Adaptive Topic Follow-up on Twitter

Abdulrahman Alsaudi*, Mehdi Sadri*, Yasser Altowim†, Sharad Mehrotra*
*University of California, Irvine
†King Abdulaziz City for Science and Technology

***Abstract*—Twitter provides a strictly limited API that makes it difficult for a simple search using pre-defined textual patterns to provide satisfying coverage of the topic of interest. This paper discusses a tweet acquisition system, that queries Twitter API using a set of key phrases, then analyzes the retrieved tweets. In order to achieve better coverage of the searched topic, the system employs an adaptive query generation mechanism that iteratively enriches the set of textual relevant patterns based on the previously collected tweets using an explore-exploit strategy. The paper also demonstrates an application called Topic Follow-up on Twitter (TFT) that is built on top of the acquisition system and aims at linking tweets with online articles. It first extracts a set of key phrases from the submitted news article and then utilizes the acquisition and analysis components of the system. Using this application, we will show how the adaptive searching mechanism of the tweet acquisition system improves the coverage of the topic of interest.**

Video: http://bit.ly/2kqkikB

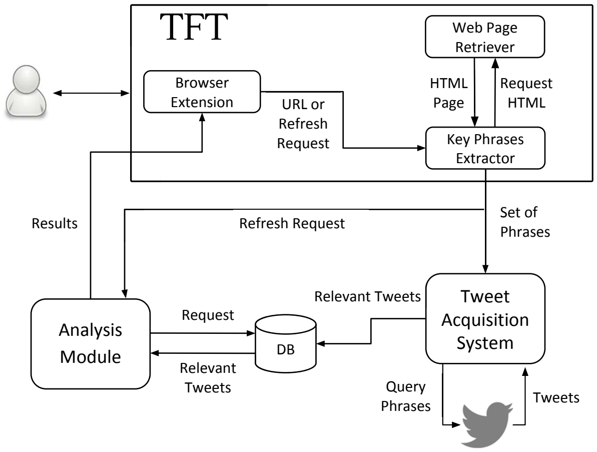


Fig. 1. System architecture.

I. Introduction

Social media applications have recently grabbed attention of millions of users around the globe. Twitter, a microblogging service, is playing a major role on how people communicate in the modern era. The vast growth of data interchanged daily in Twitter makes it an attractive venue for people to use when following up on certain news or topics of interest. For instance, [1] constructed an earthquake reporting system to track earthquake instantly through Twitter as an emergency management application. Other examples include predicting the stock market based on the Twitter mood [2] and predicting elections [3]. Such applications begin by acquiring social media content that is relevant to the topic of interest, and then perform diverse types of application-specific analysis on top of the collected data.

To help develop such applications, we have built a tweet acquisition system [4], that provides an effective method for collecting tweets relevant to a topic of interest. The acquisition system allows each user to specify an initial set of query phrases relevant to the topic of interest. Next, it queries Twitter API using those phrases and then analyzes the resulting tweets to update its internal statistics. The key concept differentiating our tweet acquisition mechanism from other systems [5], [6] is that of the *adaptive searching mechanism*. The tweet acquisition system iteratively enriches the current set of query phrases based on the previously collected tweets returned from the Twitter API. The intuition is that the current patterns existing in the incoming tweets can help retrieve more relevant tweets in the following iterations and that “repetitive co-occurrence” of terms can help identify additional patterns that may improve the quality of the acquisition process.

The tweet acquisition system has the capability of handling multiple requests, each with a set of query phrases, at the same time. It provides a mechanism to choose a set of phrases from each request to query. This is needed since the total number of all phrases cannot exceed what the Twitter API can handle in a single query.

In order to demonstrate the tweet acquisition system, we have built an application called Topic Follow-up on Twitter (TFT) that uses the underlying the tweet acquisition system to augment a web site with relevant tweets. Upon reading an online article discussing a certain topic, the user can use TFT to further explore the topic by viewing other opinions and concerns on Twitter. To this end, TFT first extracts a set of query phrases from the article, and then uses the acquisition system to query Twitter and analyze the retrieved tweets.

II. System Overview

In this section, we will provide an overview of the tweet acquisition system, the analysis module and the TFT application. All of these components have been fully developed using the Java Programming language. Figure 1 illustrates the overall system architecture.

Tweet Acquisition System. The core responsibility of the acquisition system is to adaptively query Twitter and analyze the returned tweets using the analysis module. The tweet acquisition system is implemented as a dynamically adaptive method that deals with multiple topics of interest at the same time. The set of query phrases of each article corresponds to one topic of interest. It runs on iterations to dynamically enrich the query phrases of the topics using an explore-exploit strategy. At the beginning of each iteration, the acquisition system first generates a new set of phrases per topic, where each phrase is associated with a weight representing its relevance to the

2375-026X/17 $31.00 © 2017 IEEE
DOI 10.1109/ICDE.2017.188

topic, and then queries Twitter API using the generated sets of phrases to obtain relevant tweets. The newly generated phrases are derived from the related tweets collected from the past iteration based on a *reinforcement* learning approach. The retrieved tweets will then go through a *relevance checker* to determine which tweets are relevant to which topics. The relevant tweets are then stored in the tweet database.

