# MJ no more: Using Concurrent Wikipedia Edit Spikes with Social Network Plausibility Checks for Breaking News Detection

Thomas Steiner[*]
Google Germany GmbH
ABC-Str. 19
20354 Hamburg, Germany
tomac@google.com

Seth van Hooland
Université Libre de Bruxelles
Avenue F.D. Roosevelt, 16
1050 Brussels, Belgium
svhoolan@ulb.ac.be

Ed Summers
Library of Congress
101 Independence Ave, SE
Washington, DC 20540, USA
edsu@loc.gov

## ABSTRACT

We have developed an application called *Wikipedia Live Monitor* that monitors article edits on different language versions of Wikipedia—as they happen in realtime. Wikipedia articles in different languages are highly interlinked. For example, the English article *"en:2013_Russian_meteor_event"* on the topic of the February 15 meteoroid that exploded over the region of Chelyabinsk Oblast, Russia, is interlinked with "ru:Падение_метеорита_на_Урале_в_2013_году", the Russian article on the same topic. As we monitor multiple language versions of Wikipedia in parallel, we can exploit this fact to detect *concurrent edit spikes* of Wikipedia articles covering the same topics, both in only one, and in different languages. We treat such concurrent edit spikes as signals for potential breaking news events, whose plausibility we then check with full-text cross-language searches on multiple social networks. Unlike the reverse approach of monitoring social networks first, and potentially checking plausibility on Wikipedia second, the approach proposed in this paper has the advantage of being less prone to false-positive alerts, while being equally sensitive to true-positive events, however, at only a fraction of the processing cost. A live demo of our application is available online at the URL `http://wikipedia-irc.herokuapp.com/`, the source code is available under the terms of the Apache 2.0 license at `https://github.com/tomayac/wikipedia-irc`.

## Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: Clustering

## General Terms

Algorithms

## Keywords

Breaking News Detection, Event Detection, Wikipedia, Social Networks, Internet Relay Chat

## 1. INTRODUCTION

### 1.1 Motivation

Shortly after the celebrity news website TMZ broke the premature news that the King of Pop Michael Jackson (MJ) had died,[1] the Internet slowed down.[2] Initially, Wikipedia's website administrators started noting abnormal load spikes [8]. Shortly afterwards, caching issues caused by a so-called edit war [1] led the site to go down: Wikipedia editors worldwide made concurrent edits to the Michael Jackson Wikipedia article, doing and undoing changes regarding the tense of the article, death date, and the circumstances of the (at the time) officially still unconfirmed fatality. While Wikipedia engineers have worked hard to ensure that future load spikes do not take the site down again, there is without dispute a lot of research potential in analyzing such editing activity.

### 1.2 Hypotheses and Research Questions

In this paper, we present an application that monitors article edits of different language versions of Wikipedia in realtime in order to detect concurrent edit spikes that may be the source of breaking news events. When a concurrent edit spike has been detected, we use cross-language full-text searches on social networks as plausibility checks to filter out false-positive alerts. We are led by the following hypotheses.

(H1) Breaking news events spread over social networks, independent from where the news broke initially.

(H2) If a breaking news event is important, it will be reflected on at least one language edition of Wikipedia.

(H3) The time between when the news broke first and the news being reflected on Wikipedia is considerably short.

These hypotheses lead us to the research questions below.

(Q1) Can concurrent Wikipedia edit spikes combined with social network plausibility checks capture major breaking news events, and if so, with what delay?

(Q2) Is the approach Wikipedia first, social networks second at least as powerful as the reverse approach?

In this paper, we do not answer all research questions yet, however, lay the foundation stone for future research in this area by introducing the Wikipedia Live Monitor application.

[*]This work was partially supported by the European Commission under Grant No. 248296 FP7 I-SEARCH project.

*World Wide Web Conference*, '13 Rio de Janeiro, Brazil
ACM 978-1-4503-2038-2/13/05.

