Twitter Vigilance: Modelli e Strumenti per l'Analisi e lo Studio di Dati Social Media ed il Monitoraggio in Real Time

Daniele Cenni, Paolo Nesi, Gianni Pantaleo, Irene Paoli, Imad Zaza
DISIT Lab, Distributed [Systems and internet | Data Intelligence and] Technologies Lab
Dep. of Information Engineering (DINFO), University of Florence, Italy, Fax: 0039-055-2758570
http://www.disit.dinfo.unifi.it, http://www.disit.org/tv, http://www.disit.org/rttv, paolo.nesi@unifi.it

Le tecniche e gli strumenti di analisi dei Social Media stanno diventando sempre più importanti per la previsione di eventi e tendenze, per la diagnosi precoce tramite il monitoraggio sociale e l'uso degli utenti come sensori. In questo contesto, Twitter.com è uno dei canali più interessanti per la sua diffusione e le dinamiche di risposta veloce; Twitter viene tipicamente classificato come microblog con aspetti sociali: messaggi brevi collegati e relazioni fra gli utenti. Il numero di Tweet prodotti al giorno rientra nel dominio dei big data. Anche restringendo l'insieme dei Tweet a specifiche aree tematiche, spesso la complessità dei processi di acquisizione, gestione ed analisi dei dati in tempo reale per l'analista sono difficilmente gestibili con tecniche tradizionali. A questo riguardo, dall'Aprile 2015 DISIT lab ha attivato lo strumento Twitter Vigilance per permettere a ricercatori ed analisti di effettuare analisi e ricerche su dati derivati da Twitter riferiti ad aree tematiche diverse. In questi ultimi 12 mesi, tramite Twitter Vigilance sono state sviluppate moltissime analisi negli ambiti: ambiente e meteo, disastro ambientale e resilienza, farmacologia, servizi smart city, turismo, cultura, intrattenimento e TV, grandi eventi, etc. In questo articolo si presenta una sintesi della soluzione Twitter Vigilance, con i suoi strumenti, ed alcuni informazioni che possono essere utili per comprenderne i meccanismi.

Introduzione

L'uso di strumenti che analizzano i dati provenienti da social network e/o blog è oramai diffuso. Fra le social network più diffuse come Facebook, Twitter, G+, etc., ve ne sono alcune più o meno adatte a poter essere utilizzate per fini di ricerca e di analisi. Fra queste Twitter è una delle più interessanti per le sue caratteristiche di apertura e velocità di reazione della sua utenza. In letteratura, soluzioni basate sull'analisi di dati provenienti da Twitter sono state utilizzate per: il rilevamento dell'arrivo di nuove droghe sul mercato, l'identificazione precoce di eventi e disastri, per la definizione di modelli e soluzioni di capaci di effettuare delle previsioni, per l'analisi dell'apprezzamento di prodotti e persone (in termini di sentiment negativo, positivo, neutro, etc.), per lo studio della risposta ad eventi di intrattenimento televisivo, per la stima delle dimensioni della folla e/o per le predizioni del numero di persone coinvolte in grandi eventi, per la predizione degli andamenti in borsa, ecc. In sostanza, alcuni dati estratti da Twitter, opportunamente elaborati, possono essere sfruttati per calcolare metriche e definire modelli matematici specifici che possono essere utilizzati come strumenti di previsione, diagnosi precoce e per l'analisi della risposta sociale. Ovviamente sono risultati che possono avere un grande valore, oppure un valore limitato dipendentemente dalla correlazione fra la massa delle persone e l'utenza di Twitter.

