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**
(UTICT-HPC, ENEA C.R. Portici)

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

6° Borsisti Day

24/03/2015

Roma – Consortium GARR

Borse di Studio
Orio Carlini
Consortium GARR

Borse di Studio Orio Carlini Consortium GARR

Obiettivi

Soluzioni innovative per tecniche di recupero dati dal web (*web crawling*) al fine di estrarre informazioni con strumenti avanzati (*data mining*) basati anche su aspetti semantici.

Gestione delle problematiche di archiviazione e fruizione di grandi moli di dati (*big data*) e strategie per rappresentazione di dati tipicamente non-strutturati e/o semi-strutturati (*text data).*

Aspetti chiave del progetto - aree di interesse:

- Dati provenienti dal web (web crawling e text data);
- Dati ad alta dimensionalità e numerosità (big data);
- Tecniche di analisi avanzata (data mining).

Borse di Studio Orio Carlini Consortium GARR

Aree di Ricerca

Dati provenienti dal Web

- Web Crawling
  - **Analisi dei contenuti in una rete in maniera sistematica e automatizzata.**
    - Esplorazione al fine di cercare contenuti/documenti da scaricare.
- Text Data
  - **Forma non strutturata** (doc, pdf, testi, ecc..);
  - **Forma semi-strutturata** (HTML, XML, JSON, ecc..).

Big Data

- **Dataset che richiedono strumenti non convenzionali per indicizzare, gestire e processare informazioni entro un tempo ragionevole.**
  - Numerosità dei dati;
  - Sorgenti di dati distribuite.

Data Mining

- **Task del Knowledge Discovery in Databases (KDD) process.**
  - Estrazione di pattern in maniera non supervisionata;
  - Utilizzo di analisi statistica non standard;
  - Tecniche ottimizzate di clustering con identificazione dei pattern.

Tempistiche e attività

Il progetto proposto richiederà 1 anno per la sua realizzazione e si articolerà in 4 fasi salienti.

1) Fase preparatoria [~2 mesi]:
- Studio dei prodotti software per il web crawling e individuazione delle metodologie da impiegare;
- Studio dell'infrastruttura hardware ed individuazione del tipo e della quantità delle risorse fisiche da impiegare nell'attività.

in corso

2) Fase di definizione [~4 mesi]:
- Definizione degli algoritmi di data mining per l'analisi dei dati.

3) Fase di realizzazione [~4 mesi]:
- Implementazione e produzione delle soluzioni scelte.

4) Fase finale [~2 mesi]:
- Installazione, test e collaudo;
- Produzione della documentazione tecnica e manualistica.

Fase preparatoria

Strumenti e metodologie per il web crawling

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;
- Crawling puro (download pagine web);
- Crawling di nuova generazione (download + parsing + strutturazione + preanalisi)

Problematiche

Denial of Service:

- Rallentamento dell'attività di un web server causata da una ripetuta richiesta di pagine oppure dall'esaurimento delle risorse di banda della rete.
- Può essere di due tipi:
  - accidentale (spider trap, errata configurazione);
  - intenzionale (attacco hacker singolo o distribuito).
- Soluzioni:
  - riprogettazione sitemap, filtraggio dati in arrivo, limitazioni del traffico, sistemi per riconoscimento di intrusioni.

Normative

Privacy:

- I contenuti sul Web sono di dominio pubblico;
- Informazioni aggregate su larga scala e su molte pagine;
- Due scuole di pensiero:
  - Necessità di un consenso informato *(Lin & Loui, 1998)*;
  - Consenso informato non sufficiente *(Jones, 1994)*.

Normative

Copyright:

- Copia permanente di materiale protetto (apparentemente illegale);
- Molti motori di ricerca emulano l'attività di **Internet Archive**:
  - Rispetto del protocollo *Robots Exclusion Standard*;
  - Richiesta di rimozione dall'archivio.
- In Italia:
  - Legge n. 633 del 22 aprile 1941 (Protezione del diritto d'autore);
  - Tribunale di Milano, sentenza del 4 giugno 2013 (Viaggiare s.r.l. vs. Ryanair).

Prodotti: crawling puro

Heritrix

- Scritto in Java; ✔
- Rispetta le direttive di esclusione dei META robots tags; ✔
- Raccoglie materiale a un ritmo misurato e adattivo con bassa probabilità di
- influenzare la normale attività di un sito web; ✔
- L'interfaccia è accessibile usando un web browser; ✔
- E' scalabile ma non dinamicamente scalabile. X

Nutch

- Basato su *Lucena* e *Java;* ✔
- Codificato interamente in Java, ma i dati vengono scritti in formati indipendenti
- dal linguaggio; ✔
- Architettura altamente modulare, che consente agli sviluppatori di creare plug-in
- per media-type parsing, data retrieval, querying and clustering. ✔
- Accessibile da terminale. X

Prodotti: crawling puro

Crawler4j

- Scritto in JAVA; ✔
- Permette di personalizzare e creare un proprio crawler sulla base di
- funzioni e librerie già sviluppate; ✔
  - Due funzioni principali possono essere sovrascritte:
    - *ShouldVisit:* decide se un URL deve essere visitato;
    - *Visit:* raccoglie i dati sull'URL visitato.
- Si può specificare il seme ed il numero di thread concorrenti. ✔
- Accessibile da terminale. X
- Ultima release: marzo 2013. X

Prodotti: crawling avanzato

Scrapy

- Scritto in Python; ✔
- Si può usare per estrarre dati usando API; ✔
- Restituisce output semistrutturati (JSON, XML, CSV,...); ✔
- Accessibile da terminale. X

OpenWebSpider

- Supporta il multi-threading; ✔
- Ha funzioni di search engine; ✔
- Accessibile da web browser; ✔
- Last Update: 2015-01-24. ✔

Prodotti: crawling avanzato

OpenSearchServer

- Esegue varie funzioni:
  - **Crawling:** recupero dei dati in base alle regole che sono state fornite; ✔
  - **Parsing:** estrazione dei dati da indicizzare (full-text) da ciò che è stato
  - scaricato; ✔
  - **Analisi:** applicazione di regole semantiche e linguistiche ai dati indicizzati; ✔
  - **Classificazione:** aggiunta di informazioni esterne ai documenti indicizzati; ✔
  - **Learning:** parsing dei documenti per dedurne la categoria di appartenenza. ✔

- Presenta quattro tipologie di crawling: ✔
  - *Su Web;*
  - *Su file system;*
  - *Su database;*
  - *Su file XML.*

Fase preparatoria

Infrastruttura hardware

Iniziato lo studio su infrastrutture di ENEA-GRID:

- Termini e condizioni di accesso a CRESCO;
- Configurazione di lancio;
- Sottomissione di job;
- job multi-case,
- job parallelo;
- …

E' in corso l'individuazione dei nodi fisici e dell'architettura da utilizzare.

Borse di Studio Orio Carlini Consortium GARR

Fine

Grazie per l'attenzione.

giuseppe.santomauro@enea.it