Consortium GARR

8° BORSISTI DAY
"ORIO CARLINI"
MERCOLEDÌ 14 DICEMBRE 2016 - ROMA

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)

ENEA
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: metodolgie 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.

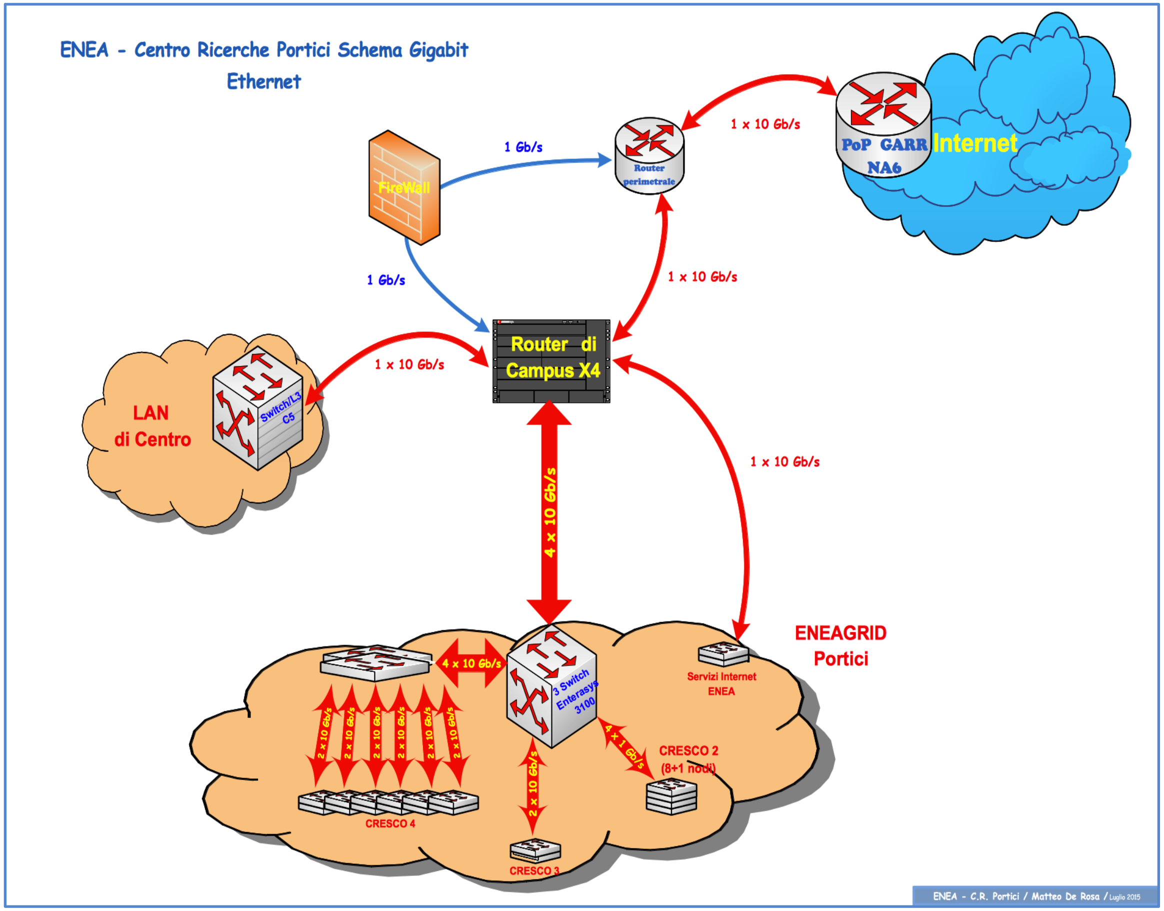
Prodotti

- 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.*

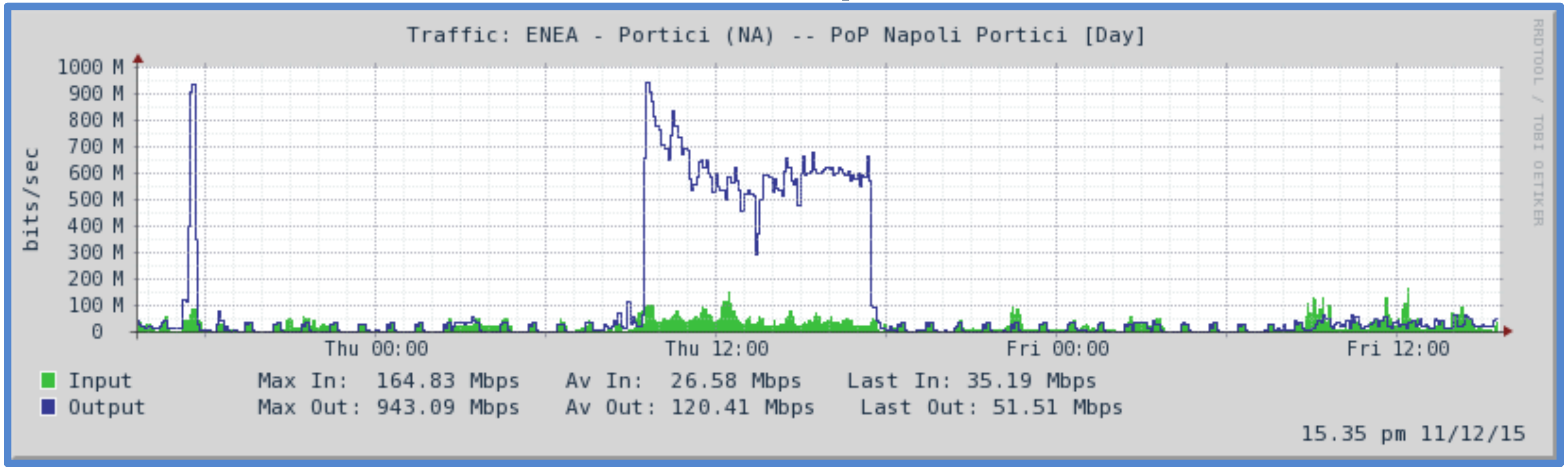


Primo Anno: risultati

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]:

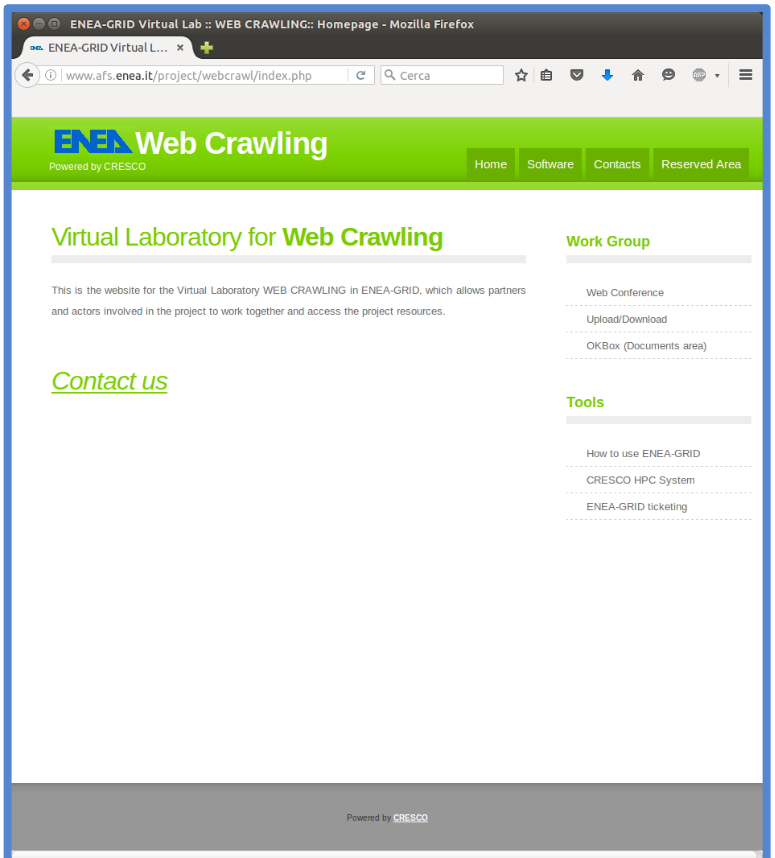
- Analisi delle prestazioni della sessione di snapshot periodici su determinate porzioni del Web a lungo termine; ✔
- Aggiunta di nuove funzionalità nella maschera grafica (setting, initial seed, analysis, display, clustering). ✔

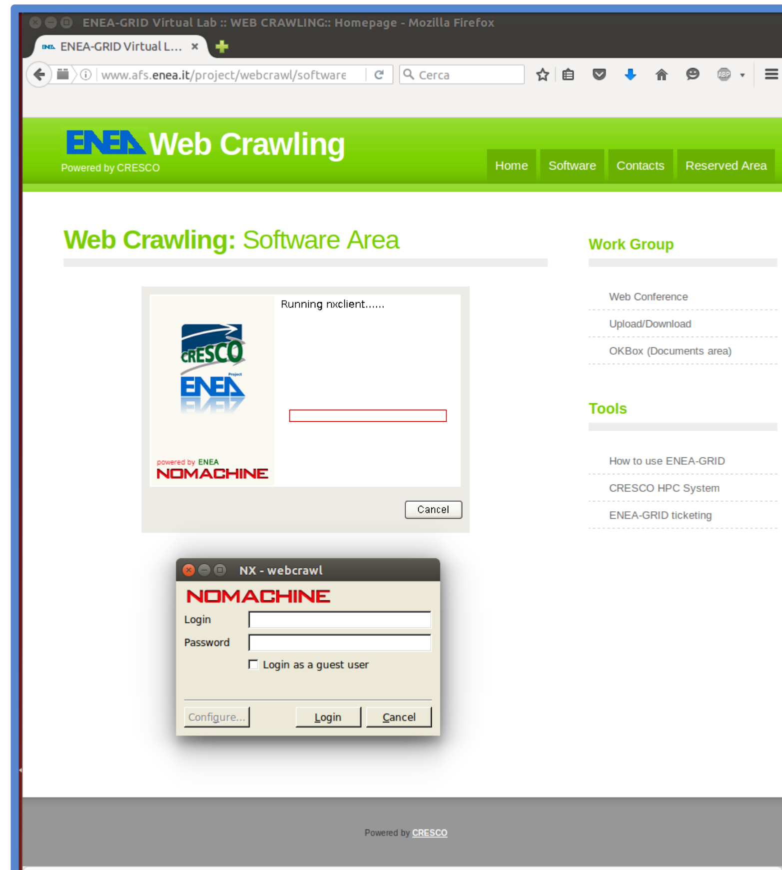
4) Quarta fase [~3 mesi]:

- Analisi dei dati scaricati.

