MultiOSN: Monitoring Multiple Online Social Networks During Real-World Events

Prateek Dewan, Mayank Gupta, Kanika Goyal, and Ponnurangam Kumaraguru

Indraprastha Institute of Information Technology, Delhi
prateekd@iiitd.ac.in, mayank.gupta@dce.edu, kanika.goyal@dce.edu, pk@iiitd.ac.in
precog.iiitd.edu.in

Abstract. The flow of information in online social media during events has been widely studied in the computer science community. It has also been shown how information picked from online social media can help to eventually aid eventful, especially, crisis situations in real life. However, most of the work has focused on utilizing a single network for monitoring such events, mostly Twitter. Given the immense popularity and diversity of various online social networks across the globe, studying multiple networks during an event can reveal much more information about the event than a single online social network. In this work, we present MultiOSN, a system which collects data from five different online social networks viz. Facebook, Twitter, Google+, YouTube, and Flickr, during real-world events, and presents real-time analytics for this data. MultiOSN can be particularly helpful to users and organizations which are directly or indirectly connected to law and order. Organizations can utilize MultiOSN to uncover the general sentiment of social media users about an event, and trace public gatherings for example, which are usually discussed and planned publicly on social networking platforms.

1 Introduction

The popularity gain and reach of Online Social Networks (OSNs) over the last few years has been exponential. Social media services like Twitter, Facebook, YouTube etc. have revolutionized the way people access information about real-world events. With over a billion monthly active users on Facebook alone, close to 20% of the world’s total population is already exposed to social media. At the same time, it is also very interesting to observe the diverse influence that social media has had on this population. During the 2011 US earthquake for example, a Twitter user tweeted “Social media is faster than seismic waves!”, when he saw a tweet from a friend (sitting in Washington DC) about an earthquake, even

1 For now, the name is only a placeholder, we may change the name in due course.
2 http://globalnews.ca/news/527329/number-of-active-users-at-facebook-over-the-years/
3 http://comcat.cr.usgs.gov/earthquakes/eventpage/pde20110823175104590_6#summary
before the earthquake actually hit his office at New York. On the other hand, a recent news article reported how Facebook users are likely to influence the elections in some states of India.

The above examples, along with countless more, depict that social media is a rich source of information, and can be used effectively to gain a lot of information in real-time, especially during an ongoing event. However, most of the existing work in the research community [1, 3, 5, 7] utilizes only a single social network to study an event; mostly Twitter. We believe that studying multiple social networks can reveal much more information about an event, than a single social network. For example, Twitter has been shown to predict elections [6], and Facebook has been used as a medium for online campaigning [4]. But to the best of our knowledge, there has been no work which leverages information from both these networks to analyze the elections. To leverage this information, we designed and developed MultiOSN, a web service which collects data about an event from 5 different online social networks, viz. Facebook, Twitter, YouTube, Google+, and Flickr, and shows real-time analytics that can be consumed directly by the end-users. The main motivation behind building MultiOSN is to collate the information present across multiple social networks, and give a unified view to the user, so that he / she can get a complete snapshot of the entire web of online social media at once. The goal of MultiOSN is to provide a platform for data collection, analysis, and visualization during real-world events, and help us to fill the gaps in the existing research work on studying real-world events on social media.

2 System Architecture

MultiOSN comprises of 3 main components which are responsible for data collection, storage, and analysis. These are the Control Panel, the database, and the Online Portal. Figure 1 represents the high-level architecture of MultiOSN. The events that are to be monitored, are controlled by the administrator. To start data collection for a particular event, the administrator enters the event name, and the relevant keywords as inputs to the MultiOSN Control Panel. The Control Panel triggers individual data collection modules for all 5 social media, i.e. Facebook, Twitter, Google+, YouTube, and Flickr, with these keywords as input. These modules then run in parallel, collect data from the respective public APIs, and store this data in the MultiOSN database in real-time. The data collection can be stopped only by the administrator from the Control Panel. Whenever a user wishes to see the analytics, she visits the MultiOSN Online Portal. Information from the MultiOSN database is then extracted by the online portal, and analytics are generated in real-time.

---

5 http://www.informationweek.in/software/13-04-12/facebook_users_can_influence_lok_sabha_elections_significantly_in_maharashtra_and_gujarat.aspx
Note that under the current setting, event and keyword selection is controlled by an administrator. However, MultiOSN can be deployed at any organization, and the end-users can be the administrators themselves, and manage events.

