This document is downloaded from DR-NTU (https://dr.ntu.edu.sg)
Nanyang Technological University, Singapore.

Social tags as news event detectors

Chua, Alton Yeow Kuan; Razikin, Khasfariyati; Goh, Dion Hoe-Lian

2011

Chua, A. Y. K., Razikin, K., & Goh, D. H. (2011). Social tags as news event detectors. Journal of Information Science, 37(1), 3-18.

https://hdl.handle.net/10356/100038

https://doi.org/10.1177/0165551510389108

© 2011 The Author(s).

Downloaded on 23 Apr 2021 02:35:26 SGT

Chua, A., Razikin, K.B. & Goh, D. H. (2011). Social tags as news event detectors. Journal of Information Science, 37(1): 3-18

Social tags as news event detectors

Alton Y.K. Chua, Razikin, K.B. and Dion H.Goh

Abstract

The objective of this study was to investigate the use of tags in iReport to detect breaking news in terms of coverage and immediacy. Coverage refers to the extent to which news reported in mainstream media can also be detected in iReport, while immediacy refers to the promptness of news reported in mainstream media vis-à-vis those detected in iReport. A total of 10 ground truth events were identified from mainstream media between 1 April 2008 and 31 December 2008. Additionally, 481,455 tags from 118,545 postings were drawn from iReport in the same period. Relative frequencies of the top 200 most frequently-used tags were analysed to check for spikes and bursts. Based on the results, four main findings emerged. First, the performance of using spikes and bursts to detect news events was found to be comparable. Next, news events detected via spikes and bursts were found to lag ranging from a few days to more than a week compared to the dates reported by mainstream media. Third, news events deemed to be significant by professional journalists did not always attract a high level of interest from iReport contributors. Finally, even though citizen journalism transcends national boundaries via the internet, news posted to iReport seemed to show a proclivity towards local context.

1. Introduction

The way news reaches audiences globally has been changed dramatically with the advent of user-generated content. While the populace in the past generally depended on mainstream media for the latest news, they can now access alternate online news sources contributed by fellow citizens. These online sources illustrate a phenomenon known as citizen journalism where ordinary people who are non-journalists collect, analyze and disseminate news pieces to the masses, not unlike what professional journalists do [1]. The rising prevalence of citizen journalism is driven by factors including the lowering cost of Internet connectivity [2], the development of user-friendly online content management tools [3] and the vast improvements made on consumer electronics such as camera phones which enable users to capture and upload pictures and videos on the Internet easily [4].

Citizen journalism gained some level of prominence in the aftermath of the 9/11 tragedy as eyewitnesses posted stories and images of the attack on the Internet. However, it was the 2005 London terrorist bombings that citizen journalism caught the attention of mainstream media. The earliest photos of the blasts captured by members of the public on their mobile phones were published on blogs and social media such as Flickr before appearing in national newspapers and television newscasts around the world the next day [5]. Thereafter, breaking news such as the 2007 Virginia Tech shooting and 2008 Sichuan earthquake further accentuated the role of citizen journalism when video footage and vivid descriptions from survivors and witnesses appeared on the Internet shortly after these incidents occurred.

Recognizing the value of grassroots' reporting, many news media companies have augmented their print and television programming with citizen journalism. In particular, CNN created a platform for anyone to share newsworthy pieces in text, images and videos through its iReport site (www.ireport.com). In iReport, the disclosure of users' real identities is optional. Also, since submitted entries are not subjected to any editorial censorship, they become available immediately once they have been posted. Readers can access the submitted entries in multiple ways including listing them by viewership popularity, or filtering them on the basis of tags used by contributors. To date, iReport has garnered more than 400,000 postings along with some 1.5 million tags associated to the postings. A tag is an uncontrolled keyword assigned by the user to annotate Web-based content. As user-generated metadata, tags are a means through which content can be succinctly described, searched and accessed [6]. The exponentially growing volume of tags available on the Internet has led to research endeavours which use tags for a variety of purposes including Web resource classification [6], community identification, ontology generation as well as user and document recommendation [7]. Another burgeoning area lies in the identification of events through tags on social media such as Flickr [8].

