessivi sovraccarichi della rete e/o di suoi utilizzi in modo improprio;
- Individuazione delle leggi che regolano il processo di crawling al fine di rispettare i diritti di privacy e/o copyright.

## Prodotti

- Utilizzo di soluzioni open source;
- Possibilità di integrazione nell'infrastruttura ENEAGRID/ CRESCO;
- Flessibilità sulla configurazione del prodotto;
- Strumenti e interfacce per il monitoring dell'esecuzione;
- Formato di archiviazione dei dati web;
- Capacità di garantire le migliori performance.

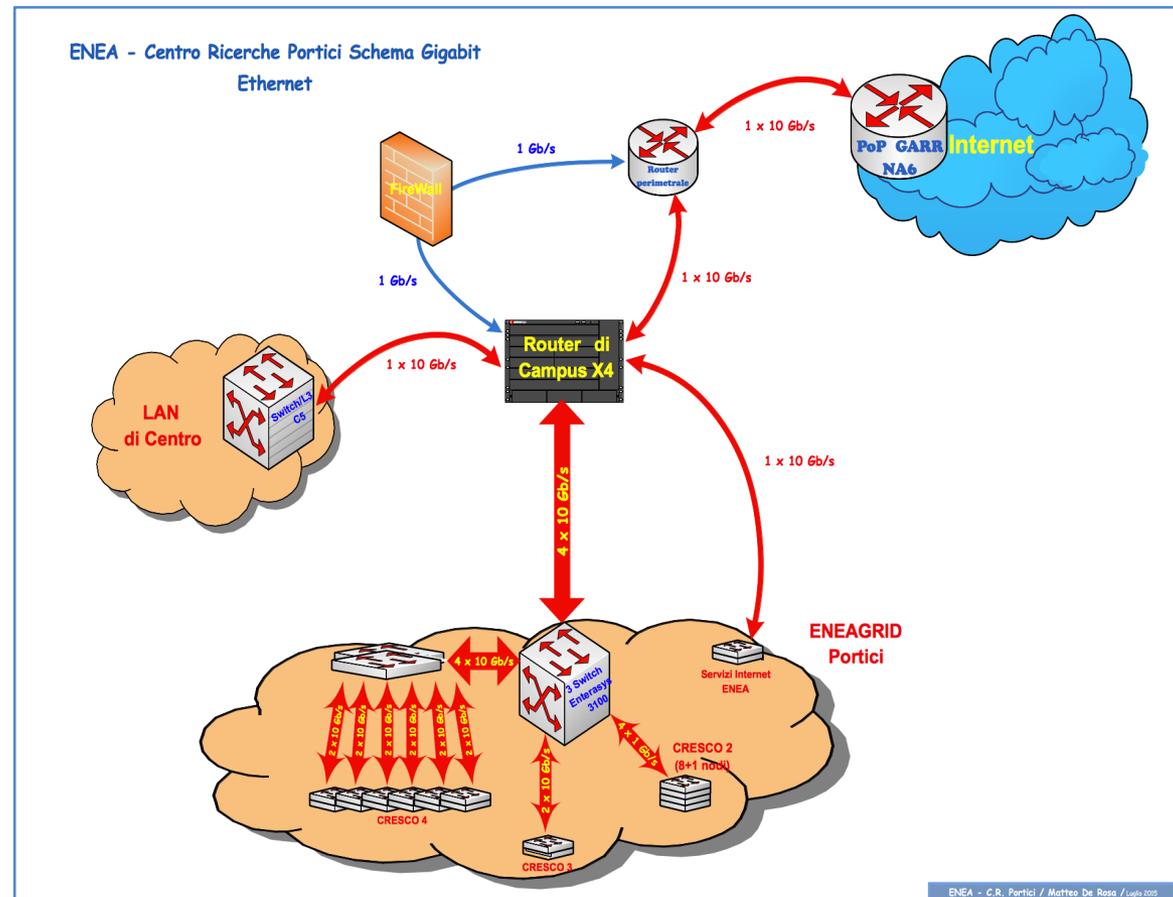


**BUBiNG**



# Primo Anno: architettura e hardware

- **8 (+1) nodi di CRESCO2:**
  - **Processore:** 2 Xeon Quad-Core Clovertown E5345;
  - **RAM:** 16 GByte;
- **Scheduler delle risorse:**  
*LSF 7.0.3;*
- **Storage:** *GPFS 4.2.2;*
- **Rete:** *1Gbs.*



