On the cutting edge of event detection from social streams – a non-exhaustive survey

Albrecht Zimmermann
LIRIS CNRS UMR 5205
albrecht.zimmermann@insa-lyon.fr

October 3, 2014

Abstract

Event detection from streams of text data has been a well-traveled research area for the last decade, resulting in a large body of work. When it comes to detecting events from social media stream, in particular breaking news, not all of these techniques are applicable, since only limited amounts of text may be available, data streams in continuously, and detection has to be performed quickly. The aim of this survey is to give the reader an overview of the current state-of-the-art, not only in terms of event detection itself but also w.r.t. event characterization, and spatial and social localization/contextualization. As such, this is very much a *non-exhaustive* survey: we discuss mainly rather recent work that has been shown to outperform earlier approaches and therefore supersedes them, and should, for now, be considered the "cutting edge".

1 Introduction

The question of how to identify events in streams of text data has been a research topic for more than a decade now, starting from e-mail, via blog posts, to Twitter messages. The motivations are manifold: disaster detection and warning, identification of newsworthy events that traditional media are slow to pick up, identification of trends that traditional media is not giving a lot of attention (yet), monitoring of brand perception etc.

The general idea underlying most of this work is identifying "bursty" topics, i.e. topics that are mentioned significantly more often during a (not too short) time period than in the period preceding it. In "retroactive event detection", this is extended to a difference to the period following it as well. For a number of reasons, this task is typically addressed by identifying bursty words. We'll summarize the state of the art w.r.t. that step in the following section.

Detecting individual terms is not enough, however: they need to be assembled into complete topics that can be made sense of, and that can be used to identify the corresponding events. This will be discussed in Section 3.

The summaries in those two sections will be informed by the assumption that event detection should happen real-time, excluding certain techniques suited for off-line work. We are making further assumptions w.r.t. the usage scenario:

1. Users are more interested in finding real-world, or "exogenous" events than in (Twitter-)endogenous ones such as tweets by celebrities.

2. Users are interested in discriminating between breaking news and other exogenous events such as (long-planned) sports events or concerts.

3. Users are interested in discriminating between local and global events.

4. Users are interested in characterizing events by certain regions and by the users involved/interested/affected by certain events.

The rationale for these assumptions is that the identified events will be used to inform users in certain cities (and certain quarters in those cities) and/or users with particular characteristics of identified events. The first two points will be discussed in Section 4. The work discussed therein also contributes to the third point yet the majority of work relating to points three and four can be found in Section 5.

2 Detecting bursty terms

The first paper that, for e-mails, proposed the idea of identifying bursty terms and using them to identify events was [7]. Notably, this paper is *badly* suited for real-time event identification: its method consists of learning finite state automata for *each word* of a dictionary over a large time-stamped corpus of text (a text stream). The reason we list it is that a) it's effectively impossible to avoid Kleinberg when discussing this topic, and b) he proposed the idea of "nested" bursts that I have not seen in any of the other work.

In recent years, more efficient proposals have been made, culminating (so far) in [6, 14]. The former uses a Fourier transform and identifies cutoff values to separate different feature classes. They model the time period of the burst explicitly. Real-time identification can be done by, for instance, sampling every ten minutes (and building on a large corpus of past Twitter messages). The latter uses a Wavelet transformation, which removes the need to model time periods explicitly. Real-time even identification would be done in the same way. Changing parameter settings leads to more finely or coarsely grained change detection at a higher/lower computational cost.

Generally speaking, the work that has been superseded by such methods used *clustering* of documents over the entire past stream to identify topics *retroactively*. Retroactive methods might still be useful to identify recurring events (such as soccer matches or flea markets) without the trouble of having to assemble the topic itself from the identified bursty terms.

3 Identifying topics/events

Because the main advantage of the clustering approaches lies in the fact that they directly group whole documents (tweets) together. Approaches that identify individual bursty terms, however, need to find a way of grouping such terms, use those groupings to query the messages, and then find a way of identifying the topic/event. This will be particularly important if we want to identify additional media related to events, e.g. links in messages, tweeted pictures etc.