Even as the credibility and legitimacy of citizen journalism Websites such as iReport have become more established, research done to exploit the ever-growing reservoir of tags from such alternate news sources has been limited. Hence, the objective of this study is to investigate the use of tags in iReport to detect breaking news in terms of coverage and immediacy. Coverage refers to the extent to which news reported in mainstream media can also be detected in iReport, while immediacy refers to the promptness of news reported in mainstream media vis-à-vis those detected in iReport. This study thus represents an effort to push the frontier of information science research using user-generated content. Beyond the context of news detection, the findings can also broaden our understanding of how tags can be used to make sense of a large body of documents.

The structure of this paper is as follows. The next section presents the literature related to citizen journalism and social tagging. Following that, the Methodology section explains the data collection and analysis procedures. The detailed results are presented and explained in the Results section. Four major findings in this study are highlighted in the Discussion section. The final section concludes the paper and provides a few possible research directions for scholars interested to study news event detection using tags.

2. Related Studies

For most of the twentieth century, news was primarily delivered by the press and television/radio broadcasting [9]. However, as the world becomes more connected in the twenty-first century through the Internet, the public's consumption pattern of news and attitude towards news reporting has changed. Not only do people who are Web-savvy expect to have access to breaking news anytime and anywhere, the once mere consumers of news have started to participate in the process of citizen journalism, helping to create a massive conversation among themselves and anyone interested. Amid the mushrooming of blogs and wikis that publish independent news-related content, several mainstream media organizations are making efforts to involve non-journalists who are keen on reporting news. For example, The Washington Post (www.washingtonpost.com) embeds live Technorati updates for each of its stories, paving the way for its readers to become citizen journalists and commentators in the web community [5]. BBC (www.bbc.com) probes its news users of their views about the news and then publishes them in a particular section of the news product while MSNBC (www.msnbc.com) makes provision for editors to suggest assignments for anyone who wish to report on specified aspects of news stories that are unfolding [10]. The response from users has hitherto been overwhelming. OhmyNews.com, a South Korean online newspaper, has more than 37,000 registered contributors; Britain's second most popular news website, Guardian.co.uk, hosts a 'News' message board to which users contributed more than 600,000 messages between 1999 and 2005 [3].

One of the major advantages of citizen journalism is that citizens with camera phones can capture images of news events more promptly than professional journalists, at least in the early minutes of the events. Also, if equipped with mobile access to the Internet, citizens can broadcast their photos, along with text content, immediately to the world [11]. Furthermore, given its open and participatory nature, diverse views of a given news event can be expected. Thus, the corpora of content created by citizen journalists could potentially be used to mine for the occurrences of major events.

Current event detection techniques generally seek to determine whether a news story contains an event by comparing the similarity between features in the current story and those in past news stories [12]. Grouping of events by their relative similarities and differences helps to track events across time. This has been introduced in text-based topic detection and tracking, which uses lexical similarity of document texts to generate coherent clusters in which elements in the same cluster share an identical topic [13]. Improvements have been made to existing event detection techniques by incorporating the event's activeness trend and adaptively adjusting the clustering threshold during the event detection process [14]. Another approach is to consider the time gap between events. The time gap between bursts of topically similar stories is often an indication of different events and the incorporation of a time window for event scoping has commonly been adopted [12; 15].

Event detection techniques can be classified into two forms: retrospective detection and online detection. Retrospective event detection (RED) seeks to discover previously undefined events from a chronologically ordered accumulation of news stories [16] while online detection strives to identify the onset of events emerging from live news feeds in real time [12]. Both forms of detection rely on

historical news stories which contain two kinds of information, namely, contents and timestamps. Many previous studies tend to focus on the exploitation of contents but the usefulness of time information has often been ignored [16]. Taking the RED approach, this study used time information in the analysis.

In a parallel line of development, social tagging has been gaining traction on the Internet [6]. Since 2004, an increasing number of Websites including del.icio.us, Flickr, YouTube as well as those dedicated to citizen journalism allow users to annotate Web resources such as Web pages, images and videos using user-defined tags [17]. The confluence of the rising popularity of social tagging and citizen journalism presents the opportunity to investigate the use of tags to detect news events. When an event of wide-spread significance occurs, a sharp rise in Web activity related to that event is observable [18]. In the case of citizen journalism Websites, breaking news invariably triggers an increase in the number of uploads and tags. Given, that a user's choice of tags is influenced in part by other users in the community [19], a sudden surge in the frequencies of similarly-themed tags is vestige of the incident [8].

