Bieber no more: First Story Detection using Twitter and Wikipedia

Miles Osborne† and Saša Petrović‡, Richard McCreadie•, Craig Macdonald•, and Iadh Ounis•
†miles@inf.ed.ac.uk ‡S.Petrovic@sms.ed.ac.uk •firstname.lastname@glasgow.ac.uk

School of Informatics
University of Edinburgh
EH8 9AB, Edinburgh, UK

School of Computing Science
University of Glasgow
G12 8QQ, Glasgow, UK

ABSTRACT

Twitter is a well known source of information regarding breaking news stories. This aspect of Twitter makes it ideal for identifying events as they happen. However, a key problem with Twitter-driven event detection approaches is that they produce many spurious events, i.e., events that are wrongly detected or simply are of no interest to anyone. In this paper, we examine whether Wikipedia (when viewed as a stream of page views) can be used to improve the quality of discovered events in Twitter. Our results suggest that Wikipedia is a powerful filtering mechanism, allowing for easy blocking of large numbers of spurious events. Our results also indicate that events within Wikipedia tend to lag behind Twitter.

Categories and Subject Descriptors: H.3.3 [Information Storage & Retrieval]: Information Search & Retrieval

Keywords: Twitter, Wikipedia, Event Detection

1. INTRODUCTION

Twitter is a popular microblogging service, which provides an easy mechanism for users to publicly and instantly post messages – known as tweets. Interestingly, Twitter has become known for breaking news stories faster than traditional newswire companies [6] and for its focus on what is happening right now [10]. For this reason, Twitter has been used as a source of evidence for real-time event detection [16, 14]. In contrast, Wikipedia is a curated user-driven encyclopedia covering a diverse range of topics. Wikipedia is also a source of information about emerging events [12]. For instance, when Michael Jackson died, his related Wikipedia page was updated 104 times that day with fresh information, with a further 641 updates on the day after.[1] These two information sources do not exist in isolation and it is intuitive that information that is reported in one may directly or indirectly be reflected in the other.

In this paper, we explore the extent that event detection, in particular *first story* detection, based on Twitter can be improved using Wikipedia. In particular, using a state-of-the-art first story detection approach [14], we identify events as they emerge within Twitter. In parallel, we track Wikipedia pages that exhibit abnormally large spikes in page views. We compare the resultant tweets and Wikipedia pages over textual and time dimensions to identify the types of information that are common across these two information streams and the latencies inherent to this form of information sharing. Apart from understanding the relationship between these two very different streams of information, our work is part of a broader research agenda looking at how user-generated content can be organised and made more useful.

[1] en.wikipedia.org/wiki/Michael_Jackson?action=history

TAIA'12, August 12–16, 2012, Portland, Oregon, USA.

2. RELATED WORK

First story detection detection was first examined as part of the Topic Detection and Tracking (TDT) task that ran as part of the Text REtrieval Conference (TREC) [2]. Under TDT, the aim of first story detection was to identify the first instance of an article that was related to a new emerging topic. At the time that TDT was run, the primary publication medium was low-volume, clean newswire, hence first story detection under TDT focused on news article content. In contrast here we investigate how to achieve high precision first story detection using multiple parallel user-generated content streams rather than (just) newswire.

First story detection/event detection using user-generated content streams is just beginning to be explored. Within the Twitter domain, Sankaranarayanan *et al.* [16] used a clustering approach to detect events using a text classifier. Meanwhile, Petrović *et al.* [14] introduced the use of locality-sensitive hashing to tackle the high-volume tweet stream for first story detection. Becker *et al.* [5] proposed an online clustering framework to group Flickr documents relating to events. They show that using a combined document feature-based similarity measure for clustering was more effective that using traditional textual similarity. In difference to these prior works, we not only combine real-time evidence from Twitter, but also live information gathered from Wikipedia. In this work we use Petrović *et al.*'s [14] event detection approach as our baseline for the Twitter stream.