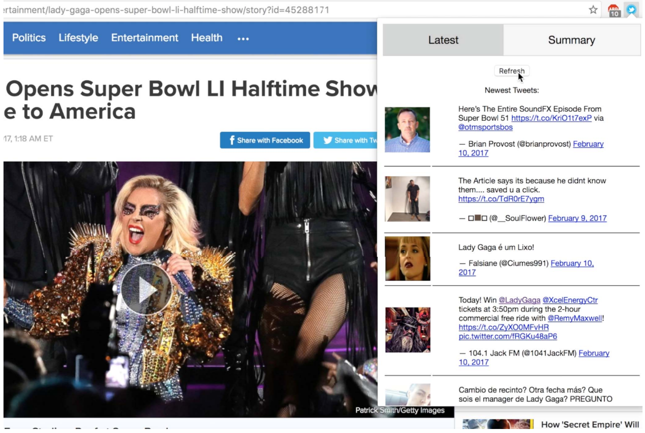
Analysis Module. The main responsibility of the analysis module is to perform various analysis operations on top of relevant retrieved tweets. It is executed by retrieving the relevant tweets for a certain topic of interest from the database and performing several algorithms to analyze and rank them based on some factors including the influentiality of the tweeter, the time and location of the tweet and how many times it has been retweeted. Such algorithms perform different types of analysis tasks such as entity extractions, sentiment analysis or top-k query answering.

Topic Follow-up on Twitter application. TFT is an application built on top of the acquisition system and aims at linking tweets with online articles. Upon reading an article, the user can send a request to TFT to fetch a set of tweets related to the topic of the article. The first request made by the user will be in the form of the URL of the web page containing the article. If the URL has been already submitted, then the user request will be directly sent to the analysis module. Otherwise, the request is handled in three steps: extracting key phrases from the article, dynamically retrieving relevant tweets from Twitter and analyzing the retrieved tweets. TFT provides the user with the ability to choose the form of analysis (e.g., summarization of tweets or sentiment analysis) to be performed on relevant retrieved tweets of the same topic.

In the first step, TFT retrieves the full content of the HTML page of the input URL and performs various operations to determine the set of key phrases that best represent the corresponding article. A pool of phrases that appear in the article document will be collected and then a score value will be assigned to each phrase based on three factors. First, the distance between the words of the phrase in the article. If the words are adjacent to each other, then a higher score will be assigned to the phrase and vice versa. Second, the location of the phrase in the HTML page of the article. Phrases that are present in the title or the headlines are given a higher score. Third, using an entity recognition tool, a phrase containing entities or people names will be assigned a higher score. The overall score of the phrase is determined using a weighted summation model on these individual scores. Then, TFT will determine the set of phrases based on a threshold value that are used to initially query Twitter.

Those phrases will be then sent to the tweet acquisition system and the analysis module which will handle the second and third steps as explained earlier. After displaying the resulting tweets, the user will have the ability to send a *refresh* request to TFT to retrieve even more relevant tweets about the article. The refresh request will be directly sent to the analysis module which will fetch more recent tweets from the database, analyze them and rank them as explained earlier.

III. Demonstration



Newest Tweets:

Here's The Entire SoundFX Episode From Super Bowl 51 https://t.co/KriO1t7exP via @otmsportsbos

— Brian Provost (@brianprovost) February 10, 2017

The Article says its because he didnt know them.... saved u a click. https://t.co/TdR0rE7ygm

— □■□ (@__SoulFlower) February 9, 2017

Lady Gaga é um Lixo!

— Falsiane (@Ciumes991) February 10, 2017

Today! Win @LadyGaga @XcelEnergyCtr tickets at 3:50pm during the 2-hour commercial free ride with @RemyMaxwell! https://t.co/ZyXO0MFvHR pic.twitter.com/fRGKu48aP6

— 104.1 Jack FM (@1041JackFM) February 10, 2017

Cambio de recinto? Otra fecha más? Que sois el manager de Lady Gaga? PREGUNTO

Fig. 2. An example illustrating the use of the tweet acquisition system and TFT in fetching tweets related to the corresponding news article.

In this demonstration, we will be using a Google Chrome Extension as the Browser Extension of TFT. The demonstration will begin by browsing an article on the web. The user will then click the "fetch tweets" button on the extension to request a Twitter report for that article. Then the system begins the retrieval of an initial set of tweets related to the topic of the article. Those tweets will be then displayed on the extension as shown in Figure 2. During that time, the acquisition system will be continually working in the background on enriching the key phrases and populating the database with more relevant tweets. The user will then click the "refresh" button to view a new set of updated tweets. The user will also explore the different types of analysis supported in our system through summary.

We will also be showing that the coverage of the topic will significantly increase as a result of each refresh request, and that the highest possible coverage would be achieved in a few number of such requests. We will also compare our system with a variation of it that lacks the adaptive searching mechanism and will be hence acquiring tweets using a static set of phrases. This experiment will capture the significance of the tweet acquisition system with regard to the dynamic expansion of topic coverage.

IV. Acknowledgments

This work was supported in part by NSF grants CNS-1527536, CNS-1545071, CNS-1450768, CNS-1450768, CNS-1059436 and CNS-1118114.

References

[1] E. S. T. Users, "Real-time event detection by social sensors takeshi sakaki," *Makoto Okazaki, Yutaka Matsuo WWW2010*.

[2] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, 2011.

[3] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment." *ICWSM*, vol. 10, pp. 178–185, 2010.

[4] M. Sadri, S. Mehrotra, and Y. Yu, "Online adaptive topic focused tweet acquisition," in *CIKM*. ACM, 2016, pp. 2353–2358.

[5] W. Guo, H. Li, H. Ji, and M. T. Diab, "Linking tweets to news: A framework to enrich short text data in social media." in *ACL (1)*. Citeseer, 2013, pp. 239–249.

[6] R. Li, S. Wang, C. Chang *et al.*, "Automatic topic-focused monitor for twitter stream," *PVLDB*, pp. 1966–1977, 2014.