For the purpose of news detection, the use of tags is preferred over titles and main texts for two main reasons. One, tags have been found to be different from words found in titles and main texts, suggesting their role in explicating the content [20]. They are also usually short and do not contain stop-words. Two, the process of generating tags appear to be similar to the process of generating search terms for subsequent retrieval [21]. Tags could thus serve as a suitable proxy of the content to which they have been associated [4]. Furthermore, while news detected from social media such as Flickr and Twitter had been attempted [8, 22], the use of tags in lieu of full text from citizen journalism websites remains relatively unexplored. This study can therefore be used as a springboard to develop more ideas and approaches for detecting events from social tagging systems.

3. Methodology

3.1 Data Collection

Data was collected through a two-step process. The first step was to build a ground truth dataset which involved manually identifying news events and the dates of their reporting from mainstream media during the study period between April 1 2008 and December 31 2008. Such an approach has been similarly adopted in previous studies [13]. A total of 10 news events were drawn from reputable news agencies including BBC (http://www.bbc.co.uk), CNN International (http://edition.cnn.com), Wall Street Journal and New York Times (http://www.nytimes.com). Each news event was selected on the basis of three criteria. One, it had to be featured among the top 10 news in 2008 in two or more mainstream media. Two, it must not be pre-announced. In other words, its emergence had to be unexpected. For example, while news about the Beijing Olympics and U.S. presidential election per se per se did not fit this criterion, the occurrence of unforeseen incidents linked to these news events could be accepted. Three, these events had to occur during the study period.

The second step was to collect tags from iReport. During the study period, there were 36,536 unique users who contributed a total of 118,545 postings. The average daily number of postings was 478 (S.D. = 461.52). The postings were annotated with 481,455 tags out of which 28,823 were unique. Among the unique tags, the top 200 most frequently-used tags (henceforth simply known as tags) were extracted for further analysis.

3.2 Methods of Analysis

The unit of measurement for detecting news events used in this study was relative frequency. Relative frequency was preferred over absolute frequency as it would provide normalized baselines across time for all tags and tag pairs. The relative frequency of a tag on a given day was computed by dividing the frequency of the tag by the total number of tags on that day [23, 24].

The relative frequencies of the tags were computed daily throughout the study period. To detect news events, two metrics, namely spike and burst were used [25]. Both metrics were included since neither had shown to out-perform the other in previous work [23]. A spike in the relative frequency of a tag on a given day is defined as having at least five times the average relative frequency of the tag on all the previous days. This sudden rise in relative frequency could indicate the occurrence of a

news event to which that tag was associated. A burst in the relative frequency of a tag is defined as a timeframe during which the minimum relative frequency on any given day is at least three times higher than the average relative frequency of that tag on all the previous days. Informed by an earlier work [23], a burst is defined using three timeframes in this study, namely, a three-, five- and nine-day window. Thus, bursts in different timeframes provide clues to the length over which the level of interest on a news event had been sustained.

4. Results and Analysis

4.1 Ground truth collection

Shown in Table 1 are 10 news events which represent the ground truth in this study. All the news events fulfilled three criteria, namely, they had been featured in the top 10 news events in 2008 by at least two or more mainstream media, they were unanticipated, and had occurred during the study period. The themes of these news events range from natural catastrophes to war, terrorism and issues related to politics and the economy.

Table 1. Ten events selected as ground truth

No.	Event	Date reported
1.	Chaos on Olympic torch's global route	April 6
2.	Cyclone Nargis in Myanmar	May 2
3.	Earthquake in China's Sichuan Province	May 12
4.	Betancourt's rescue from Colombian guerrilla fighters	July 2
5.	Oil price boost	July 11
6.	Russia-Georgia war	August 7
7.	Lehman Brothers' bankruptcy	September 15
8.	Obama's presidential victory	November 4
9.	Terrorist attacks in Mumbai, India	November 26
10.	Israel air strike on Gaza Strip	December 27

