Expert Systems With Applications 55 (2016) 351–360

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

# A rule dynamics approach to event detection in Twitter with its application to sports and politics

Mariam Adedoyin-Olowe[a], Mohamed Medhat Gaber[a,*], Carlos M. Dancausa[a], Frederic Stahl[b], João Bártolo Gomes[c]

[a] School of Computing Science and Digital Media, Robert Gordon Univeristy, Aberdeen, AB10 7GJ, UK
[b] School of Systems Engineering, University of Reading, PO Box 225, Whiteknights, Reading, RG6 6AY, UK
[c] DataRobot Inc., Singapore

ARTICLE INFO

*Keywords:*
Twitter
Hashtags
Associations rules
Rule matching
Rule mapping

ABSTRACT

The increasing popularity of Twitter as social network tool for opinion expression as well as information retrieval has resulted in the need to derive computational means to detect and track relevant topics/events in the network. The application of topic detection and tracking methods to tweets enable users to extract newsworthy content from the vast and somehow chaotic Twitter stream. In this paper, we apply our technique named *Transaction-based Rule Change Mining* to extract newsworthy hashtag keywords present in tweets from two different domains namely; sports (The English FA Cup 2012) and politics (US Presidential Elections 2012 and Super Tuesday 2012). Noting the peculiar nature of event dynamics in these two domains, we apply different time-windows and update rates to each of the datasets in order to study their impact on performance. The performance effectiveness results reveal that our approach is able to accurately detect and track newsworthy content. In addition, the results show that the adaptation of the time-window exhibits better performance especially on the sports dataset, which can be attributed to the usually shorter duration of football events.



## 1. Introduction

There has being a surge in Twitter activities (Li, Vishwanath, & Rao, 2014) since its launch in 2006 as well as steady increase in event detection awareness on the network (Lau, Collier, & Baldwin, 2012) in recent times. With over 645 million subscribers (Beaupierre, 2014), Twitter has continued to grow both in size and activity. The network is rapidly changing the way the global audience sources for information and thus influences the process of journalism (Lasorsa, Lewis, & Holton, 2012; Newman, 2009). Twitter is becoming an information network rather than just a social media when compared with other social networks such as Facebook and Tumblr. This explains why traditional news media follow activities on Twitter to enhance their news reports and updates. News media like *BBC* or *CNN* that contain full story they broadcast on their Twitter page thereby giving readers the opportunity of reading the full story. Breaking news are sometimes posted on Twitter before they are published by traditional news media due to users' nearness to the location of events (Castillo, Mendoza, & Poblete, 2011; Cataldi, Di Caro, & Schifanella, 2010). An example of such a situation is the news of the death of America's female pop star Whitney Houston, which was posted on Twitter before its broadcast on news media (Whiting, Zhou, Jose, Alonso, & Leelanupab, 2012) as a breaking news. The dynamic and streaming nature of Twitter data (known as tweet) also includes noise on the network resulting in the difficulty of manually obtaining meaningful contents from Twitter. Where some tweets are relevant to specific real life events and are worthy of being broadcasted, others constitute noise (Naaman, Boase, & Lai, 2010). It shows the need for filtering in order to extract relevant tweets from Twitter. According to (Allan, 2002a), a topic as defined in Topic Detection and Tracking (TDT) context can be "*a set of news stories that are strongly related by some similar events*". Events often trigger topics; for instance, breaking news about the sudden death of a public figure will trigger other related news such as police investigations into the cause of death, trail of suspects, arrest and trials of suspect. All these unfolding events will generate news updates and result in the evolvement of related topics. TDT methods are currently used

* Corresponding author. Tel.: +44 7400402591.
*E-mail addresses:* m.a.adedoyin-olowe@rgu.ac.uk (M. Adedoyin-Olowe), m.gaber1@rgu.ac.uk (M.M. Gaber), c.j.martin-dancausa@rgu.ac.uk (C.M. Dancausa), F.T.Stahl@reading.ac.uk (F. Stahl), joao@datarobot.com (J.B. Gomes).

http://dx.doi.org/10.1016/j.eswa.2016.02.028
0957-4174/

to detect and track trending events on Twitter over time (Aiello et al., 2013).

In our previous work (Adedoyin-Olowe, Gaber, & Stahl, 2013; Gomes, Adedoyin-Olowe, Gaber, & Stahl, 2013) we analysed hashtag keywords in tweets on the same topic at 2 consecutive time periods using Association Rule Mining *(ARM)* and Transaction-based Rule Change Mining *(TRCM)*. Our TRCM method was able to identify 4 temporal Association Rules *(ARs)* relating to evolving concept of tweets. The identified *ARs* are namely; "**New rules**", "**Emerging rules**", "**Unexpected Consequent/Conditional rules**" and "**Dead rules**". The results of our previous experiments (Adedoyin-Olowe et al., 2013; Gomes et al., 2013) relates the identified *ARs* to evolving events in real life. To maintain coherence in this paper, *ARM, ARs and TRCM* concepts will be explained in subsequent sections.

In this paper we automate the detection of real life topics generated in 3 Twitter datasets from 2 different domains; sports (the English FA Cup Final 2012) and politics (US Presidential Elections 2012 and US Super Tuesday 2012). We map all hashtag keywords extracted by our system during training process to related topics from carefully chosen ground truth to ascertain a match and subsequently to validate our system's performance. A match is said to have occurred if the time-slot of an extracted hashtag keyword correlates with the time of event occurrence in the ground truth. We evaluate how the dynamics of each dataset affects our experimental results. For performance effectiveness analysis of our method, we consider precision over recall. This is because we are more concerned with generating relevant hashtag keywords (precision) that are related to targeted real life topics/events. As far as we are aware of, *TRCM* is the only method that detects topics from Twitter using hashtags and *ARM*.

