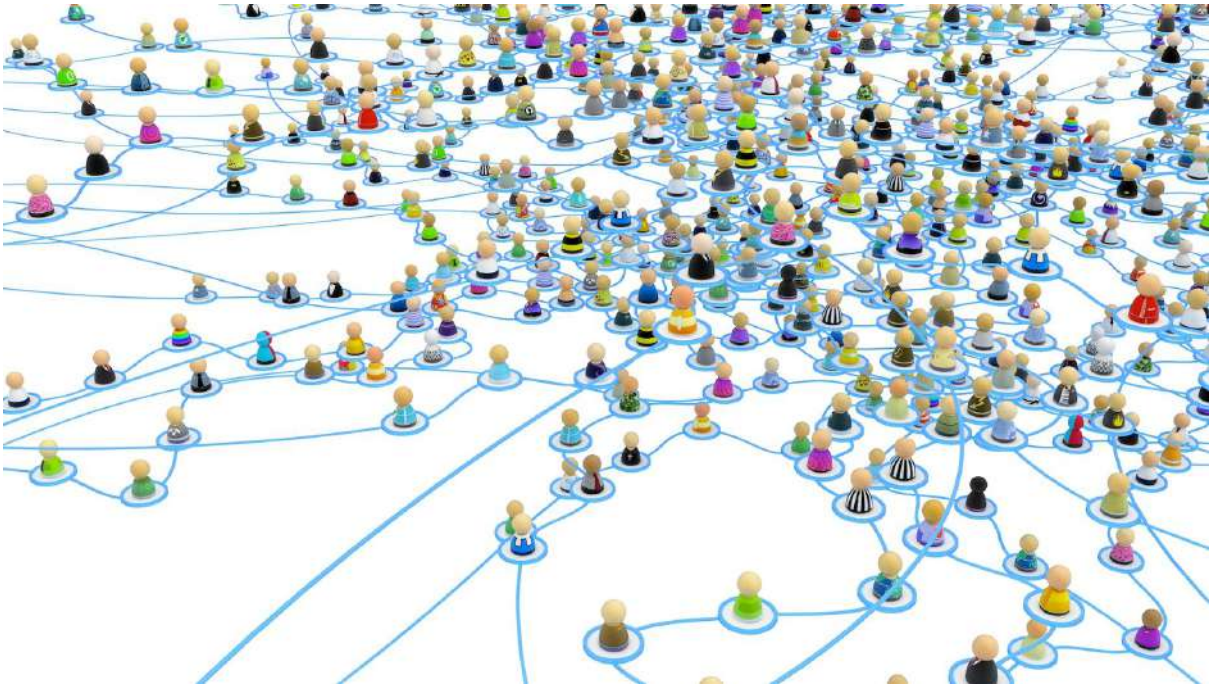
Alumni Mathematica



Instant Trend Detection Tool

come si costruisce un tool per individuare i trend della rete partendo dai social network

Algoritmo, API e sviluppi

anteprima

si ringrazia per il supporto

Autori
Dr Stefano Franco
Dr Pierpaolo Basile

Quali sono i trend della rete e come individuarli?

È questa la domanda a cui abbiamo cercato di rispondere. Il presente lavoro è stata la base per la realizzazione dell'Instant Trend Detection Tool, uno strumento che analizza i dati dei social network e calcola i trend della rete individuando gli hashtag e attribuendo a essi un punteggio a seconda della previsione del loro successo nei successivi sei giorni.

L'Instant Trend Detection Tool risulta quindi uno strumento altamente innovativo e particolarmente indicato per i social media manager che potranno in questo modo trovare i migliori hashtag da utilizzare nell'ottica del real-time marketing. Inserendo una parola chiave o, semplicemente, selezionando uno o più social network, l'algoritmo applica tecniche predittive e di analisi dati che danno una lista di hashtag e uno score a seconda della futuribilità attesa nei successivi sei giorni. L'Instant Trend Detection Tool non è, dunque, un semplice strumento che permette di visualizzare i dati dei social network. L'Instant Trend Detection Tool analizza tali informazioni e ne estrae valore, attraverso l'applicazione di tecniche algoritmiche e modelli matematici e statistici. Uno strumento che, in questo senso, risulta molto interessante anche per i blogger e, più in generale, per gli utenti dei social network che potranno in questo modo individuare le tendenze della rete e utilizzare gli hashtag più idonei per i propri aggiornamenti.