The first of those steps has been addressed already: [6] uses Kullback-Leibler divergence to calculate the correlation between feature terms, while [14] describe how to use cross correlation. In particular, both those methods allow to identify meaningless terms, i.e. stop-words and terms that are at no point in time bursty, and reduces later computation time. Especially in the latter work, however, the authors show an example of two bursty terms grouped together that refer to unrelated events but have very similar burstiness profile. Such a mistake can be discovered rather easily when querying messages including the combination of terms and coming up empty. The authors of the latter work followed up with an approach that models topics explicitly via hidden variables [5], which requires pre-defining the number of topics, however.

But how to get from a grouping of terms/messages to a semantically meaningful event automatically is still pretty much unsolved (especially in the case of messages of only 140 characters). The best solution for this problem remains using humans that collate messages and decide on a meaning (ideally more than one to get some failure prevention).[1] This could be supported by using the same search terms for collecting articles from the web but particularly for breaking news this will not really work. The authors of [9] have proposed querying Wikipedia pages instead of news articles and tagging only those topics as events of which viewing numbers spike.

4 Characterizing events

To support humans in labeling events, it is necessary to reduce the amount of potential events they have to wade through. The work in [6] takes a step in this direction by identifying thresholds that allow them to separate bursty terms into those relating to

- Little reported aperiodic events – e.g. locally important breaking news, such as a surprise concert on a square.
- Important aperiodic events – e.g. the death of a celebrity.
- Important periodic events – e.g. national soccer matches.

and separate those in term from noisy terms and stop words. Hence, this method could be used to group types of events directly, and for instance ignore periodic events during interpretation.

Other authors have proposed learning classifiers (see e.g. [3]) that separate events from non-events. Unfortunately, this idea suffers from several bottle necks: 1) one needs to define meaningful attributes, 2) one needs to have a collection of both event and non-event instances (requiring curation of an existing corpus), and 3) it still does not separate different types of events. A well-founded exploration of especially the first and third issues can be found in [8]. The authors define three discrimination settings that are of interest to us:

1. Exogenous vs. endogenous trends: the former come from "outside" the Twitter universe, a.k.a. the real world, the latter from inside, e.g. controversial statements from people with lots of followers.

[1] Incidentally, this is the method chosen in basically all work on event detection when evaluating the quality of proposed approaches.

2. Breaking news vs. other exogenous events: this includes, for instance, the discrimination between soccer matches and earth quakes.

3. Local events vs. other exogenous events.

and formulate hypothesis w.r.t. differences in a number of characteristics related to content, interaction (e.g. retweets, mentions), time, authorship, and social networks. While they do not find all of their hypotheses borne out, there seem to be clear enough differences to separate those different categories.

There are additional ways of characterizing events: [10], for instance, proposed identifying how *controversial* certain events are, which can help to fine tune recommendation for individual users, or offers to corporate users (which can be expected to be controversy-averse).

5 Geospatial/social localization

The final building block consists of identifying who might be interested in recommendations in the first place, i.e. identifying the area or social groups for whom recommendations were relevant. The case study for such a system can be found in [11] – the authors describe an automated system for earthquake warnings based on Twitter messages. In addition to the temporal and content information of tweets, they use GPS and/or location information of Twitter profiles to triangulate the epicenter, identify the likely trajectory, and target accounts in its path. It should be obvious that without location information, such an approach will not work, which means that at the latest at this point, this becomes a data problem in addition to everything else.

How to acquire and use location information is somewhat up to choice:

1. In addition to using GPS/profile information, one could attempt to tag profiles with their current location based on tweet contents. This adds an additional layer of text (or image) analysis.

2. Event detection could be limited to tweets originating in a certain area (country, city, quarter).

3. Events could be spatially situated based on tweet contents in addition to profile information.

4. Events could be spatially situated based solely on tweet contents and used to inform people of goings-on in their home town, for example (or any other place they have expressed interest in, for instance because they intend to travel there).

This is very much an on-going research topic, and newer/complementary approaches can be found (e.g. [12]).

In a similar vein, events can be *socially* located. The additional data needed to do this is easier to collect: twitter users have followers and are following other twitter users themselves, they retweet, and they can address other twitter users directly. This aids, among other things, topic identification – users that are connected to each other are more likely to discuss the same topics, meaning that bursts in a social network are more likely to belong together. This gets exploited in [2], a return to clustering messages for event detection, in which

similarity involves not only the content of messages but also the involved users. The paper proposes an on-line algorithm, getting around the issues with earlier clustering approaches, and identifies bursts by the creation of new clusters (and the removal of "stale" ones). This is a potentially very attractive approach since it delivers the combined terms and the social component for free – it suffers from having to specify the number of clusters, however.

