A Search and Summary Application for Traffic Events Detection Based on Twitter Data

Meiling Liu[1+3], Kaiqun Fu[2], Chang-Tien Lu[2], Guangsheng Chen[1], Huiqiang Wang[3],

[1]Northeast Forestry University, P.R.China

[2]Department of Computer Science, Virginia Tech, USA

[3]Harbin Engineering University, P.R.China

{meiling,fukaiqun,ctlu}@vt.edu

ABSTRACT

As a form of social media, Twitter records real life events in our cities as they happen. Huge numbers of tweets under the heading of transportation or metro are published every day. This paper presents an application for Traffic Events Detection and Summary (TEDS) based on mining representative terms from the tweets posted when anomalies occur. The proposed ensemble application contains an efficient TEDS search engine with multiple indexing, ranking, and scoring schemes. Spatio-temporal analysis and a novel wavelet analysis model are applied for traffic event detection. This application could benefit both drivers and transportation authorities. Users can search transportation status and analyze traffic events in specific locations of interest. Utilizing the proposed signal processing technology, we demonstrate the system's effectiveness by examining traffic and metro travel in the Washington D.C. area. As the collaboration between a citizen's life and social media becomes ever greater, this could have a significant impact on the prediction of traffic flow, travel selection, and other city computing functions.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data Mining*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval — *Search Process*

Keywords

Wavelet, traffic events search, signal analysis

1. INTRODUCTION

In today's information age, the rapid development of social media means that everything the public sees or hears is immediately shared and spreads rapidly in many directions. "If you see something, say something" [1] is an ad slogan promoted by the New York City Metropolitan Transportation Agency (MTA). One way of doing this is to use Twitter, which allows users to publish short tweets (messages with a 140-character limit) about "what's happening." In March 2011, the estimated number of Twitter users worldwide was 200 million [2], with 177 million tweets sent on a single day, March 11, 2011. This rapidly growing online social medium allows people to post information (tweets) that reflect what they are seeing, hearing, and feeling as they go about their daily lives, effectively making them human sensors reporting events in the physical world [3]. This motivated us to find a way to retrieve and utilize relevant information from these human sensors to identify traffic anomalies on crowded urban road systems. To this end we chose to develop and implement a convenient application to perform such searches, Traffic Event Detection and Summary (TEDS), which leverages crowd intelligence by allowing users to search locations of interest such as specific traffic nodes and analyze what is happening in near real time.

This paper focuses on detecting traffic events and provides a useful summary of our understanding of what users are actually encountering and discussing on Twitter. Event detection has long been a fruitful research topic [4] and the underlying assumption of most studies is that related words show an increase in usage when an event is occurring. An event is therefore conventionally represented by a number of keywords, where a burst in appearance count signals a significant event [4,5]. Weng and Lee [6] described this as finding a way to identify useful information among Twitter's characteristic flood of meaningless "babble". Wavelet analysis is often used to process time frequency and the current study has also adopted this approach, applying it to traffic event detection and to improve the search platform performance.

The major contributions of the new TEDS application proposed here are as follows:

- **TEDS based on wavelet time domain analysis:** The proposed integrated framework for TEDS utilizes wavelet processing technology for spatio-temporal-textual analysis to enhance the efficiency and effectiveness of the knowledge discovery process.
- **Spatial MMR (Maximal Marginal Relevance) on edge judgment:** The new MMR analytical algorithms automatically and efficiently reduce marginal redundancy for tweets signals. Uncovering redundancy and developing novel ideas for ranking and calculating the score of edge tweets enables the algorithm to determine appropriate values.
- **NLP documents summarization modeling:** Utilizing the document analysis algorithm in NLP field allows us to model the event summary for the processed tweets data. Users can understand the real event content by querying a particular traffic intersection or traffic land mark.
- **Search platform:** The new TEDS visualized search platform is designed to rapidly process users' queries by combining different search methods. Several functions help both traffic enforcement agencies and visitors identify and analyze high-congestion areas, as well as types of traffic questions. The

SIGSPATIAL '14, Nov 04-07 2014, Dallas/Fort Worth, TX, USA

intelligent platform combines spatial locations and data mining methods to facilitate traffic event detection.