All'interno del paper abbiamo analizzato e riportato tutte le analisi e lo studio effettuato per la realizzazione dell'Instant Trend Detection Tool. Il lavoro è suddiviso in due parti:

- nella prima parte vengono analizzati i social network maggiormente diffusi e le banche date particolarmente utili per l'individuazione dei trend emergenti. Tali risorse vengono suddivise in due categoria (social principali - Facebook, Twitter, Instagram - e social minori) a seconda del peso dato a ciascuno di essi per l'individuazione dei trend all'interno dell'algoritmo dell'Instant Trend Detection Tool. Per ogni social network (e banca dati) sono descritte le principali funzionalità e vengono passate in rassegna le API e gli strumenti per gli sviluppatori messi a disposizione dalla piattaforma;

- nella seconda parte vengono analizzate le caratteristiche tecniche dei modelli predittivi attualmente utilizzati per l'individuazione dei trend. Vengono passate in rassegna le tecniche e i modelli matematici per la trend detection e, infine, viene disegnata una architettura tipica per un modello di event detection. Viene effettuata un'analisi anche su quali sono i gap delle attuali tecniche di trend detection rispetto alle esigenze del mercato e, una volta individuate, vengono proposti degli approcci che sono poi diventati il *technical core* dell'Instant Trend Detection Tool.

Il presente lavoro può anche essere interpretato come un interessante tentativo di utilizzare le ultime novità in ambito tecnologico e di ricerca con le esigenze reali del mercato. Oggigiorno sempre più aziende sono presenti sul web e le stime individuano nel mercato digitale uno dei principale attori dei PIL nazionali. Il web diventa quindi croce e delizia di tali aziende che devono necessariamente approfondire le dinamiche dello stesso per una promozione adeguata della propria attività. Web Marketing, Social Media Marketing, Digital Marketing diventano così non più termini

astratti ma canali attraverso i quali poter presentare al meglio i propri prodotti e servizi e penetrare nuovi mercati.

A livello consulenziale, il paper rappresenta, inoltre, l'attività degli autori per la validazione dell'esigenza di un ipotetico cliente nella realizzazione di una attività di ricerca in outsourcing. All'interno del paper vengono, infatti, evidenziate le attività da svolgere per la costruzione di un prodotto tecnologico altamente innovativo, partendo da un'analisi tecnologica dei mezzi da utilizzare, passando per le tecniche analitiche e scientifiche e lo stato dell'arte nel dominio di intervento, giungendo, infine, alla prototipazione del tool da implementare.

Un approccio altamente performante, che delinea la linea operativa utilizzata da Alumni Mathematica per rispondere alle esigenze dei propri clienti e svolta in collaborazione con partner industriale altamente specializzati (in questo caso **Pigreek**[1] e **QuestionCube**[2]).

Tutti i proventi derivanti dalla distribuzione del presente documento saranno donati ad Alumni Mathematica ed utilizzati per migliorare le performance dell'Instant Trend Detection Tool.

Gli Autori

[1] http://www.pigreek.com/
[2] https://www.questioncube.com/

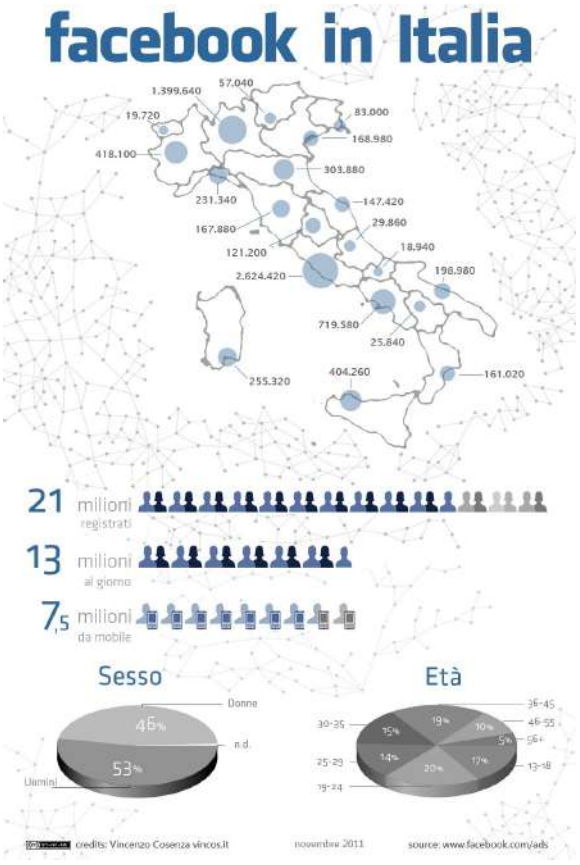
Api Social Network

Facebook

Introduzione

Facebook è il social network più diffuso nel pianeta. Nato nel 2004 con l'obiettivo di rendere il mondo più aperto e connesso, ad oggi raggruppa 1,71 miliardi di utenti attivi mensilmente con un tasso di crescita annuale pari al 15%. Ogni giorno 1,13 miliardi di utenti si scambiano contenuti, 1,03 miliardi dei quali lo fa da attività su dispositivi mobile. Approssimativamente l'84,5% degli utenti attivi giornalmente risiede fuori dai confini americani [3]

Per quanto riguarda l'Italia, in un giorno mediamente sono 20 milioni gli italiani che popolano il Social Network (erano 17 milioni ad agosto 2013 e 15 milioni ad aprile 2013[4]). In questo caso la penetrazione sale al 92%, considerando che gli utenti in collegamento ogni giorno sono 21,7 milioni (fonte Audiweb). Nella cornice sulla destra è possibile avere una panoramica della suddivisione degli utenti in Italia per fascia d'età e sesso.