[1] MJ dead: http://www.tmz.com/2009/06/25/michael-jackson-dies-death-dead-cardiac-arrest/, accessed 02/18/2013

[2] Internet slow-down: http://news.bbc.co.uk/2/hi/technology/8120324.stm, accessed 02/18/2013

## 2. RELATED WORK

We refer to an event as breaking news, if the event is of significant importance to a considerable amount of the population. Petrović et al. define [5] the goal of new event detection (or first story detection) as "given a sequence of stories, to identify the first story to discuss a particular event." They define an event as "something that happens at some specific time and place." Classic streaming analysis of social network microposts so far has been mainly focused on Twitter, a microblogging social network that provides access to a sampled stream of generated microposts by means of its Streaming API.[3] Petrović et al. explain [5]: "in the streaming model of computation, items arrive continuously in a chronological order, and have to be processed in bounded space and time." In the referenced paper, the authors report on a system for streaming new event detection applied to Twitter based on locality sensitive hashing. Hu et al. provide an analysis of how news break and spread on Twitter [3]. The task of linking news events with social media is covered by Tsagkias et al. in [7]. With this paper, we stand on the shoulders[4] of Osborne et al. [4], who use Wikipedia page view statistics[5] as a means to filter spurious events stemming from event detection over social network streams. Our approach reverses theirs, however, instead of the only hourly updated page view statistics, we use realtime change notifications, as explained in Subsection 3.1. Our Wikipedia Live Monitor is partly based on an application called Wikistream, developed by Ed Summers et al., which was described in [6].

## 3. IMPLEMENTATION DETAILS

### 3.1 Wikipedia Recent Changes

As described earlier, our application monitors concurrent edit spikes on different language versions of Wikipedia. In the current implementation, we monitor 42 different Wikipedias, 5 with $\geq$ 1,000,000 and 37 with $\geq$ 100,000 articles.[6] Changes to any single one article are communicated by a chat bot over Wikipedia's own Internet Relay Chat (IRC) server (irc.wikimedia.org),[7] so that parties interested in the data can listen to the changes as they happen. For each language version, there is a specific chat room following the pattern "#" + language + ".wikipedia". For example, changes to Russian Wikipedia articles will be streamed to the room #ru.wikipedia. A special case is the room #wikidata.wikipedia for Wikidata [9], a platform for the collaborative acquisition and maintenance of structured data to be used by Wikimedia projects like Wikipedia. A sample chat message with the components separated by the asterisk character '*' announcing a change can be seen in the following. "[[Juniata River]] http://en.wikipedia.org/w/index.php?diff=516269072&oldid=514659029 * Johanna-Hypatia * (+67) Category:Place names of Native American origin in Pennsylvania". The message components are (i) article name, (ii) revision URL, (iii) Wikipedia editor handle, and (iv) change size and change description.

### 3.2 Article Clusters

We cluster edits of articles about the same topic, but written in different languages, in article clusters. The example of the English "en:2013_Russian_meteor_event" and the Russian article "ru:Падение_метеорита_на_Урале_в_2013_году" that are both in the same cluster illustrate this. We use the Wikipedia API to retrieve language links for a given article. The URL pattern for the API is as follows. `http://$LANGUAGE.wikipedia.org/w/api.php?action=query&format=json&prop=langlinks&titles=$ARTICLE`.

### 3.3 Comparing Article Revisions

The Wikipedia API provides means to retrieve the actual changes that were made during an edit including additions, deletions, and modifications in a `diff`-like manner. The URL pattern is as follows. `http://$LANGUAGE.wikipedia.org/w/api.php?action=compare&torev=$TO&fromrev=$FROM&format=json`. This allows us to classify edits in categories, like, *e.g.*, negligible trivial edits (punctuation correction) and major important edits (new paragraph for an article), which helps us to disregard seemingly concurrent edits in order to avoid false-positive alerts.

