Università degli Studi di Milano
Facoltà di Scienze Matematiche Fisiche e Naturali
Corso di Laurea in Informatica per la Comunicazione

PROFILE MATCHING IN ONLINE SOCIAL NETWORKS

Relatore: Prof. Sabrina GAITO

Correlatore: Dr. Matteo Zignani

Tesi di:
Mattia Dimauro
Matricola: 808902

Anno Accademico 2013-2014

dedicato a ...

Prefazione

Prefazione lalala.

Organizzazione della tesi

La tesi è organizzata come segue:

- nel Capitolo 1 ....

Ringraziamenti

Ringrazio, grazie.

Indice

Capitolo 1

Introduzione

Ciao volevo bla bla lba

Capitolo 2

Pattern comportamentali e costruzione di features

Gli individui mostrano spesso pattern comportamentali nella scelta dei loro usernames. Questi pattern, risultanti in ridondanza di informazione, possono essere utili per identificare individui su diversi social networks. I soggetti potrebbero evitare queste ridondanze selezionando usernames su siti diversi in modo che risultino completamente diversi dai loro altri usernames. Cosi presentati gli usernames risultano essere cosi differenti che dato uno username, nessuna informazione riguardante gli altri username può essere estratta. Idealmente per raggiungere questo stato di indipendenza, l'individuo dovrebbe scegliere uno username che presenti entropia massima. Ovvero uno username composto da una lunga sequenza di caratteri, lunga quanto il massimo consentito dal sistema, senza ridondanze - una sequenza di caratteri completamente casuale. Sfortunatamente o no, tutti questi requisiti non vengono incontro alle abilità umane. Gli umani hanno difficoltà a memorizzare lunghe sequenze, con la possibilità della memoria a breve termine di ricordare 7 ± 2 elementi. Queste limitazioni risultano condurre gli individui a selezionare generlamente usernames *non lunghi*, *non casuali* e che presentano *abbondante ridondanza*. Queste proprietà possono essere colte utilizzando features specifiche.

2.1 Patterns due to human limitation

Exogenous factors

2.2 Exogenous factors

Capitolo 3

Collezione dataset

Esistono più modalità attraverso le quali è possibile collezionare dati riguardanti stessi individui presenti su diversi social networks. Un modo semplice consisterebbe nel organizzare un questionario dove si richiede agli utenti di elencare i propri profili. Questo metodo permette di raccogliere una quantità di dati spesso limitata. Esistono compagnie, alcune risultano essere gli stessi social networks, che richiedono queste informazioni ai propri utenti ma queste non sono disponibili pubblicamente. Fortunatamente esistono servizi di *social network aggregation* che permettono di collezionare contenuti da diversi *social network services* in una presentazione unificata. Il compito svolto da un *social network aggregator*, che raggruppa insieme informazioni in un singolo luogo, supporta i propri utenti a riunire molteplici profili di social network in un singolo profilo e aiuta a tenere traccia delle attività che avvengono sui diversi profili semplificando la *social networking experience* dell'utente. La maggior parte di questi servizi non permette l'accesso pubblico alle informazioni aggregate, sono per lo più uno strumento per l'utilizzatore per seguire i propri profili, permettendo di avere tutte le notifiche relative in unico luogo, o di pubblicare lo stesso contenuto in più profili in una volta sola.Questa tipologia di aggregatori non sono frutto di interesse per la collezione di informazioni che cerchiamo. Prenderemo in considerazione invece quegli aggregatori che permettono di condividere queste informazioni aggregate con altre persone. Attraverso questi è possibile visualizzare tutte le attività di un utente sui diversi social network che questo a deciso di far seguire all'aggregatore. Ovviamene è possibile, ed è l'informazione che andremo a estrarre, risalire su quale social network è stata eseguita tale azione e avere cosi un elenco di profili di social network services appartenenti alla stessa persona.

3.1 Scraping Alternion

Alternion - All your social web and email in one place - è un aggregatore i cui servizi e features lo fanno rientrare tra la seconda classe di *social network aggregator* descritta. Attraverso *Alternion* è in fatti possibile per un utente condividere con altri contatti le informazioni riguardanti i propri profili di diversi online social networks. Il numero di utenti iscritti al servizio non è noto. Alternion dichiara di permettere ai propri utenti di aggregare un grosso numero di online social networks, dai più popolari come *Facebook*, *Twitter*, *Google+*, *LinkedIn*, *Flickr* fino a più di 220 online social networks. Di seguito spiegheremo l'approccio utilizzato al fine di ottenere i profili degli utenti iscritti al servizio e di recuperare da questi le informazioni riguardati i profili dei loro social networks. Alternion non dispone di un servizio per recuperare i dati attraverso un'interfaccia *web (API)*, con richieste e risposte documentate. Dovremo dunque procedere analizzando le pagine web restituite svolgendo un'attività conosciuta come *web scraping*. Il *web scraping* o *web data extraction* è una tecnica di estrazioni di dati da pagine web. L'estrazione di dati viene automatizzata attraverso un software che simula un utente umano nell'esplorazione di documenti presenti sulla rete internet. Il software deve quindi implementare il protocollo HTTP, fondamento per la comunicazione di dati per il World Wide Web per recuperare il documento attraverso la rete internet. Questi documenti, tipicamente descritti attraverso un linguaggio di markup, sono pensati per rappresentare un'interfaccia grafica per l'utente. Ottenuto il documento il *web scraper* si occupa di analizzarne i dati strutturati ricercandone quelli di interesse. La pagina non viene interpretata visivamente ma esaminandone il contenuto descritto con il linguaggio di markup, tipicamente HTML. Questa ricerca può essere effettuata con diversi approcci: dal più semplice, seppur potente, *text grepping* combinato con *regular expression matching* o può prevedere un'analisi più strutturata della pagina attraverso una tecnica di *DOM parsing*. Per i nostri scopi utilizzeremo entrambe queste tecniche. Esistono altre tecniche e metodologie per eseguire *web scraping*, ma i dettagli esulano dallo scopo di questa tesi.

