Event detection in Tweets

Andrei-Bogdan Baran
"Alexandru Ioan Cuza" University,
Faculty of Computer Science
General Berthelot, No. 16
andrei.baran@info.uaic.ro

Adrian Iftene
"Alexandru Ioan Cuza" University,
Faculty of Computer Science
General Berthelot, No. 16
adiftene@info.uaic.ro

ABSTRACT
Twitter is among the fastest-growing online social networking services, with more than 140 million users producing over 400 million tweets per day. It enables users to post status updates (tweets) about a huge variety of topics to a network of followers using various communication services such as cell phones, e-mails, Web interfaces, or other third-party applications. Monitoring and analyzing this rich and continuous user-generated content can lead to obtaining valuable information about local and global news and events, because virtually, any person witnessing or involved in any event is nowadays able to disseminate realtime information, which can reach the other side of the world as the event unfolds.

Having a rich data set, we are going to show how to process the tweets in order to obtain valuable information in real time, based on user preferences and different search criteria.

Author Keywords
Twitter; Event detection; Hadoop Ecosystem; Map-Reduce.

ACM Classification Keywords
H.5.2. Information interfaces and presentation (e.g., HCI): User Interfaces. H.3.2. Information Storage and Retrieval: Information Storage.

General Terms
Human Factors; Design.

INTRODUCTION

Unlike other media sources, Twitter messages provide timely and fine-grained information about any kind of event, reflecting, for instance, personal perspectives, social information, conversational aspects, emotional reactions, controversial opinions or even natural hazards. Tweets can be seen as a dynamic source of information enabling individuals, corporations and government organizations to stay informed of "what is happening now".

In contrast to conventional data streaming, event detection from Twitter streams brings new challenges, as it contains large amounts of meaningless messages, which makes the recognition of relevant events even more difficult to achieve.

Event detection in Tweets (EDETE) is an application that extracts various tweets, based on different search criteria (hashtags, locations, etc.) in order to get information about a specific topic of interest. This application focuses on collecting and processing messages from Twitter for detecting relevant events based on a user profile. For collecting data, complex queries are run over Twitter stream (and Twitter API). The data is processed using advanced techniques and tools.

STATE OF THE ART

Despite the fact that the problem of event recognition is social media networks has been studied by many teams of researchers, it is still difficult to find an effective way of extracting relevant information from this type of texts. In order to approach this problem, researchers used techniques from various fields, such as machine learning, natural language processing, data mining, information extraction and retrieval or text mining.

Sankaranarayanan et al. [1] proposed a news processing system based on Twitter (called TwitterStand), to capture tweets that correspond to late breaking news. They employ a naive Bayes classifier to separate news from irrelevant information and an online clustering algorithm based on weighted term vector and use hashtags to reduce clustering errors.

Phuvipadawat and Murata [2] presented a method to collect, group, rank, and track breaking news from Twitter. They first sample tweets using predefined search queries and index their content with Apache Lucene. Messages that are similar to each other are then grouped together to form a news story. They use users' reliability and popularity of tweets, represented by a weighted combination of number of followers and the number of retweeted messages. An application based on the proposed method called Hot-streams has been developed.

Petrovic et al. [3] adapted the online NED (Named-Entity Disambiguation) approach proposed for news media, which is based on cosine similarity between documents to detect new events that have never appeared in previous tweets. Results have shown that ranking according to the number of users is better than ranking according to the number of tweets and considering entropy of the message reduces the amount of spam messages in output.

Becker et al. [4] focused on online identification of realworld event content. They use an incremental clustering algorithm, based on a support vector machine (SVM) classifier. In addition to traditional preprocessing steps such as stop-word elimination and stemming, the weight of hashtag terms is doubled because they are considered a strong indication of the message content.

Cordeiro [5] proposed a continuous wavelet transformation based on hashtag occurrences combined with a topic model inference using latent Dirichlet allocation (LDA). Instead of individual words, hashtags

are used for building wavelet signals. An abrupt increase in the number of a given hashtag is considered a good indicator of an event that is happening at a given time. Therefore, all hashtags were retrieved from tweets and then grouped in intervals of 5 minutes. Wavelet peak and local maxima detection techniques are used to detect peaks and changes in the hashtag signal.