Introduzione alle Api

Facebook rilascia una piattaforma per costruire applicazioni che poi diventeranno disponibili agli utenti del Social Network. Le API permettono alle applicazioni di sfruttare le connessioni sociali degli utenti e le informazioni del profilo, nonché gli interessi, il feed della home e le pagine seguite, sempre compatibilmente ai permessi sulla privacy che ogni utente concede. Tramite le API Facebook è possibile realizzare contenuti e immagazzinare informazioni disponibili nei profili personali, nelle pagine, nei gruppi, negli eventi, nelle amicizie e nei file multimediali condivisi sulla piattaforma.

Le API di Facebook utilizzano il *protocollo REST* [5] e le risposte sono in *formato JSON*[6].
La piattaforma rilasciata si chiama *Facebook for Developer* e deve diventare il riferimento assoluto per ogni programmatore che vuole interfacciarsi con le API di Facebook.

Dal 2007 Facebook ha introdotto il **Social Graph**[7]. Come presentato dal manager Facebook Dave Morin nel 2007, il Social Graph "è la rete di legami (connessioni) che esistono nel modo in cui ogni soggetto (nodo) comunica e si scambia informazioni con gli altri nodi". In buona sostanza, tutto ciò che è dentro Facebook (utenti, contenuti multimediali, pagine...) è inserito all'interno di un grosso *grafo.*

[3] https://newsroom.fb.com/company-info/
[4] http://vincos.it/2013/06/01/state-of-the-net-2013-parte-1-lo-scenario-italiano/
[5] il paradigma REST (REpresentational State Transfer) è basato su un protocollo di comunicazione stateless, client-server, cacheable e scalabile, tipicamente HTTP (ma non solo)
[6] JavaScript Object Notation, è un formato adatto all'interscambio di dati fra applicazioni *client-server*
[7] https://developers.facebook.com/docs/graph-api/

In matematica, un grafo è la configurazione formata da un insieme di punti (*vertici* o *nodi*) e un insieme di linee (*archi*) che uniscono coppie di nodi. In pratica, tramite questa nozione adattata a Facebook, è possibile percorrere il Social Graph riuscendo a passare da un oggetto all'altro tramite le connessioni che ciascuno di questi oggetti possiede. Sintetizzando, il Social Graph è una rappresentazione delle informazioni presenti su Facebook composta da:

- *Nodi*: elementi come utente, foto, pagina o commento;
- *Segmenti*: le connessione tra i Nodi;
- *Campi*: informazione sugli elementi dei Nodi.

Anche le API di Facebook sono all'interno del Social Graph e, in particolare, lo strumento che Facebook mette a disposizione degli sviluppatori per facilitare nell'utilizzo delle API è l'**API Graph**. Si tratta del metodo principale tramite cui le app Facebook possono leggere e scrivere nel Social Graph. Si tratta di un'API di basso livello basata su *HTTP* che è possibile usare per richiedere dati, pubblicare notizie, gestire inserzioni, caricare foto e per altre operazioni necessarie alle app Facebook. Nella maggior parte delle richieste API Graph, è necessario usare i token d'accesso, che l'app può generare implementando Facebook Login.

Facebook mette a disposizione degli sviluppatori una guida molto dettagliata in cui vengono illustrate tutte le modalità con cui utilizzare l'API Graph[8]. L'API Graph è uno strumento gratuito.

Oltre all'API Graph, Facebook rilascia altre due modalità di API, purtroppo entrambe per il momento non disponibili:

- *Api Public Feed*[9]*:* offre notizie sugli aggiornamenti di stato di un utente o di una Pagina nel momento in cui sono pubblicati su Facebook. Sono inclusi solo gli aggiornamenti di stato con privacy configurata su "Tutti". Le notizie non sono trasmesse tramite un endpoint API HTTP, mentre gli aggiornamenti sono inviati al server tramite una connessione *HTTPS* dedicata. Inoltre, le notizie includono solo i dati principali relativi a un determinato post;
- *Chat Api*[10]*:* dismesso dal 30 Aprile 2014.

Rassegna delle API

Facebook rilascia 214 API divise in 17 categorie. Una categoria è una macro area che contiene API che appartengono alla stessa tipologia di oggetti nel Social Graph. Le categorie sono[11]: Search (1), Album (9), Application (28), Checkin (6), Comment (5), Event (14), Friendlist (14), Group (8), Link (6), Note (6), Page (34), Photo (10), Post (6), Status (6), User (53), Notification (1), Video (7).

