Twit-Digest: An Online Solution for Analyzing and Visualizing Twitter in Real-Time

Aditi Gupta, Mayank Gupta, Ponnurangam Kumaraguru
Indraprastha Institute of Information Technology, Delhi
{aditig,pk}@iiitd.ac.in, mayankgupta2005@gmail.com
precog.iiitd.edu.in

ABSTRACT

Twitter, is one of the popular micro-blogging website, which has also emerged as a prominent news media in last few years. Large volume of content generated on Twitter, makes manual monitoring and analyzing of data expensive, in terms of time and resources required. The aim of this project is to develop a Twitter search tool for extraction, analysis and visualization of content from Twitter, with special emphasis on security aspects. Twit-Digest, is an easy-to-use, free-for-all web based tool, which can be used to analyze feed from Twitter in real-time, either we can pass live feed from Twitter or from our archived database. The tool extracts data (tweets and user information) from Twitter and presents them to the user in real-time, along with various analysis outputs, like spam / phishing detection, credibility detection, social network analysis and query expansion. The link to YouTube video of the screen-cast is http://youtu.be/vjc9SVpktK0.

1. INTRODUCTION

Twit-Digest is an online service, developed to provide an interface over the live Twitter stream. In addition to providing tweets from Twitter based on a search query requested by the user, Twit-Digest also provides a variety of analysis of the Twitter data. Analyses are aimed at providing a user with quick inferences and big-picture about the activity around the search query. The tool can be effectively used by security analysts, organizations, and professional agencies that wants to monitor content of its interest on Twitter. One of the major contributions of this work, are the credibility and spam detection of all the tweets displayed to the users. The algorithms for credibility and spam assessments compute the score for the data in real-time and presents the results to the user. To the best of our knowledge, this is the first such system to provide real-time analysis for credibility of content on Twitter.

Currently Twit-Digest is developed in two versions of the application. In the *Real Time* version tweets are obtained from Twitter using the streaming API. In the *Data Base* version, we display tweets stored in our database. Twi-Digest system has been live and used by multiple users over past few months.

2. TWIT-DIGEST FEATURES

In this section, we discuss the features in Twit-Digest. We provide analytical results for the Twitter data at two levels: First at the tweet level, and second at the topic level. At the tweet level, we analyze the content and the user who posted a tweet to draw inferences regarding the credibility and spam / phishing nature of the tweet. At the topic level, we perform analysis of the cumulative data of all tweets obtained for that topic.

2.1 Tweet Level Features

2.1.1 *Credibility Score*

For each of the tweet, extracted from Twitter containing the query word, Twit-Digest provides the credibility score. We compute the credibility of a tweet based on features such as the tweet content and user who tweeted it. This score calculation is done based on the algorithm developed and evaluated by us [1]. We adopted a supervised machine learning and relevance feedback approach, to rank tweets according to their credibility score. The value obtained from the above algorithm is mapped to a score of 1-5 and visually presented to user as shown in Figure 1. Less the number of stars, less is the credibility.

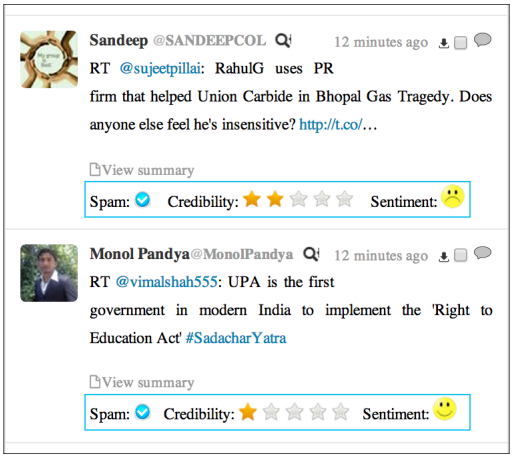

Figure 1: For each tweet credibility score is computed and tweets are marked as spam or not spam.

2.1.2 *Spam / Phishing Detection*

The tool validates whether a tweet is spam or phishing using two different techniques. Firstly, this is done with the help of GoogleSafeBrowsing and PhishTank APIs. [1] [2] Secondly, we use some of the results obtained from previous research

[1]http://www.phishtank.com/developer_info.php
[2]https://developers.google.com/safe-browsing/

works regarding spam and phishing detection on Twitter, for example, presence of more than three hashtags [2] [4] [3]. Based on the above results, each Tweet is marked with *Green Tick* if it is not a Spam or a *Red Cross*, which indicates Spam / Phishing.