### 3.4 Breaking News Criteria

Our application *Wikipedia Live Monitor* puts detected article clusters in a monitoring loop in which they remain until their time-to-live (240 seconds) is over. In order for an article cluster in the monitoring loop to be identified as breaking news candidate, the following breaking news criteria have to be fulfilled.

$\geq$ 5 Occurrences: An article cluster must have occurred in at least 5 edits.

$\leq$ 60 Seconds Between Edits: An article cluster may have at maximum 60 seconds in between edits.

$\geq$ 2 Concurrent Editors: An article cluster must have been edited by at least 2 concurrent editors.

$\leq$ 240 Seconds Since Last Edit: An article cluster's last edit may not be longer ago than 240 seconds.

The exact parameters of the breaking news criteria above were *determined empirically* by analyzing Wikipedia edits over several hours and repeatedly adjusting the settings until major news events happening at the same time were detected. The resulting dataset split into three chunks has been made publicly available.[8]

### 3.5 Social Network Plausibility Checks

When a breaking news candidate has been identified, we use cross-language full-text social network searches on the social networks Twitter, Facebook, and Google+ as a plausibility check. As the *article titles* of all language versions of the particular article's cluster are know, we use these very article titles as search queries for cross-language searches, as can be seen in Figure 1. This approach greatly improves the recall of the social network search, however, requires either automatic translation, or an at least basic understanding of the languages being searched in. Currently the plausibility checking step is not yet fully automated, as the search results are for the time being meant to be consumed by *human evaluators*. Driven by (H1), we assume breaking news events are

[3] Twitter Streaming API: https://dev.twitter.com/docs/api/1.1/get/statuses/sample, accessed 02/18/2013

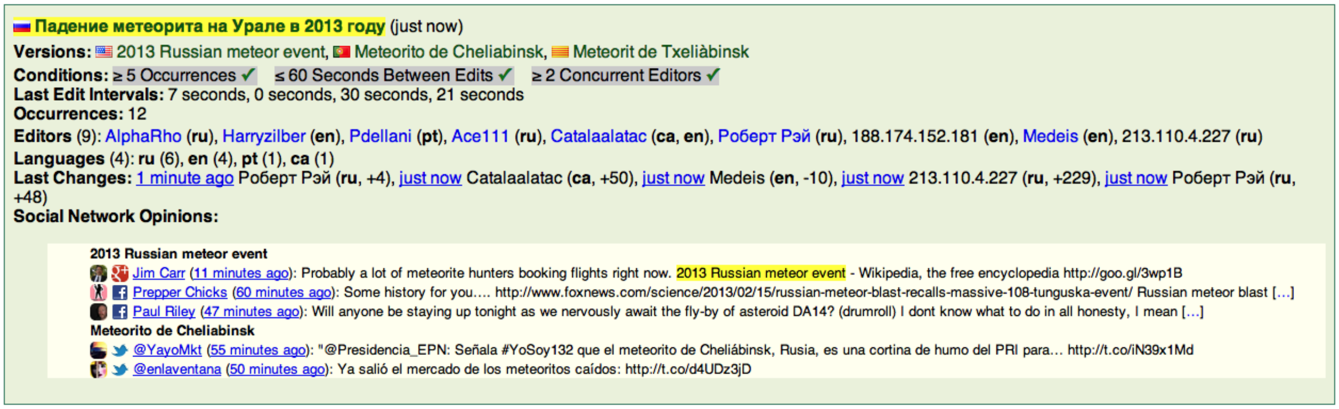
[4] Hence the title of this paper.