Le API rilasciate da Facebook sono di 3 tipi: **GET**, **POST** e **DELETE** a seconda che la funzione sia di leggere, scrivere, cancellare i contenuti multimediali presenti sul social. Negli scopi del nostro lavoro abbiamo utilizzato le API di tipo GET e solo in un caso l'API di tipo POST.

Passiamo in rassegna le API selezionate, indicando la categoria di appartenenza e la funzione, evidenziandone le potenzialità per le finalità del nostro lavoro.

- **Search**

cerca tutti gli oggetti pubblici nel social graph di Facebook

Senza dubbio l'API più importante di tutte. Permette di navigare nell'API Graph di Facebook.

- **Checkin**[12]

rappresenta una singola visita di un utente in una location

[8] https://developers.facebook.com/docs/graph-api/using-graph-api/
[9] https://developers.facebook.com/docs/public_feed
[10] https://developers.facebook.com/docs/chat
[11] in parentesi il numero di API contenute in ogni categoria
[12] documentazione: https://developers.facebook.com/docs/graph-api/reference/v2.7/checkin

GET {checkin}/likes

Utenti a cui piace un certo checkin

API interessanti per riuscire ad individuare gli utenti che fanno checkin in alcuni luoghi e che di conseguenza sono interessati a ciò che succede in quei posti.

- **Event**[13]

GET {event}

Definisce tutte le informazioni circa un evento, inclusa la location, il nome dell'evento e quali invitati parteciperanno

GET {event}/feed

Tutti i post scritti sull'evento

GET {event}/invited

Tutti gli utenti che sono stati invitati in un dato evento

GET {event}/attending

Tutti gli utenti che parteciperanno ad un dato evento

Interessante per gli stessi motivi della categoria dei Checkin. Si possono individuare gli utenti che parteciperanno ad alcuni eventi. Si possono utilizzare queste informazioni per riuscire ad allenare l'algoritmo e capire gli interessi degli utenti.

- **FriendList**[14]

GET {friendlist}/members

Tutti gli utenti che sono membri di una certa lista

API che permette di categorizzare gruppi di utenti che appartengono ad una stessa lista e che quindi condividono gli stessi interessi. L'utilizzo dell'API si interfaccia con le impostazioni sulla privacy dei singoli utenti che appartengono alla categoria e quindi non tutte le informazioni risultano accessibili.

- **Group**[15]

GET {group}

Restituisce un gruppo Facebook

GET {group}/feed

Il feed di un dato gruppo

Tramite queste API possiamo capire i trend che ci sono in un determinato gruppo. Potrebbe essere interessante monitorare periodicamente dei gruppi di interesse e precedentemente individuati per riuscire a tracciare le novità.

GET {group}/members

Tutti i membri di un gruppo

Anche in questo caso è possibile riuscire a categorizzare gli utenti in base ai gruppi in cui sono presenti.

- **Link**[16]

GET {link}

Un link condiviso sulla bacheca di un utente

GET {link}/likes

Utenti a cui piace un certo link

Categoria di API molto utile per monitorare l'andamento di alcuni link di nostro interesse (magari quelli presenti nei nostri contest) per capire come sta evolvendo la rete e pensare a delle attività collaterali per aumentare l'engagement e le performance.

[13] documentazione: https://developers.facebook.com/docs/graph-api/reference/event
[14] documentazione: https://developers.facebook.com/docs/graph-api/reference/friend-list
[15] documentazione: https://developers.facebook.com/docs/graph-api/reference/v2.7/group
[16] documentazione: https://developers.facebook.com/docs/graph-api/reference/v2.7/link

- **Page**[17]

GET {page}

Restituisce una Pagina

GET {page}/feed

La bacheca di una Pagina

GET {page}/photos

Le foto contenute su una Pagina

GET {page}/statuses

Gli aggiornamenti di stato della Pagina

GET {page}/videos

I video contenuti su una Pagina

GET {page}/posts

Tutti i post pubblicati dalla Pagina

GET {page}/events

Gli eventi a cui la Pagina partecipa

GET {page}/checkins

Tutti i checkin fatti dall'utente corrente su una Pagina (se la Pagina è individuata come località) e dagli amici dell'utente corrente

Categoria di API molto popolosa con una serie di funzioni esclusive per la stessa. Come per la categoria Gruppi, anche qui può risultare significativo monitorare particolari pagine in modo da capire come evolve la rete, finalità tanto più efficace quanto più la pagina monitorata è quella di un *influencer*[18]. Dato il numero elevato di API interessanti per questa categoria, tale categoria si appresta ad essere un riferimento costante per chiunque intenda creare servizi basati su API Facebook.

- **Photo**[19]

GET {photo}

Restituisce una singola foto

GET {photo}/comments

Tutti i commenti su una foto

GET {photo}/likes

Utenti a cui piace una determinata foto

GET {photo}/tags

Tutti i tags presenti su una certa foto

Categoria interessante per realizzare un ipotetico sviluppo futuro come integrazione dell'*Instant Trend Detection Tool*. L'idea è quella di monitorare e fare degli analytics alle foto dei contest lanciati (sia quelle pubblicate dall'azienda che propone il contest, sia quelle pubblicate dall'utente che vi partecipa).

