Analisi Spazio-Temporale di Messaggi Twitter per l'Identificazione di Eventi

Paolo Arcaini (*), Gloria Bordogna (**, ***), E. Mangioni (***), Simone Sterlacchini (***)

(*) Università di Bergamo, via Marconi 9, 2044 Dalmine (BG) Italy, arcaini@unibg.it
(**) CNR IREA, via Bassini, 15, 20131 Milano (MI) Italy, bordogna.g@irea.cnr.it
(***) CNR IDPA, c/o Univ. di Milano Bicocca, piaz.le della Scienza, 1, Milano (MI) Italy, name.surname@idpa.cnr.it

Abstract
Density based clustering is proposed as an effective way to perform geographic and temporal exploration of messages freely generated within social contexts, in order to reveal and map their latent spatio-temporal structure. The approach is exemplified to identify geographic regions where many geotagged Twitter messages about a given event have been created, possibly in the same time period in the case of aperiodic event, or at regular timestamps in the case of periodic events.

Riassunto Esteso
L'uso diffuso delle reti sociali da dispositivi smart dotati di sensori GPS sta promuovendo una nuova era di applicazioni che sfruttano l'analisi del contesto spazio-temporale dell'utente, ad esempio per fornire raccomandazioni per il tempo libero, per la salute e la sicurezza, la gestione delle catastrofi, e l'identificazione di crisi periodiche.
Adam et al. in [1] riportano che la croce rossa statunitense ha valutato che i cittadini americani fanno sempre più affidamento sui social networks e sui dispositivi mobili per ottenere informazioni su situazioni critiche in corso, come ingorghi stradali , diffusione di pandemie , e per chiedere assistenza e informazioni sulla sicurezza durante o dopo le emergenze
Tali servizi possono sfruttare i contenuti informativi forniti dagli utenti delle reti sociali, che possono essere in forma di testo libero, immagini e video, accoppiati con la data, l'ora e la geolocalizzazione acquisita dal sensore GPS del dispositivo mobile, per identificare gli eventi che si sono verificati in particolari regioni in date specifiche e, in funzione del tipo di evento, restituire segnali di allarme o informazioni utili per la pianificazione territoriale.
In questa proposta riportiamo il nostro lavoro di ricerca per l'identificazione di eventi di interesse sulla base di analisi spazio-temporale di informazioni create mediante Twitter [2]. La nostra proposta potrebbe essere una base per lo sviluppo di servizi context aware.
In particolare , abbiamo progettato e sviluppato un algoritmo di clustering spazio-temporale che può essere personalizzato in modo da identificare gli eventi aperiodici e periodici su scala globale e locale attraverso un'analisi spazio-temporale dei messaggi Tweet con un determinato hashtag [3].
L'algoritmo sfrutta nuove metriche spazio-temporali per filtrare i messaggi isolati e per raggruppare i messaggi inviati da località vicine nello spazio e nel tempo relative allo stesso argomento.
Tale algoritmo è stato applicato per l'identificazione di eventi periodici e aperiodici.
In particolare, per gli eventi periodici per due mesi abbiamo collezionato 78.718 Tweets, usando le API di Twitter, relativi a ingorghi stradali, vale a dire contenenti gli hashtags #trafficjam, #stau, #engarrafamento, #traffico, etc. che esprimono ingorghi stradali in diverse lingue: Tedesco, Inglese, Francese, Olandese, Greco, Italiano, Giapponese, Tailandese, Coreano, Portoghese, Russo, Spagnolo, Turco, indondazioni e tornado (vedi tabella 1), e 5156 tweets relativi al campionato di tennis US OPEN 2013 con hashtag *#usopen*.

Il primo obiettivo è stato quello di identificare le fasce orarie di maggior traffico a livello globale. Come evidenziato in figura 1 sono state identificate due fasce orarie, una mattutina (dalle 7:41 alle 9:35) e una pomeridiana (dalle 16.24 alle 20:13). Tuttavia la fascia oraria pomeridiana è più ampia della fascia mattutina, probabilmente a causa del fatto che c pi variabilit nell orario in cui si termina la giornata di lavoro rispetto a quando la si inizia. Abbiamo successivamente effettuato diversi raggruppamenti spazio-temporali individuando ingorghi a livello locale come quello a Jacarta in Indonesia riportato in Figura 2.
Relativamente ai Tweets dell US OPEN 2013 abbiamo ottenuto clusters temporali corrispondenti alle fasce orarie in cui sono state effettuate le partite del torneo:
Abbiamo poi calcolato la purezza dei cluster ottenuti applicando l algoritmo di clustering con diverse definizioni della distanza spazio-temporale all intera collezione di Tweets. La purezza media dei cluster generati è uguale a 0.74%. Il miglior risultato ottenuto con purezza media uguale a 0.96, minima uguale a 0.83 e massima uguale a 1, si ottiene applicando la distanza combinata spazio temporale con periodo di 1 giorno. Ciò è dovuto al fatto che i fenomeni di cui i Tweets raccolti trattano sono principalmente ingorghi stradali. Lo strumento ha bisogno di valutaioni su più vasta scala ma i risultati ottenuti fino ad ora sono incoraggianti per qualificarlo come metodo utile per l esplorazione spazio-temporale e l identificaione di eventi periodici e aperiodici.

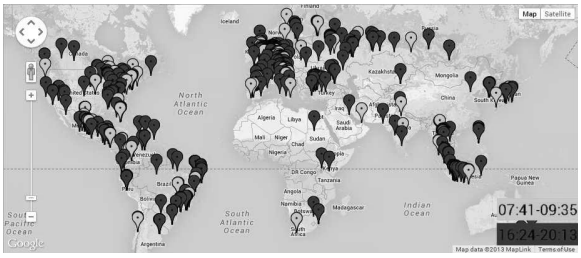

Figura 1: Fasce orarie di maggior traffico stradale

Figura 2: Ingorgo stradale a Jacarta

[1] Adam, N., Shafiq B., Staffin R., Spatial computing and social media in the context of disaster management, IEEE Intelligent Systems, vol 12, 1514-1672, 2012
[2] Arcaini P., Bordogna G., Sterlacchini S., Geo-temporal density based clustering for exploring periodic and aperiodic Events reported in Social networks, submitted to Information Science, 04-2014.