La maggior parte delle metriche basate su dati provenienti da Twitter si fondano sul conteggio del numero di tweet, del numero di retweet, del numero di follower/amici, il numero di commenti, le relazioni fra utenti, e molti altri parametri che possono essere ottenuti con svariate, e più e o meno complesse elaborazioni. Su questa base, DISIT lab dell'Università degli studi di Firenze ha sviluppato la famiglia di strumenti Twitter Vigilance che oramai sono attivi 24 ore su 24 dall'aprile 2015. In questo periodo sono stati raccolti e analizzati oltre 200 milioni di Tweet per scopi di ricerca. Twitter Vigilance è uno strumento che offre servizi di "intelligence" per la creazione di cruscotti e viste personalizzate per lo studio di eventi e tendenze tramite metriche derivate da Twitter e consente la creazione di nuovi modelli per la previsione, la diagnosi precoce, la valutazione e il monitoraggio, in svariati domini applicativi, definibili dall'utente stesso.

Twitter Vigilance colleziona in modo automatico i dati e su questi effettua operazioni di data mining del contenuto. In accordo alla terminologia di Twitter Vigilance, l'utente può creare dei "Canali" di ascolto,

dove ogni Canale di Twitter Vigilance può essere configurato per monitorare un gruppo di chiavi di ricerca su Twitter.com con una sintassi espressiva ed efficace. Dall'interfaccia utente è possibile ottenere direttamente l'andamento di alcune metriche di volume collegate a queste ricerche e molte altre informazioni e andamenti. Alcuni dei Canali che sono attivi su Twitter Vigilance, o che lo sono stati in passato per un certo periodo di tempo, sono accessibili e sono a disposizione del pubblico tramite la pagina web http://www.disit.org/tv/. E' proprio l'utente che definisce il Canale che può decidere se renderlo accessibile per il pubblico o meno. Inoltre, la pagina di riferimento per informazioni e news su Twitter Vigilance è http://www.disit.org/6693.

Twitter Vigilance viene utilizzato per il monitoraggio dei servizi della città Firenze, per molti aspetti anche a livello regionale, nazionale e/o internazionale; sempre per il controllo della risposta dell'utenza rispetto a eventi critici reali e potenziali, per la valutazione dei servizi di mobilità e di trasporto, per la risposta alle problematiche ambientali e meteo, per la valutazione dei canali e modelli di comunicazione, etc. Twitter Vigilance fornisce una serie di strumenti di analisi e di soluzioni di base ed avanzati per il controllo di metriche basate su dati che provengono da Twitter. In particolare, Twitter Vigilance è in uso in svariati contesti come ad esempio, nel:

- Progetto Sii-Mobility Smart City Nazionale http://www.sii-mobility.org per lo studio degli aspetti di mobilità e trasporti: per la valutazione della qualità del servizio, per lo studio di eventi;
- Progetto RESOLUTE H2020 http://www.resolute-eu.org per gli aspetti di resilienza, la valutazione della riposta a eventi critici in città e/o lo studio di modelli per la diagnosi precoce di eventi critici: per esempio le bombe d'acqua in città nel 2015, l'evento di Lungarno Torrigiani, le ondate di calore, etc.
- Progetto REPLICATE H2020 per il monitoraggio della comunicazione relativa a servizi innovativi in città a supporto della Control Room della città, per gli eventi di intrattenimento, etc.
- Corso di Master in Big Data Analytics and Technologies for Management, MABIDA, http://www.disit.org/mabida a supporto delle sperimentazioni; e in
- svariati progetti di più piccole dimensioni.

Servizi accessibili per gli utenti di Twitter Vigilance

Gli utenti di Twitter Vigilance possono:

