The Citizen IS the Journalist - Automatically Extracting News from the Swarm

João Marcos de Oliveira[1], Peter A. Gloor[2]

[1]FederalUniversityof Juiz de Fora Minas Gerais, Brazil, jmarcosdo@gmail.com
[2]MIT Center for Collective Intelligence Cambridge MA, pgloor@mit.edu

Abstract User generated content has become a major trend in today's journalistic ecosystem, where in many cases news arrive on social media platforms even before they reach mainstream media. Due to today's hyperconnected society this type of event is becoming more frequent and “news-like” information is being produced all over the Internet on blogs, posted on Facebook or Twitter, Wikipedia or any other platform that allows users to share their ideas and experiences. In this paper we describe Swarmpulse, a system that extracts news by combing through Wikipedia and Twitter to extract newsworthy items. We measured the accuracy ofSwarmpulse comparing it against the Reuters and CNN RSS feeds and the Google News feed. We found precision of 83% and recall of 15 % against these sources.

1. Introduction

Wikipedia is one of the top ten websites, with more than 37 millions articles in more than 250 languages. It lists approximately 27 millions registered editors among whom 114,000 are listed as active contributors. Those users are spread all over the world, creating a 24 by 7 online community. This community quickly creates articles based on news coming from various news sources, with some articles even written by Wikipedians involved into the actual events (Iba et al. 2010).

Twitter is another website on the top ten list. It has 340 million active users every month; those users make on average 6,000 tweets per second. Twitter also allows users to retweet and share links from external sources. Twitter provides various APIs that allow researchers to easily access its contents. In previous work (Petrovic et al. 2013)the author shows that Twitter includes information from all the main breaking news related major online journals.

Earlier studies found that Wikipedia is in some cases faster than conventional news channels (Becker et al. 2011).These observations formed the foundation of the Wikipulse project (Futterer et al. 2013) and prompted (Fuehres et al. 2012)to propose the use of Wikipedia content to find “latest trends based on the analysis of recent edits on Wikipedia articles.” Others studies had found that Twitter has become a popular news channel. The Swarmpulse project introduced in this paper combines these ideas by generating latest news based on Wikipedia article edits and the most recent tweets, presenting them in an user friendly news format.

The main contributions of this paper are the SwamPulse algorithm, which combines Twitter and Wikipedia to generate news automatically, the description of a first implementation, and an algorithm for measuring the accuracy of Swarmpulse.

2. Related work

In earlier research studying the collaborative behavior of Wikipedia editors, (Bayer et al.) found that unlike just many eyes having a look at an article, the experience of the editors is important – they should have worked on many other articles for the quality of their articles to be good. It was also found that a high number of editorial events contribute positively to a page's quality. In other earlier work (Becker et al. 2011), it was found that entertainment and sports news appeared on average about two hours earlier on Wikipedia than on CNN and Reuters online. Wikirage, another Wikipedia-based news system, tracks the pages in Wikipedia which are receiving the most edits over various periods of time(Wood).While this site does a good job collecting the edits, it does not process the results further and as evidenced in (Bayer et al. 2012) edits alone are not enough to justify newsworthiness. Nevertheless, Wikirage delivers a good benchmark to validate against the results of our news generation algorithm.

Twitter has also been used as a news detection tool, many research projects have proven its value to find relations between its data and real world events. (Sakaki et al. 2010) uses tweets to build an earthquake detection system based on the frequency of tweets and hashtags in specific locations. Another project using twitter to find breaking news events is discussed in (Petrovic et al. 2013), in this work the authors build a FSD (First Story Detection) system using the tweets retrieved from specific users employing hashtags like “#breakNews”, “#News” and so on, subsequently ranking and clustering those tweets into groups. The problem with this approach is the limitation by the number of users and that the system loses most of the comments that come from users outside of that list. Twitterstand is another application that uses twitter to identify breaking news. It increases its accuracy by only showing trusted sources. Also, in other research it was found that although most of the breaking news can be found on twitter, twitter users mostly are not creating or contributing to news but comment on them (Subašić and Berendt 2011).