The IR literature has also investigated the identification and presentation of content relating to events within Twitter. For instance, Mathioudakis and Koudas [11] proposed the TweetMonitor system, that identifies trending topics for users to further interact with. Prior works have also examined the detection of specific types of events. For example, Sakaki *et al.* [15] proposed to use classification to identify Twitter tweets relating to earthquakes as they happen. Similarly, Okazaki and Matsuo [13] proposed to use support vector machines to classify regional tweets relating

to earthquakes such that warnings could be propagated before the earthquake arrived in neighbouring regions. These approaches to event detection focus primarily on tracking changes in the volume and popularity of rare and informative terms over time [7, 11, 13, 15]. Our approach differs from these prior works in that we automatically identify previously unknown events as they happen using multiple streams rather than any single stream.

Wikipedia has also seen some investigation with regard to current events. In particular, Ciglan and Norvag [8] proposed the WikiPop service. WikiPop detects unexpected increases in the popularity of topics related to information needs expressed by users using Wikipedia page views. Relatedly, Ahn *et al.* [1] clustered spiking page views into topically-related groups for event detection. We also use Wikipedia's page views in this work, with the difference that we use it as a source of supporting evidence for events first identified in Twitter.

Becker *et al.* [4] examined how to identify related content to *predictable* events from Twitter, YouTube, and Flickr. They demonstrate how documents from these social media streams can be used to enhance document retrieval from other related streams for the same events. We similarly exploit multiple user-generated content streams in this work. However, we tackle the task of detecting both predictable and unpredictable events using real-time evidence from Twitter and Wikipedia.

3. FIRST STORY DETECTION USING TWITTER AND WIKIPEDIA

We improve the quality of events detected using a state-of-the-art first story detection (FSD) system leveraging a live stream of tweets, by exploiting parallel event reporting within Wikipedia. In this section, we describe our approach in detail.

3.1 First Story Detection from Twitter

To achieve effective first story detection, we build upon a prior streaming FSD system described by [14], deployed upon the live tweet stream. This FSD system processes each tweet as it arrives, dynamically building up tweet clusters representing events. In particular, for each incoming tweet, it compares that tweet against the stream of previously seen tweets using a fast hashing strategy. If the current tweet is sufficiently (textually) dissimilar from its nearest neighbour, it is flagged it as being a potential first story. The system attempts to reduce false positives by waiting for a short deferral period, such that it can collect all follow-up posts, producing clusters of closely related tweets. For an event cluster, the tweet closest to the centroid of the cluster (using a standard vector space) is emitted. This approach is state-of-the-art for streaming FSD and produces a stream of time-stamped tweets, each one corresponding to a (potential) first story.

3.2 Identifying Event-Related Wikipedia Pages

Wikipedia is a rich, constantly evolving information source. To improve the quality of the potential events from our Twitter FSD system, we need to similarly identify events in parallel from Wikipedia, i.e. identify event-related pages. Within Wikipedia, events are reflected in terms of edits, page views and new page creation. Here we focus upon page views as a proxy for interest and in particular equate spikes in page views with it.

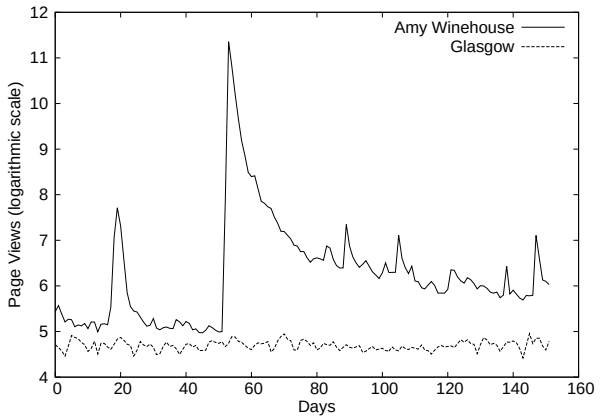

Figure 1: Raw page view counts for the pages *Amy Winehouse* (top) and *Glasgow* (bottom). Note logarithmic scale.

For illustration, Figure 1 shows the raw page view requests (on a logarithmic scale) for the pages *Amy Winehouse* and *Glasgow* from 1st June 2011 to 31st November 2011. Each point represents the per-day hourly average. The large spikes for Amy Winehouse correspond to real-world events (e.g. arrest, her death). The viewing counts for Glasgow (which is presumably less newsworthy) by contrast are much flatter. Notably, we only consider page titles rather than full web pages - future work will take into account actual pages.