POST {photo}/tags

Crea un tag su una certa foto

La segnalazione dell'API di tipo POST è dovuta al fortissimo impatto che ha la parola *tags* sulle finalità del Tool. Si è ritenuto opportuno segnalare tutte le API (anche questa, che è l'unica di questo tipo riportata nel presente report) che riportano tale parola.

- **Post**[20]

GET {post}

Restituisce un post di Facebook

GET {post}/likes

Utenti a cui piace un certo post

Sono validi tutti gli stessi ragionamenti fatti per la categoria Photo.

[17] documentazione: https://developers.facebook.com/docs/graph-api/reference/page
[18] si definisce influencer un utente/pagina con un seguito molto maggiore della media capace di influenzare il proprio pubblico indirizzandolo su determinati temi di proprio interesse
[19] documentazione: https://developers.facebook.com/docs/graph-api/reference/photo
[20] documentazione: https://developers.facebook.com/docs/graph-api/reference/v2.7/post

- **Status**[21]

GET {status}

Restituisce uno stato di Facebook

GET {status}/likes

Gli utenti a cui piace un certo stato

Valgono i ragionamenti espressi per le categorie precedenti. Tramite la lettura degli stati degli utenti, individuati per interesse integrando le API della categoria FriendList, possiamo individuare i trend che stanno seguendo.

- **User**[22]

GET {user}

Restituisce il profilo di un utente

GET {user}/activities

Le attività elencate sul profilo di un utente

GET {user}/checkins

I luoghi in cui un utente ha effettuato un checkin

GET {user}/events

Gli eventi a cui questo utente sta partecipando

GET {user}/feed

La bacheca di questo utente

GET {user}/friends

La lista degli amici dell'utente

GET {user}/groups

I gruppi a cui l'utente appartiene

GET {user}/home

Gli aggiornamenti nella home dell'utente

GET {user}/inbox

Le discussioni dei messaggi dell'utente

GET {user}/interests

Gli interessi elencati nel profilo dell'utente

GET {user}/likes

Tutte le pagine a cui l'utente ha cliccato mi piace

GET {user}/links

I link postati dall'utente

GET {user}/posts

I post dell'utente

GET {user}/statuses

Lo stato dell'utente aggiornato

Questa categoria risulta molto importante in quanto ci sono tutte le API che permettono di capire l'attività degli utenti Facebook. Si potrebbe procedere in un duplice modo per sfruttare a pieno le potenzialità di questa categoria:

1. analizzare a campione le informazioni di alcuni utenti e trarre conclusioni in maniera statistica su quali sono i trend della rete;
2. analizzare le informazioni di un numero di utenti particolari segnalati a priori per la loro influenza, in modo da partire da questi per tutte le considerazioni inserite nell'algoritmo.

- **Video**[25]

GET {video}

Restituisce un singolo video

GET {video}/likes

Gli utenti a cui piace questo video

Restano valide tutte le considerazioni effettuate per la categoria Photo, Status e Post.

[21] documentazione: https://developers.facebook.com/docs/graph-api/reference/v2.7/status
[22] documentazione: https://developers.facebook.com/docs/graph-api/reference/user
[25] documentazione: https://developers.facebook.com/docs/graph-api/reference/video

Ulteriori risorse

Oltre alle API, che rimangono lo strumento principale con cui gli sviluppatori lavorano utilizzando le risorse e i dati di Facebook, è bene sapere che la società di Zuckerberg mette a disposizione degli sviluppatori diversi altri strumenti. In particolare, merita una nota la piattaforma *Facebook Open Source*[23]. Si tratta di una piattaforma rilasciata da *Facebook Engineering* dove è possibile trovare una serie di tool che agevolano nella scrittura del codice che si interfaccia con Facebook e che consente agli sviluppatori di integrare il proprio codice con le risorse Facebook interfacciandolo al meglio. Tali risorse sono differenziate per sistemi Android, sistemi IOS, sistemi Web, sistemi Hardware e sistemi Backend.