2. SYSTEM ARCHITECTURE

This section describes the system architecture of TEDS. As shown in Figure 1, at the highest level there are three main components: data processing, wavelet modeling, and the search platform itself.

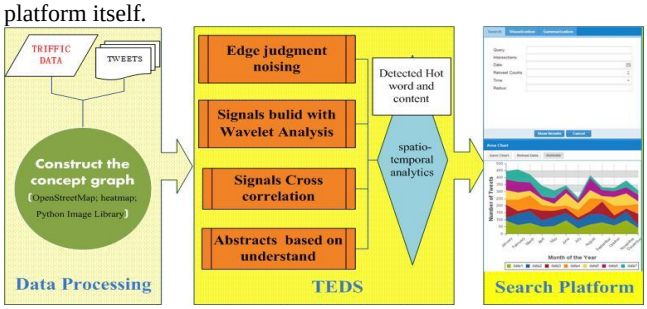

Figure 1: System Architecture

Data Processing: This component pre-processes traffic data sets and transforms them into search records that can then be indexed by the search engine. It combines all the tweet location information to construct the concept graph, and at the same time filters incorrect data such as invalid dates and locations. MongoDB is used to store both the traffic data and tweet data. The system presents a map of the tweet distribution for each traffic node using OpenStreetMap [9,10]. A density based method is also used to generate a heatmap for traffic incidents in each month. The Python Image Library (PIL) is used for visualization.

TEDS: This is the core server component of the TEDS application. Utilizing a hybrid index to store text and traffic data enables us to calculate signal weights and hence display the top-ranked words. The data is filtered and useful words derived by building signals to obtain a hot word set that describes a particular event occurring at an individual traffic node. This signal processing method is known to have a good performance on temporal analytics, so here it is applied to identify the resulting hot word set based on the type of traffic anomaly. A reduced redundancy algorithm for rank is used to improve the quality of the search results. In this system, Natural Language Processing methods are applied to the spatial MMR ranking method to optimize the edge weights. The summarization generated will enable users to detect traffic events and/or discover traffic-related knowledge.

Search platform: This is the primary user interface of the application. It acts as a client for the core search, reporting, calculation, and analytical features of the application as well as providing indexing and searching functionalities. It also provides access to inquiries regarding real time traffic information for travel path selection. The summary displayed will show the main events in a way that is easily understood by the user. By utilizing crowd sourcing, wavelet transform regional differentiating, and clustering analytics in cooperation with the information been published by official, it will also promote effective and efficient use of traffic data by individual citizens and local government bodies.

3. FEATURES

TEDS allows users to search for traffic information near a specific target location and summarize new traffic events in the vicinity. We have developed a novel application for traffic data, namely wavelet time domain analytics. Once a location is specified, our approach ranks and extracts tweets' top words around the location center according to signal weights and cross correlations in the local neighborhood.

3.1 TRAFFIC RELATED TWEETS DATA

Upon launching the search feature of the application, the user can specify specific search criteria (keyword, intersection, date retweet counts, time and radius), as shown in Fig 1. The application presents the results in a map of relevant transportation tweets (with location information) sent by pre-selected influential users such as *WTOPtraffic* and *VaDOT* who are active under the topic of transportation in the Washington DC area. Figure 2 shows a search page and a map of the relevant tweets identified.

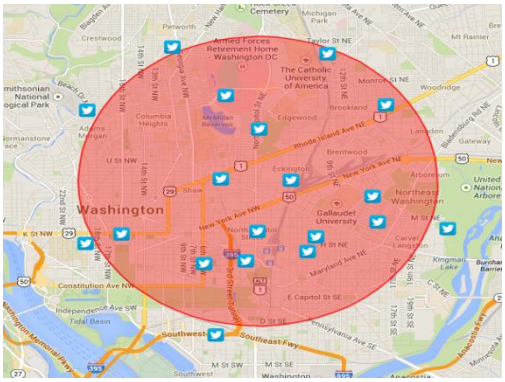

Figure 2: Search Map

The search screen provides a map view such as the one shown in Figure 2 (native Google Maps) of the search results after submitting the inquiry. The circle depicted in the figure 2 represents the area selected by the user, who is then presented with the search results organized as a concept graph to present the tweets their distribution on the map and the details of each node (Figure 3). Clicking on each tweet icon brings up a balloon with the user name, posted time, latitude and longitude, and the tweet content. The user can click on any item on the page to display additional details. Detailed traffic information for a selected intersection or traffic land mark are provided and users can examine the tweet distribution at different times.