We track per-hour page views for all English-language pages.[2] At each hour for each page i with page views w^i, we maintain a moving window of k hours over previous page view counts $w^i_{j-k}, \ldots, w^i_j$. When we move into a new hour, we update the moving windows for all pages and then apply Grubb's test to each moving window, determining if the latest page view number is an outlier with respect to previously seen views for the page in question [9]. If $\overline{x}^i$ is the mean page views for page w^i at time j, with standard deviation σ^2, then we declare reading w^i an outlier if $(w^i_j - \overline{x}^i)/\sigma^2 > z$ for some fixed z (here 3.5 standard deviations). This is a very simple test, with clear assumptions which need exploring. The test it two-sided (an outlier can either be abnormally large or small) but we only consider abnormally large readings. All page views are normalised using the total number of views for the hour and are in log-scale.

The outcome of this is a stream of time-stamped outlier pages, corresponding to abnormally large page views. Unlike the Twitter FSD stream, at times we might emit the same page more than once in succession. This happens when a page is rapidly increasing over a few hours.

3.3 Multi-stream FSD

To increase FSD quality, we combine the two streams together as follows. For each potential event identified from Twitter and represented by a tweet, using a simple vector space model over Wikipedia pages, we find the closest page for that tweet (also represented as a vector). The intuition here is that a first story in Twitter is reflected in a spike in page views for an associated Wikipedia page. Because many of our potential Twitter first stories are false

[2]We obtain page view requests from Wikipedia public logs.

positives, we would expect them to have no close matches in the Wikipedia stream. For genuine Twitter first stories, we would expect them to match (at least partially) against Wikipedia titles. Wikipedia can therefore filter-out spurious stories in Twitter.

We round tweet timestamps to the nearest hour to be consistent with our Wikipedia polling interval. We remove stop words, hashtags, user mentions, and URLs from tweets.

4. EXPERIMENTAL SETUP

We focus upon the time period from June 30^{th} to July 24^{th} 2011. For Twitter, we collected the data from the public streaming API (`http://stream.twitter.com/`). This gave us about 2 million tweets per day. We ran the FSD system on these tweets and performed single-link clustering to form candidate clusters as in [14].[3] We then keep the clusters that had over 30 tweets in them, which produced 235 candidate first stories (clusters). For each cluster, we selected a single tweet (the centroid) as being representative of the cluster.

Events in this period include: the death of Amy Winehouse; telephone hacking scandals in the UK; a Tsunami warning following an earthquake in New Zealand; Casey Anthony released from prison; Yao Ming retiring. Approximately 80% of the proposed Twitter first stories are spurious

Each English-only Wikipedia hourly dump contained approximately 10 million page requests.[4] We used a moving window of 48 hours for each Wikipedia page and produced 625 thousand Wikipedia outliers for the time period. Varying the window size produces different outliers and in preliminary experiments, sizes greater than 48 did not yield significantly different results.

5. RESULTS

In this section, we examine the performance of our first story detection approach that combines both Twitter and Wikipedia evidence. In particular, within each of the following three sub-sections, we investigate a specific research question, namely:

1. Is there a latency between the two streams?
2. Are newly created (requested) Wikipedia pages useful?
3. Can Wikipedia be used to improve the quality of events detected in Twitter?

5.1 Latency

We varied the temporal alignment of Twitter and Wikipedia, simulating when it lags and when it leads. For example, a latency of -3 means that Wikipedia lags Twitter by three hours; a latency of 1 means that Wikipedia leads Twitter by one hour. We evaluate performance in terms of the average distance between each time-aligned Twitter first story and the corresponding nearest neighbour Wikipedia page title. This gives us an indication of the extent to which the Twitter first story stream matches the Wikipedia outlier stream. The lower this figure is, the better the alignment. We also measured the variance of these distances. If we assume that spurious Twitter stories will always be far away from any Wikipedia page –irrespective of alignment– and that actual events will have a corresponding close Wikipage –when correctly aligned– then the better the alignment between our two streams, the higher the variance will be.