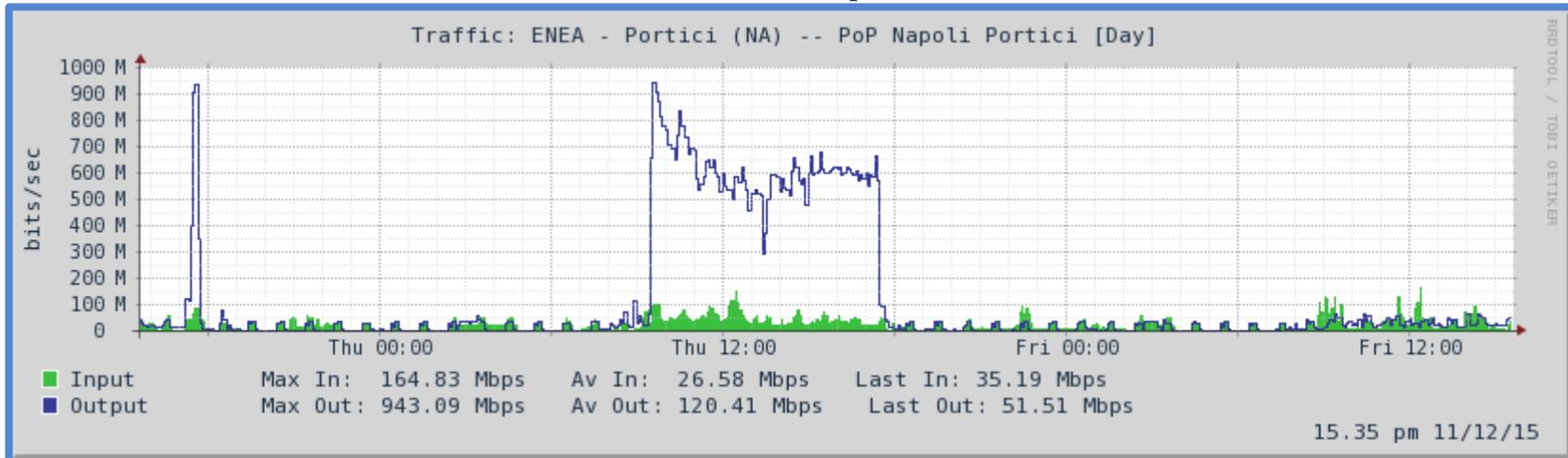


# Primo Anno: risultati

## Test del 10/12/2015

- Numero di agenti: 16
- Tempo di esecuzione: ~8 h
- Quantità di dati scaricati: ~2,94 TB
- Quantità di risorse scaricate: 66.806.790 Pagine
- Velocità di dati scaricati: ~850 Mbps
- Velocità di risorse scaricate: ~2305 Pagine/Sec.

## Traffico PoP Napoli-Portici





# Secondo Anno: sintesi

## Integrazione di strumenti di analisi per contenuti provenienti dal web

**Web Data Analysis: Fruizione dei contenuti; Qualità dei dati; Analisi.**

### Obiettivi:

- Studio e installazione di opportuni strumenti per l'indicizzazione, l'archiviazione e la visualizzazione dei dati;
- Integrazione di strumenti per data aggregation e individuazione di topic;
- Investigazione sulla qualità dei dati scaricati.



# Secondo Anno: sintesi

## Attività svolte

- Creazione di un Laboratorio Virtuale sulla tematica del web crawling per una community di lavoro integrato e collaborativo;
- Sviluppo di un'interfaccia grafica (GUI) per il Laboratorio Virtuale;
- Esecuzione di snapshot periodici;
- Integrazione di *Apache Solr* per queryng e analisi dei dati;
- Integrazione di *OpenWayback* per la visualizzazione dei dati;
- Integrazione di *Carrot2* per clustering dei dati;



# Secondo Anno: tempistiche

## 1) Prima fase [~3 mesi]:

- Creazione di un sito internet sul Web Crawling per il Laboratorio Virtuale; ✓
- Creazione di un'interfaccia grafica (sottomissione snapshot singoli e periodici). ✓

## 2) Seconda fase [~3 mesi]:

- Miglioramento funzionalità maschera grafica (monitoring, statistics); ✓
- Avvio snapshot periodici su determinate porzioni del Web a lungo termine. ✓

## 3) Terza fase [~3 mesi]:

- Analisi delle prestazioni della sessione di snapshot periodici su determinate porzioni del Web a lungo termine; ✓
- Aggiunta di nuove funzionalità nella maschera grafica (setting, initial seed, analysis, display, clustering). ✓