- Creare uno o più Canali di vigilanza/monitoraggio. Ogni Canale può essere impostato per monitorare una o più query di ricerca su Twitter, che raccolgono un numero variabile di tweet. La query più semplice può essere la singola parola chiave, hashtag o il singolo utente, quelle più complesse compongono in AND parole, hashtag, citazioni, etc.;
- Creare e attivare Canali multipli che possono utilizzare query di ricerca (nuove o già definite). L'attivazione a posteriori di una nuova chiave di ricerca attiva una procedura di rianalisi automatica su tutti i dati passati in modo da aggiungere nuove metriche a tutte le viste;
- Accedere ai dati di andamento di metriche del Canale e delle ricerche, per volume di tweet o di retweet, ed informazioni su metriche e distribuzioni relative agli utenti;
- Effettuare analisi, andamenti nel tempo e distribuzioni su canali e ricerche circa:
  - volume di tweet e retweet nel tempo;
  - attività degli utenti, popolazione, relazioni;
  - hashtag, parole chiave, aggettivi, sostantivi, citazioni, verbi;
  - sentiment analysis a livello canale, ricerca, etc.
- Scaricare i valori numerici in vari formati relativi alle metriche come serie di dati per il raffinamento dell'analisi;
- Produrre viste grafiche ed esportarle in formati grafici diversi;
- Fornire l'accesso al pubblico ad una visione di alto livello di goni singolo Canale, queste informazioni saranno accessibili senza registrazione;
- Condividere i propri canali con altri utenti;

- effettuare ricerche complesse full text multilingua e sfaccettate (faceted, https://en.wikipedia.org/wiki/Faceted_classification) su tutto lo storico dei tweet di Twitter Vigilance, ed in particolare per: canale, ricerca, utenti, citazioni, hashtag, lingua, tweet/retweet, etc.
- attivare Canali e pertanto ricerche specifiche da effettuare in tempo reale per inserire pannelli di controllo sugli andamenti all'interno di strutture di controllo come Control Room per la protezione civile, amministrazioni pubblici, operatori della città e commerciali, operatori pubblicitari, etc.

Gli Strumenti di Twitter Vigilance

Gli strumenti della soluzione integrata Twitter Vigilance sono accessibili via WEB e sono adatti per lo studio, la ricerca ed il monitoraggio di social media via Twitter. In particolare, i loro punti accesso sono:

- **Twitter Vigilance main** tool: http://www.disit.org/tv/
- **Real Time Twitter Vigilance**: http://www.disit.org/rttv/
- **Twitter Vigilance Advanced Search** facility based on SOLR: http://tvsolr.disit.org/search/?collection=1

Sono pubblicamente accessibili e offrono, anche senza registrazione, un set di canali liberamente consultabili su cui è possibile effettuare analisi mediante tutte le metriche descritte. Oltre a questi strumenti, il team Twitter Vigilance di DISIT è in grado di sviluppare campagne di analisi specifiche o di fornire semplicemente supporto formativo e di *big data analytic* come per esempio gli studi effettuati su EXPO2015, XFactor, Pechino Express, Caldo in Toscana, bombe d'acqua a Firenze, Mugnone2016, Lungarno Torrigiani, etc., anche in collaborazione con importanti enti come LAMMA, e CNR IBIMET.

Twitter Vigilance main tool permette di (i) gestire, attivare e configurare Canali e ricerche, (ii) seguire le principali tendenze di Twitter tramite i propri canali, (iii) analizzare i trend delle metriche computate in automatico, (iv) eseguire analisi sugli utenti e relazioni, e su richiesta (v) eseguire elaborazioni di natural language processing, NLP, e sentiment analysis, SA, ecc. La maggior parte di queste valutazioni sono calcolate/aggiornate con cadenza oraria e/o giornaliera, permettendo in questo modo di effettuare previsioni e valutazioni su eventi/accadimenti che hanno dinamiche lente o che sono avvenuti nel passato. In particolare, l'analisi relativa all'NLP e alla SA integra in maniera innovativa tecniche allo stato dell'arte come il parsing sintattico automatico del testo, soluzioni di Part-Of-Specch (POS) tagging, nonché algoritmi di Named Entity Recognition e disambiguazione, insieme a risorse esterne semantiche annotate per la determinazione della polarità di Sentiment. Da questo strumento si possono scaricare informazioni e dati che possono essere rielaborati per creare modelli statistici predittivi.

