RAPID: Real-time Analytics Platform for Interactive Data Mining

Kwan Hui Lim[*‡], Sachini Jayasekara[*], Shanika Karunasekera[*], Aaron Harwood[*], Lucia Falzon[†], John Dunn[†] and Glenn Burgess[†]

[*]The University of Melbourne and [†]Defence Science and Technology, Australia
[‡]Singapore University of Technology and Design
{kwan.lim@,w.jayasekara@student.,karus@,aharwood@}unimelb.edu.au,
{FirstName.LastName}@dst.defence.gov.au

Abstract. Twitter is a popular social networking site that generates a large volume and variety of tweets, thus a key challenge is to filter and track relevant tweets and identify the main topics discussed in real-time. For this purpose, we developed the Real-time Analytics Platform for Interactive Data mining (RAPID) system, which provides an effective data collection mechanism through query expansion, numerous analysis and visualization capabilities for understanding user interactions, tweeting behaviours, discussion topics, and other social patterns.

Keywords: Twitter, Social Networks, Real-time, Topic Tracking

1 Introduction

Social networking sites, such as Twitter, have become a prevalent communication platform in our daily life, with discussions ranging from mainstream topics like TV and music to specialized topics like politics and climate change. Tracking and understanding these discussions provide valuable insights into the general opinions and sentiments towards specific topics and how they change over time, which are useful to researchers, companies, government organizations alike, e.g., advertising, marketing, crisis detection, disaster management. Despite its usefulness, the large volume and wide variety of tweets makes it challenging to track and understand the discussions on these topics [2, 5]. To address these challenges, we proposed and developed the **R**eal-time **A**nalytics **P**latform for **I**nteractive **D**ata mining (RAPID) for topic tracking and analysis on Twitter (Figure 1). RAPID offers a unique topic-tracking capability using query keyword and user expansion to track topics and related discussions, as well as various analytics capabilities to visualize the collected tweets, users and topics, and understand tweeting and interaction behaviours.

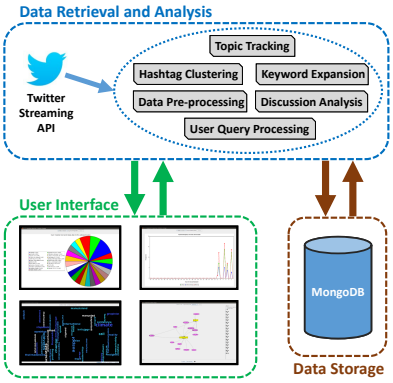

Fig. 1. Overview of RAPID System

Related Systems and Differences. There has been a number of interesting Twitter-based systems developed for specific application domains such as politics [10], crime and disasters [4], diseases [3], recommendations [11], and they typically utilize a mention/keyword-based retrieval of tweets relating to each domain. Others focus on specific capabilities on Twitter such as a SQL-like query language [6], clustering tweets into broad topics [8], detecting events based on keyword frequency [7]. While these systems provide many interesting capabilities, our RAPID system differs in the following ways: (i) Instead of targetting specific domains, RAPID is designed to be generalizable to any application domain, topic or event; (ii) Many earlier systems retrieve tweets based on user-provided keywords, which may not adequately represent the topic of interest. In contrast, RAPID provides a unique query expansion collection capability that allows for the expansion of seeding keywords and users for a broader collection coverage; (iii) In addition, RAPID allows its users to interact with and control the data stream in real-time, as well as perform a wide and in-depth range of analysis and visualizations techniques, which we further describe in this paper; and (iv) RAPID is highly scalable to the growing volume of tweets generated, by utilizing real-time distributed computing technologies like Apache Storm and Kafka, compared to earlier systems that do not utilize such technologies.

2 System Architecture

RAPID is developed to perform real-time analysis and visualization, as well as post-hoc analysis and visualization on previously collected data. Communication between the client and server are facilitated through Kafka queues, based on the publish-subscribe model where researchers are able to specify their various information requirements. We now describe the main components of RAPID.

Data retrieval and analysis component. This component performs two main tasks, which are:

- **Data Retrieval**. For real-time retrieval, RAPID interfaces with the Twitter Streaming API and collects information such as tweets related to a particular topic, posted by specific users or are within a geo location subscribed by the user, Twitter user details such as the list of followers, profile information and timeline information. For post-hoc processing, RAPID retrieves information stored in the data storage unit based on the researcher's requests. The researcher is able to access all functionalities of the real-time retrieval and in addition, is able to further drill-down on the data by filtering the collected tweets based on specific topics, time periods, locations and set of hashtags. Unlike many earlier systems, RAPID is designed with an integrated data retrieval and analysis capability such that the data retrieval is continuously expanded for better coverage based on real-time analysis of collected tweets, which we discuss next.

- **Data Analysis**. This includes the sub-tasks of: (i) tweet pre-processing, i.e., tokenizing, topic labelling, extraction of geo-location and other tweet features; (ii) topic tracking via keywords, usernames or bounding boxes, and an enhanced query expansion capability that automatically track topics and related discus-

sions through dynamic expansion of keywords; (iii) user query processing, such as filtering and drilling down the collected data for further analysis based on topics, time periods and/or locations; and (iv) data statistics and analysis, such as updating data storage with latest collection statistics and performing advanced analytics like analyzing hashtags and inferring relationships between hashtags, analyzing word-to-word pairs and word clusters of tweets, tracking discussions through pro-actively fetching tweets replies related to discussions.