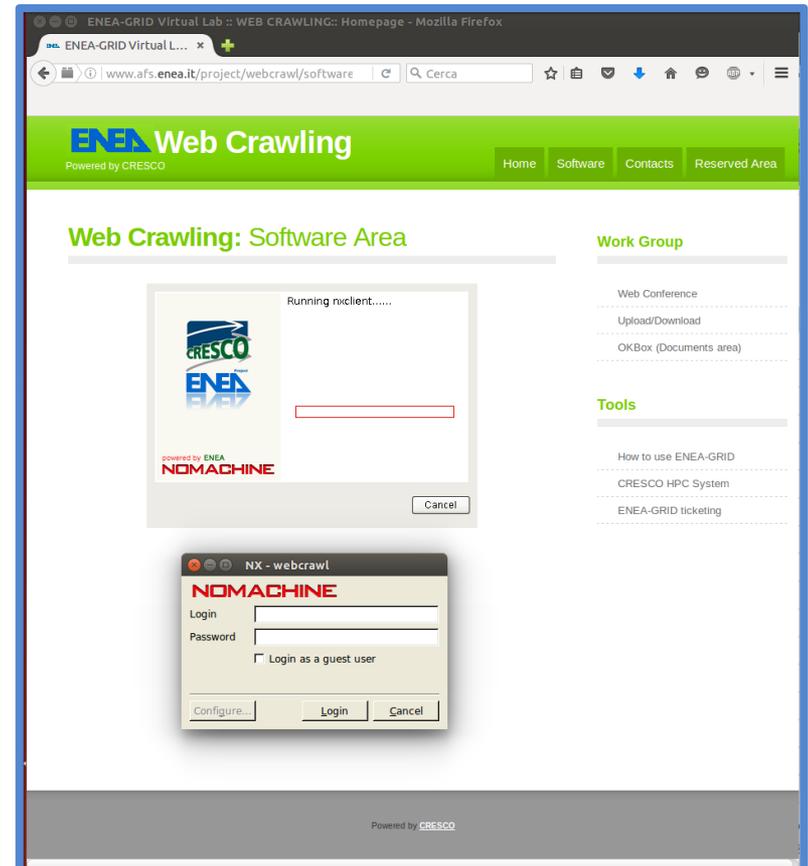
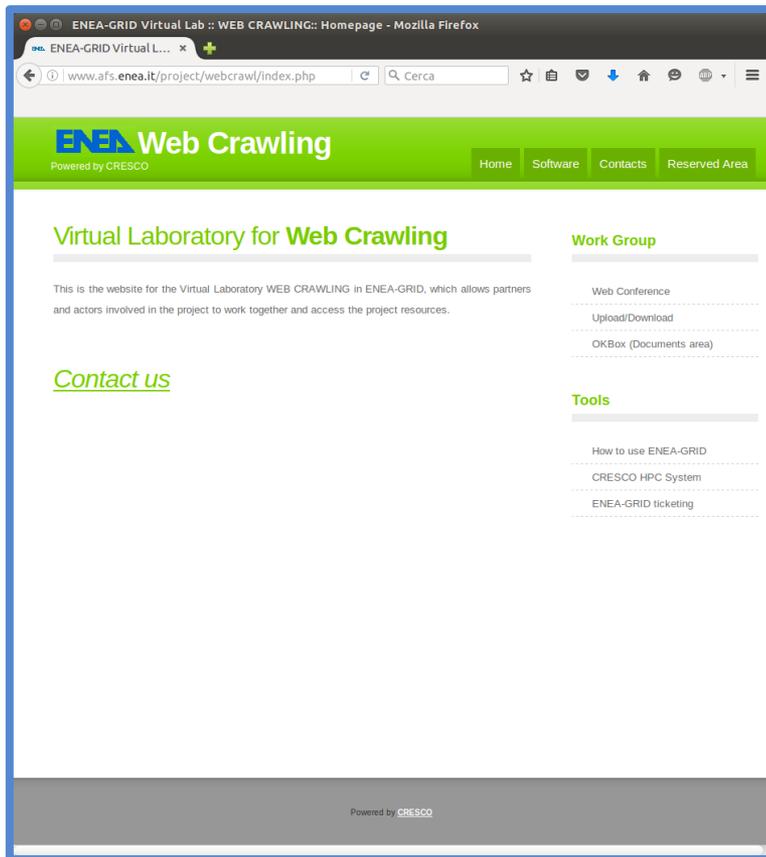
## 4) Quarta fase [~3 mesi]:

- Analisi dei dati scaricati.