Popescu and Pennacchiotti [6] focused on identifying controversial events that provoke public discussions with opposing opinions in Twitter, such as controversies involving celebrities. Their detection framework is based on the notion of a Twitter snapshot, a triplet consisting of a target entity, a given period and a set of tweets about the entity from the target period.

Benson et al. [7] present another approach to identify Twitter messages for concert events using a factor graph model, which simultaneously analyzes individual messages, clusters them according to event type, and induces a canonical value for each event property. Clustering is guided by term popularity and the idea is to uncover rare event messages that are dominated by the popular ones.

Lee and Sumiya [8] present a geosocial local event detection system based on modeling and monitoring crowd behaviors via Twitter, to identify local festivals. They rely on geographical regularities deduced from the usual behavior patterns of crowds using geotags.

Sakaki et al. [9] exploited tweets to detect specific types of events such as earthquakes and typhoons. They formulated event detection as a classification problem and trained an SVM on a manually labeled Twitter data set comprising positive events (earthquakes and typhoons) and negative events (other events or nonevents).

PROPOSED SYSTEM

EDETE system is able to notify the user in real time regarding different event categories: natural hazards, social, sports or political events based on his/her preferences.

The input data set will consist in tweets messages from Twitter Stream API (https://dev.twitter.com/streaming/overview) and the processing step will be performed using Map-Reduce jobs and other tools from the Hadoop Ecosystem (https://hadoopecosystemtable.github.io/).

Architecture

In Figure 1 is described a high level deployment diagram and the main modules of the system.

Data retrieving and storage

The input data will be collected by using the Twitter streaming API. This stream is limited to receive 1% of the total tweets posted on Twitter. However, 1% is between 3 to 5 million tweets in a day, more than enough for performing different statistics and analysis.

By using the filtered stream, the collecting service is able to filter the tweets by track (kind of keywords), locations and users. A track cannot contain more than 60 characters. Also, the track filter type supports logical operators AND & OR as shown below.

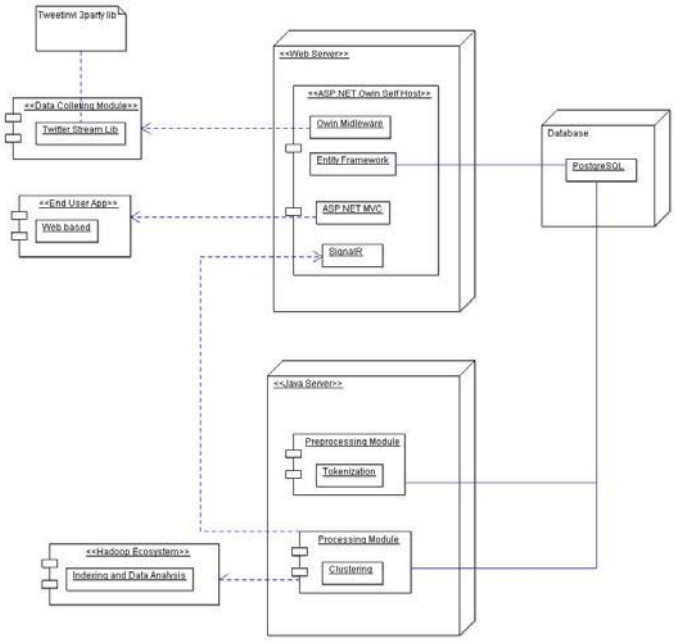

Figure 1. Deployment Diagram.

For example, a track that is containing multiple keywords, separated by space is evaluated with the AND operator: e.g. “apple iPhone”, means “apple” AND “iPhone”. If multiple tracks are added to a stream, then the OR operator is used instead.

Another type of filter is by location. A location filter represents an area (4 coordinates), and the stream will be filtered only the tweets tagged within the specified location. Moreover, the stream can be filtered by specific users, and this feature will allow to follow the activity of a specific group of users.

The preliminary results showed that there are a lot of tweet messages that don’t seem to be relevant to our app. The messages contain a lot of grammar mistakes, abbreviations, shortcuts, missing letters or not even a relevant content based on the attached hashtag.