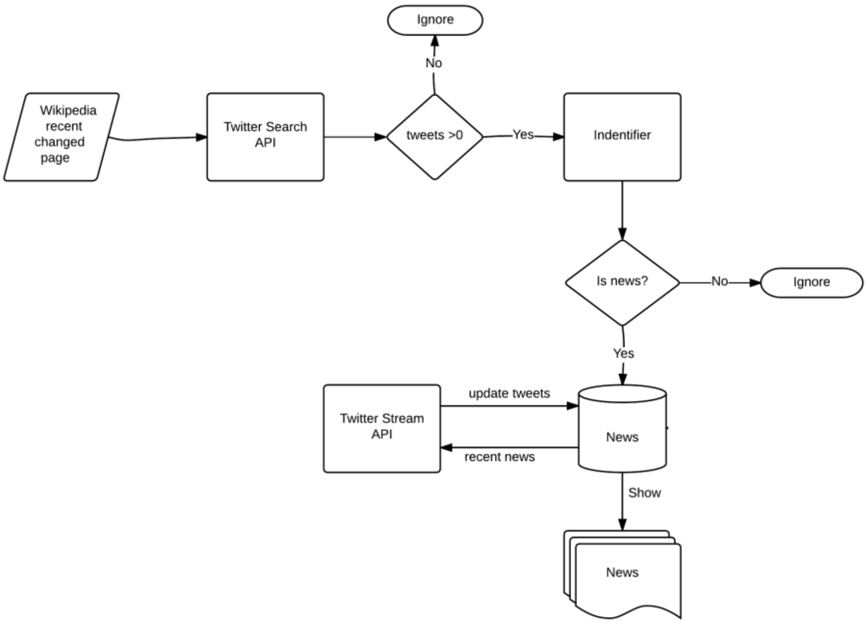
Another system using Wikipedia and Twitter for breaking news detection can be found in (Osborne et al. 2012). In this system the authors analyze the use of Wikipedia as a possible filter for news extracted on Twitter. Wikipedia news detection is done by counting the number of views of a Wikipedia page in a time period then analyzing this data to identify the increase in page views. A similar analysis is made on Twitter. As a result it was found that Wikipedia seems to lagging behind Twitter by about two hours. In our system we analyze the Wikipedia text information without looking at the frequencies of edits and news. While most of the pages identified by the system had an increase in the number of changes in the time period, we are using Twitter as our news filter.

3. Swarmpulse Algorithm

In order to create our Swarmpulsenews-reader using data from Twitter and Wikipedia we follow the steps as shown in figure 1.

1. Extract data from Wikipedia
2. Search recent tweets
3. Test their newsworthiness
4. Rank the article
5. Display the news

Fig. 1. Swarmpulse algorithm



We will now describe these steps in detail.

1. Extract data from Wikipedia

The first step to build our news-reader is to extract the most recently edited articles from Wikipedia using Mediawiki. Mediawiki is another project under the umbrella of the Wikipedia Organization. It gives access to the Wikipedia metadata. With this API it is possible to request the links, categories, editors and even the edit history of an article. Mediawiki also provides specific queries to collect the most recently edited articles from Wikipedia, which is exactly what we need. Once we have collected those articles we have a list of potential news candidates.

To identify breakings news events, some studies suggest to analysis the connection between editors or the frequency of edits in the article (Ciglan and Nørvåg 2010). Although those methods had been proven to work in Futterer (Futterer et al. 2013)Swarmpulse gives this task to the end users, by using Twitter as a first news filter.

2. Search recent tweets.

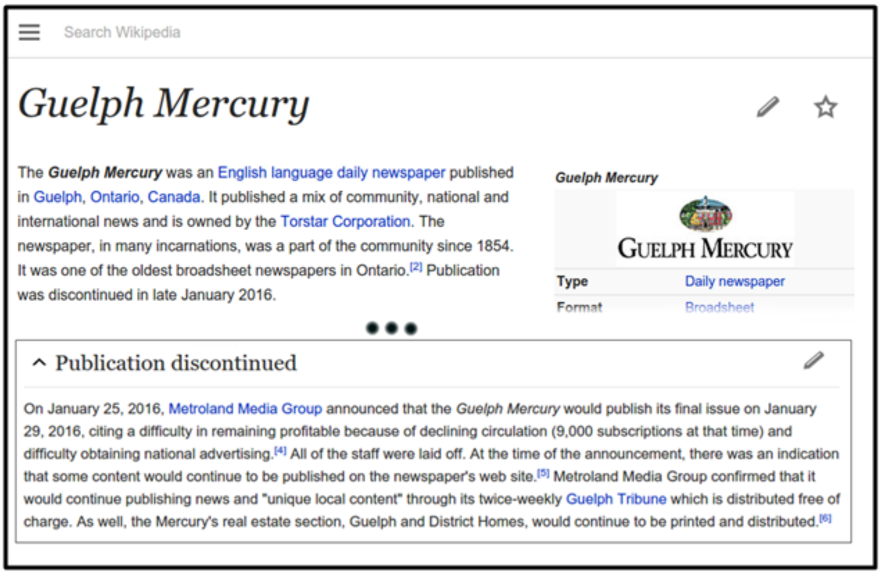
The next step is to combine the Wikipedia article with the tweets. The Twitter API allows developers to have access to some of its data. In this project we use two APIs: The Search API and the Streaming API. As its name suggests, the search API is used to make specific searches for tweets. It enables search for specific words, users, dates and so on. However, it has some limitations. The main current limitation comes from Twitter's API in that it allows only 450 queries per 15 minute window, with a limit of 100 results per query. In our implementation we had to work around this restriction to ensure that it did not block processing of the data and put a limit on the number of Wikipedia pages the system can evaluate. The Streaming API gives access to the most recent tweets, it does not have a tweets limit. It depends of how fast one can process the data.