# Laboratorio Virtuale

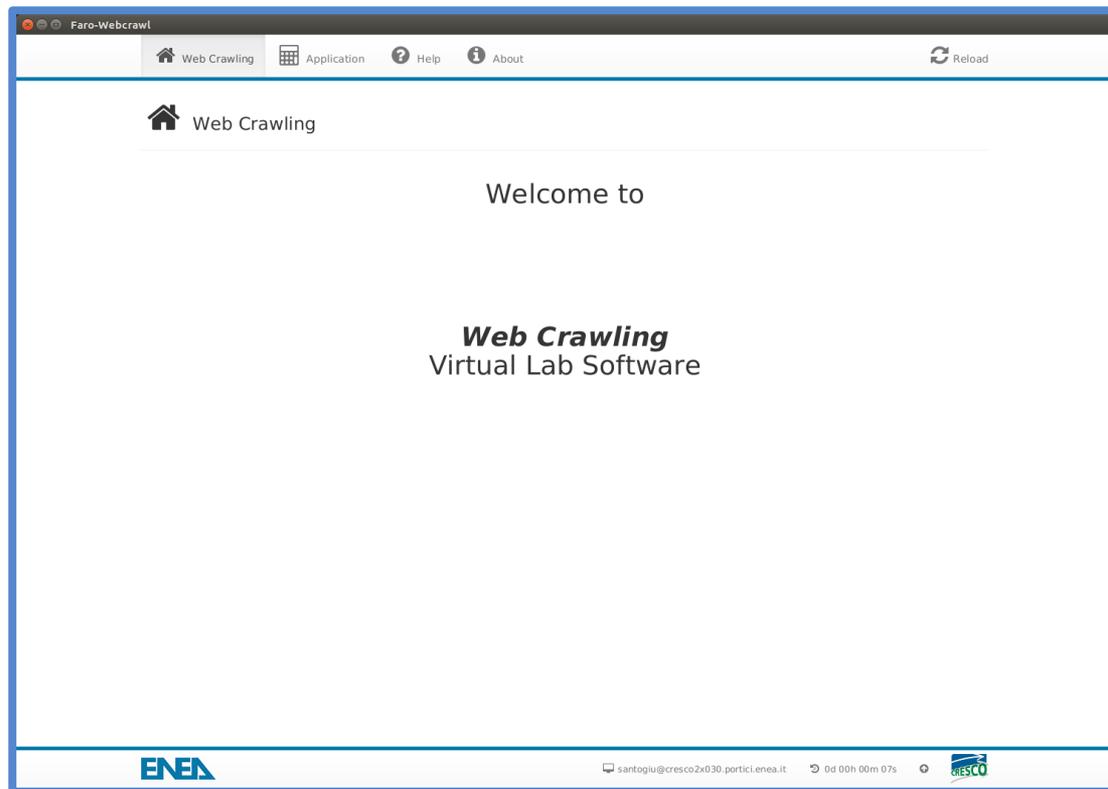
**Metodologia per consentire l'utilizzo remoto di servizi di ricerca e caratterizzazione erogati da grandi impianti e/o apparecchiature, unici per caratteristiche e/o costo.**





# Interfaccia grafica

Web application, sviluppata in *JavaFX*, che permette all'utente di interagire facilmente con le risorse hardware e software dedicate all'attività di web crawling.



Schermata di Benvenuto.



# Interfaccia grafica: snapshot

The screenshot shows the 'FARO-Webcrawl' application interface. At the top, there are navigation tabs: 'Web Crawling', 'Application', 'Help', and 'About'. A 'Reload' button is also present. The main content area is titled 'Application' and contains several sub-sections: 'Snapshot', 'New Configuration', 'New Initial Seed', 'Monitoring', 'Statistics', and 'Analysis & Display'. The 'Snapshot' section is active and contains the following fields:

- 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

A 'Submit Session' button is located at the bottom of the form. The footer of the application shows the ENEA logo, the email address 'santogiu@cresco2x030.portici.enea.it', a timer '0d 00h 00m 06s', and the CRESO logo.

## Funzionalità:

- Sottomissione di snapshot singoli o periodici;
- Possibilità di scegliere la durata, la data e l'ora di esecuzione, la configurazione e il seme iniziale.

Tab la sottomissione di snapshot.