Latency (Hours)	Mean Distance	Standard deviation
Lagging		
-3	0.81	3.0
-2	**0.78**	**3.2**
-1	0.83	2.8
Equal		
0	0.83	2.6
Leading		
1	0.84	2.6
2	0.84	2.6

Table 1: Mean and standard deviation of distance between Twitter first-stories and nearest Wikipedia page titles.

As can be seen from Table 1, the best results (in bold) are obtained when Wikipedia lags Twitter, by around two hours. Even if we took into account the fact that we sample Wikipedia every hour and also round the Twitter timestamps, it will still be unlikely that Wikipedia will be fully time-synchronous with Twitter. Hence, we conclude that there is a delay between events breaking on Twitter and the time when users start to search Wikipedia for information about it.

5.2 Using newly created Wikipages

Wikipedia has a great number of new pages requested each hour and it is conceivable that these new pages coincide with breaking news. A page is considered new if we do not have a history for it and 'dead' if it has no page views over a 24 hour period. New pages present problems as we have no history information for them. They are potentially interesting as they might coincide with new events.

We re-ran the previous experiments but also emitted any new Wikipedia pages encountered (pages with no history). However, the inclusion of these new pages greatly increased the number of Wikipedia outlier pages identified per hour (from approximately 4k to about 200k pages per hour, with 2.4 billion page requests for the entire duration). Note that determination of a 'new' page is fully automatic and does not imply knowledge of the future: a page is deemed new if we do not have a history for it at the current hour. Our results suggest that this massive increase in pages (most of which are spurious) made finding useful constraining information a lot harder.

5.3 Can Wikipedia Improve the Quality of Events Detected in Twitter?

One way to judge the usefulness of Wikipedia for event detection is to take the original first stories over the entire time interval (which correspond with clusters of closely related tweets) and rank them. We can rank them by cluster size (which is a metric based only on tweets) or we can rank them by their distance to the closest Wikipedia page.

To evaluate whether Wikipedia outliers can improve the quality of our Twitter FSDs, we conducted the following experiment. We reranked our first stories according to the Wikipedia score and had two judges label the top 100 ranked tweets. The two judges also labelled first stories when sorted

[3]Recording every single possible first story in the interval would generate a vast number of candidates, making analysis harder.

[4]We treat a page request as a page title. These can also include typos as well as genuine Wikipedia pages.

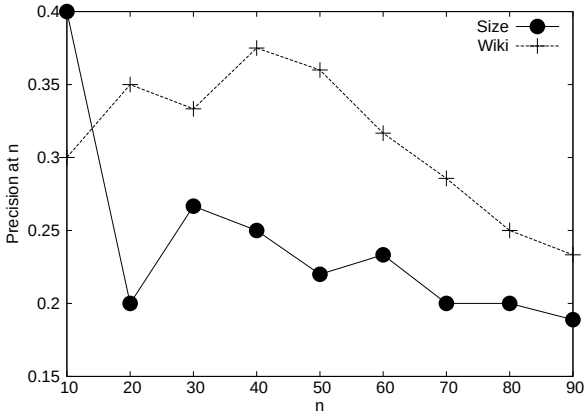

Figure 2: Precision at n for two different ranking strategies.

Rank	Event Tweet
1	I love Seth meyers! #ESPYs
2	@tanacondasteve amy whinehouse is dead
3	RT @katyperry: HAPPY 4TH OF JULY!!!!!!!!!!!!!! ...
4	Yao Ming retired
5	Derek jeter 3000 hits.

Table 2: Top five highest-scoring detected events using Twitter and Wikipedia.

by their associated cluster size alone. When labelling the clusters, only the centroid tweet in the cluster was shown to the judges. Showing all the tweets in the cluster to the judges was just not feasible here as some of these clusters had thousands of tweets in them. Nevertheless, we manually inspected the centroids and found them to be a good representation of the clusters. The two judges labelled the centroid tweets as being about an event or not. The kappa agreement between the judges was 0.74 on the set of clusters returned by the Wikipedia score, and 0.79 on the 100 biggest clusters, indicating good agreement between them. For the set of clusters where the two judges agreed on the label, we computed the precision at n measure. We show this precision in Figure 2. First, we can see that just sorting by size performs better at a very low value of $n = 10$. This means that truly huge clusters (very popular stories) on Twitter are indeed about real events. However, we see that ranking using Wikipedia outperforms the size-based ranking for all other stories.