[5] Page view statistics for Wikimedia projects: http://dumps.wikimedia.org/other/pagecounts-raw/, accessed 02/18/2013

[6] List of Wikipedias by size: http://meta.wikimedia.org/wiki/List_of_Wikipedias, accessed 02/18/2013

[7] Raw IRC feeds of recent changes: http://meta.wikimedia.org/wiki/IRC/Channels#Raw_feeds, accessed 02/18/2013

[8] Wikipedia Live Monitor dataset: https://www.dropbox.com/sh/2qsg1zhb8p35fxf/Dghn55y0kh, accessed 02/18/2013



**Падение метеорита на Урале в 2013 году** (just now)
**Versions:** 2013 Russian meteor event, Meteorito de Cheliabinsk, Meteorit de Txeliàbinsk
**Conditions:** ≥ 5 Occurrences ✓ ≤ 60 Seconds Between Edits ✓ ≥ 2 Concurrent Editors ✓
**Last Edit Intervals:** 7 seconds, 0 seconds, 30 seconds, 21 seconds
**Occurrences:** 12
**Editors** (9): AlphaRho (**ru**), Harryzilber (**en**), Pdellani (**pt**), Ace111 (**ru**), Catalaalatac (**ca**, **en**), Роберт Рэй (**ru**), 188.174.152.181 (**en**), Medeis (**en**), 213.110.4.227 (**ru**)
**Languages** (4): **ru** (6), **en** (4), **pt** (1), **ca** (1)
**Last Changes:** 1 minute ago Роберт Рэй (**ru**, +4), just now Catalaalatac (**ca**, +50), just now Medeis (**en**, -10), just now 213.110.4.227 (**ru**, +229), just now Роберт Рэй (**ru**, +48)
**Social Network Opinions:**

**2013 Russian meteor event**
Jim Carr (11 minutes ago): Probably a lot of meteorite hunters booking flights right now. 2013 Russian meteor event - Wikipedia, the free encyclopedia http://goo.gl/3wp1B
Prepper Chicks (60 minutes ago): Some history for you.... http://www.foxnews.com/science/2013/02/15/russian-meteor-blast-recalls-massive-108-tunguska-event/ Russian meteor blast [...]
Paul Riley (47 minutes ago): Will anyone be staying up tonight as we nervously await the fly-by of asteroid DA14? (drumroll) I dont know what to do in all honesty, I mean [...]
**Meteorito de Cheliabinsk**
@YayoMkt (55 minutes ago): "@Presidencia_EPN: Señala #YoSoy132 que el meteorito de Cheliábinsk, Rusia, es una cortina de humo del PRI para... http://t.co/iN39x1Md
@enlaventana (50 minutes ago): Ya salió el mercado de los meteoritos caídos: http://t.co/d4UDz3jD

Figure 1: Screenshot with an article cluster of four concurrently edited articles (ru, en, pt, ca). All breaking news criteria are fulfilled, the cluster is a breaking news candidate. Cross-language social network search results for en and pt can be seen.

being discussed on social networks. We will show arguments for this assumption in Section 4. For now, we expect social networks to be a short period ahead of Wikipedia. In consequence, if the human rater can find positive evidence for a connection between social network activities and Wikipedia edit actions, the breaking news candidate is confirmed to indeed be breaking news.

### 3.6 Application Pseudocode

The *Wikipedia Live Monitor* application has been implemented in Node.js, a server side JavaScript software system designed for writing scalable Internet applications. Programs are created using event-driven, asynchronous input/output operations to minimize overhead and maximize scalability. Listing 1 shows the pseudocode of the two main event loops of the *Wikipedia Live Monitor* application. The actual implementation is based on Martyn Smith's Node.js IRC library[9] and the WebSockets API and protocol [2], wrapped by Guillermo Rauch's library Socket.IO.[10]