# Interfaccia grafica: new configurations

The screenshot shows the 'Faro-Webcrawl' application interface. At the top, there is a navigation bar with 'Web Crawling', 'Application', 'Help', and 'About' tabs, and a 'Reload' button. Below this is a main menu with buttons for 'Snapshot', 'New Configuration' (which is highlighted), 'New Initial Seed', 'Monitoring', 'Statistics', and 'Analysis & Display'. The 'New Configuration' form is divided into two sections: 'General Options' and 'Software parameters'. Each section has a dropdown menu to select from saved configurations and a 'Reload' button. The 'General Options' section includes fields for 'H/W environment', 'Title', 'Comment', 'Nodes', and 'Agents per node'. The 'Software parameters' section includes fields for 'Software', 'rootDir', 'maxUrlsPerSchemeAuthority', 'parsingThreads', 'dnsThreads', 'fetchingThreads', 'fetchFilter', and 'scheduleFilter'. The bottom of the interface shows the ENEA logo, a contact email 'santogiu@cresco2x030.portici.enea.it', a timer '0d 00h 01m 31s', and the CRESO logo.

## Funzionalità:

- Creazione di nuovi settaggi per il software;
- Possibilità di caricare vecchie configurazioni, modificarle e salvarle.

Tab la gestione dei settaggi.



# Interfaccia grafica: new initial seed

The screenshot shows the 'Faro-Webcrawl' application window. The top navigation bar includes 'Web Crawling', 'Application', 'Help', and 'About'. The main interface has a sidebar with 'Application' and a top menu with 'Snapshot', 'New Configuration', 'New Initial Seed' (highlighted), 'Monitoring', 'Statistics', and 'Analysis & Display'. The 'New Initial Seed' section contains an 'Upload Seed' field with 'it.seed' and a 'Reload' button. Below is a 'URLs list' section with a 'Title' field and a list of URLs. At the bottom, there is a 'Save Initial Seed' button. The footer shows the ENEA logo, a user email 'santogliu@cresco2x030.portici.enea.it', a timestamp '0d 00h 02m 23s', and the CRESCO logo.

## Funzionalità:

- Creazione di nuove liste di semi di url iniziali;
- Possibilità di caricare vecchi liste, modificarle e salvarle sovrascrivendo o creando nuovi elenchi.

Tab la gestione delle liste di url iniziali.



# Interfaccia grafica: monitoring

The screenshot shows the 'Faro-Webcrawl' application interface. At the top, there are navigation tabs: 'Web Crawling', 'Application', 'Help', and 'About'. The 'Application' tab is active. Below the navigation, there are several buttons: 'Snapshot', 'New Configuration', 'New Initial Seed', 'Monitoring' (highlighted), 'Statistics', and 'Analysis & Display'. The 'Monitoring' section contains a 'Running snapshots:' area with a dropdown menu showing 'snapshot-2016-12-06-16-45-00' and a 'Reload' button. Below this are three buttons: 'Pause Snapshot', 'Resume Snapshot', and 'Stop Snapshot'. Further down, there are three buttons: 'Start Monitoring', 'Stop Monitoring', and 'Hide Monitoring'. A table displays the status of eight agents and a total summary.

|               |               |
|---------------|---------------|
| Agent 01:     | 153.25MB      |
| Agent 02:     | 154.75MB      |
| Agent 03:     | 141.50MB      |
| Agent 04:     | 156.25MB      |
| Agent 05:     | 123.25MB      |
| Agent 06:     | 114.25MB      |
| Agent 07:     | 116.75MB      |
| Agent 08:     | 108.25MB      |
| <b>TOTAL:</b> | <b>1.04GB</b> |

At the bottom of the interface, there is an ENEA logo, a contact email 'santogiu@cresco2x030.portici.enea.it', a timer '0d 00h 39m 42s', and a CRESO logo.

## Funzionalità:

- Monitoraggio della quantità di dati scaricata in tempo reale.
- Possibilità di selezionare tra gli snapshot in esecuzione, metterli in pausa, riavviarli o stopparli.

Tab per il monitoring di snapshot.