Figure 3: Search Results

3.2 TEDS

TEDS is targeted specifically at users who are interested in traffic information. By applying the wavelet analysis method, which is commonly used in the field of signal processing, it is possible to process the tweets using a standard signal analysis. TEDS builds signals for individual tweet words by applying wavelet analysis [6] to the frequency-based raw signals of the words via the four steps listed below. Note that although this method was presented in [6], it was not applied to traffic analysis. Step A. **Edge Noise Judgment**: Applying the edge noise judgment method to traffic

detected events ensures the division of space borders by calculating meaningful relationships. Step B**. Construction of Signals with Wavelet Analysis**: A TEDS signal for each individual word (unigram) is built. Step C**. Cross correlation computation**: Trivial words are filtered out by examining their corresponding signal auto-correlations. The remaining words are then clustered to form events with a modularity-based graph partitioning technique. Step D**. Summary based on the NLP method**: An abstracts approach based on understanding [8] is used for the summary.

We demonstrate the system by applying it to examine traffic and metro travel in the Washington D.C. areas to evaluate the effectiveness of this approach.

A. Edge Noise Judgment

A reduced redundancy algorithm for the ranker is used to improve the quality of the query results. In this system, we apply a natural language processing method that incorporates an improved ranking method to optimize the query results. Applied spatial MMR for tweets reduces the marginal redundancy further. Redundancy and novelty are punished or rewarded, respectively, to calculate the score of edge tweets to determine their relevance. The redundancy parameter gives a lower score to old records and the novelty parameter gives a higher score to new records.

A weighted function is applied as follows. Based on the traditional MMR algorithm, equation 1 gives the basic function,

$$Weight(tweets\_i) = \sum_{j-1}^{total} Sim_{ij} \quad (1)$$

where *Weight (tweets_i) (0<i<count)* represents the weight of tweet *i*, *total* represents the total number of tweets in the selected area, and Sim_{ij} represents the similarity of two tweets.

We provide an improved function to identify edge tweets. Equation 2 gives their weighting calculation,

$$Weight(tweets\_i)^{(1)} = Weight(tweets\_i)^{(0)} - a * Sim_{ij} + \beta * \frac{count(tweets\_j)}{total} \quad (2)$$

where *Weight(tweets_j)(0)* represents the status of a traditional weighted function shown in Equation 1, *Weight(tweets_j)(1)* represents the status of adding punishment and reward, count represents the number of tweets blurred at the edges, α is a redundancy parameter and β a novelty parameter.

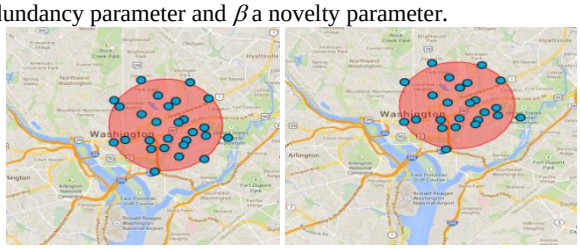

Figure 4(a) MMR **Figure 4(b) Spatial MMR**

B. Construction of Signals with Wavelet Analysis

Figure 4 compares the MMR (Figure 4 (a)) and the spatial MMR (Figure 4 (b)) scores. Both redundancy and novelty for the punishment and reward method can provide an effective and comprehensive way to judge the edge tweet distribution. As shown in the figure, the spatial MMR has a better performance at the division edges and in removing redundancy. Comparing Figs. 4(a) and 4(b) by clicking on each deleted node confirms the good performance of the spatial MMR.

The signal for each individual word (unigram) is built in two stages. Assuming *Tc* is the current time, in the first stage the signal for a word w at *Tc* can be written as a sequence:

$$S_w = [S_w(1), S_w(2), \dots, S_w(T_c)] \quad (3)$$

Sw(t) at each sample point *t* is given by its DF-IDF score, which is defined as:

$$S_w(t) = \frac{N_w(t)}{N(t)} * log \frac{\sum_{i=1}^{N_c} N(i)}{\sum_{i=1}^{T_c} N_w} \quad (4)$$

The first component of the right hand side (RHS) of Eq. (4) is DF (document frequency), *Nw(t)* is the number of tweets which contain word *w* and appear after sample point $t - 1$ but before *t*, and *N(t)* is the number of all the tweets in the same period of time. DF is the counterpart of TF in TF-IDF (Term Frequency-Inverse Document Frequency), which is commonly used to measure words' importance in text retrieval [7].