Laboratorio Virtuale

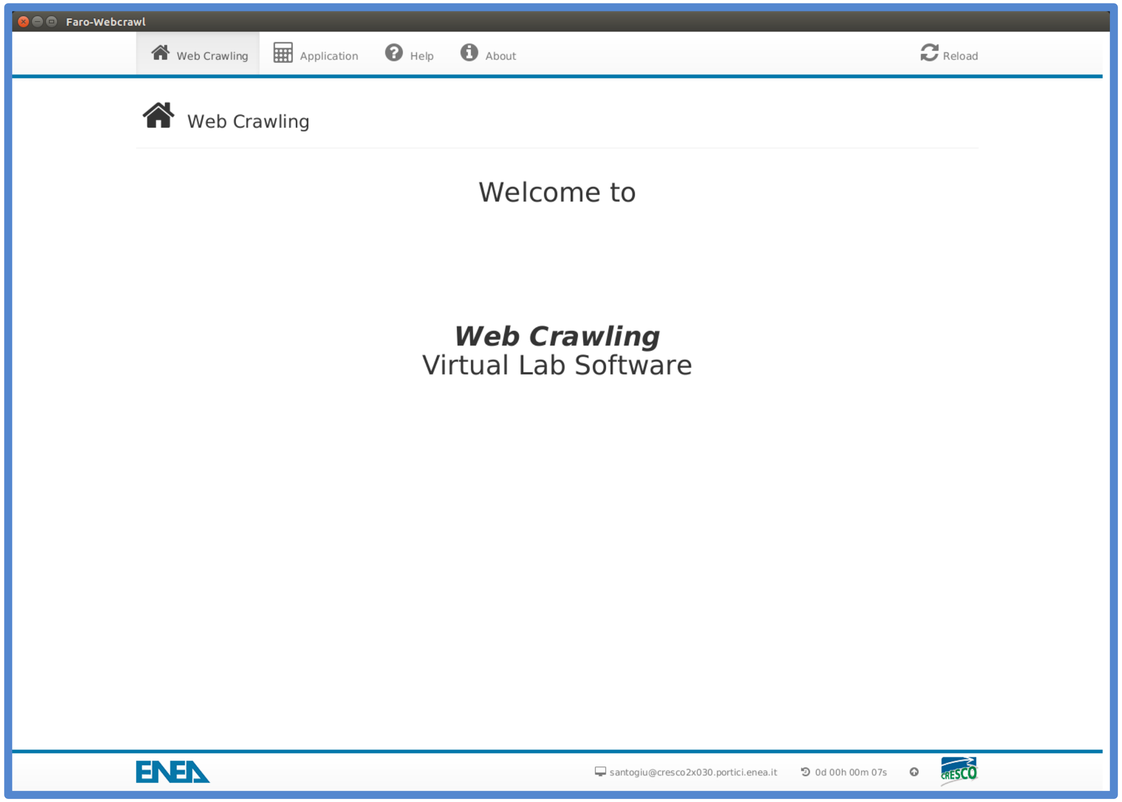
Metodologia per consentire l'utilizzo remoto di servizi di ricerca e caratterizzazione erogati da grandi impianti e/o apparecchiature, unici per caratteristiche e/o costo.





