Graph-Based Methods for Clustering Topics of Interest in Twitter

Hugo Hromic(✉), Narumol Prangnawarat, Ioana Hulpuş, Marcel Karnstedt, and Conor Hayes

Insight Centre for Data Analytics, National University of Ireland Galway (NUIG), Galway, Ireland
{hugo.hromic,narumol.prangnawarat,ioana.hulpus,marcel.karnstedt, conor.hayes}@insight-centre.org
http://www.insight-centre.org

Abstract. Online Social Media provides real-time information about events and news in the physical world. A challenging problem is then to identify in a timely manner the few relevant bits of information in these massive and fast-paced streams. Most of the current topic clustering and event detection methods focus on user generated *content*, hence they are sensible to language, writing style and are usually expensive to compute. Instead, our approach focuses on mining the *structure* of the graph generated by the interactions between users. Our hypothesis is that bursts in user interest for particular topics and events are reflected by corresponding changes in the structure of the discussion dynamics. We show that our method is capable of effectively identifying event topics in Twitter ground truth data, while offering better overall performance than a purely content-based method based on LDA topic models.

1 Introduction

Twitter is possibly today the most widely used *microblogging* system in the world, allowing for real-time broadcasting of short messages (or *Tweets*) among friends and followers. This vast stream of content, despite containing a large amount of noisy data, also contains relevant and updated information [6,7]. Being able to timely identify these topical "gold nuggets" within busy social streams becomes essential to help users discover potentially interesting content.

The majority of existing approaches for topic finding in Twitter focus on [1]: (a) *textual features*, e.g. using topic models [9], keeping track of bursty words [3] or clustering trending Tweets [5], and (b) *activity dynamics*, e.g. monitoring keyword usage patterns [10] or analysing Tweet/Retweet interactions [2,12]. On the one hand, text-based approaches are inherently sensitive to the writing style and language (e.g. English, Chinese), where colloquial expressions dominate and are often expensive to process. On the other hand, activity-based approaches are faster but less effective in presence of noise and mostly dependant on the a priori chosen seed keywords or terms.

In this paper we propose to instead focus on a more efficient *structure-based* approach, which can be less expensive to process than text-based techniques,

P. Cimiano et al. (Eds.): ICWE 2015, LNCS 9114, pp. 701–704, 2015.
DOI: 10.1007/978-3-319-19890-3_61

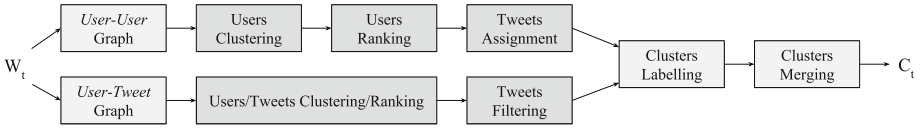


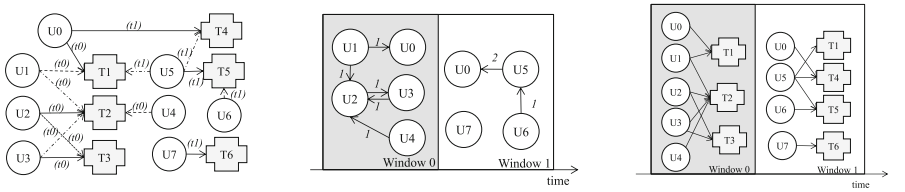
Fig. 1. Per-window processing pipeline for topics clustering. $W_t = \{tw_0, tw_1, \cdots, tw_n\}$ is a window of n Tweets at a time t and C_t is a resulting topics clustering configuration for the same time.

thanks to a simpler graph model; and less sensitive to noise compared to an activity-based approach that ignores the connection between user interactions.

It is well understood that events and topics generate bursts in Twitter activity [2]. Thus, we hypothesise that by modelling this activity as edges in a *Twitter interaction graph* we can capture bursty topics and events by using graph analysis methods. Therefore, we mine for groups of tightly connected users and Tweets under the assumption that these groups represent an emerged topic or event.

2 Graph-Based Pipeline for Topics Clustering

Our topic clustering approach is built around a processing pipeline (Figure 1) where an incoming stream of Tweets is received from Twitter and aggregated using a sliding window approach to generate two alternative graph representations of Twitter interactions (Figure 2(a)). The first is a **User-to-User (UU)** perspective (Figure 2(b)), where edges represent links between the author of a Tweet and all the mentioned, retweeted or replied users in it. The second is a bipartite **User-to-Tweet (UT)** view (Figure 2(c)), where edges link users to their posted Tweets, replies, Retweets, or posts in which they have been mentioned.



(a) Twitter Interactions (b) Homogeneous Graph Representation (UU) (c) Heterogeneous Graph Representation (UT)

Fig. 2. Graph models for Twitter interactions across processing windows. U-nodes denote users, T-nodes denote Tweets. Interaction types (edges): solid → *tweeted*, dashed → *mentionedIn*, dotted → *repliedIn* and dashed/dotted → *retweeted*.

Topic clusters are extracted from the above networks. For the UU Graph approach, we use the OSLOM community finding algorithm [8] to produce a set of user communities based on tightly interacting users. The users inside each community are ranked using the PageRank algorithm and their latest Tweets are selected to form clusters of Tweets. For the heterogeneous UT Graph, we use

the RankClus algorithm [14] to build clusters of ranked Tweets. Tweets ranked lower than a given threshold are removed.