3 System design

Collecting data for multiple events from five different sources, and storing it in a manageable form is a challenging task. Given that one of the main objectives of MultiOSN is to provide real-time analytics, our main focus was to make the analytics efficient, and at the same time, maintain simplicity in the design. We thus decided to implement two-layered design at the database level. All the five OSM services have one table each, forming the bottom layer. Most of the analytics have one table each, forming an intermediate “cache” layer. The cache layer stores only recent data for all the events, and provides the ability to generate real-time analytics quickly, and efficiently.

The front end of MultiOSN has been built using technologies like PHP \(^6\), and jQuery \(^7\), and is dynamic (figure 2). The frontend generates and presents

---

\(^6\) http://php.net/
\(^7\) http://jquery.com/
all the analytics on the home page by querying the intermediate layer cache tables from the database backend, which have been implemented using MySQL. These cache tables store data for the last 24 hours from the latest post in our database, and are updated at regular intervals. As mentioned previously, apart from these cache tables, the database also contains different tables for storing data from the five social media services that MultiOSN currently tracks. Each event being tracked, has one data collection module running per social media, except Twitter, which has two modules running per event. MultiOSN collects Twitter data from both, searching and streaming APIs. The search API is utilized to get historical data for the period between the time of the event occurring, and the administrator starting data collection for that event from the Control Panel. For example, the Boston Marathon blasts took place on April 15, at 2:45 pm local time. However, we did not come to know of the blasts until about 9:00 pm local time, when the day broke in India. Using only the streaming API would have caused loss of data for this duration of over 6 hours. It is also important to note that for such events, the activity on social media is the highest during the first few hours of the event taking place. This makes it all the more important to find ways to collect this data. Since all other OSM services being tracked by MultiOSN provide only search APIs, we did not encounter this issue with any services other than Twitter.

Thus, to collect data for the Indian Premiere League (IPL) for example, there are currently 6 data collection modules running in parallel in the backend, each collecting data from the public API of the respective social media service, and feeding into the respective table in the database. Each table contains an “event” field, and a “keyword” field, to demarcate data corresponding to one event, from another. Each of the data collection modules run independent of each other, query the respective APIs for each keyword corresponding to an event, serially. Repeated results returned by the APIs are ignored. This process continues in an infinite loop, until the administrator “stops” the data collection process from the Control Panel.

The Control Panel is a secured area, which needs valid log in credentials to access. Figure 3 shows the Control Panel interface. As seen in the figure, this page gives the administrator, the ability to start data collection for a new event, by supplying an event name, and keywords relevant to that event. The interface also provides the administrator with an option to select which all social media services does the administrator want to track for a particular event. In addition, the administrator can also stop data collection for an ongoing event, or resume data collection for a previous event.

MultiOSN is publicly available online at http://precog.iiitd.edu.in/tools/beta/multiosnportal.

Please note that the system is still in it’s early stages of design and development,

---

8 http://www.mysql.com/
10 http://www.iplt20.com/
4 Analytics

In the current version, MultiOSN provides basic real-time analytics for ongoing events, and static analytics for events that have been tracked in the past. The events can be selected from a drop-down menu located at the top left on the home page (figure 2). The total number of posts for each social media service, pertaining to an event, are displayed at the top the page. For ongoing events, this number refreshes every 5-7 seconds. Just below the number of posts are the total number of URLs that are present in these posts. This number is the total count, and not the number of unique URLs. Apart from these numbers, other analytics include sentiment analysis, geographical analytics, activity, and tag clouds showing the most frequently discussed topics for an event, over the last 24 hours. We discuss our current dataset, and each of these analytics in more detail.

4.1 Data description

Currently, MultiOSN has tracked data for 6 events in total, viz. Boston Marathon Blasts (April, 2013), Texas fertilizer plant explosion (April, 2013), Earthquakes...
in India and the middle-east (April 16, and May 1, 2013), the Indian Premiere League (IPL) (April 3 - May 26, 2013), and the Bangalore Blasts (April, 2013). Out of these, data collection for the IPL is currently still going on. The biggest source of information in MultiOSN in terms of number of posts is Twitter, which accounts for 59.41% of the total number of posts in our current dataset of over 12 million posts.

4.2 Sentiment Analysis

The top left of the MultiOSN homepage (figure 2) shows a sentiment graph depicting the sentiment of the posts on each social media during the past 24 hours. The sentiment for all posts collected for an event during an hour are extracted separately, and aggregated to get the final sentiment value for each social media. Hence, 5 sentiment values corresponding to the 5 social media being tracked, are calculated each hour, and plotted on a percentage scale. Understandably, this value is also influenced by the number of posts on a particular social media during a particular hour. For example, if there is only one post on Google+ about an event during an hour, and the sentiment associated with this post is positive, the graph would show a value of +100%. Figure 4 shows a screenshot of page containing the latest positive and negative tweets for an event being visualized on MultiOSN. This feature is part of the detailed analysis page.
Fig. 4. The detailed analysis page containing samples of recent tweets that have been classified as positive and negative. The sentiment analysis is done using the Sentiment140 tool.