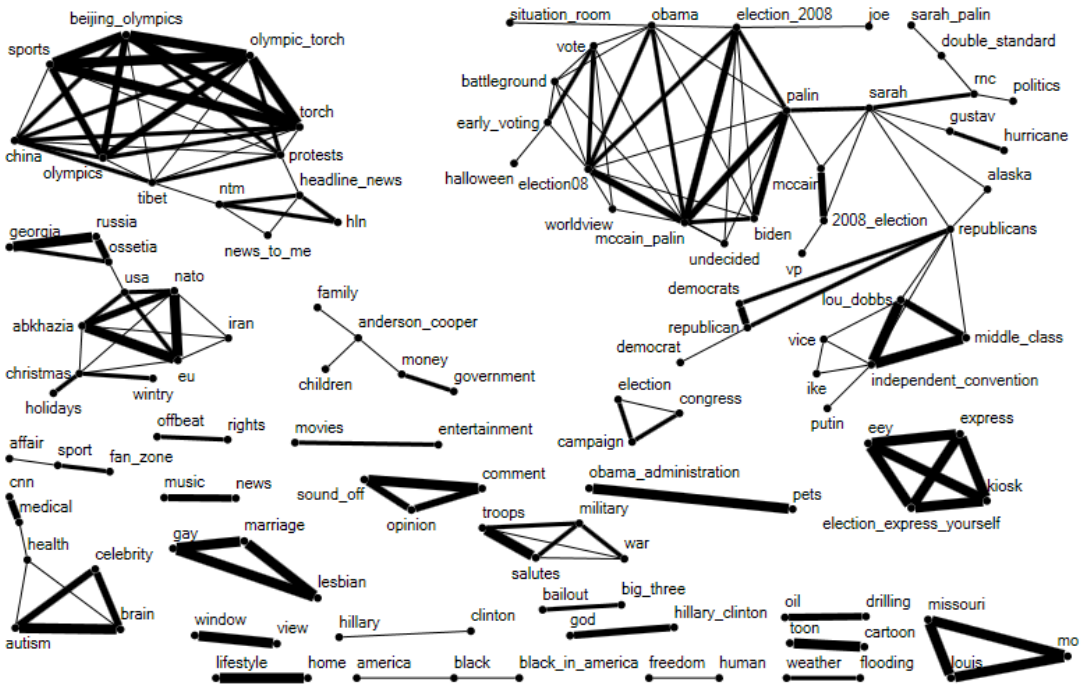
4.2 Tags

The tags are listed in Table 2. Pair-wise correlation analysis was performed among the tags. Due to space constraints, only correlated tags with a value of more than 0.6 and are statistically significant ($p < 0.05$) are shown in Figure 1. The thickness of the edges denotes the strengths of correlation between the tags.

Table 2. Tags in iReport during the study period

election08, obama, mccain, economy, ireport_for_cnn, 2008_election, politics, palin, election, opinion, travel, vp, sound_off, russia, weather, comment, georgia, obama_biden, debate, china, offbeat, situation_room, war, mccain_palin, money, hurricane, news_to_me, gas_prices, military, bailout, storm, black_in_america, olympics, prep, sarah_palin, biden, campaign, headline_news, republican, democrats, election_express_yourself, undecided, troops, usa, obama_administration, cnn, gay, ossetia, bush, health, marriage, news, barack, president, clinton, cartoon, ike, lesbian, tibet, iraq, double_standard, government, salutes, america, sarah, entertainment, world, 2008, vote, holidays, express, flooding, toon, barack_obama, eey, environment, hillary, kiosk, sports, beijing_olympics, family, pets, olympic_torch, rights, autism, earthquake, protests, view, republicans, morning_express, cutting_costs, energy, window, cnn_money, torch, oil, fall, celebrity, anderson_cooper, wintry, terrorism, larry_king, congress, gustav, gas, democrat, freedom, big_three, political, black, race, elections, john, photography, early_voting, religion, election_2008, wildfires, eu, christmas, worldview, rnc, ireport_kit, movies, human, racism, india, children, fan_zone, issue_1, american, house, lifestyle, nato, university, battleground, crisis, hln, women, brain, putin, education, weather_fx, cars, protest, ntm, policy, prices, middle_class, music, space, green, severe_weather, united, prime_news, joe, lou_dobbs, iran, change, college, media, peace, alaska, california, food_costs, florida, god, san, snow, independent_convention, free, community, missouri, message_for_obama, affair, fun, hillary_clinton, sport, corruption, halloween, vice, fuel, abkhazia, louis, drilling, army, washington_dc, cnn_heroes, white, home, africa, presidential, stories, dnc, medical, israel, tornado, life, mo, energy_fix