# Interfaccia grafica: statistics

Done snapshots: snapshot-2016-08-31-21-00-00 **Reload**

**View Statistics** **Compute Statistics** **Hide Statistics** **Delete Snapshot**

| Name         | Time        | Store            | Resource               |
|--------------|-------------|------------------|------------------------|
| Agent01      | 01h:01m:05s | 29.87 GB         | 693994 Pages           |
| Agent02      | 01h:02m:16s | 29.57 GB         | 664591 Pages           |
| Agent03      | 01h:02m:22s | 31.22 GB         | 684284 Pages           |
| Agent04      | 01h:02m:16s | 30.35 GB         | 684152 Pages           |
| Agent05      | 01h:02m:21s | 29.94 GB         | 680600 Pages           |
| Agent06      | 01h:02m:34s | 32.03 GB         | 703445 Pages           |
| Agent07      | 01h:01m:14s | 28.84 GB         | 644660 Pages           |
| Agent08      | 01h:02m:15s | 28.39 GB         | 642974 Pages           |
| Agent09      | 01h:01m:40s | 31.66 GB         | 717954 Pages           |
| Agent10      | 01h:03m:17s | 30.38 GB         | 703600 Pages           |
| Agent11      | 01h:02m:36s | 31.03 GB         | 679730 Pages           |
| Agent12      | 01h:01m:34s | 30.04 GB         | 687203 Pages           |
| Agent13      | 01h:04m:16s | 28.75 GB         | 652359 Pages           |
| Agent14      | 01h:03m:30s | 28.72 GB         | 656992 Pages           |
| Agent15      | 01h:03m:11s | 30.76 GB         | 701935 Pages           |
| Agent16      | 01h:02m:49s | 30.49 GB         | 695443 Pages           |
| <b>TOTAL</b> |             | <b>482.02 GB</b> | <b>10893916 Pages</b>  |
| <b>SPEED</b> |             | <b>1.00 Gb/s</b> | <b>2825.19 Pages/s</b> |

ENE A santogiu@cresco2x030.portici.enea.it 0d 00h 04m 54s CRESO

## Funzionalità:

- Visualizzazione delle statistiche relative ad uno snapshot;
- Possibilità di caricare vecchie esecuzioni, ricalcolare le statistiche, oppure cancellare completamente uno snapshot.

Tab per la visualizzazione delle statistiche.



# Interfaccia grafica: analysis & display

Faro-Webcrawl

Web Crawling Application Help About Reload

Application

Snapshot New Configuration New Initial Seed Monitoring Statistics Analysis & Display

Analysis & Display

Done snapshots: snapshot-2016-08-31-21-00-00 Reload

Start Analysis Stop Analysis Start Display Stop Display Start Clustering Stop Clustering

ENE A santogiu@cresco2x030.portici.enea.it 0d 00h 05m 24s CRESCO

## Funzionalità:

- Avvio degli strumenti di analisi, visualizzazione e clustering sugli snapshot;
- *Solr* (querying);
- *OpenWayback* (display);
- *Carrot2* (clustering).

Tab per l'analisi, la visualizzazione e il clustering dei dati.



# Analisi dei dati: *Solr*

The screenshot shows the Solr Admin interface with the following details:

- Request-Handler (qt):** /select
- Request:** q=text:"terremoto"
- Response (JSON):**