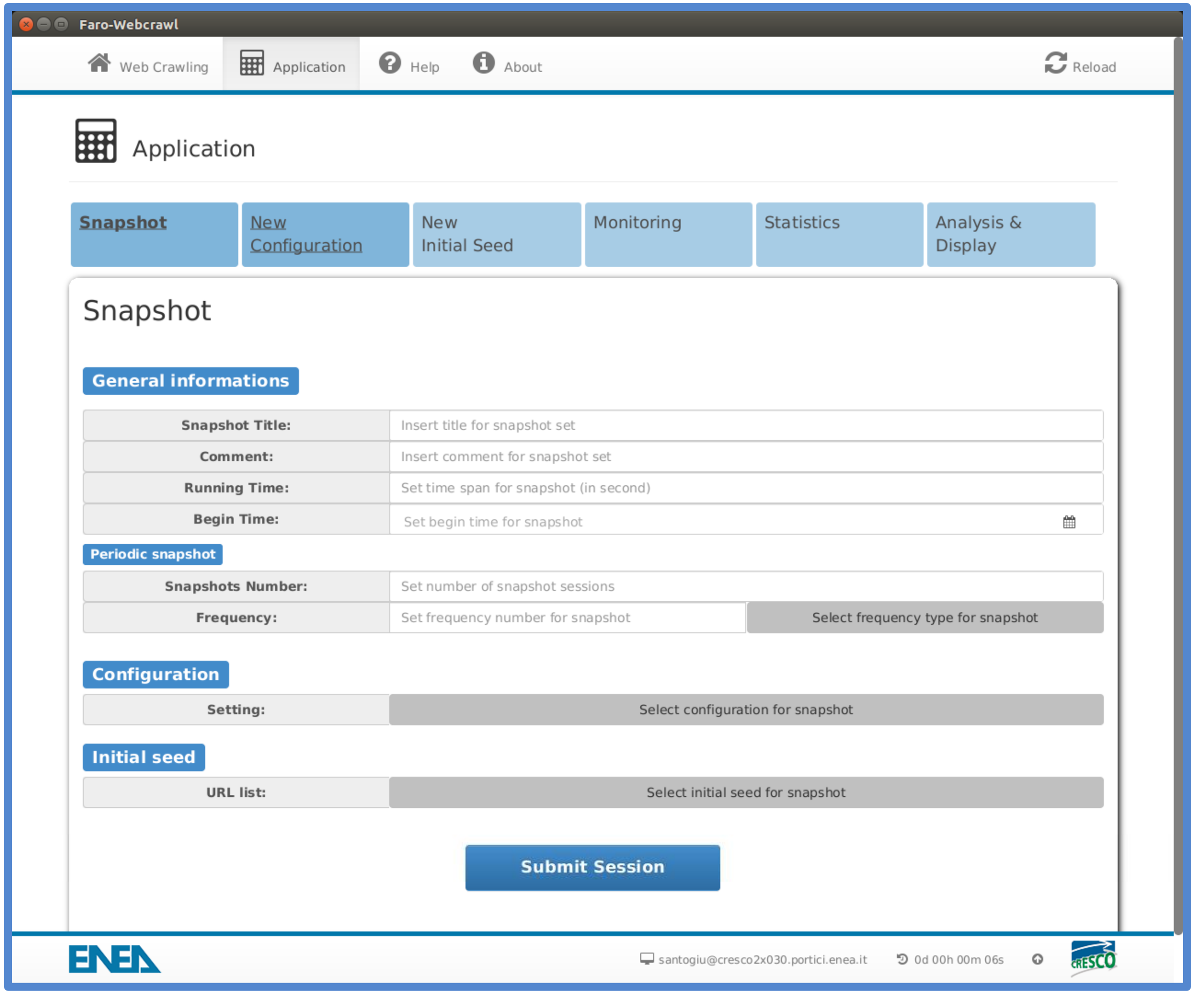
Interfaccia grafica

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**

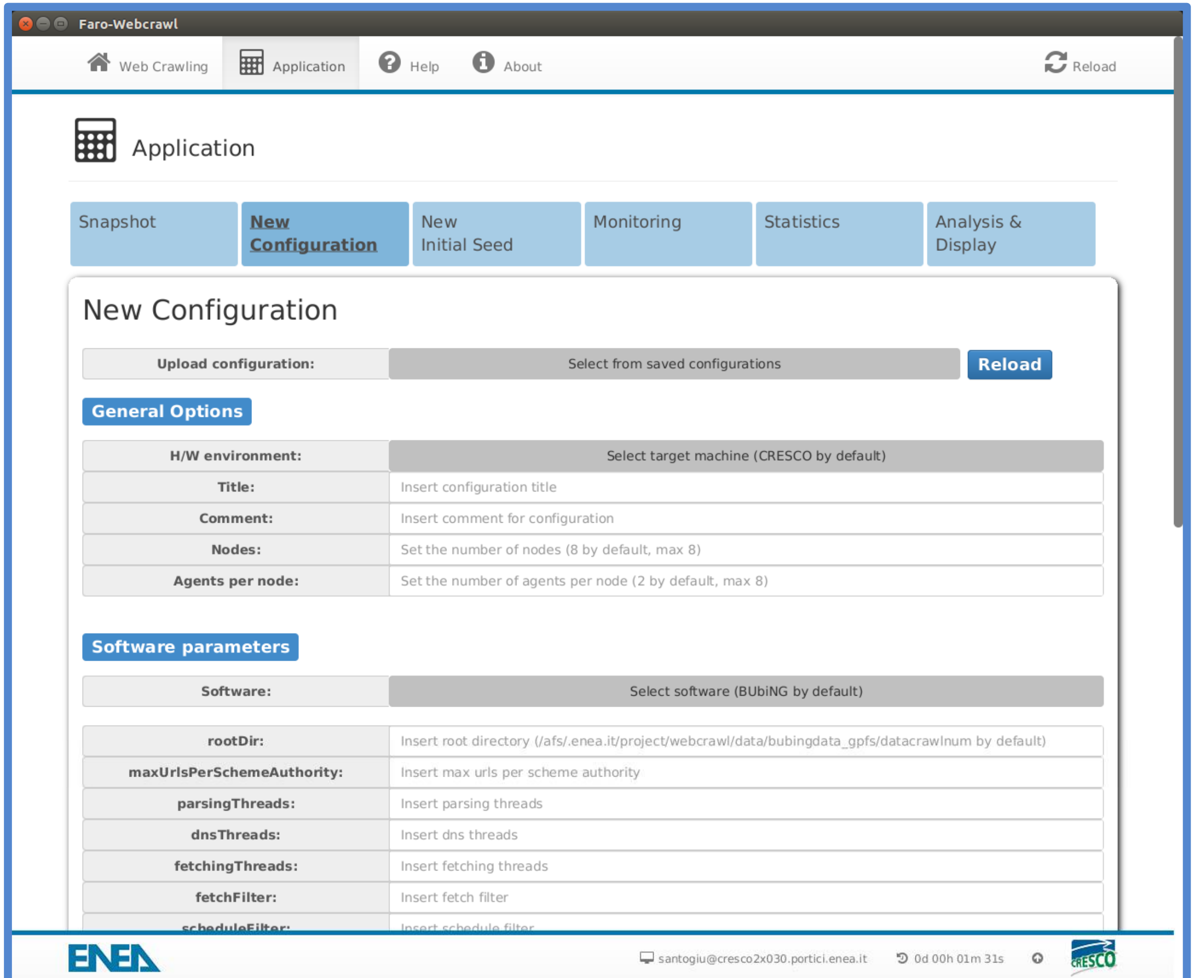


Tab la sottomissione di snapshot.

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.