Now that we have explained how the Twitter API works we can continues with the news filter. Using the news candidates that we found using the Mediawiki, we can search for recent tweets using the Twitter Search API. For each news candidate from Wikipedia, we search for related recent tweets. If we find a keyword match we hypothesize that something is probably happening in the topic of the article. Figure 2 shows a news candidate article in Wikipedia.

3. Identify newsworthiness

For to the news identification algorithm we take advantage of the Wikipedians' writing style. Wikipedia by definition will give a history of events, that implies telling when important events happened or are going to happen. So, using this information we can use a natural language processing algorithm to identify dates in the article and match those dates with sentences or paragraphs. In this step, we also process other data from the Wikipedia page. For instance we collect the Wikipedia categories for a page and then use specific tags to map those categories into 5 categories and 19 subcategories (table 1). If the date is equal or close to the current date, we have found an event related to the present one where we already have tweets about it. Figure 2 shows the distribution of different categories in Swarmpulse.

Fig. 2. Wikipedia News Example



Guelph Mercury

The ***Guelph Mercury*** was an English language daily newspaper published in Guelph, Ontario, Canada. It published a mix of community, national and international news and is owned by the Torstar Corporation. The newspaper, in many incarnations, was a part of the community since 1854. It was one of the oldest broadsheet newspapers in Ontario.[2] Publication was discontinued in late January 2016.

Guelph Mercury

Type	Daily newspaper
Format	Broadsheet

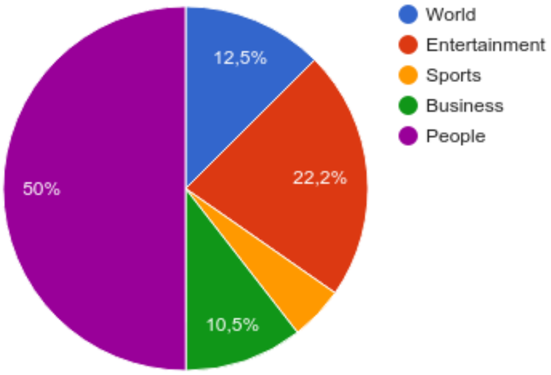
Publication discontinued

On January 25, 2016, Metroland Media Group announced that the *Guelph Mercury* would publish its final issue on January 29, 2016, citing a difficulty in remaining profitable because of declining circulation (9,000 subscriptions at that time) and difficulty obtaining national advertising.[4] All of the staff were laid off. At the time of the announcement, there was an indication that some content would continue to be published on the newspaper's web site.[5] Metroland Media Group confirmed that it would continue publishing news and "unique local content" through its twice-weekly Guelph Tribune which is distributed free of charge. As well, the Mercury's real estate section, Guelph and District Homes, would continue to be printed and distributed.[6]

Table 1. Categories and subCategories

Categories	Subcategories
Business	Companies
Entertainment	Games, Movies, Music
People	Actors-Actresses, Businesspeople, Players, Politicians, Singers, Writers
Sports	Sports, Baseball, Basketball, Football
World	Africa, Asia, Europe, North America, South-America

Fig. 3. Percentage of each News category in Swarmpulse



4. Rank the article

Finally we need to rank the articles found in step 3. Again, we leave that task to the Twitter users. Using the Twitter Streaming API to collect the most recent tweets about the most recent events identified in the previous step, we then add those tweets to the article, ranking the news by the number of associated tweets. Table2 shows the top ranked news on January 20 2016, We can see in this table that the Sarah Palin endorsement of Donald Trump's presidential campaign was the news with the highest reflection on Twitter. Twitter also gives more information about the articles such as the reactions of people and links to other news sources