In the second stage, the signal for word w at current time *T'c* is again represented as a sequence:

$$S'_w = [S'_w(1), S'_w(2), \dots, S'_w(T_c)] \quad (5)$$

Note that *t* in the first stage and *t'* in the second stage are not necessarily in the same units. For example, the interval between two consecutive *t's* in the first stage could be 10 minutes, while that in the second stage could be one hour, making $\Delta = 6$.

Wavelet Energy, Entropy, and H-Measure, the concept of energy derived from Fourier theory, can also be applied here. The Shannon wavelet entropy (SWE) of signal S measures the signal energy distribution at different scales; H-Measure provides a normalized value of SWE(S).

C. Cross correlation computation

In signal processing, cross correlation is a common measure of similarity between two signals. Representing two signals as functions *f(t)* and *g(t)*, the cross correlation between the two is defined as:

$$(f * g)(t) = \sum f * (\Gamma) g(t + \Gamma) \quad (6)$$

TEDS first computes the median absolute deviation (MAD) for each:

$$MAD(S^\tau) = median(|A_i^\tau - median(A_i^\tau)|) \quad (7)$$

MAD is a statistically robust measure of the variability of a sample of data in the presence of "outliers". In the case of TEDS, we are interested in those "outliers" with outstandingly high correlations. Filtering away those signals with $A_i{}^\tau < \theta$, gives the following:

$$\theta_1 = median(A_i^\tau) + \gamma * MAD(S^\tau) \quad (8)$$

Denote the number of the remaining signals as K. The cross correlation is then computed in a pair-wise manner between all the remaining K signals. Denote the cross correlation between $S_i{}^\tau$ and $S_i{}^\tau$ as X_{ij}.

Applying another threshold θ_2 on X_{ij}, this is defined as follows:

$$\theta_2 = medians_i^\tau \epsilon S^{\tau(X_{ij})} + \gamma * MAD\ MAD_j^\tau \epsilon\ S^{\tau(X_{ij})} \quad (9)$$

The Quantification of Event Significance is then:

$$\varepsilon = \left(\sum w_{ij}^c\right) * \frac{e^{1.5n}}{(2n)!},\ n = |V^c| \quad (10)$$

Equation 10 contains two parts. The first part sums up all the cross correlation values between the signals associated with an event, while the second part discounts the significance if the event is associated with too many words.

D. Summary base on the NLP method

This part shows a tweet summary concept based on the NLP method. Taken together with the algorithm for the statistical analysis, we consider a long twitter text content to be useful for this summarization because it can help the user to understand the main content, not simply the topic word. The system applies the multi-document summarization method [8] to extract the data in compliance with the length conditions. According to the dynamic time presentation, the text content is scored to generate a readable abstract.

Based on this analysis, the main computation task in this component [6] is the pair-wise cross correlation computation, which has a time complexity of$O(n^2)$, where n is the number of individual signals involved in the computation; n is generally very small after filtering with θ_1 (in Eq. (8)). In the experimental studies, less than 5% of all the words remained after filtering withθ_1. Figure 5 shows the filtering and clustering performance.

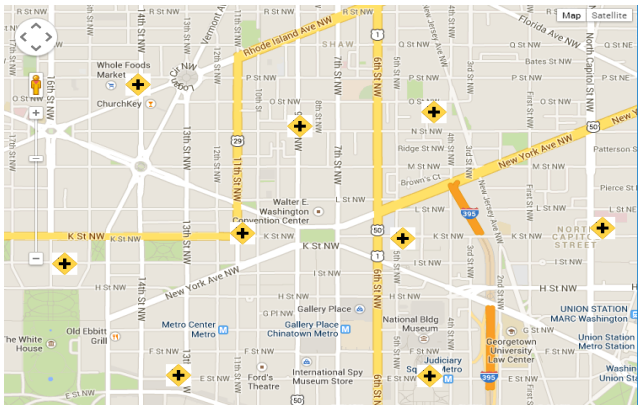

Figure 5: Hotspots

Figure 5 shows a rapid and good inquiry result. To address the interests of most users, this example shows the important road intersections in the Washington D.C. area using the TEDS methods.