Data Storage. The data storage component uses MongoDB for storing meta-data as well as the processed tweets, which can be used later for further post-hoc processing and visualization. RAPID also allows users the freedom to decide the type of processed data that should be persisted in the storage. Meta-data stored in the database includes the details of the users, details of user activities such as commands given by users to the RAPID system and the topics users are subscribed to. In addition to the meta-data, tweets processed by the system, discussions occurred related to tweets can also be stored in the database. One major advantage of having this useful capability is that users can reprocess and visualize the tweets later if such requirement arises, e.g., further drill-down to filter and analyze crisis-related tweets posted on 20 Nov 2017 in Melbourne CBD.

User Interface Component. This component performs three main tasks, namely:

- **User Input**. For topic tracking, researchers can specify a set of keywords, users and/or geo-bounding boxes associated with the topic as the input. The interface also allows users to modify or delete existing tracked topics, with a detailed log of these activities.

- **Real-time Visualization**. Key information and statistics of the tracked topics are visualized using a set of predefined charts, which are updated in real-time as new tweets related to the topic are analyzed by the RAPID system. Screenshots and descriptions of selected charts are shown in Figure 2.

- **Workbench**. The workbench allows users to visualize tweets that have been stored in the storage component for further analysis. For more flexibility in post-collection analysis, users are able to define a specific time period the tweets have occurred and then the workbench retrieves the related tweets and visualizes them using the same charts used for real-time visualization. Moreover, the workbench summarizes the key statistics of the retrieved tweets including the number of tweets fetched, number of unique authors, unique hashtags, unique mentions and unique replies.

3 Target Users and Demonstration

We presented the RAPID system for real-time topic tracking and analysis on Twitter, where RAPID offers a unique and effective collection approach via query expansion, numerous analysis capabilities to understand user interactions, tweeting behaviours and discussion cascades, and various visualizations of these types of information. RAPID has been used by researchers from both the Army Research Laboratory in the USA and Defence Science and Technology

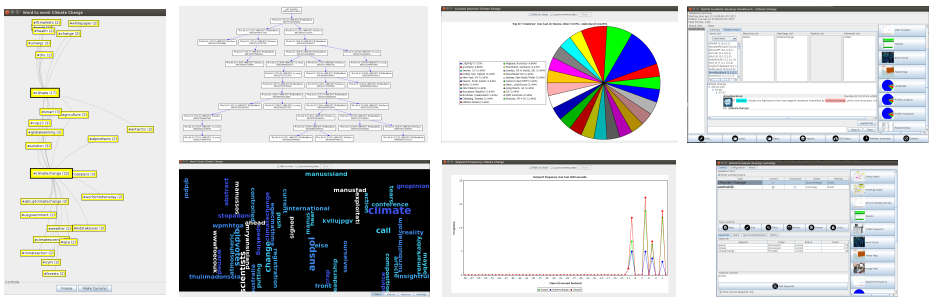

Fig. 2. Screenshots of selected RAPID functionalities, which include (clockwise from left): word-to-word network graph, discussion tree visualization, profile of tweeting locations, overview of RAPID workbench, workbench view on query expansion, real-time tracking of keyword frequency, word cloud of collected tweets.

in Australia [1, 9], and will also be of interest to any user interested in tracking, analysing and visualizing topics on Twitter. We will demonstrate the various capabilities of RAPID via use cases of political campaign analysis, monitoring of crises and incidents, in-depth analysis of tweets and users. A demonstration video of RAPID is available at `https://youtu.be/1APLeLT_t8w`.

Acknowledgments. This research is supported by Defence Science and Technology.

References

1. Falzon, L., McCurrie, C., Dunn, J.: Representation and analysis of twitter activity: A dynamic network perspective. In: Proc. of ASONAM'17 (2017)
2. Kumar, S., Morstatter, F., Liu, H.: Twitter Data Analytics. Springer (2013)
3. Lee, K., Agrawal, A., Choudhary, A.: Real-time disease surveillance using twitter data: demonstration on flu and cancer. In: Proc. of KDD'13 (2013)
4. Li, R., Lei, K.H., Khadiwala, R., Chang, K.C.C.: TEDAS: A Twitter-based Event Detection and Analysis System. In: Proc. of ICDE'12 (2012)
5. Liao, Y., Moshtaghi, M., Han, B., Karunasekera, S., Kotagiri, R., Baldwin, T., Harwood, A., Pattison, P.: Mining micro-blogs: Opportunities and challenges. In: Social Networks: Computational Aspects and Mining (2011)
6. Marcus, A., Bernstein, M.S., Badar, O., Karger, D.R., Madden, S., Miller, R.C.: Tweets as data: demonstration of TweeQL and TwitInfo. In: SIGMOD'11 (2011)
7. Mathioudakis, M., Koudas, N.: TwitterMonitor: trend detection over the Twitter stream. In: Proc. of SIGMOD'10. pp. 1155–1158 (2010)
8. O'Connor, B., Krieger, M., Ahn, D.: TweetMotif: Exploratory Search and Topic Summarization for Twitter. In: Proc. of ICWSM'10 (2010)
9. Vanni, M., Kase, S.E., Karunasekara, S., Falzon, L., Harwood, A.: RAPID: real-time analytics platform for interactive data-mining in a decision support scenario. In: Proc. of SPIE 10207 (2017)
10. Wang, H., Can, D., Kazemzadeh, A., Bar, F., Narayanan, S.: A system for real-time twitter sentiment analysis of 2012 US presidential election cycle. In: Proc. of ACL'12. pp. 115–120 (2012)
11. Wang, J., Feng, Y., Naghizade, E., Rashidi, L., Lim, K.H., Lee, K.E.: Happiness is a Choice: Sentiment and Activity-Aware Location Recommendation. In: Proc. of WWW'18 Companion. pp. 1401–1405 (2018)