3.1.1 Ottenere i profili

Siamo interessanti a ottenere un considerevole numero di profili di utenti che utilizzano l'aggregatore in analisi. Alternion non prevede una funzionalità per mostrare l'elenco completo dei profili iscritti al proprio servizio. Permette invece di reperirne un sotto insieme di questo attraverso una funzionalità di ricerca, divisibile in due classi: la ricerca parametrizzabile secondo alcuni criteri o la presentazione casuale di profili. La prima permette di interrogare il sistema per estrarne i profili corrispondenti ad alcuni parametri di ricerca come Nome, Sesso, Età, Paese di provenienza, Educazione, Interessi, etc... La seconda funzionalità consente di ottenere ad ogni richiesta un

profilo selezionato casualmente.[1] Ho deciso di accantonare questa seconda via per due ragioni. La selezione casuale potrebbe potenzialmente portare a richiedere più volte lo stesso profilo, dipendentemente dalla bontà (che non è stata testata) della casualità con cui il profilo viene estratto. Vogliamo inoltre limitare l'estrazione di profili non appartenenti a persone fisiche: alcuni profili presenti fanno infatti riferimento a prodotti o società e aziende. Concludo di optare per la ricerca parametrizzata, in particolare, la ricerca per nome. Come lista di parametri per l'interrogazione del sistema useremo una lista di nomi propri [2], formata da 4275 nomi femminili e 1219 maschili. Utilizzando strumenti messi a disposizione da Google[3] per analizzanre i pacchetti HTTP scambiati tra il server di Alternion e il client Google Chrome, viene identificata la richiesta HTTP da eseguire per interrogare il server per farsi restituire la pagina web contenente la lista di utenti corrispondenti al parametro di ricerca. Eseguiremo quindi una richiesta per ogni nome presente nelle nostre liste di nomi. Da questa, tramite tecniche di scraping descritte precedentemente, estrapoliamo per ogni utente l'*URL* idendificativo della risorsa. Una volta collezionati tutti gli URL, dove ogni indirizzo corrisponde a un utente di Alternion, potremo recuperare la pagina profilo di questi utenti e cercare al suo interno le informazioni riguardanti i loro social networks.

MONGODB PER STORE

3.1.2 Profili recuperati

Sono stati recuperati 15341 profili di Alternion, di cui 11274 presentano almeno due profili di social networks. Il numero di profili non è ingente, in quanto la funzionalità ricerca permette di recuperare un massimo di 30 profili per richiesta, ad esempio solo trenta profili delle persone che si chiamano 'Anna'. Ad ogni modo si è notato che è possibile scalare, se non verticalmente per numero di users, orizzontalmente per numero di profili per users. Ogni utente ha in media ∼4,6 OSNS collegati, su un totale di 168 online social network services diversi presenti. Effettuando una combinazione a 2 elementi sull'insieme dei social networks, per ogni profilo, possiamo espandere il numero di coppie di usernames. Ad esempio, se di un utente U abbiamo 4 profili di 4 social network diversi, con uno username u per profilo, poniamo u-Alternion, u-Facebook, u-Twitter, U-Instagram, una combinazione a 2 elementi ci permette di avere le seguenti coppie di usernames: (u-Alternion, u-Facebook), (u-Alternion, u-Twitter), (u-Alternion-uInstagram), (u-Facebook,u-Twitter), (u-Facebook,u-Instagram), (u-Twitter,u-Instagram)

[1] Più probabilmente, pseudo-casualmente!

[2] Lista reperita da census.gov

[3] Google Chrome DevTools

OSNS presenti Lista comprensiva 168 social network.

DATI OTTENUTI - QUANTITATIVAMENTE utenti vs n social networks PLOT GRAFI DISTANZA NAIVE

3.2 Duolingo

3.3 Campagna Lays e Twitter

Capitolo 4

Risultati

Performance classificatore.

Bibliografia

[1] M. Gotti, I linguaggi specialistici, Firenze, La Nuova Italia, 1991.

[2] R. Wellek, A. Warren, Theory of Literature , 3rd edition, New York, Harcourt, 1962.

[3] A. Canziani et al., Come comunica il teatro: dal testo alla scena. Milano, Il Formichiere, 1978.

[4] Ministry of Defence, Great Britain, Author and Subject Catalogues of the Naval Library, London, Ministry of Defence, HMSO, 1967.

[5] H. Heine, Pensieri e ghiribizzi. A cura di A. Meozzi. Lanciano, Carabba, 1923.

[6] L. Basso, "Capitalismo monopolistico e strategia operaia", Problemi del socialismo, vol. 8, n. 5, pp. 585-612, 1962.

[7] L. Avirovic, J. Dodds (a cura di), Atti del Convegno internazionale "Umberto Eco, Claudio Magris. Autori e traduttori a confronto" (Trieste, 27-28 novembre 1989), Udine, Campanotto, 1993.

[8] E.L. Gans, "The Discovery of Illusion: Flaubert's Early Works, 1835-1837", unpublished Ph.D. Dissertation, Johns Hopkins University, 1967.

[9] R. Harrison, Bibliography of planned languages (excluding Esperanto). `http://www.vor.nu/langlab/bibliog.html`, 1992, agg. 1997.