3.3 PLATFORM PAGE

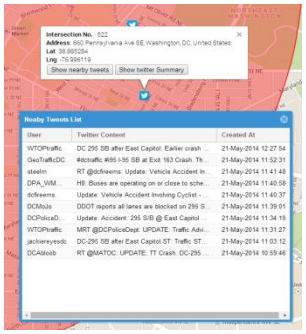

Figure 6 (a)Detected Lists

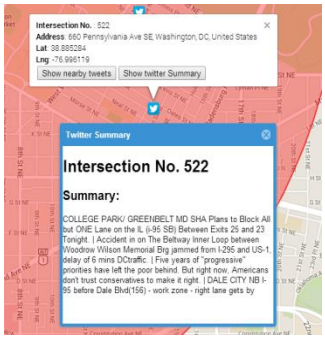

Figure 6(b) Summary

Figure 6(a) shows how clicking on the icon for each important road intersection shows a list of the tweets sent from the vicinity, giving relevant information including the user, Twitter content, time of creation. A balloon shows the intersection number, detailed street address, latitude and longitude. Clicking on the button shows the nearby tweet distribution.

A density based method is then applied to generate a heatmap showing the number of traffic events in each time interval, utilizing the Python Image Library (PIL) for visualization. Multiple resolution versions of the same heat map are generated. Clicking on the button provides a traffic information summary such as the one shown in Figure 6(b).

4. SUMMARY

This new application could help local transportation departments detect hidden traffic events by accessing the information embedded in social media. It could also enable traffic engineers to identify high-congestion areas, types of traffic problems, and locations where they occur frequently. Visitors to a city can be informed of intersection hotspots and thus plan a fast and convenient travel itinerary that avoids them based on a summary of recent tweets. Researchers may also find the application useful: in-depth information on events with a high occurrence density will be readily available and they can then utilize appropriate inquiries to gather data for the analysis of different traffic events and mine other valuable information. TEDS enhances the performance of the entire search and analysis process and provides a better understanding of traffic flow through today's congested urban centers.

Acknowledgements: Supported by the "Fundamental Research Funds for the Central Universities", Northeast Forestry University, P.R.China. 2572014CB26.

5. REFERENCES

[1] De Longueville, Bertrand, et al, *"OMG, from here, I can see the flames!": a use case of mining location based social networks to acquire spatio-temporal data on forest fires.* In Proceedings of the 2009 International Workshop on Location Based Social Networks. ACM, NY, USA, 2009

[2] Shiels, Mark. *BBC News: Twitter co-founder Jack Dorsey rejoins company.* http://www.bbc.co.uk/news/business-12889048 (2011), [Online; accessed 01-November-2011]

[3] Kosala, Raymondus, and Erwin Adi. *"Harvesting real time traffic information from Twitter."* Procedia Engineering 50 (2012): 1-11.

[4] Yang, Yiming, Tom Pierce, and Jaime Carbonell. *A study of retrospective and online event detection.* In SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 28–36, New York, NY, USA, 1998. ACM.

[5] Kleinberg, Jon. *Bursty and hierarchical structure in streams.* In KDD `02: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 91–101, New York, NY, USA, 2002. ACM.

[6] Weng, Jianshu, and Bu-Sung Lee. *"Event detection in Twitter."* ICWSM 11 (2011): 401-408.

[7] Salton, Gerard, and Christopher Buckley. *"Term-weighting approaches in automatic text retrieval".* Information Processing & Management, 24(5):513 – 523, 1988.

[8] Liu, Mei-Ling, et al. *"Research on dynamic multi-document summarization by topic detection and tracking technology."* Journal of Harbin Institute of Technology 11 (2010): 020.

[9] Wang, Bingsheng, et al. *"An integrated framework for spatio-temporal-textual search and mining."* In Proceedings of the 20th International Conference on Advances in Geographic Information Systems. ACM, 2012.

[10] Shah, Sumit, et al. *"Crowdsafe: crowd sourcing of crime incidents and safe routing on mobile devices."* In Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, 2011.