Figure 1. Highly correlated tags



A number of tags in Figure 1 could be traced to significant news events identified in Table 1. Specifically, the largest cluster comprises tags such as ‘election08’, ‘obama’, ‘mccain’, ‘palin’, ‘biden’ that pertain to the U.S. presidential election. Smaller clusters that were contextually related to the event include tags such as ‘election’, ‘campaign’, ‘congress’, ‘election_express_yourself’, ‘obama_administration’, ‘hillary_clinton’, ‘hillary’ and ‘clinton’. The second largest cluster comprises tags such as ‘china’, ‘olympics’, ‘tibet’, ‘sports’, ‘beijing_olympics, ‘olympic_torch’, ‘protests’ and ‘torch’ which are related to the protests during the Olympic torch relay. A third cluster comprises tags such as ‘georgia’, ‘russia’, ‘ossetia’, ‘nato’, ‘abkhazia’. These tags are linked to political issues, and in particular, the conflict between Russia and Georgia in South Ossetia.

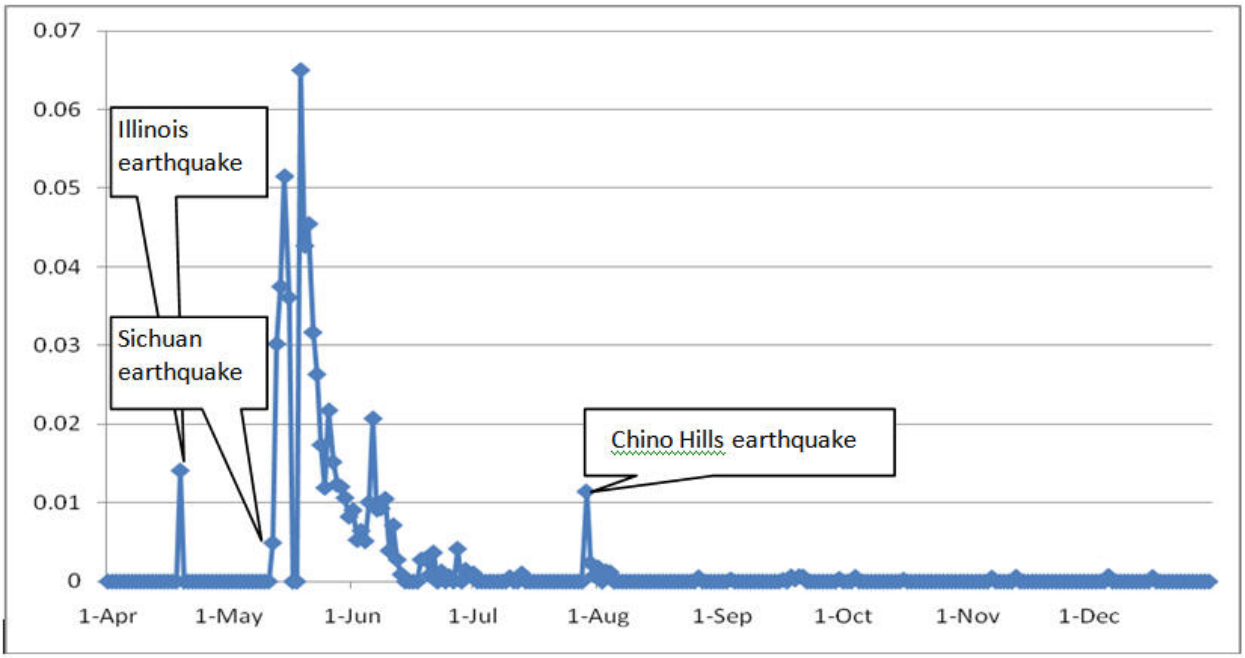
In line with their social nature, tags such as ‘opinion’, ‘comment’ and ‘view’ explicitly point to personal viewpoints but the absence of specific contexts prohibited further interpretation. More interestingly, personalized tags including ‘sound_off’, ‘news_to_me’, ‘black_in_america’, ‘fan_zone’, ‘hln’, ‘ntm’ are found, reflecting social proof, a phenomena describing how people look to others to guide their own actions, and in this case, the choice of tags [26].