Wikipedia allows us to filter-out spurious first stories in Twitter. When not using new pages, we see that the vast majority of spurious Twitter first stories are clearly down-weighted and that the closest matching first stories often coincide with real news. For example, sorting by the Wikipedia distance, the top five events are shown in Table 2.

In contrast, when Wikipedia is not used as a filter and we instead sort by the size of the corresponding cluster, we obtain the results shown in Table 3.

6. CONCLUSION

In this paper we presented an approach that combines two streams, Twitter and Wikipedia, for the purpose of improving event detection. We first explored the latency between the two streams and on average found that, when it comes to real-world events, Wikipedia seems to be lagging behind Twitter by about two hours. This means that, for truly real-time event detection, the usefulness of Wikipedia may be limited. However, there are many cases when this lag is acceptable and we thus investigated whether Wikipedia can be used as a filter to remove noise in events detected in Twitter. Our results indicate that the quality of detected events can be substantially improved when considering this additional source of information. Future work will look at better ways of intersecting the two streams.

Rank	Event Tweet
1	get your free $1000 bestbuy giftcard now! #iloveshopping
2	RT @SkyNewsBreak: Sky Sources: 27-year-old singer Amy Winehouse found dead at her flat in North London
3	Do you think caylee got justice? #caseyanthony
4	Tweeting from my new iPad2!! thank you!! #freestuff
5	how dumb are you?-take this quiz and retweet your score

Table 3: Top five highest-scoring detected events just using Twitter.

Acknowledgements

The authors acknowledge financial support from EPSRC grant EP/J020664/1.

7. REFERENCES

[1] B. G. Ahn, B. Van Durme and C. Callison-Burch. WikiTopics: what is popular on Wikipedia and why In *Proceedings of ACL-HLT'11*.

[2] J. Allan. Introduction to topic detection and tracking. In *Topic Detection and Tracking*, pages 1-16, 2002.

[3] H. Becker, M. Naaman, L. Gravano. Beyond trending topics: real-world event identification on Twitter. In *Proceedings of ICWSM'11*.

[4] H. Becker, D. Iter, M. Naaman and L. Gravano. Identifying content for planned events across social media sites. In *Proceedings of WSDM'12*.

[5] H. Becker, M. Naaman and L. Gravano. Learning similarity metrics for event identification in social media. In *Proceedings of WSDM'10*.

[6] T. O'Brien. Twitter breaks news of plane crash in the Hudson. `http://www.switched.com/2009/01/15/twitter-breaks-news-of-plane-crash-in-the-hudson/`. 2009.

[7] M. Cataldi, L. Di Caro and C. Schifanella. Emerging topic detection on Twitter based on temporal and social terms evaluation. In *Proceedings of MDMKDD'10*.

[8] M. Ciglan and K. Norvag. WikiPop: personalized event detection system based on Wikipedia page view statistics. In *Proceedings of CIKM'10*.

[9] F. Grubb. Procedures for detecting outlying observations in samples. *Technometrics*, 11(1), pages 1-21, 1969.

[10] H. Kwak, C. Lee, H. Park and S. Moon. What is Twitter, a social network or a news media. In *Proceedings of WWW'10*.

[11] M. Mathioudakis and N. Koudas. TwitterMonitor: Trend detection over the Twitter stream. In *Proceedings of SIGMOD'10*.

[12] R. McCreadie, C. Macdonald and I. Ounis. Insights on the horizons of news search. In *Proceedings of SSM'10*.

[13] M. Okazaki and Y. Matsuo. Semantic Twitter: analyzing tweets for real-time event notification. In *Recent Trends and Developments in Social Software*, pages 63-74, 2011.

[14] S. Petrović, M. Osborne and V. Lavrenko. Streaming First story detection with application to Twitter. In *Proceedings of NAACL'10*.

[15] T. Sakaki, M. Okazaki and Y. Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of WWW'10*.

[16] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman and J. Sperling. TwitterStand: news in tweets. In *Proceedings of GIS'09*.

[17] Y. Zhai and M. Shah. Tracking news stories across different sources. In *Proceedings of Multimedia'05*.