Quando gli eventi variano velocemente nel tempo, come per esempio per la diagnosi precoce di condizioni critiche, per seguire l'evoluzione di una trasmissione televisiva, per seguire la risposta dell'uditorio rispetto a un dibattito; è necessario attivare delle elaborazioni mirate che seguono il flusso dati in tempo reale. Questo tipo di elaborazione è possibile tramite **Real Time Twitter Vigilance (**http://www.disit.org/rttv/**)** in cui ogni Canale, acquisisce i dati da Twitter e gli elabora in tempo reale effettuando valutazioni statistiche, NLP e SA; fornendo in questo modo direttamente i risultati in forma di grafico e di lista dei Tweet su base temporale e della risposta emotiva agli stimoli ed accadimenti.

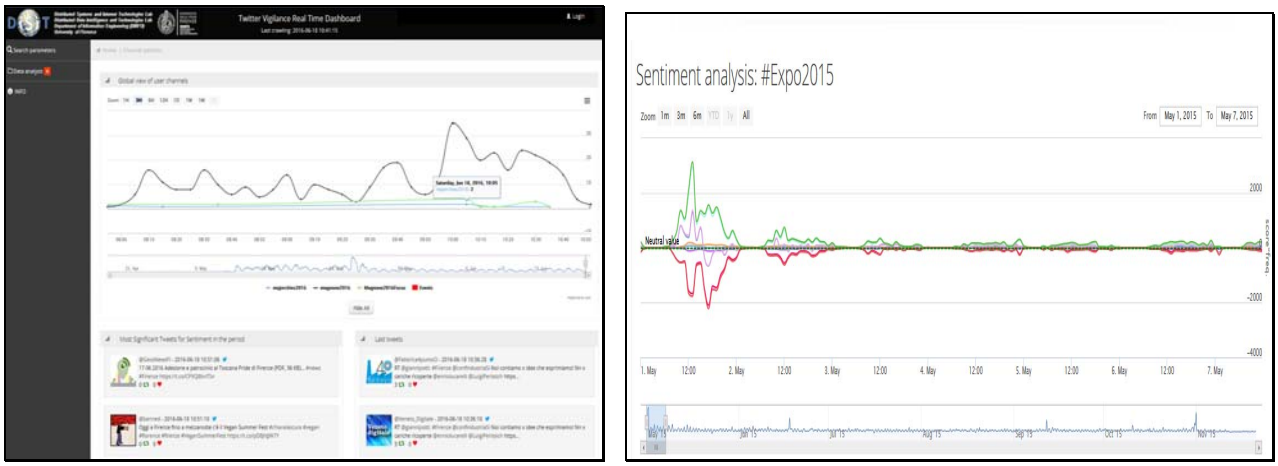

Figura 1 – Real Time Twitter Vigilance (a sinistra), una vista sulla Sentiment Analysis (a destra)

In alcuni casi, gli analisti di dati social media hanno la necessità di ricercare all'interno dei tweet e dati collezionati combinazioni particolari incrociando canali, ricerche, lingue, citazioni, utenti, periodi temporali, luoghi, etc., effettuando varie combinazioni AND/OR e/o faceted/sfaccettate. A questo fine è stato sviluppato lo strumento **Twitter Vigilance Advanced Search** che permette di navigare nel pool di tutti i tweet collezionati con un indice SOLR, per esempio, http://tvsolr.disit.org/search/?collection=1. Anche in questo caso si possono configurare e concordare la creazione di viste specifiche, o anche l'ampliamento del modello di ricerca attuale.