4.3 Tag with spikes

In general, tags associated with an unexpected event would experience a sharp spike in their relative frequencies at the onset of that event [27]. The occurrence of a natural disaster such as an earthquake offers a good illustration. Figure 2 shows the spikes in the relative frequency for the tag ‘earthquake’. The three spikes coincided with earthquakes that occurred in Illinois on April 19 2008, the Sichuan earthquake in China on May 12 2008 and the Chino Hills earthquake in California on July 29 2008 respectively.

Chua, A., Razikin, K.B. & Goh, D. H. (2011). Social tags as news event detectors. Journal of Information Science, 37(1): 3-18

Figure 2. Spikes for tag 'earthquake'



The relative frequencies r(d) and average frequencies ave(r(d)) for all tags were compiled. Thereafter, the spike magnitude of each tag was computed by dividing r(d) by ave(r(d)). Of the 200 tags, 197 saw at least a spike (i.e. spike magnitude > 5) during the study period. While all these tags were included in the analysis, only the top 50 tags with the highest spike magnitude, arranged chronologically, are shown in Table 3 due to space constraints.

Table 3. Top 50 tags with the highest spike magnitude during the study period

Date	Tag	r(d)	ave r(d)	r(d)/ ave r(d)
9-Apr-08	Tibet	0.043	0.0021	20.4
9-Apr-08	protests	0.036	0.0019	19.1
10-Apr-08	olympic_torch	0.076	0.0072	10.6
29-Apr-08	opinion	0.057	0.0037	15.3
29-Apr-08	comment	0.057	0.0038	15
29-Apr-08	sound_off	0.057	0.0038	15
12-May-08	food_costs	0.067	0.0027	24.7
19-May-08	earthquake	0.065	0.0036	18.2
23-May-08	energy	0.073	0.0003	241.8
4-Jun-08	gas_prices	0.158	0.008	19.8
16-Jul-08	Cars	0.074	0.0002	377
18-Jul-08	movies	0.037	0.0007	55.2
23-Jul-08	severe_weather	0.041	0.0001	359.5
24-Jul-08	black_in_america	0.178	0.0041	43.4
28-Jul-08	entertainment	0.037	0.002	18.3
9-Aug-08	Affair	0.104	0	3406
17-Aug-08	georgia	0.108	0.0038	28.6
17-Aug-08	Russia	0.1	0.0037	27.4
19-Aug-08	college	0.063	0.0006	103.3
23-Aug-08	obama_biden	0.167	0.0001	1614.3
23-Aug-08	Vp	0.169	0.0003	644.9
25-Aug-08	cnn_money	0.057	0.0001	492.9
29-Aug-08	2008_election	0.176	0.005	35.2

29-Aug-08	mccain	0.168	0.0063	26.7
1-Sep-08	gustav	0.069	0.0006	123.6
1-Sep-08	hurricane	0.071	0.0024	29.5
4-Sep-08	politics	0.131	0.0115	11.4
3-Oct-08	debate	0.133	0.0028	47.6
4-Oct-08	Palin	0.044	0.0033	13.4
7-Oct-08	express	0.036	0.0005	67.7
7-Oct-08	election_express_yourself	0.04	0.0006	63.3
28-Oct-08	situation_room	0.079	0.0038	20.8
4-Nov-08	Prep	0.142	0.0001	1146.5
6-Nov-08	obama_administration	0.111	0	24420
6-Nov-08	Pets	0.105	0.0004	288.8
7-Nov-08	message_for_obama	0.048	0	2121.6
9-Nov-08	window	0.143	0.0003	549.8
9-Nov-08	View	0.144	0.0003	509.7
10-Nov-08	sarah_palin	0.059	0.0016	36.8
11-Nov-08	military	0.045	0.0033	13.5
15-Nov-08	lesbian	0.062	0.0015	41.8
15-Nov-08	marriage	0.071	0.0018	40.5
15-Nov-08	Gay	0.071	0.0019	38.1
16-Nov-08	wildfires	0.056	0.0013	42.9
19-Nov-08	bailout	0.05	0.0022	22.6
20-Nov-08	big_three	0.048	0.0004	130.6
3-Dec-08	India	0.039	0.0003	114.7
22-Dec-08	wintry	0.067	0.0009	72.4
30-Dec-08	holidays	0.044	0.0014	32.4
31-Dec-08	Israel	0.043	0.0005	88.9

The dates on which tags spot high relative frequencies provide clues to the time dimension of a number of significant news events identified in Table 1. For example, tags related to the events surrounding the Beijing Olympics include ‘tibet’ and ‘protests’ which were observed to spike on April 9. This represents the culmination of a chain of events starting with the tumult in Tibet against Beijing’s rule and the subsequent bloody military crackdown reported on mainstream media about a month earlier. Thereafter, protests led by pro-Tibet human rights activists disrupted the Olympics torch’s route in several cities. The tag ‘olympic_torch’ saw a spike on April 10, which was a day after the start of the U.S. leg of the Olympic torch relay.

