UNIVERSITÀ DEGLI STUDI GUGLIELMO MARCONI
FACOLTÀ DI SCIENZE E TECNOLOGIE APPLICATE

CORSO DI LAUREA MAGISTRALE IN
INGEGNERIA INFORMATICA

Estrazione di informazione strutturata dai Big Data guidata da un'ontologia di riferimento

Relatore
Prof.ssa Maria Teresa Pazienza

Candidato
Francesco Pagliarulo

Matricola: STA07117/LM32

ANNO ACCADEMICO
2017/2018

Abstract

Il contesto in cui si inserisce il progetto di tesi è l'estrazione di informazione strutturata dai Big Data. Per Big Data si intende l'analisi di quantità incredibilmente grandi di informazioni che si presentano in varie forme. Richiedono tecnologie e metodi analitici specifici che possono portare all'estrazione di valori di interesse.
Nello specifico del progetto, l'estrazione di informazione dai Big Data sarà guidata da un'ontologia di riferimento. Si procederà ad individuare ed estrarre informazioni provenienti da un social network (Facebook) ed appartenenti ad un gruppo di soggetti target inclusi come istanze dell'ontologia di riferimento.
I soggetti target sono nel nostro caso profili ricavati da record aziendali che hanno come caratteristiche (Nome, Cognome, Ruolo in azienda, Luogo di residenza e di lavoro). L'ontologia di riferimento metterà a disposizione una serie di questi profili da cui partire.
L'estrazione di queste informazioni viene svolta da due sottosistemi sviluppati utilizzando Node.js (un runtime javascript basato su eventi asincroni). Uno consiste in uno *script di ricerca*, che attraverso le api graph di Facebook, cerca i profili degli utenti a partire da Nome e Cognome del profilo target ricavati dall'ontologia di riferimento.
Il secondo è costituito da un *crawler* che si occupa di analizzare le pagine Facebook dei profili trovati e di estrarre da queste, tutte le informazioni pubbliche degli utenti, inclusi la lista di amici dell'utente stesso. Queste informazioni vengono poi salvate in un unico file per ogni profilo target dell'ontologia di riferimento.
La parte centrale del progetto è la realizzazione di un *annotatore UIMA*. UIMA, Unstructured Information

Management Architecture, comprende sistemi software in grado di analizzare grandi volumi di informazioni non strutturate al fine di scoprire le conoscenze rilevanti per un utente finale. Nel nostro caso, partendo dall'antologia descritta in precedenza, l'annotatore sarà in grado di registrare ulteriori informazioni provenienti dai profili individuati nel social network di riferimento. In particolare l'annotatore analizzerà tutti i documenti estratti dal crawler ed annoterà le informazioni di interesse.
Le problematiche affrontate sono legate:

- ai possibili casi di omonimia (più profili appartenenti a soggetti diversi potrebbero appartenere ad uno dei profili target),
- alla mancanza di informazioni (profili con informazioni di base insufficienti),
- alla presenza di informazioni errate o non aggiornate.

Le annotazioni oltre ad individuare le informazioni di interesse indicheranno anche il livello di confidenza con il quale possiamo ritenere un profilo collegato ad un soggetto target. Tra le annotazioni sarà quindi presente una che indica il livello di confidenza (un valore compreso tra [0,1]) con il quale si ritiene che il profilo trovato possa realmente appartenere al soggetto cercato.
Tale livello di confidenza è calcolato da un algoritmo a partire dalle annotazioni dei profili ricevuti e quindi dalle features (proprietà) di tali annotazioni.
Per il calcolo del livello di confidenza, verranno forniti all'algoritmo le ontologie di riferimento da confrontare con le annotazioni estratte dall'annotatore UIMA, a partire da questi l'algoritmo confronta le features estratte dall'annotatore con quelle presenti nell'ontologia di riferimento. Considerando un peso diverso per ogni feature, in base all'importanza che la feature possa avere nel

determinare l’appartenenza di un profilo estratto al profilo target (ad esempio una corrispondenza della feature ‘luogo di lavoro’ potrebbe avere un peso maggiore e quindi prevalere sulla feature ‘luogo di residenza’), l’algoritmo, calcola il livello di confidenza dell’annotazione con il profilo dell’antologia di riferimento.
Infine l’applicazione restituisce come risultato per ogni utente dell’ontologia di riferimento una lista di profili Facebook che più si avvicinano a quello del profilo target, ordinati in base al livello di confidenza più alto riscontrato.

Bibliografia

A. REZZANI, *Big data. Architettura, tecnologie e metodi per l'utilizzo di grandi basi di dati*, Apogeo Education, 1 edizione (20 novembre 2013), pp. 19-24, pp. 42-43, pp. 61-64

K. JANOWICZ, F.VANHARMELEN, J. A. HENDLER, P. HITZLER, *Why the Data TrainNeedsSemanticRails*, SPRING 2015, pp. 5-13

D. PEZZATINI, *Introduzione a Node.js*, Server Side Programming / MMM 2012

W3C OWL Working Group, *OWL 2 Web Ontology Language DocumentOverview (Seconda edizione),* W3C dell'11 dicembre 2012

ALGARNI, ABDULLAH, XU, YUE & CHAN, *Susceptibility to social networking sites: The case of Facebook*, December 2015

Olga Peled, Michael Fire, Lior Rokach, and Yuval Elovici, *Matching Entities Across Online Social Networks*, Novembre 2014.