At this stage, the resulting clusters from both of our graph models do not have any topic information. For this we perform two post-processing steps for labelling and merging topically similar groups. First, we label each Tweet cluster by the top-k most frequent hashtags occurring in its Tweets. If no hashtag is found, then we extract the top-k most frequent named entities using the Python NLTK library[1]. Second, to minimise topic redundancy among clusters, i.e. those with similar labels, we combine the clusters that have the same first-n labels, ordered by their usage frequency.

3 Experiments and Results

To test our approach we use a third-party human annotated ground truth Twitter dataset containing a number of known public events, that we consider as *topics*, and their associated Tweets [11]. We constructed 28 day-long sliding windows and extracted labelled topic clusters for each. These were generated by our proposed pipeline, as well as a topic model approach using LDA [9] which serves as our text-based baseline. Our evaluation measured two aspects: clustering quality (using the F_1-Measure of Precision and Recall, see Figure 3) and runtime performance (Figure 4). In both experiments, our structure-based approach outperforms the baseline.

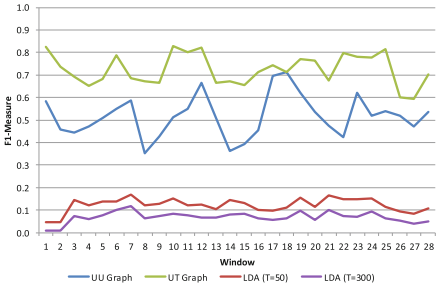


Fig. 3. Per-window F_1-Measure for both of our network types and the LDA topic models method (using two settings)

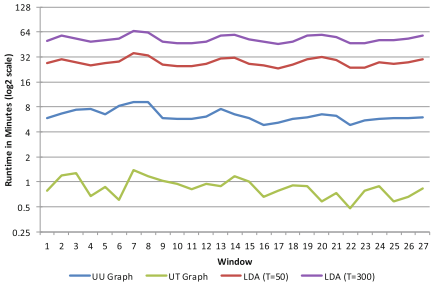


Fig. 4. Per-window runtime performances – in log scale – for all the studied methods (graph- and text-based)

4 Conclusions and Future Work

We proposed a graph-based processing pipeline for clustering topics of interest in Twitter. For this, we presented two different types of graphs for modelling interactions between users, one that only represents User-to-User actions, and another that captures the relations between users and Tweets. Based on the homogeneous

[1] Available in http://www.nltk.org

model, in 2012 we successfully developed *Whassappi* [4], a prototype mobile application for topic finding aimed at the visitors of the final leg of the Volvo Ocean Race 2012 in Ireland. We experimented with two state of the art network clustering algorithms, one for each type of graph. Our experiments and results support our hypothesis that analysis of user interactions through graph mining reveal discussions that ultimately correlate with human annotated events. Moreover, our approach outperformed a baseline text-based LDA topic model technique. Our study opened some interesting research questions that we plan to address in the future. For example, we noticed distinctive graph patterns for various types of events: the properties of the graph clusters and their nodes might potentially be used for classifying events. In this regard, one of our future objectives is to devise methods able to describe events, for instance to distinguish between local or personal topics, and events of world-wide interest [13].

Acknowledgements. This work was supported by Science Foundation Ireland (SFI) partly under Grant No. 08/SRC/I1407 (Clique) and partly under Grant No. 12/RC/2289 (Insight Centre for Data Analytics).

References

1. Atefeh, F., Khreich, W.: A survey of techniques for event detection in twitter. In: Computational Intelligence. Wiley Online Library (2013)
2. Chierichetti, F., et al.: Event detection via communication pattern analysis. In: Proc. of ICWSM, pp. 51–60. AAAI (2014)
3. Fung, G.P.C., et al.: Parameter free bursty events detection in text streams. In: Proc. of VLDB, pp. 181–192. VLDB Endowment (2005)
4. Hromic, H., et al.: Event panning in a stream of big data. In: Knowledge Discovery and Machine Learning Workshop in LWA (2012)
5. Hu, Y., et al.: Whoo.ly: facilitating information seeking for hyperlocal communities using social media. In: Proc. of SIGCHI, pp. 3481–3490. ACM (2013)
6. Hurlock, J., et al.: Searching twitter: separating the tweet from the chaff. In: Proc. of ICWSM, pp. 161–168. AAAI (2011)
7. Kwak, H., et al.: What is twitter, a social network or a news media? In: Proc. of WWW, pp. 591–600. ACM (2010)
8. Lancichinetti, A., et al.: Finding statistically significant communities in networks. In: PloS one, vol. 6, p. e18961. Public Library of Science (2011)
9. Lau, J.H., et al.: On-line trend analysis with topic models: #twitter trends detection topic model online. In: Proc. of COLING, pp. 1519–1534. Citeseer (2012)
10. Marcus, A., et al.: TwitInfo: aggregating and visualizing microblogs for event exploration. In: Proc. of SIGCHI, pp. 227–236. ACM (2011)
11. McMinn, A.J., et al.: Building a large-scale corpus for evaluating event detection on twitter. In: Proc. of CIKM, pp. 409–418. ACM (2013)
12. Popescu, A.M., et al.: Detecting controversial events from twitter. In: Proc. of CIKM, pp. 1873–1876. ACM (2010)
13. Prangnawarat, N., et al.: Event analysis in social media using clustering of heterogeneous information networks. In: Proc. of FLAIRS. AAAI (2015)
14. Sun, Y., et al.: RankClus: integrating clustering with ranking for heterogeneous information network analysis. In: Proc. of EDBT/ICDT, pp. 565–576. ACM (2009)