Table 2. Top news on 2016-01-20

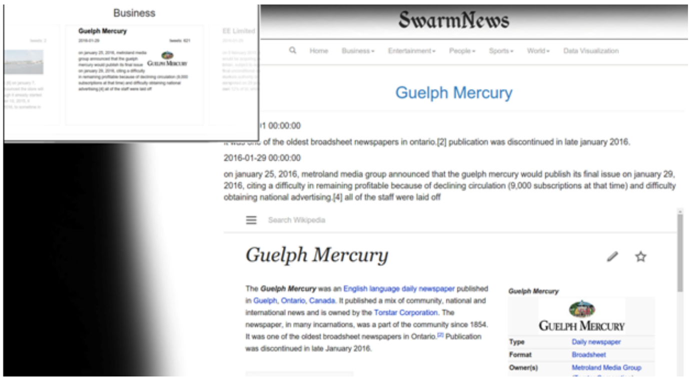
Total Tweets	page id	Title	News
127917	2941511	Sarah Palin	onjanuary 19, 2016, palin endorsed donald trump's campaign to become preside...
28412	175537	Netflix	reedhastings admitted that netflix's china expansion could take "many years" on ...
18181	536880	Glenn Frey	souther,[30] jack tempchin,[31] irvingazoff,[32] lindaronstadt,[33] don felder,[34] ...
16137	40884573	@midnight	note: minimum of 3 wins or 5 appearances, updated on january 26, 2016
7227	43630864	Jihadi John	on 19 january 2016 in the isil magazine dabiq, the group confirmed that emwazi ...

4682	16175	Jennifer Lopez	ryanseacrest will also produce the series.[176] it premiered on january 7\
4633	28944259	Jorge Sampaoli	on 19 january, 2016 sampaoli and the chilean federation mutually decided to ...
3952	33804	Wellington	retrieved 19 january 2016.
3808	46299779	Legends of Tomorrow	the series airs on the cw and premiered on january 21, 2016
2119	3861139	Steven Naismith	* senior club appearances and goals counted for the domestic league only and correct...
1817	23184259	4Minute	onjanuary 20, 2016, it was announced that the group would release their seventh mini album ...
1787	4106856	Jim Schwartz	onjanuary 19, 2016, schwartz was hired by the philadelphia eagles to be their defensive coordinator.[23]
1124	26903	Solar System	onjanuary 20, 2016 astronomers at the california institute of technology announced a possible ninth planet...
1063	24831215	Zika virus	on 15 january 2016, cdc issued a level 2 travel alert for people traveling to regions and...

5. Display the news

Once we have collected all this information, we display it on a specific website, with the event as a news header, including part of the Wikipedia page for more information as well as the tweets to allow users to search for more links and to look at reactions about the topic.

Fig. 4. Swarmpulse page of Guelph Mercury



Guelph Mercury

1 00:00:00

of the oldest broadsheet newspapers in ontario.[2] publication was discontinued in late january 2016.

2016-01-29 00:00:00

on january 25, 2016, metroland media group announced that the guelph mercury would publish its final issue on january 29, 2016, citing a difficulty in remaining profitable because of declining circulation (9,000 subscriptions at that time) and difficulty obtaining national advertising.[4] all of the staff were laid off