Data processing

Before starting the data processing, there is an intermediary step called data preprocessing. This involves filtering out stop-words (*on, of, are*, etc.) and applying words stemming and tokenization techniques. As shown above, the data preprocessing already started in the data retrieval module, by applying different filters such as keywords using the logical operators, location, language, etc.

After the preprocessing step, the data is ready to be processed using the Hadoop Ecosystem. Having a rich data set, by using map-reduce jobs we can apply a series of transformations and operations for detecting changes in the data stream.

The proposed approach monitors the evolution of the hashtags over the time from the Twitter stream by using wavelet analysis [10, 11]. The premise that we are starting from is that if an important event is about to happen, there will be a significant increase of the same hashtags during a specific period that will describe the same topic. This approach is based on peak analysis of individual hashtags and the local maxima detection for detecting when an actual change occurred.

The entire processing workflow for detecting event can be summarized as follow:

- For each hashtag, there will be generated a series of signals during a specific period of time.
- The period of time will be divided in smaller time intervals (e.g. 10 minutes), where the hashtag occurrence will be counted for each time slot.
- After, for each hashtag, there will be a list of a key value pairs that will contain the timestamps and the number of the hashtag occurrence in that specific time.
- Having these hashtag signals, by applying the wavelet analysis, we can detect the peeks and the changes of a possible event that may occur at a given time.

Because the Twitter hashtags can introduce a lot of noise, to overcome this issue, we can apply the Kolmogorov-Zurbenko (KZ) [12] filter to reveal the hidden signals or to reconstruct the signals because of the poor preliminary results.

This method of event detection is focused only on the hashtags, without any semantic processing. For future work, we are going to add some knowledge driven elements that can improve the accuracy and the level of trust of the event, with a more relevant details of the event.

After processing, for the end user, a series of data visualization and features will be available, that are described briefly in the next section.

End user app

The user interface for EDETE consists in a web application that will have different functionalities described below:

- Each user will have a profile and a set of preferences: such as the *topic of the events*, *a location* and *a set of keywords* that will be used to perform queries over the Twitter stream.
- When the system will identify an event based on user preferences, it will notify the user via a push message mechanism or email.
- Different options of data visualization of the events: *graphs*, *charts* and *maps*.
- Different data analysis views.
- Google maps integration.

CASE STUDY

Bellow we will describe a general case study for event detection of a natural hazard like an earthquake.

A user, Komura, is interested in monitoring *Tokyo*, *Japan*, an area of high seismic risk, for any possible earthquake. His parents are very old and are living in an old house, at the margin of the city. Komura is worried because he cannot be close to his parents, because he is living in USA with his family, and lately a high seismic activity happened close to Tokyo. By using the EDETE, he can be notified by any possible event that might happen before it will appear on the radio or TV, based on the Twitter data stream.

Komura will create a profile with some relevant keywords (hashtags) and a location near his parents leaving area.

The keywords he set are: #earthquake, #quake and #tsunami. In Figure 2 is captured the screen for the user profile where he can set the hashtags and the location of interest.

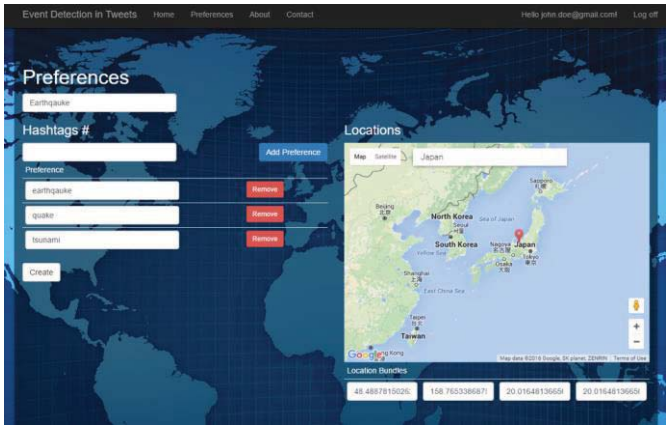

Figure 2. EDETE - User Preferences.