Related to the Sichuan disaster is the tag ‘earthquake’ which saw a spike on May 19. This catastrophe was reported by the mainstream media on May 12, representing almost a week’s delay between the occurrence of the event and the surge in postings.

Tags related to the oil price boost include ‘food_costs’, ‘energy’, ‘gas_prices’ and ‘oil’. Oil prices breached the $100 per barrel mark at the end of February which in turn drove up food prices as reported by the mainstream media on April 15. Thereafter it became a heated topic, as evidenced by a spike in the tag ‘food_costs” on May 12. The tag ‘energy’ was observed to spike on May 23, which coincides with the day mainstream media reported that oil prices rose above $135 per barrel for the first time. The magnitude of the spike (i.e. 241.8) is also indicative of the high level of interest generated. From May 28 till the beginning of June, there were numerous protests by lorry drivers against the hike in fuel price from various countries such as UK, France, Spain and Thailand. This accounts for the spike of the tag ‘gas_prices’ observed on June 4. The price of oil prices hit an all-time high of $147 per barrel on July 11 which triggered a debate evidenced by spike (i.e. 7.59) in the tag ‘oil’ on July 16.

Tags related to the Russia-Georgia war include ‘russia’ and ‘georgia’, both of which saw spikes on August 17. This was some ten days after the conflict was first reported on mainstream media.

Even though tags related to Lehman's Brothers' bankruptcy could not be detected, those related to the economic recession such as 'cars' and 'economy' could be found. The tag 'cars' saw a spike of magnitude 377 on July 16, arising from postings on the declining car sales amidst the financial crisis that plagued the U.S. economy. The tag 'economy' was observed to spike on September 29 with a magnitude of 8.8. A deeper analysis revealed that the postings were made in response to the government's announcement on September 16 to bail out American International Group (AIG) after the insurance company suffered $62 billion quarterly losses.

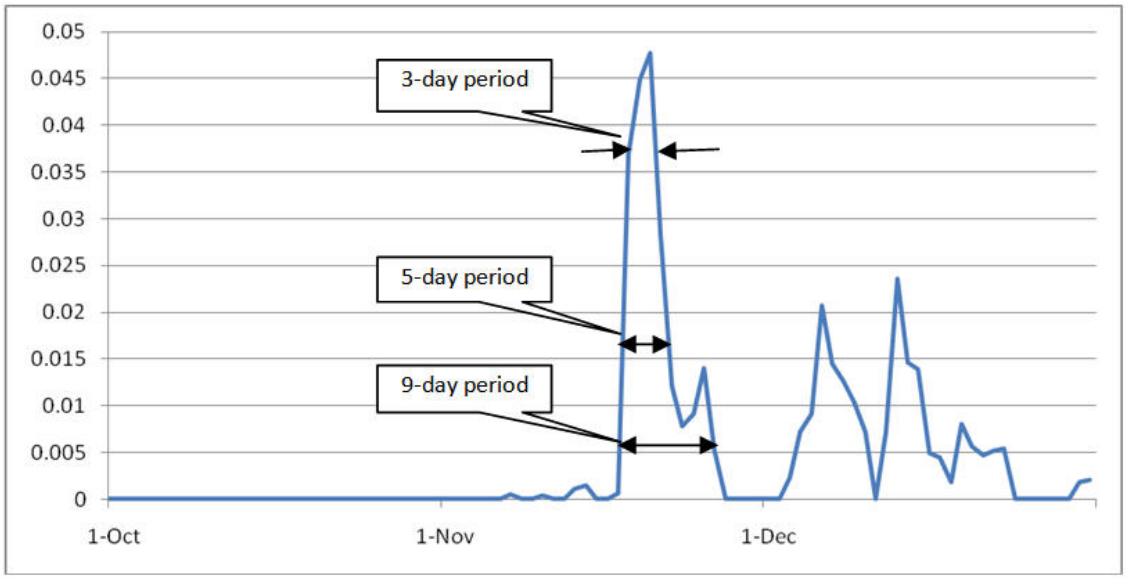
Tags related to Obama's presidential victory include those surrounding the U.S. election such as 'vp' and 'obama_biden' which were observed to spike on August 23, immediately after Obama's announcement of Biden as his running mate. The magnitude of the spikes (i.e. 644.9 and 1614.3) suggests that the topic attracted highly vigorous attention. Likewise, tags 'mccain' and '2008_election' saw spikes on August 29, the same day McCain announced Palin as his running mate. While 'election08' saw a spike on November 4, the day Obama was elected president, this tag per se is not indicative of the election result. Tags 'obama_administration' and 'message_for_obama' spiked on November 6 and November 7 respectively, in the wake of Obama's victory. The news drew intense interests, as seen from the high magnitude of the spikes (i.e. 24420 and 2121.6 respectively).