Guelph Mercury

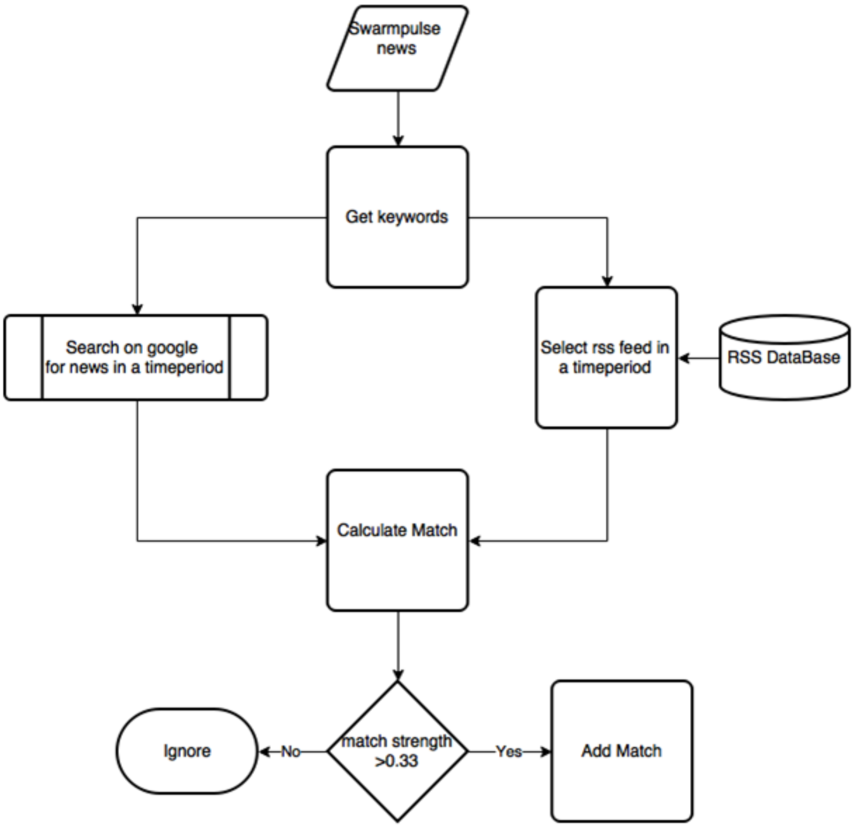
The *Guelph Mercury* was an English language daily newspaper published in Guelph, Ontario, Canada. It published a mix of community, national and international news and is owned by the Torstar Corporation. The newspaper, in many incarnations, was a part of the community since 1854. It was one of the oldest broadsheet newspapers in Ontario.[2] Publication was discontinued in late January 2016.

Guelph Mercury	
Type	Daily newspaper
Format	Broadsheet
Owner(s)	Metroland Media Group

4. Data Analysis

In order to measure the performance of the news identification algorithm, we compared the news found in two different ways: comparing the news with the RSS feed from CNN and Reuters online, and using the Google news page to get news for a specific search term, obtaining its results as RSS feeds. The main challenge is to match two news items reported from two different websites because these two news items might be about the same events using completely different words. To overcome this obstacle we developed a keyword based matching heuristic (figure 5).

Fig. 5. Data Analysis Process



We first remove all stopwords from the RSS feeds and the news found in Swarmpulse. Then we process the content of each news item, creating two list of keywords from the Wikipedia article, one for the title and the other for the content. We repeat this process for the CNN, Reuters, and Google News RSS feeds. The news items from Wikipedia are then compared with the news items from the RSS feed using the keywords. A match is expressed in terms of 'match strength' - a fraction between 0 and 1 with 0 indicating no match and 1 indicating a perfect match. To calculate the match we use a weighted arithmetic mean using the formula:

$$F = \frac{\alpha t + \beta c}{\alpha T + \beta C}$$

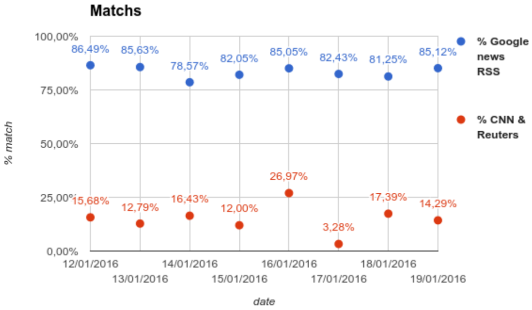
where αis the weight for a match on a 'title" tag and βis the weight for a content tag. 'c' and 't' represent the number of tags from the content or the title that were found in the RSS feed. 'T' and 'C' are the total number of tags in the list of keywords. Based on experimental testing we found that a coefficient of 0,33 or higher indicates a good match. this approach is similar to the one used in Wikipulse which also found by human evaluation the same score (Fuehres et al. 2012).

With this heuristic we found that on average 14.85% of the news reported on Wikipedia were present on the CNN and Reuters rss feed. Comparing Swarmpulse with the Google news search page, we found that 83,32% of the news we found were reported on the days we ran the analysis. This analysis was made from 01/12/2016 to 01/19/2016 during this time Swarmpulse was able to indentify over 680 possible news items. This illustrates that most of the content retrieved from Swarmpulse are in fact news, and although many News outlets did not report it, those news were found in other places. Table 3 shows an example of the match strength for the Wikipage "Daniel Holtzclaw", with the news that "on December 10, 2015, an all-white jury convicted him on 18 of 36 charges, and on January 21, 2016 he was sentenced to 263 years in prison."