In the Table 1, there are captured the tweets count using the keywords described above. The data captured was made on 29th May 2016, between the 13:00 and 22:00, EEST time zone. The results were queried using a Hive (https://hive.apache.org/) script and the grouping filter was by user time zone.

User time zone	Tweets count
Pacific Time (US and Canada)	1347
Eastern Time (US and Canada)	359
Tokyo	287
Central Time (US and Canada)	235

Table 1. Results of counting the tweets by user time zone

As you can see, in Tokyo there were a total of 278 tweets that were related with the hashtags mentioned. Moreover, between these periods, there has been an earthquake of 4.5 magnitude that stroked 162 km ESE of Hasaky, Japan.

Moreover, by using the wavelet analysis, the EDETE application will be able to detect the peeks where most tweets messages are about earthquake in Tokyo and will send a notification to the user, with a brief description based on the tweets.

CONCLUSION

EDETE is a solution that can have a real impact on our life by helping us to take fast some actions in case any "negative" event happens. Also it can have a huge potential in our social life or other fields by keeping us informed on topics of our interests. It's true that Twitter contains a lot of noise and irrelevant data, but with the help of a tool that can extract relevant info, it can be turned into a powerful analytic system with a lot of potential usages in different domains.

The first version of the EDETE application will consist only in a hashtag analysis, without any text processing. But as future work, we are planning to introduce semantic text analysis in order to extract more relevant information from tweets. The next version of the application will take into consideration the context of the event, the location and other important elements that can improve the user experience. As a result, more details about the processing step, text analysis algorithms and strategies will be provided in the second version of the application.

ACKNOWLEDGMENTS

We thank to our collaborators that help us during the project developing phases.

REFERENCES

1. Sankaranarayanan, J., H. Samet, B. E. Teitler, M. D. Lieberman and Sperling, J. TwitterStand: News in tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '09, ACM, New York, NY, (2009), 42–51.
2. Phuvipadawat, S. and Murata, T. Breaking news detection and tracking in Twitter. *In IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* (WI-IAT), 3, Toronto, ON, (2010), 120–123.
3. Petrovic, S., Osborne, M. and Lavrenko, V. Streaming first story detection with application to Twitter. *In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, (2010), 181–189.
4. Naaman, M., Becker, H. and Gravano, L. Hip and trendy: characterizing emerging trends on Twitter. Journal of the American Society of Information Science and Technology, 62(5), (2011), 902–918.
5. Cordeiro, M. Twitter event detection: Combining wavelet analysis and topic inference summarization. *In Doctoral Symposium on Informatics Engineering*, DSIE'2012. 2012.
6. Popescu, A. M. and Pennacchiotti, M. Detecting controversial events from Twitter. In Proceedings of the 19th ACM international Conference on Information and Knowledge Management, CIKM '10, ACM, New York, NY, (2010), 1873-1876.
7. Benson, E., Haghighi, A. and Barzilay, R. Event discovery in social media feeds. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Volume 1 of HLT '11, Association for Computational Linguistics, Stroudsburg, PA, (2011), 389-398.
8. Fujisaka, T., Lee, R. and Sumiya, K. Discovery of user behavior patterns from geo-tagged microblogs. *In Proceedings of the 4th International Conference on Uniquitous Information Management and Communication*, ICUIMC '10, ACM, New York, NY, (2010), 36:1–36:10.
9. Sakaki, T., Okazaki, M. and Matsuo, Y. Earthquake shakes Twitter users: Realtime event detection by social sensors. *In Proceedings of the 19th International Conference on World Wide Web*, WWW '10, ACM, New York, NY, (2010), 851–860.
10. Lee, D. and Yamamoto, A. *Wavelet analysis: Theory and applications*. Hewlett-Packard Journal, 1994.
11. Cordeiro, M. Twitter event detection: combining wavelet analysis and topic inference summarization. *In Doctoral Symposium on Informatics Engineering*, DSIE. 2012.
12. Yang, W., Zurbenko, I. kzft: Kolmogorov-Zurbenko *Fourier Transform and Applications*. R-Project. 2007.