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 1,
    "params": {
      "q": "text:\u0022terremoto\u0022",
      "indent": "true",
      "fl": "source_file, url",
      "sort": "source_file asc",
      "rows": "5",
      "wt": "json",
      "_": "1478275598653"
    }
  },
  "response": {
    "numFound": 32,
    "start": 0,
    "docs": [
      {
        "source_file": "snapshot-2016-10-06-18-22-00-bkp-store01.warc.gz",
        "url": "http://www.ilfattoquotidiano.it/2016/08/25/terremoto-centro-italia-labbraccio-tra-renzi-e-il"
      },
      {
        "source_file": "snapshot-2016-10-06-18-22-00-bkp-store01.warc.gz",
        "url": "http://www.cittadellascienza.it/futuroremoto/2016/calendario/"
      },
      {
        "source_file": "snapshot-2016-10-06-18-22-00-bkp-store01.warc.gz",
        "url": "http://iononrischio.protezionecivile.it/terremoto-io-non-rischio/sei-preparato-terremoto/"
      },
      {
        "source_file": "snapshot-2016-10-06-18-22-00-bkp-store01.warc.gz",
        "url": "http://www.lauracima.it/categoria/donne/"
      },
      {
        "source_file": "snapshot-2016-10-06-18-22-00-bkp-store01.warc.gz",
        "url": "http://webtv.esercito.difesa.it/Detail/Dettaglio?ChannelId=02c9a5c5-8b45-46b7-94f4-1b15a732d"
      }
    ]
  }
}
```

## Funzionalità:

- Creazione di indici per collezioni di dati;
- Ricerca testuale tra i documenti;
- Analisi dei contenuti.

Schermata di *Solr*.



# Display dei dati: *OpenWayback*

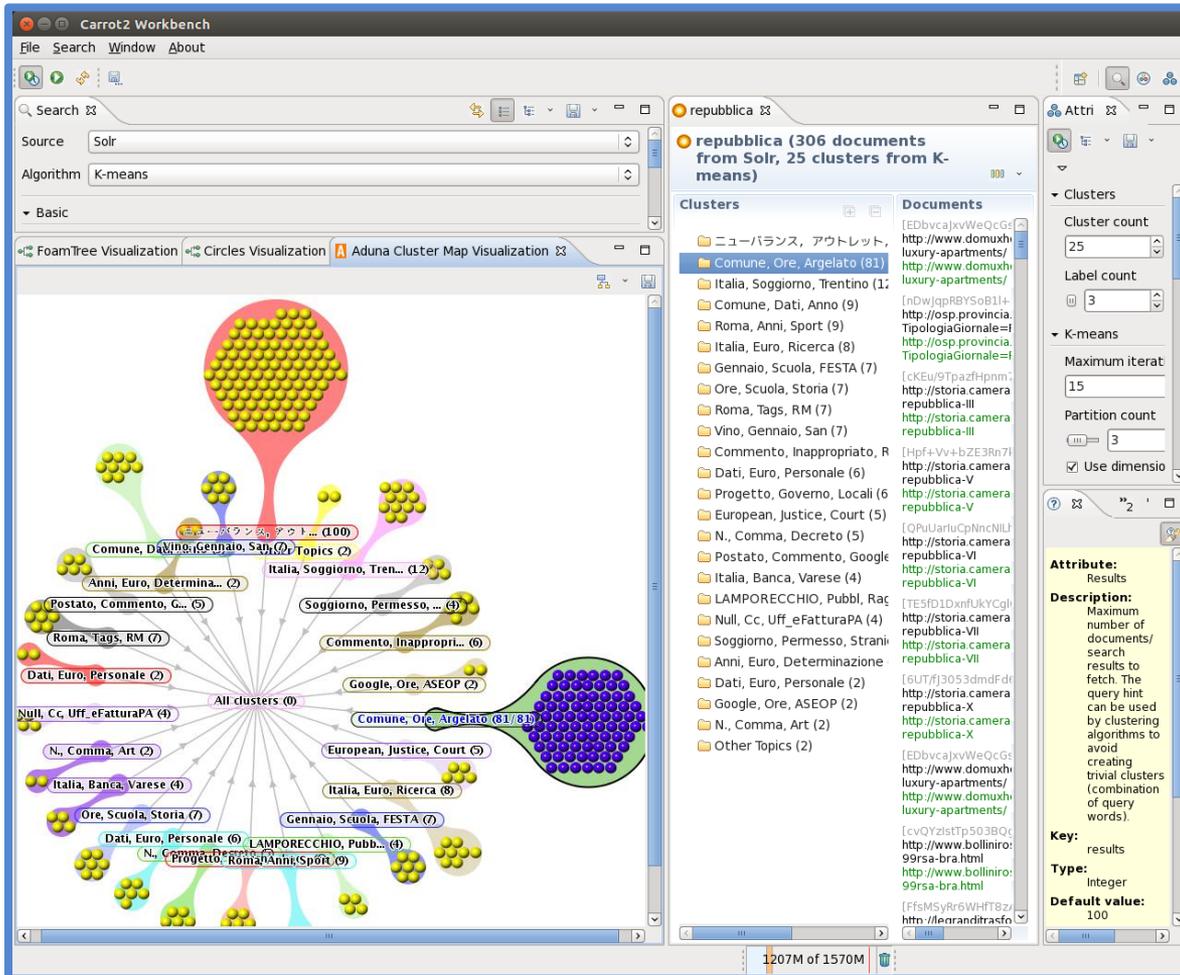
Schermata di *OpenWayback*.

## Funzionalità:

- Creazione di indici per collezioni di dati;
- Ricerca per url e data;
- Visualizzazione dei contenuti testuali.