Interfaccia grafica: **new configurations**

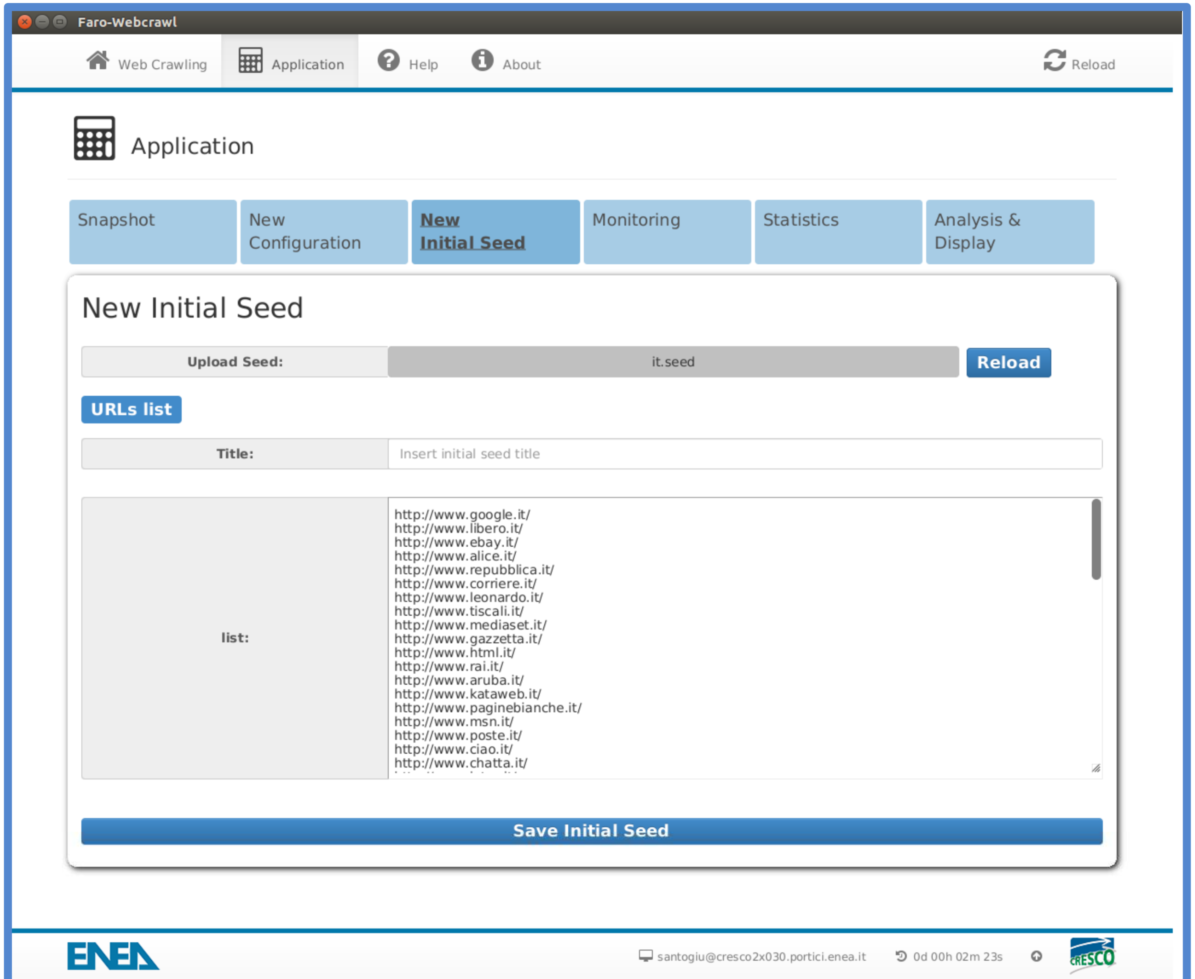


Tab la gestione dei settaggi.

Funzionalità:

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

Interfaccia grafica: **new initial seed**

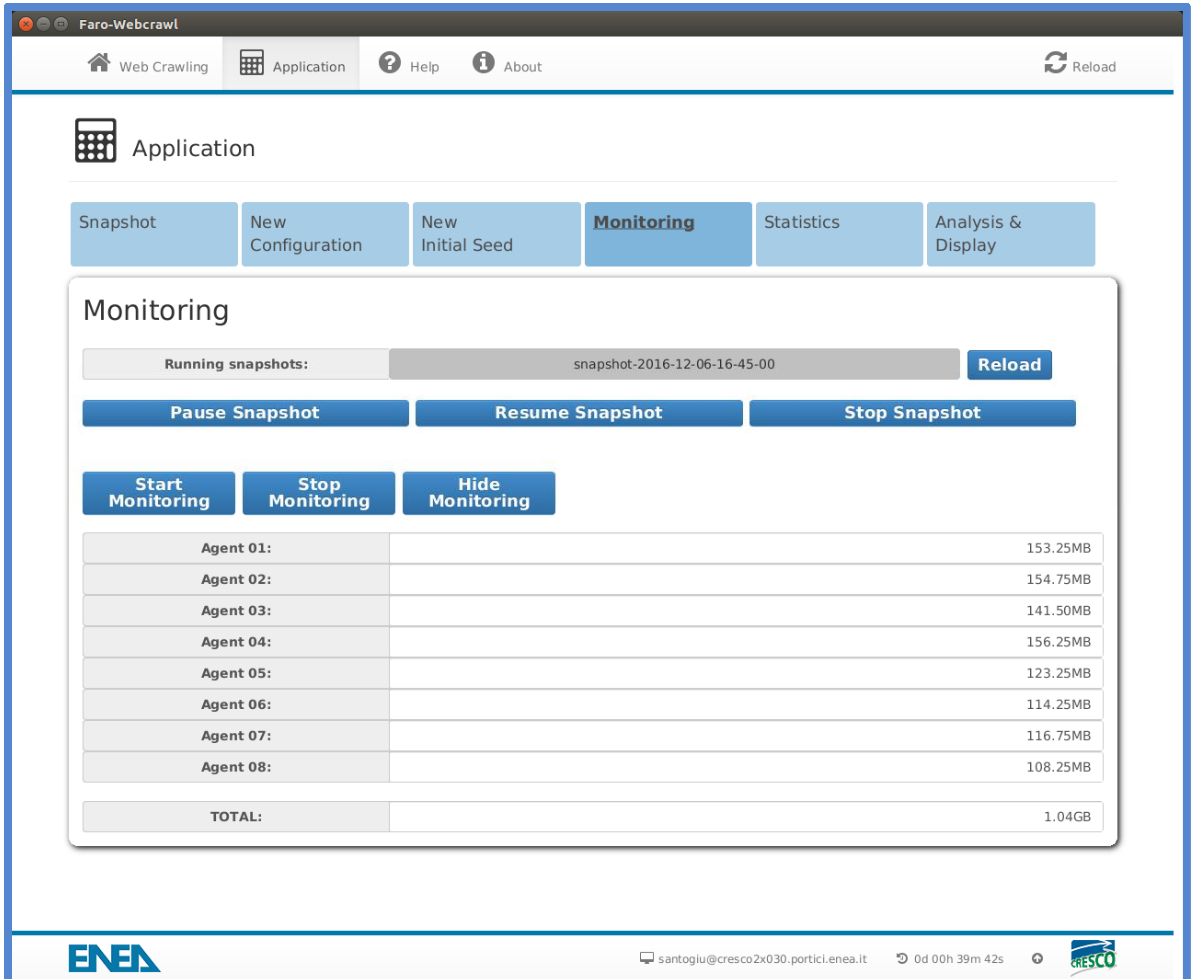


Tab la gestione delle liste di url iniziali.

Funzionalità:

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

Interfaccia grafica: **monitoring**

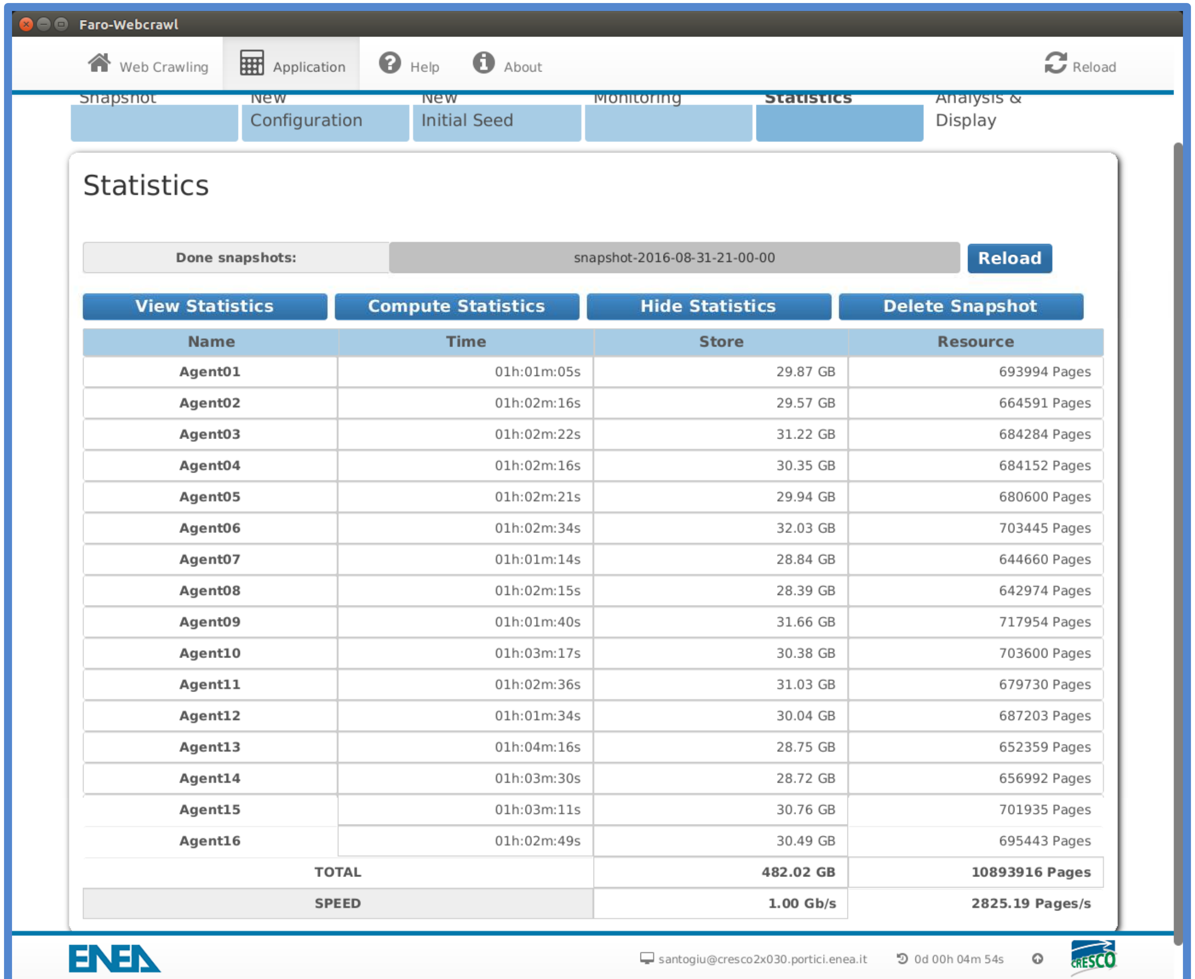


Tab per il monitoring di snapshot.

Funzionalità:

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

Interfaccia grafica: **statistics**



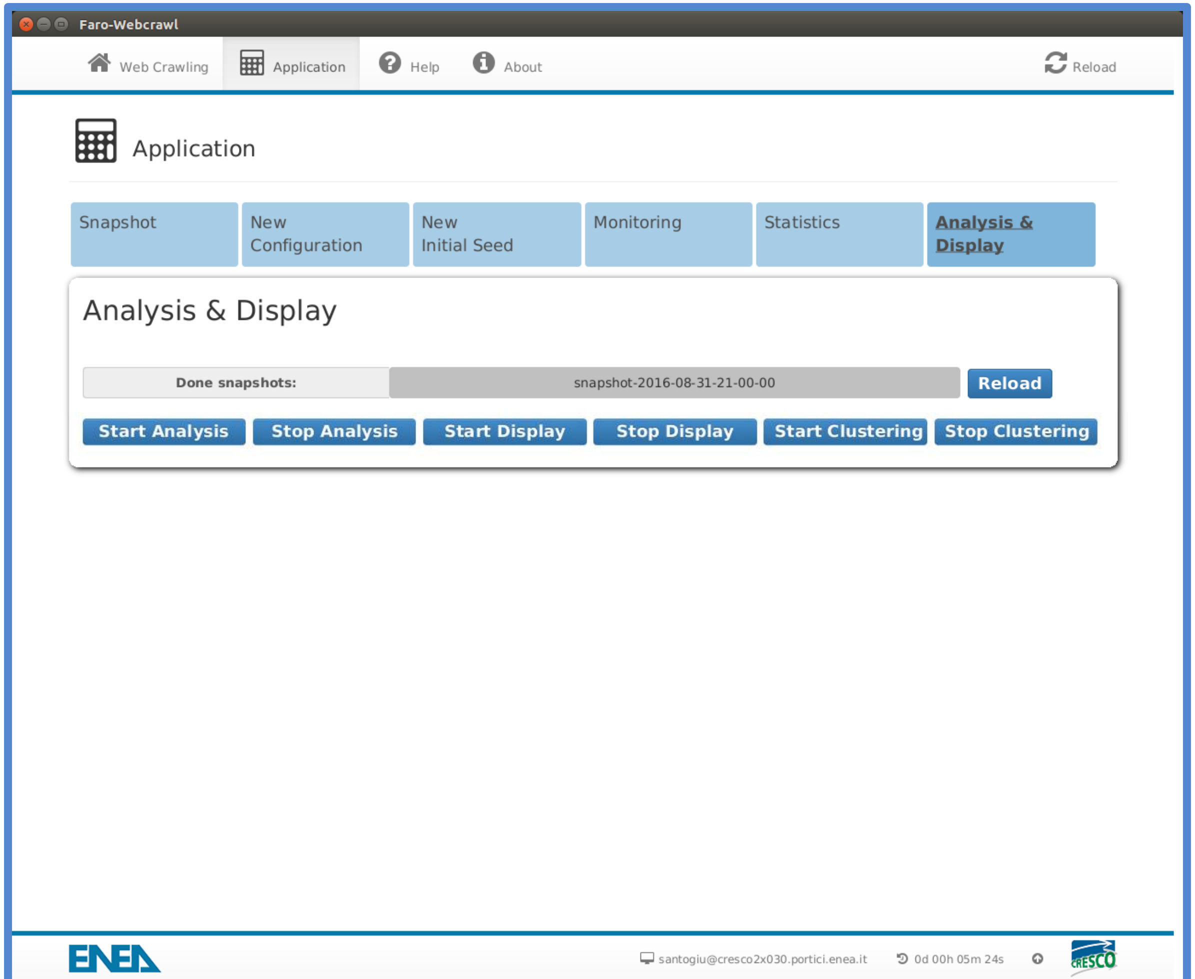
Name	Time	Store	Resource
Agent01	01h:01m:05s	29.87 GB	693994 Pages
Agent02	01h:02m:16s	29.57 GB	664591 Pages
Agent03	01h:02m:22s	31.22 GB	684284 Pages
Agent04	01h:02m:16s	30.35 GB	684152 Pages
Agent05	01h:02m:21s	29.94 GB	680600 Pages
Agent06	01h:02m:34s	32.03 GB	703445 Pages
Agent07	01h:01m:14s	28.84 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
TOTAL		482.02 GB	10893916 Pages
SPEED		1.00 Gb/s	2825.19 Pages/s