Tutti gli aggiornamenti per gli sviluppatori vengono sempre presentati ed inseriti sulle seguenti risorse che quindi è buona norma consultare periodicamente:

- https://code.facebook.com/
- https://research.facebook.com/
- https://github.com/facebook

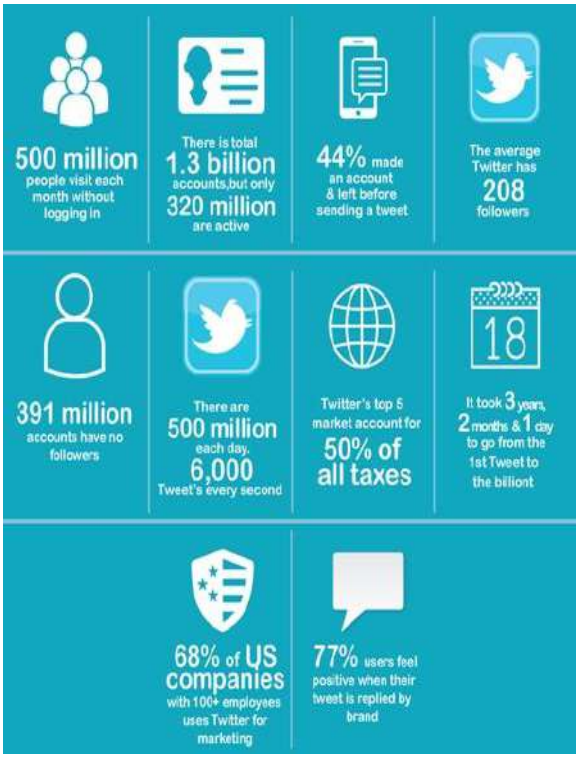
Concludiamo segnalando che il modo migliore per approfondire le logiche del Social Graph di Facebook è quello di documentarsi al meglio sul **protocollo Open Graph**[24], originalmente creato da Facebook che lo ha poi rilasciato in maniera *Open Source* e che quindi ha seguito uno sviluppo autonomo. Il Social Graph di Facebook rispetta il protocollo Open Graph.

Twitter

Introduzione

Twitter è un servizio gratuito di micro-blogging, spesso considerato un semplice social network, ma in realtà possiede delle peculiarità uniche nel settore che lo rendono esclusivo. Creato nel 2006 a San Francisco, su Twitter nasce l'hashtag "#", come simbolo utilizzato per etichettare post per evidenziare particolari parole chiave che si vogliono condividere con gli altri utenti. Attualmente sono 313 milioni gli utenti attivi ogni mese su questo social network, l'82% dei quali tramite dispositivi mobile[26].

In Italia Twitter è in una fase nuova, in quanto nei 2015 il numero di utenti attivi sul social network è stato di 6,8 milioni di utenti, con un calo addirittura del 28% rispetto all'anno precedente (fonte Audiweb). Purtroppo non ci sono dati ufficiali rilasciati da Twitter sulle statistiche in Italia per il 2016 quindi non si possono fare valutazioni oggettive su questo aspetto.



[23] https://code.facebook.com/projects
[24] http://ogp.me/
[26] https://about.twitter.com/it/company

Stato dell'arte e Gap Analysis

In questo capitolo verranno illustrate le moderne tecnologie in essere nella *trend detection analysis* con l'obiettivo di individuare quali saranno i principali strumenti e le principali tecniche che daranno vita all'output del progetto.

Introduzione

Comprendere il comportamento di gruppi di utenti in maniere dipendente dal tempo può aiutarci ad identificare e predire nuovi importanti trend del mondo reale.

Che tipo di eventi del mondo reale e *trend* si possono riflettere sui dati sociali? E quali proprietà di questi eventi siamo interessati a conoscere?

Supponiamo che un influente analista finanziario scriva una tweet con una forte opinione su un particolare titolo azionario e che questo tweet diventi **virale**. Oppure, supponiamo che un grande numero di utenti utilizzi Twitter per lamentarsi di un particolare nuovo prodotto. In entrambi i casi, la prima domanda corretta da porsi è: *quando sono accaduti questi eventi, o quando il loro trend ha avuto inizio*? Successivamente, potremmo chiederci: *quanto è significativo il cambiamento nel trend*? *Quanto grande è l'incremento o il decremento di questo trend*? Ma più importante è chiedersi *quanto sia grande questo cambiamento rispetto ai cambiamenti che avvengono tipicamente su Twitter*. Twitter in questo scenario è da considerarsi solo un esempio, questo tipo di analisi possono essere effettuate su ogni social media, più genericamente su ogni stream di dati sui quali possiamo costruire una serie temporale.

Quantificare il cambiamento permette agli analisti di distinguere un comportamento atipico da uno tipico, ma permette anche di confrontare tra loro eventi atipici. E' lecito chiedersi: ci sono delle caratteristiche comuni a più eventi atipici che ci permettano di suddividerli in gruppi ad esempio trend stagionali, o eventi connessi a particolari particolare scelta di un modello quantitativo per quel determinato tipo di evento? Se l'identificazione di un evento atipico nel tempo può essere quantificato, questo può essere fatto automaticamente? E' può essere utilizzato per predire comportamenti futuri?

Questa analisi vuole dare gli strumenti di base per costruire un sistema in grado di scoprire, misurare, confrontare e discutere i cambiamenti in una serie temporale di dati provenienti dalle interazioni sociali.