Table 3. Match strength example

News Source	Best news match	match strength
cnn	danielhotlzclaw the ex oklahoma city officer convicted of rape ...	0,425
reuters	to access the newsletter click on the link http share thomsonreuters ...	0,225
google news	the latest ex officer convicted of rape won t get new trial ...	0,450

Fig. 6. Percentage of Matches in Google News and CNN and Reuters RSS data feed



The reader can try out the prototype version of swarmpulse at swarmpulse.galaxyadvisors.com.

5. Limitations

Although a user generated news portal opens up many new opportunities, it will have limitations with regards to the content generated. Wikipedia's main focus is on global or national events, which makes its data less useful for local news, except for local news of national interest. Another key question is the trustworthiness of the Wikipedia pages. While Wikipedia is geared towards discovering fake news quickly, for a limited amount of time a fake Wikipedia page can show up once we have some tweets covering the subject (there was for instance a case where for a short period of time the Ebola page claimed that Ebola was caused by gays). A possible next step might be to test the resulting SwarmPulse news articles with human readers to determine its usefulness. It may be necessary to have a human editor edit the cacophony of such a unique combination of text to present a "journalist's" coherent written story.

Another open issue is the dependency of our system on externals APIs, while Wikipedia is free and open source, Twitter is run by a commercial company which over the past years has repeatedly changed availability and terms of use of their Twitter stream.

6. Future Work and Conclusions

The extraction of news from Twitter and Wikipedia opens the door to many exciting possibilities. Among those possibilities are the connection between articles, the sentiment analysis of the tweets, data visualizations and others.

We believe that this information can be used in many different fields. For instance it can be used by conventional News media to identify breaking news through data collected from social media and, to gauge the relevance of news to users. Also, due to the information from the tweets, we have direct links to different sources of information that can be applied to create an index to news articles.

We also believe that data visualization is a big opportunity to create cybermaps that represent the network structure between different news items, giving us a first impression of how events might be connected with each other.

The main contribution of our work is a novel news reader that combines Wikipedia articles and Twitter data. It opens new windows of opportunities to different types of analysis, leveraging the power of the "citizen journalist" as a trusted provider of late breaking news.

7. References

Bayer T, Ford H, Tar D, Romanesco Quantifying quality collaboration patterns, systemic bias, POV pushing, the impact of news events, and editors'reputation. http://en.Wikipedia.org/wiki/Wikipedia:Wikipedia_Signpost/2011-11-28/Recent_research.

Becker H, Naaman M, Gravano L (2011) Beyond trending topics: Real-world event identification on twitter.

Ciglan M, Nørvåg K (2010) WikiPop: Personalized Event Detection System Based on Wikipedia Page View Statistics. Proceedings of the 19th ACM International Conference on Information and Knowledge Management. ACM, New York, NY, USA. doi:10.1145/1871437.1871769

Fuehres H, Gloor PA, Henninger M, Kleeb R, Nemoto K (2012) Galaxysearch - Discovering the Knowledge of Many by Using Wikipedia as a Meta-Searchindex. ArXiv e-prints

Futterer T, Gloor PA, Malhotra T, Mfula H, Packmohr K, Schultheiss S (2013) WikiPulse - A News-Portal Based on Wikipedia. ArXiv e-prints

Iba T, Nemoto K, Peters B, Gloor PA (2010) Analyzing the Creative Editing Behavior of Wikipedia Editors: Through Dynamic Social Network Analysis. Procedia - Social and Behavioral Sciences 2:6441-6456. doi:http://dx.doi.org/10.1016/j.sbspro.2010.04.054

Osborne M, Petrovi Sa, McCreadie R, Macdonald C, Ouni I (2012) Bieber no more: first story detection using Twitter and Wikipedia. Proceedings of the SIGIR Workshop in Time-aware Information Access. Association for Computing Machinery,

Petrovic S, Osborne M, McCreadie R, Macdonald C, Ounis I (2013) Can twitter replace newswire for breaking news?

Sakaki T, Okazaki M, Matsuo Y (2010) Earthquake shakes Twitter users: real-time event detection by social sensors. Proceedings of the 19th international conference on World wide web.

Subašić I, Berendt B (2011) Peddling or creating? investigating the role of twitter in news reporting. Advances in Information Retrieval. Springer,

Wood C wikirage. http://www.wikirage.com/topedits/.