# Clustering dei dati: Carrot2



## Funzionalità:

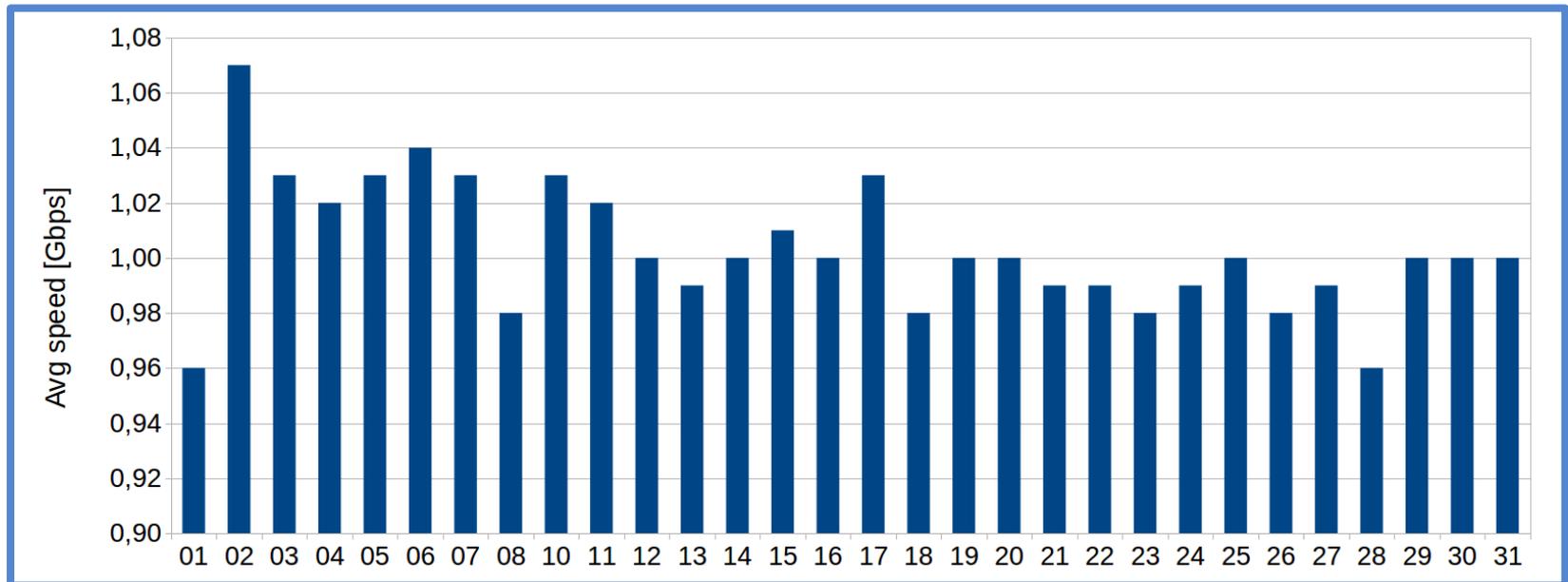
- Creazione di cluster sui di dati.
- Utilizzo di varie tecniche di clustering (*K-means*, *Lingo*, *STC*, ...);
- Visualizzazione dei risultati in diverse forme.

Schermata di Carrot2.



# Snapshot periodici

**Sessione di snapshot a cadenza giornaliera della durata di un'ora ciascuno (dalle 21:00 alle 22:00 circa), considerando solo pagine web appartenenti al dominio italiano *.it*.**



Velocità media di download per ogni snapshot durante agosto 2016.



# *Demo*



# Ultime attività

## Entro il termine della borsa si prevede di:

- Migliorare il processo di indicizzazione dei documenti (correzione dei bug, parallelizzazione,...);
- Analizzare i dati contenuti nel dataset di snapshot periodici;
- Applicare algoritmi di clustering per individuare raggruppamenti e evidenziare topic sulla base di query e parole chiave.



# Fine

Grazie per l'attenzione.

[giuseppe.santomauro@enea.it](mailto:giuseppe.santomauro@enea.it)



# Primo Anno: tempistiche

## 1) Prima fase [~2 mesi]:

- Studio e individuazione delle metodologie per il web crawling;
- Analisi e individuazione dei prodotti software.

## 2) Seconda fase [~4 mesi]:

- Studio dell'infrastruttura ENEAGRID/CRESCO;
- Individuazione del tipo e della quantità delle risorse fisiche da impiegare nell'attività di crawling.

## 3) Terza fase [~4 mesi]:

- Installazione, configurazione e prime esecuzioni di test dei prodotti software;
- Tuning dei parametri e individuazione di configurazioni ottimali.

## 4) Quarta fase [~2 mesi]:

- Esecuzione di crawling di grandi dimensioni;
- Analisi prestazionale dei risultati.