## 4. PREMATURE EVALUATION

As noted earlier, in this paper, we do not answer all research questions yet. Nevertheless, we extract trends from our experiments so far. In Subsection 1.2, we have set up three hypotheses. (H1) has been proven by Hu *et al.* in [3] for Twitter. We argue that it can be generalized to other social networks and invite the reader to have a look at our dataset, where the lively discussions about breaking news candidates on the considered social networks Twitter, Facebook, and Google+ support the argumentation. It is hard to prove (H2), as the concept of *important breaking news* is vague and dependent on one's personal background, however, all evidence suggests that (H2) indeed holds true, as, to the best of our knowledge and given our background, what the authors consider *important breaking news* is represented on at least one language version of Wikipedia. (H3) has been examined by Osborne *et al.* in [4]. In the paper, they suggest that Wikipedia lags about two hours behind Twitter. It has to be noted that they look at hourly accumulated page (article) *view* logs, where we look at realtime article *edit* log streams. Our experiments suggest that the lag time of two hours proposed by Osborne *et al.* may be too conservative. An educated guesstimation at this stage is that the lag time for breaking news is more in the range of 30 minutes, and for global breaking news like celebrity deaths in the range of five minutes and less, albeit the edits by our experience will be small and iterative (*e.g.*, "X is a" to "X was a", or the addition of a death date), followed by more consistent thorough edits.

The (at time of writing) recent breaking news event of the resignation of *Pope Benedict XVI* helps respond to (Q1). The three first edit times after the news broke on February 11, 2013 of the Pope's English Wikipedia article[11] are as follows (all times in UTC): 10:58, 10:59, 11:02. The edit times of the French article[12] are as follows: 11:00, 11:00, 11:01. This implies that by looking at only two language versions (the actual number of monitored versions is 42) of the Pope article, the system would have reported the news at 11:01. The official Twitter account of Reuters announced[13] the news at 10:59. Vatican Radio's announcement[14] was made at 10:57:47.

Not all breaking news events have the same global impact as the Pope's resignation, however, the proposed system was shown to work very reliably also for smaller events of more regional impact, for example, when *Indian singer Varsha Bhosle* committed suicide[15] on October 8, 2012. A systematic evaluation of (Q1) compulsorily can only be done by random samples, which has turned out positive results so far. Again, we invite the reader to explore our dataset and to conduct own experiments. A systematic evaluation of (Q2) requires a commonly shared dataset, which we have provided, however, at this point in time, we do not have access to the system of Osborne *et al.*

Regarding *Wikipedia Live Monitor*'s scalability, we could scale the monitoring system up to currently *all* 284 Wikipedias on a standard consumer laptop (mid-2010 MacBook Pro, 2.66 GHz Intel Core 2, 8 GB RAM), which once more proves the efficiency of

[9] Node.js IRC library: https://github.com/martynsmith/node-irc, accessed 02/18/2013
[10] Socket.IO library: http://socket.io/, accessed 02/18/2013
[11] Edit history en: http://en.wikipedia.org/w/index.php?title=Pope_Benedict_XVI&action=history, accessed 02/18/2013
[12] Edit history fr: http://fr.wikipedia.org/w/index.php?title=Beno%C3%AEt_XVI&action=history, accessed 02/18/2013
[13] Reuters announces Pope resignation: https://twitter.com/Reuters/status/300922108811284480, accessed 02/18/2013
[14] Vatican Radio announces Pope resignation: http://de.radiovaticana.va/Articolo.asp?c=663810, accessed 02/18/2013
[15] Varsha Bhosle suicide: http://en.wikipedia.org/wiki/Varsha_Bhosle, accessed 02/18/2013

the Node.js architecture for this kind of event-driven applications. In practice, however, the majority of the smaller Wikipedias being very rarely updated, we limit ourselves to the Wikipedias with $\geq$ 100,000 articles at no remarkable loss of recall.