Trend ed Eventi

E' possibile *misurare il comportamento degli utenti nel tempo analizzando qualsiasi quantità che può essere* ***contata*** *temporalmente*, e questo conteggio può essere fatto ad intervalli di tempo regolari (1 minuto, 1 ora, 1 giorno...). A titolo esemplificativo, potremmo contare gli hashtag, i follower, le amicizie, i link, i like, o qualsiasi altra quantità che può essere conteggiata in funzione del tempo. Se la quantità che misuriamo è definita da una parola o sequenza di parole allora potremmo dare a questa quantità la definizione di topic. Quando parliamo di cambiamenti in una serie temporale dobbiamo aver chiaro a che tipo di cambiamento siamo interessati. Ad esempio, siamo interessati a come cresce una certa quantità nel tempo, oppure siamo interessati a cambiamenti ciclici (stagionali). Oppure siamo interessati a cambiamenti emergenti, in cui un cambiamento parte da un punto trascurabile per raggiungere un punto significativo. Oppure siamo interessati a cambiamenti strutturali in cui la serie temporale cambia brutalmente andamento da un momento all'altro.

festività? Nel caso di una risposta affermativa, queste suddivisioni suggeriscono una

Identificare incrementi, cicli e cambiamenti emergenti e strutturali è abbastanza complesso. La principale difficoltà risiede nel fatto che spesso non conosciamo in anticipo la scala o la dimensione del cambiamento. L'intervallo di tempo nel quale un cambiamento è significativo potrebbe variare da una frazione di secondo a anni. In aggiunta, la dimensione del cambiamento potrebbe variare da 10 a un miliardo in funzione di ciò che stiamo contando. Il gruppo di utenti interessato dal cambiamento potrebbe andare da un singolo utente ad una comunità di centinaia di milioni di utenti. E' difficile definire algoritmi in grado di soddisfare questa varietà di dati.

La *dimensione dei dati* da analizzare è enorme e questo introduce ulteriori difficoltà. Alcuni segnali di interesse sono difficili da intercettare perché relativamente piccoli rispetto all'enormità di tutti i dati. La grande dimensione dei dati implica l'esistenza di molti *pattern*[27] atipici che sono interamente dovuti a variazioni statistiche piuttosto che riflettere eventi nel mondo reale. Questo rende difficile l'identificazione di cambiamenti veramente esistenti nel mondo reale da quelli che si verificano solamente nella piattaforma social che si sta analizzando. In genere si preferisce prendere in considerazione qualsiasi tipo di cambiamento. Inoltre va considerato anche il contrario, ovvero che un evento che si verifica prima sui social potrebbe poi successivamente riflettersi nel mondo reale.

Il primo passo per capire questo tipo di fenomeni è **quantificare i trend nei dati sociali,** ossia riuscire a trovare una certa metrica o dei criteri per valutare l'impatto dei trend. Passiamo in rassegna le principali tecniche utilizzate per tale scopo.

Analisi Trade-off

I tentativi di quantificare i cambiamenti nei dati sociali sono soggetti a dei *trade-off*[28]. Alcune fluttuazioni casuali nei dati potrebbero essere identificati come trend. Invece dei trend reali potrebbero non essere identificati.

Possiamo definire tre misure di performance che prendono in considerazione questo tipo di errori:

- la prima è la *time-to-detection*, ovvero il tempo tra l'evento nel mondo reale e la sua identificazione nei dati sociali;
- la seconda è la precisione (*precision*), calcolata come la frazione di trend correttamente identificati che non sono dovuti alla semplice fluttuazione statistica;
- infine il richiamo (*recall*), ovvero la frazione di trend che sono identificati rispetto al totale degli eventi reali.

Due metriche simili alla precisione e al richiamo sono la frazione di veri positivi e quella di falsi positivi. Queste metriche non possono essere ottimizzate simultaneamente. Ad esempio, se vogliamo conoscere velocemente e con elevata accuratezza dei cambiamenti emergenti per forza di cose saremo costretti a tralasciare alcuni eventi andando ad inficiare il richiamo. Contrariamente potremmo avere il richiamo massimo identificando ogni possibile cambiamento, ma in questo modo saremo poco precisi in quanto identificheremo un elevato numero di falsi positivi, ovvero di cambiamenti che in realtà non sono trend. Per questo dobbiamo scegliere un trade-off in funzione del tipo di analisi che vogliamo realizzare e del contesto applicativo.

[27] Un pattern è un termine inglese utilizzato per indicare una certa regolarità in un insieme di dati
[28] *Trade-off*, letteralmente *compromesso,* è una tecnica che cerca di trovare una via di mezzo tra gli svantaggi e i benefici di una certa situazione, in modo da creare una configurazione che tende a quella ottimale

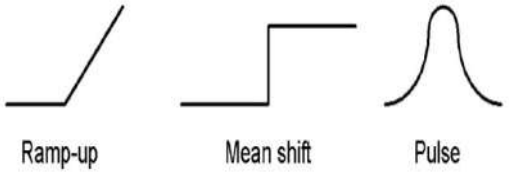
Classificazione

Una volta che lo schema di identificazione dei topic è definito, gli analisti devono interpretarlo e agire di conseguenza ogni volta che un evento anomalo è osservato.