Tab per la visualizzazione delle statistiche.

Funzionalità:

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

Interfaccia grafica: **analysis & display**

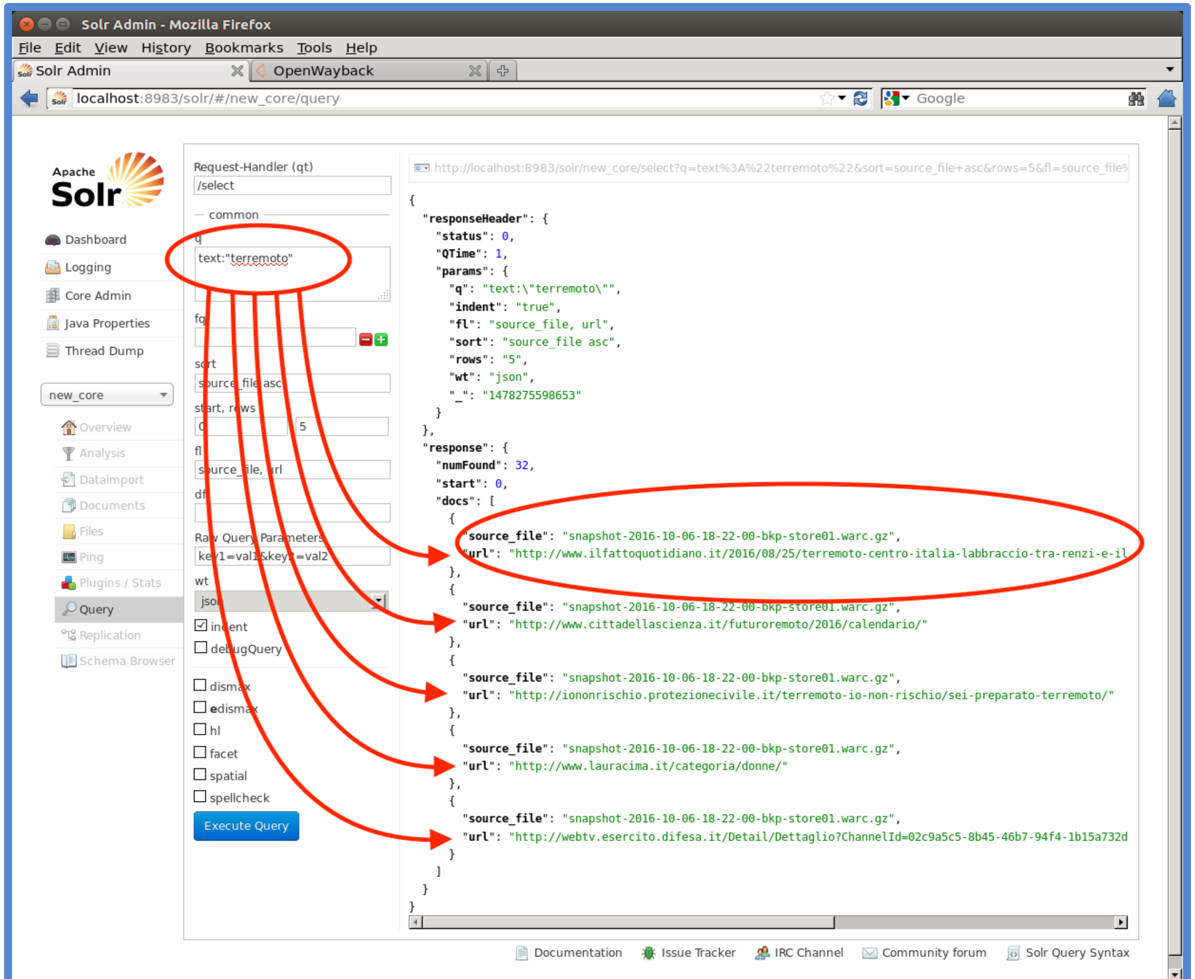


Tab per l'analisi, la visualizzazione e il clustering dei dati.

Funzionalità:

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

Analisi dei dati: ***Solr***



Schermata di *Solr.*

Funzionalità:

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

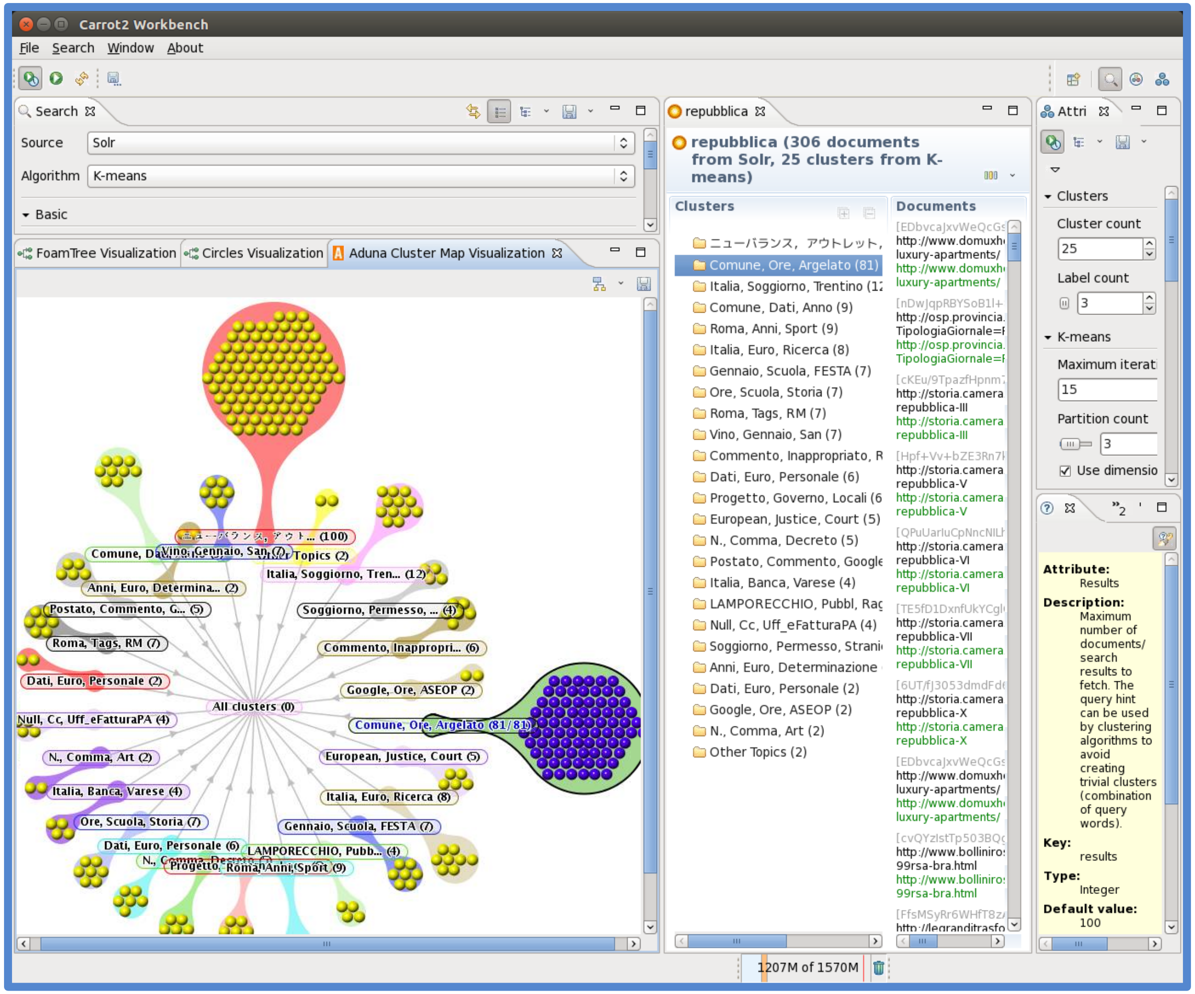
Display dei dati: *OpenWayback*

Schermata di *OpenWayback.*

Funzionalità:

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

Clustering dei dati: *Carrot2*



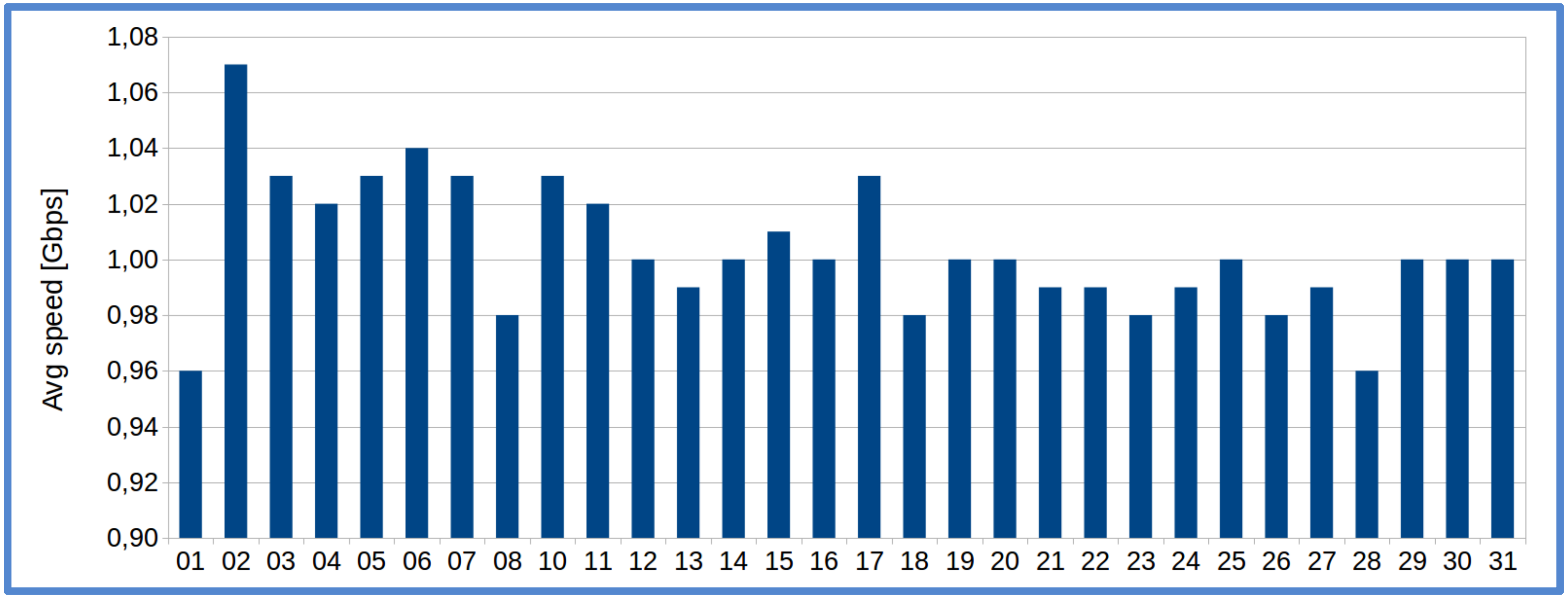
Schermata di *Carrot2.*

Funzionalità:

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

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 dowload 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

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.