2.2 Topic / Cumulative features

2.2.1 Geographical Analysis

Location of tweets and users are marked on a map to show the impact of an event on a particular country or continent using GoogleMaps API. On Twitter there are two primary ways in which it can be shared: firstly, as an user profile attribute called location; and secondly, as a geographic co-ordinate (latitude and longitude) associated with geographically tagged tweet of a user. A sample geo-location output by Twit-Digest are shown in Figure 2.

Figure 2: Sample geographical distribution of the tweets and users corresponding to a query.

2.2.2 Social Network Analysis

In user based social network analysis, Twit-Digest presents a graph, with users as nodes and presence of retweets or @-mentions in tweets (i.e. re-tweeting, replies and mentions) as edges between two user nodes. Such graphs visualization helps in identifying the influencers and communities of users who are posting messages about the topic and driving the discussions on the Twitter. For social movements and campaigns organized via Twitter, we often observed very dense graphs, with a few nodes driving the discussion, and they are retweeted or replied by all other nodes. Figure 3 shows the snapshot of the network graph from Twit-Digest.

2.2.3 Query Expansion

Since many a times, the user may not be aware of the exact query she wants to analyze or all the components of a news event, we provide a query expansion feature in Twit-Digest. When a user searches for a word, we extract all the hashtags from the tweets containing that query word, and present to the user with a tag-cloud of 50 top hashtags related to her based on the TF-IDF score. [3] The size of the word in the tag-cloud indicates its TF-IDF score. Figure 4 represents the tag-cloud for a query.

2.2.4 Statistical and Popular Link Analysis

We also provide the user with cumulative graphs representing various statistical analysis, such as: piechart to represent percentage of tweets versus retweets, scatter chart to show

[3]http://nlp.stanford.edu/IR-book/html/htmledition/tf-idf-weighting-1.html

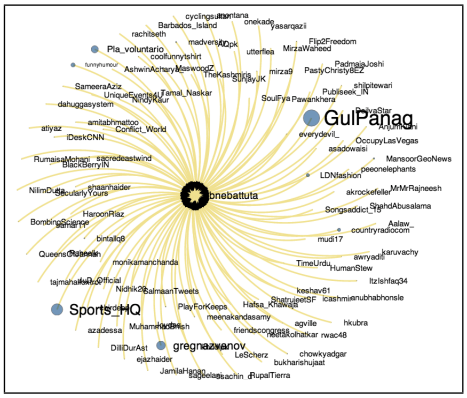

Figure 3: The network graph, with users as nodes and edges representing the retweets and replies by one user to another.

Figure 4: The tag-cloud of hashtags present in the tweets containing the query word.

the relation between number of retweets and follower count, column chart between users and their follower count and lastly, bar chart to show frequency of various URLs present in tweets. In the popular link analysis tab, we provide the link and preview of top URLs (based on frequency of occurrence) posted about the event and also classify them as images, videos and news links.

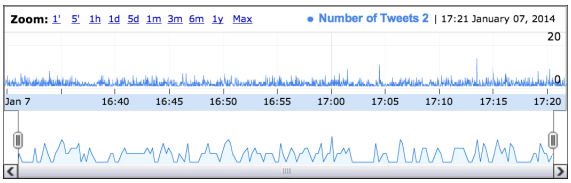

Figure 5: Statistical analysis shows temporal analysis for the activity on Twitter for a query word.

3. REFERENCES

[1] Aditi Gupta and Ponnurangam Kumaraguru. Credibility ranking of tweets during high impact events. In *Workshop on Privacy and Security in Online Social Media*, WWW, 2012.

[2] Anupama Aggarwal, Ashwin Rajadesingan, and Ponnurangam Kumaraguru. Phishari: Automatic realtime phishing detection on twitter. *7th IEEE APWG eCrime Researchers Summit (eCRS)*, 2012.

[3] Fabrício Benevenuto, Tiago Rodrigues, Virgílio Almeida, Jussara Almeida, and Marcos Gonçalves. Detecting spammers and content promoters in online video social networks. In *ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, 2009.

[4] Sidharth Chhabra, Anupama Aggarwal, Fabricio Benevenuto, and Ponnurangam Kumaraguru. Phi.sh/$ocial: the phishing landscape through short urls. In *Collaboration, Electronic messaging, Anti-Abuse and Spam Conference*, CEAS, 2011.