Le azioni potrebbero essere le seguenti:

- *allertare*: porre l'attenzione su qualcosa di nuovo o urgente;
- *informare*: notificare lo stato delle cose;
- *scoprire*: raffinare iterativamente l'analisi in modo da scoprire nuovi trend o l'origine/causa dell'evento identificato;
- *costruzione del modello*: consentire l'analisi a valle del segnale per la costruzione di nuovi modelli.

Definite queste sfide e considerazioni è possibile organizzare l'analisi attorno a tre classi di anomalie.



Nonostante l'analisi dei decrementi anomali in serie temporali sia interessante in questo documento prenderemo in esame solo lo studio degli incrementi.

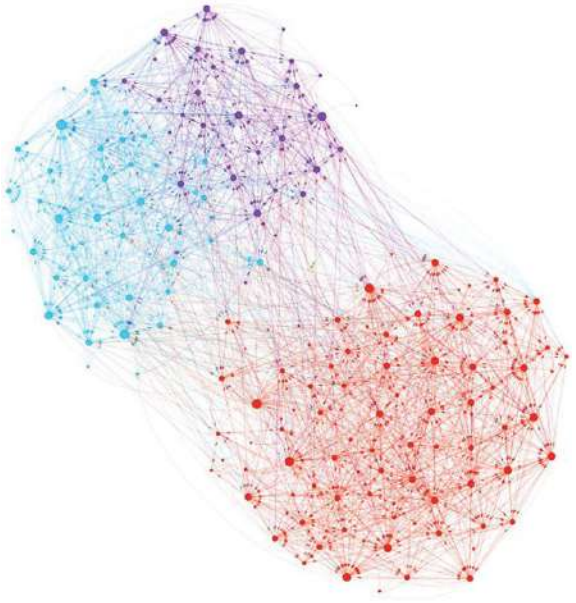
Possiamo identificare tre tipi di incrementi anomali:

1. **Ramp-up**: da un punto ben definito la serie temporale esibisce un incremento costante che perdura per un numero consistente di intervalli di tempo;
2. **Mean shift**: da un certo punto in poi la serie temporale subisce un cambiamento repentino rispetto al suo cambiamento medio, e questo cambiamento perdura nel tempo in maniera consistente, almeno per un intervallo più lungo dell'intervallo di analisi utilizzato;
3. **Pulse**: in un certo punto ben definito della serie storica l'incremento è repentino per poi successivamente ritornare rapidamente al suo andamento tipico.

Ci sono delle correlazioni tra questo tipi di anomalie di base. Per esempio, un *pulse* potrebbe essere considerato come una coppia di cambiamenti *mean shift* o *ramp-up/ramp-down*. O, ancora, un ciclo potrebbe essere pensato come una sequenza di questo tipo di anomalie di base.

L'ultima sfida è quella di associare i cambiamenti rilevati ad eventi del mondo reale. Possiamo in maniera *naif*, ossia in maniera istintiva e non ragionata, definire un evento come un cambiamento nel mondo reale a cui possiamo associare un nome, un'etichetta (ad esempio Superbowl), ma esso potrebbe anche essere un cambiamento nella serie temporale abbastanza atipico al quale non siamo capaci di associare un nome.

In questo documento faremo riferimento ad evento, solo nel caso in cui siamo capaci di assegnare un nome a qualcosa che accade sia nel mondo reale che nel mondo sociale. Durante la progettazione di un algoritmo di event-detection bisogna attentamente considerare che tipo di cambiamento vogliamo identificare, quale tipo di anomalia nella serie è significativa rispetto al contesto in cui stiamo lavorando. Solo così è possibile realizzare degli strumenti efficaci rispetto al dominio di utilizzo.



Le informazioni raccolte nella presente pubblicazione sono state realizzate per l'associazione Alumni Mathematica. Il documento è concesso gratuitamente in forma ridotta e in versione anteprima. Per maggiori informazioni su come ricevere il documento integrale scrivere a *staff@alumnimathematica.org*

Non è possibile utilizzare i presenti contenuti per la diffusione in rete senza il consenso di Alumni Mathematica e degli autori. Per autorizzazioni: *zero@pigreek.com*

Autori

Stefano Franco è ricercatore specializzato in algoritmi, trasferimento tecnologico e modelli matematici. Co-fondatore di Alumni Mathematica e di altre startup, molto attivo nella diffusione del sapere scientifico e nella ricerca scientifica indipendente. Fondatore della società di consulenza Pigreek.

Pierpaolo Basile è ricercatore presso il Dipartimento di Informatica di Bari. Specializzato nello studio di metodi e tecniche per l'estrazione di informazioni da documenti di testo, nella semantica e nell'intelligenza artificiale. Co-fondatore della software house QuestionCube.