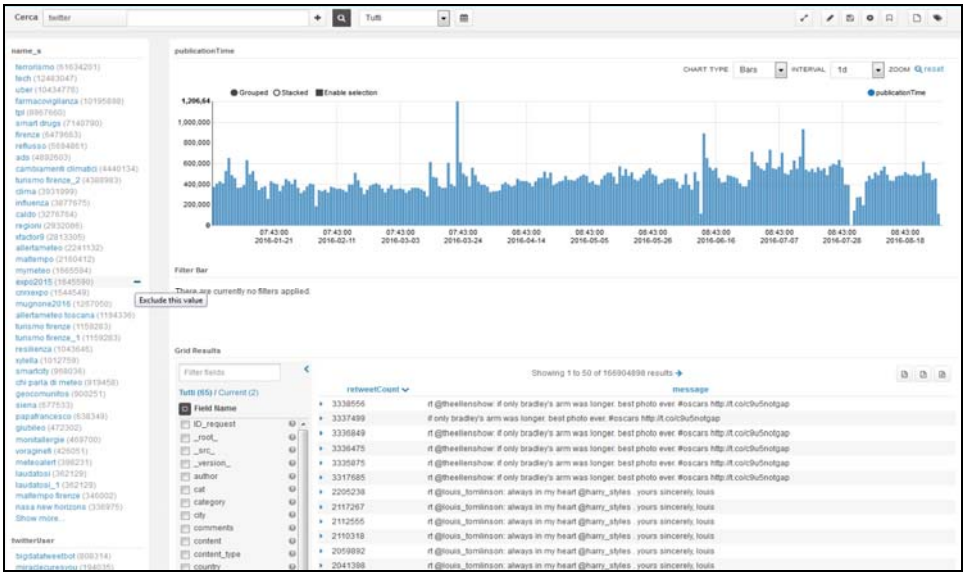

Figura 2 – Twitter Vigilance Advanced Search, una vista.

Conclusioni

La famiglia di strumenti di Twitter Vigilance offre soluzioni per l'automazione di analisi del canale social media Twitter e permette di effettuare la diagnosi precoce di problemi, lo sviluppo di modelli predittivi, lo studio dei social media, in vari domini, ecc. Le varie sperimentazioni hanno permesso alla piattaforma Twitter Vigilance di crescere e diventare uno strumento efficace e semplice da utilizzare. Questi risultati si fondano su più di un anno di rilevamento e analisi. Gli strumenti presentati non sono la risposta definitiva all'analisi del social media o di dati Twitter ma permettono di collezionare i dati in automatico, produrre delle deduzioni semplici anche in automatico, produrre indagini semplici in modo pressoché standardizzato, automatizzato ed assistito, e allo stesso tempo lasciano spazio per lo sviluppo di analisi dettagliate e approfondite dei dati collezionati, capire se vi sono margini per la definizione di modelli più complessi e raffinati, anche con il supporto del team Twitter Vigilance del DISIT lab, se necessario.

Riferimenti

- V. Grasso, Zaza I, Zabini F, Pantaleo G, Nesi P, Crisci A. (2016) Weather events identification in social media streams: tools to detect their evidence in Twitter. PeerJ Preprints 4:e2241v1 https://doi.org/10.7287/peerj.preprints.2241v1
- Valentina Grasso, Alfonso Crisci, Alice Cavaliere, Simone Menabeni, Paolo Nesi, Un dialogo costruito anche grazie a Twitter, in IL CONSUMO DI SUOLO: STRUMENTI PER UN DIALOGO, Laura Cremonini Ed., Istituto di Biometeorologia IBIMET-CNR, Italia, Bologna, ISBN 978889559724, 2015
- Alfonso Crisci, Valentina Grasso, Simone Menabeni, Paolo Nesi, Gianni Pantaleo, "Predicting Number of Visitors and TV programme Audience by Using Twitter Based Metrics", submitted
- Paolo Nesi, Alice Cavaliere, Gianni Pantaleo (University of Florence), Alfonso Crisci (IBIMET-CNR), Valentina Grasso (Lamma Consortium), Simone Menabeni (University of Florence) Monitoring Public Attention on Environment Issues with Twitter Vigilance