Since more than half of the data in our dataset is Twitter data, we decided to use the Sentiment140 sentiment analysis tool [2], which has been built specially for detecting sentiment in tweets. We understand that the overall sentiment value obtained by using this tool for analyzing data from social media services other than Twitter, may be slightly inaccurate. However, the objective of MultiOSN is just to give a sense to the end users about the overall sentiment. We believe that the data being Twitter-heavy, would not hamper the accuracy of the overall sentiment value too much. We intend to test this hypothesis, and if needed, come up with a better solution to this problem in the near future.

4.3 Geographical Analytics

The geographical analytics help the user to visualize the geographical source of the information on a world map. We pull the last 1000 posts from each social media service, which contain location information in the form of a geographical co-ordinate (latitude, longitude), and plot these co-ordinates on a world map using Google Maps API. Each point plotted on the map is color-coded according to the social media service it has originated from. Clicking on any point reveals the post, and it’s source.

https://developers.google.com/maps/
4.4 Activity

The activity graph is an hour-wise plot of the number of posts on each social media against time, for the past 24 hours. As evident from the graph presented in Figure 2, the number of tweets overpowers the number of Flickr images, Google+ posts, and YouTube videos by an enormous margin. The number of Facebook posts are still comparable with the number of tweets. MultiOSN also shows the live feeds from all the 5 social networks at the lower half of the homepage. These feeds contain the latest tweets, Facebook posts, Flickr images, Google+ posts, and YouTube videos that are being collected, in real time. Each feed has a feature to show more content, and to refresh the feed to load the most recent content. To speed up the process of getting latest posts from our constantly growing database, we implemented B-Tree indexing on the timestamp column in our database. This speeds up the process of sorting posts according to time, and fetching the latest ones.

4.5 Tag clouds

The detailed analysis page contains the term-frequency based tag clouds for all social media services, in addition to the sample positive and negative sentiment tweets, as discussed previously. The textual data present across all posts from each social media service is accumulated, and subjected to tokenization, followed by

![Fig. 5. Weighted tag-clouds for the content posted during the past 24 hours. These tag clouds are dynamic; i.e. the user can bring a term of her interest to the foreground by hovering over it.](image-url)
by frequency distribution calculation using PHP. The top 50 most frequently occurring words are then used to generate a weighted tag-cloud. MultiOSN presents 6 tag clouds in all; i.e. 5 corresponding to the 5 social media services being tracked, and 1 tag cloud which is generated by collating data from all the 5 social networks together. This tag cloud depicts the most talked-about topics across all the 5 social media clubbed together.

5 Discussion and future work

MultiOSN is a first step towards building real-time social media monitoring systems to visualize real-world events while they are happening. To the best of our knowledge, this is the first attempt towards studying real-world events on multiple online social networks. The system is in its early stages of development, and a lot of features are being continuously worked upon. One of the major issues that we foresee, and would like to address in the near future, is the increasing size of the database, which might become hard to manage under our current techniques of implementation. We intend to implement a distributed collection and processing architecture to achieve greater efficiency and improve the performance of MultiOSN. In addition, we also intend to remove our dependency on sentiment140’s API for sentiment analysis, and implement a sentiment analysis library on our own infrastructure. However, we do not intend to generate our own training dataset for this purpose in the near future, since sentiment analysis is outside the scope of this work at present.

We understand that there exist other similar systems which can be used to monitor social media for specific topics, but we would like to emphasize on the fact that our endeavor is to collect and analyze data collected during real-world events, and use it to address some broader research goals. Some of these research goals are as follows:

- Identifying and extracting good quality information during real-world events.
- Detecting multiple identities of the social media users across multiple networks.
- Characterizing spam and phishing content posted during real-world events.
- Understanding the privacy concerns of users posting content during real-world events.
- Characterizing the spread of malicious content posted across multiple social networks during real world events.

Although some of these questions have been addressed previously, our emphasis stays with the “real-world events” part, which, to the best of our knowledge, has not been answered yet.

---

12 https://sheriframadan.com/2012/05/creating-a-word-cloud-with-php/
13 http://www.goat1000.com/tagcanvas-weighted.php
14 http://help.sentiment140.com/api
References