Notably, [8] also use social network information for discriminating types of events, and [4] proposes to exploit users' "authority" to give additional weight to certain bursty terms. Finally, to augment the collection of geospatial/social information, detecting temporally correlated events [13] could prove useful. By using the correlation of events, one can tap the location/social information of non-English speakers, and potentially make recommendation to such users as well (for instance with prepared templates for planned events).

6 Summary

In this short survey, we have discussed the current cutting edge of event detection from social streams. We have focused on recent work and on work that either has been shown to outperform earlier approaches, or to address issues that had been not or not very much tackled so far. We first discussed the start-of-the-art in detecting bursty terms, and then progressed to the question of how to identify bursty topics/events, pointing out some of the problems with finding the actual meaning of collections of terms, and potential solutions for supporting end users. Related to this, we continued with how to characterize events, e.g. discriminating between offline events that are *reflected* and online events *originating* in Twitter, and pointed towards work attempting to do this automatically. Finally, we discussed how to localize detected events spatially and/or socially – to identify who might benefit from being informed about them.

Acknowledgments

This survey was compiled in the context of the European Project GRAISearch, FP7- PEOPLE-2013-IAPP (612334).

References

[1] L. A. Adamic, R. A. Baeza-Yates, and S. Counts, editors. *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011*. The AAAI Press, 2011.

[2] C. C. Aggarwal and K. Subbian. Event detection in social streams. In *Proceedings of the Twelfth SIAM International Conference on Data Mining, Anaheim, California, USA, April 26-28, 2012.*, pages 624–635. SIAM / Omnipress, 2012.

[3] H. Becker, M. Naaman, and L. Gravano. Beyond trending topics: Real-world event identification on twitter. In Adamic et al. [1].

[4] M. Cataldi, L. Di Caro, and C. Schifanella. Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining*, MDMKDD '10, pages 4:1–4:10, New York, NY, USA, 2010. ACM.

[5] Q. Diao, J. Jiang, F. Zhu, and E. Lim. Finding bursty topics from microblogs. In *The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, July 8-14, 2012, Jeju Island, Korea - Volume 1: Long Papers*, pages 536–544. The Association for Computer Linguistics, 2012.

[6] Q. He, K. Chang, and E. Lim. Analyzing feature trajectories for event detection. In W. Kraaij, A. P. de Vries, C. L. A. Clarke, N. Fuhr, and N. Kando, editors, *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007*, pages 207–214. ACM, 2007.

[7] J. M. Kleinberg. Bursty and hierarchical structure in streams. *Data Min. Knowl. Discov.*, 7(4):373–397, 2003.

[8] M. Naaman, H. Becker, and L. Gravano. Hip and trendy: Characterizing emerging trends on twitter. *JASIST*, 62(5):902–918, 2011.

[9] M. Osborne, S. Petrovic, R. McCreadie, C. Macdonald, and I. Ounis. Bieber no more: First story detection using twitter and wikipedia. In *Proceedings of the Workshop on Time-aware Information Access. TAIA*, volume 12, 2012.

[10] A. Popescu and M. Pennacchiotti. Detecting controversial events from twitter. In J. Huang, N. Koudas, G. J. F. Jones, X. Wu, K. Collins-Thompson, and A. An, editors, *Proceedings of the 19th ACM Conference on Information and Knowledge Management, CIKM 2010, Toronto, Ontario, Canada, October 26-30, 2010*, pages 1873–1876. ACM, 2010.

[11] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In M. Rappa, P. Jones, J. Freire, and S. Chakrabarti, editors, *Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010*, pages 851–860. ACM, 2010.

[12] A. Stefanidis, A. Crooks, and J. Radzikowski. Harvesting ambient geospatial information from social media feeds. *GeoJournal*, 78(2):319–338, 2013.

[13] X. Wang, C. Zhai, X. Hu, and R. Sproat. Mining correlated bursty topic patterns from coordinated text streams. In P. Berkhin, R. Caruana, and X. Wu, editors, *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, California, USA, August 12-15, 2007*, pages 784–793. ACM, 2007.

[14] J. Weng and B. Lee. Event detection in twitter. In Adamic et al. [1].