We list the contributions of this paper as follows:

- Automation of event detection and tracking in Twitter in one cohesive computational framework, different compared with earlier work that separated event detection and tracking;
- Application of the proposed methods on datasets of different nature of dynamism (from the very dynamic in sports to the slow unfolding events in politics);
- Providing proof of universality of our proposed methods in a number of application domains;
- Concluding insightful application-oriented guidelines as to the importance of the different types of the rules to the application domain.

The rest of the paper is organised as follows: Section 2 discusses other topic detection methods already employed on Twitter data. Section 3 presents the notation of terms used in the paper, while Section 4 gives an overview of the development of *TRCM* architectural framework. Section 5 explains trend analysis of rules in tweets hashtags. Section 6 describes the methodology used in this work, while Section 7 presents our experimental set-up. Section 8 evaluates the experimental results and the paper is concluded with a discussion in Section 9.

## 2. Related work

TDT methods can be used to extract interesting topics from Twitter streaming data and present patterns that demonstrate a representation of specific real life topics. This is achieved by mapping detected results to real life news/events and subsequently tracking the evolvements of such topics. Since Twitter streams high volume of data very rapidly, it is important to apply TDT to Twitter data in order to organise this large volume of data in a meaningful way. There is very limited work on the application of *ARM* as TDT method on Twitter data. Diverse TDT methods are being used to detect relevant events and news topics embedded in online tweets. Events tweets are robust ranging from sports (Guzman & Poblete, 2013; van Oorschot, van Erp, & Dijkshoorn, 2012), politics (Ausserhofer & Maireder, 2013), stock market (McCreadie, Macdonald, Ounis, Osborne, & Petrovic, 2013). The N-grams method effectively captures intricate combination of tweets' keywords in real life topics of diverse composite and time scale by recognising the trend in the topics (Aiello et al., 2013). Other TDT methods are applied to tweets to analyse real life events and occurrences such as sparsely reported events (Agarwal, Vaithiyanathan, Sharma, & Shroff, 2012), differentiating between real world events and non-event tweets (Becker, Naaman, & Gravano, 2011). These methods are also capable of monitoring topic trends (emerging topics) on Twitter in real time (Mathioudakis & Koudas, 2010). Our method not only monitors the trend of emerging topics in real life, it also detects and tracks any changes in the flow of the detected topic or event. Scalable distributed event detection (McCreadie et al., 2013) as well as characterising emerging trends (Naaman, Becker, & Gravano, 2011) have also been conducted on Twitter data. Similarly, TDT methods detect and track breaking news (Phuvipadawat & Murata, 2010) and first mention of story often referred to as *first story* (Osborne, Petrovic, McCreadie, Macdonald, & Ounis, 2012) on Twitter. Our method is holistic in that it detects and tracks different types of topics/events either breaking news or emerging stories. TDT methods are also trained to predict the outcome of national elections (Tumasjan, Sprenger, Sandner, & Welpe, 2010) and to detect local events posted on Twitter (Watanabe, Ochi, Okabe, & Onai, 2011).

Becker, Iter, Naaman, and Gravano (2012) used an online clustering and filtering framework to distinguish between messages about real life events and non-events. The framework clusters subsequent tweet-based messages using their similarity with existing clusters. On the other hand, graph-based approaches can detect keyword clusters in tweets based on their pairwise comparison (Aiello et al., 2013; Inouye & Kalita, 2011). This can be a term unison graph with nodes clustered and the use of community detection algorithm based on betweenness centrality (Sayyadi, Hurst, & Maykov, 2009). Graph-based methods can also be applied to evaluate the effectiveness of topic extraction from tweets (Meng et al., 2012). Jackoway, Samet, and Sankaranarayanan (2011) used a clustering technique to detect events using a text classifier. Phuvipadawat and Murata (2010) proposed a method for collecting, grouping, ranking and tracking breaking news in Twitter. They built a framework named 'Hotstream' to enable users to discover breaking news from Twitter timeline. Other approaches considered first story detection on the network. First story detection structures are created on the basis of documents as vectors within a duration using term frequencies (Allan, 2002b; Yang & Honavar, 1998). Distance measurement is used to detect first story, this is obtained by comparing new documents to their nearest neighbour by measuring their distance gap. Documents with distance that exceed a pre-defined maximum value are considered as first story. This method collects all document term frequencies in memory and detects the nearest neighbour for in-coming documents (Indyk & Motwani, 1998). Tweets pertaining to a planned real life event are distinguished from the stream of non-event tweets (Becker et al., 2011) using an incremental online clustering algorithm. This scalable algorithm clusters a huge volume of Twitter messages without prior knowledge of the number of clusters. An incremental clustering algorithm is applied during training phase to place each message in a related existing cluster. Any new message that is not similar to the ones in an existing cluster forms a new cluster (Becker et al., 2011). An improved Locality Sensitive Hashing (LSH) was proposed by Petrović, Osborne, and Lavrenko (2010) to search for nearest neighbour enhancement that satisfies the data stream mining prerequisites using constant size buckets.