## 5. FUTURE WORK

Future work towards a thorough scientific evaluation will mainly address two areas. First, the *automatic categorization of edits on Wikipedia* needs to be more fine-grained. In the context of breaking news detection, not all edits are equally useful. An image being added to an article is an example of an edit that usually will not be important. In contrast, the category “Living people” being removed from an article is a strong indicator of breaking (sad) news. Second, the *connection between social network search and Wikipedia edits* needs to be made clearer. In an initial step, the concrete changes to an article, as detailed in Subsection 3.3, can be compared with social network microposts using a cosine similarity measure. More advanced steps can exploit the potential knowledge from Wikipedia edits (*e.g.*, category “Living people” removed implies a fatality).

```
Input: irc, listening on Wikipedia recent changes
Output: breakingNewsCandidates, breaking news candidates

monitoringLoop = {}
articleClusters = {}
breakingNewsCandidates = {}

# Event loop 1:
# When a new message arrives
irc.on.message do (article)
  langRefs = getLanguageReferences(article)
  articleRevs = getArticleRevisions(article)
  cluster = clusterArticles(article, langRefs)

  # Create new cluster for previously unseen article
  if cluster not in monitoringLoop
    monitoringLoop.push(cluster)
    articleClusters.push(cluster)
    updateStatistics(cluster)
    emit.newCluster(cluster, articleRevs)
  # Update existing cluster, as the article was seen before
  else
    updateStatistics(cluster)
    emit.existingCluster(cluster, articleRevs)
    # Check breaking news criteria
    if cluster.occurrences >= 5
      if cluster.secsBetweenEdits <= 60
        if cluster.numEditors >= 2
          if cluster.secsSinceLastEdit <= 240
            socialNetworks.search(langRefs)
            breakingNewsCandidates.push(cluster)
            emit.breakingNewsCandidate(cluster)
          end if
        end if
      end if
    end if
  end if
  return breakingNewsCandidates
end do

# Event loop 2:
# Remove too old clusters regularly
timeout.every.240 seconds do
  for each cluster in monitoringLoop
    if cluster.secsSinceLastEdit >= 240
      monitoringLoop.remove(cluster)
      articleClusters.remove(cluster)
    end if
  end for
end do
```

Listing 1: The two main event loops of the application

## 6. CONCLUSIONS

In this paper, we have shown an application called *Wikipedia Live Monitor* and released its source code under the Apache 2.0 license. This application monitors article edits on 42 different language versions of Wikipedia. It detects breaking news candidates according to well-defined breaking news criteria, whose exact parameters were determined empirically and the corresponding dataset made available publicly. We have shown how cross-language full-text social network searches are used as plausibility checks to avoid false-positive alerts. In a first step towards a full evaluation, we have shown promising preliminary results and actionable next steps in future work for improving the application.

## 7. REFERENCES

[1] C. Beaumont. Michael Jackson’s death sparks Wikipedia editing war, June 2009. `http://bit.ly/Michael-Jacksons-death-sparks-Wikipedia-editing-war`, accessed 02/18/2013.

[2] I. Hickson. The WebSocket API. Candidate Recommendation, W3C, Sept. 2012.

[3] M. Hu, S. Liu, F. Wei, Y. Wu, J. Stasko, and K.-L. Ma. Breaking News on Twitter. In *Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems*, CHI ’12, pages 2751–2754. ACM, 2012.

[4] M. Osborne, S. Petrović, R. McCreadie, C. Macdonald, and I. Ounis. Bieber no more: First Story Detection using Twitter and Wikipedia. In *Proceedings of the SIGIR Workshop on Time-aware Information Access*, 2012.

[5] S. Petrović, M. Osborne, and V. Lavrenko. Streaming First Story Detection with Application to Twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT ’10, pages 181–189. Association for Computational Linguistics, 2010.

[6] E. Summers. An ode to node, Nov. 2011. `http://inkdroid.org/journal/2011/11/07/an-ode-to-node/`, accessed 02/18/2013.

[7] M. Tsagkias, M. de Rijke, and W. Weerkamp. Linking Online News and Social Media. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM ’11, pages 565–574. ACM, 2011.

[8] B. Vibber. Current events and traffic spikes, June 2009. `http://blog.wikimedia.org/2009/06/25/current-events/`, accessed 02/18/2013.

[9] D. Vrandečić. Wikidata: A New Platform for Collaborative Data Collection. In *Proceedings of the 21st International Conference Companion on World Wide Web*, WWW ’12 Companion, pages 1063–1064. ACM, 2012.