Related to the terrorist attack in Mumbai is the tag 'india' which was observed to spike on December 3 with a magnitude of 114.7. The attack made headline news on mainstream media eight days earlier on November 26. The tag 'israel' saw a surge in postings on December 31 with a spike magnitude of 88.9. This is attributable to Israeli's air strike on Gaza Strip reported on December 27.

4.4 Tags with bursts

When the minimum relative frequency of a tag over a three-, five- or nine-day window is at least three times the average relative frequency of that tag on all the previous days, a burst is said to have been observed. A burst thus indicates a period of at least three, five or nine consecutive days in which an event was actively discussed. For example in Figure 3, the relative frequency of tag 'big_three' increases steadily after November 17. A burst in a three-, five- and nine-day window can be observed from November 18 and 20; November 18 and 22; and November 18 and November 26 respectively, in the wake of a congressional hearing on November 17 where senior executives of Detroit's big three automakers, namely, GM, Ford and Chrysler presented their $25 billion bailout plans in Washington.

Figure 3. Burst for tag 'big_three'



Of the 200 tags, the number of tags which saw at least a burst during a three-, five- and nine-day window was 173, 99 and 54 respectively. While all tags with at least a burst were included in the analysis, only the top 50 tags with the highest minimum spike magnitude over a three-, five- and nine-

day window throughout the study period, arranged chronologically, are shown in Table 4, 5 and 6 respectively due to space constraints.

Table 4. Top 50 tags which saw a burst in a three-day window

Date	Tag	Min r(d,3)	Ave r(d)	Min r(d,3)/ Ave r(d, 3)
8-Apr-2008 to 10-Apr-2008	china	0.053	0.0125	4.2
8-Apr-2008 to 10-Apr-2008	torch	0.032	0.0078	4.1
19-May-2008 to 21-May-2008	earthquake	0.043	0.0055	7.8
9-Jun-2008 to 11-Jun-2008	storm	0.033	0.0079	4.2
18-Jun-2008 to 20-Jun-2008	flooding	0.041	0.0053	7.7
18-Jul-2008 to 20-Jul-2008	entertainment	0.015	0.0018	8.3
24-Jul-2008 to 26-Jul-2008	black_in_america	0.145	0.0068	21.3
15-Aug-2008 to 17-Aug-2008	georgia	0.078	0.0038	20.5
15-Aug-2008 to 17-Aug-2008	russia	0.075	0.0037	20.3
15-Aug-2008 to 17-Aug-2008	ossetia	0.016	0.0006	26.7
19-Aug-2008 to 21-Aug-2008	hurricane	0.028	0.0014	20.0
19-Aug-2008 to 21-Aug-2008	severe_weather	0.028	0.0012	23.3
23-Aug-2008 to 25-Aug-2008	obama_biden	0.021	0.0017	12.4
26-Aug-2008 to 28-Aug-2008	ireport_kit	0.016	0.0005	32.0
29-Aug-2008 to 31-Aug-2008	Vp	0.041	0.0038	10.8
31-Aug-2008 to 2-Sep-2008	gustav	0.015	0.001	15.0
2-Sep-2008 to 4-Sep-2008	mccain	0.102	0.01	10.2
2-Sep-2008 to 4-Sep-2008	2008_election	0.092	0.0085	10.8
3-Sep-2008 to 5-Sep-2008	politics	0.06	0.0123	4.9
15-Sep-2008 to 17-Sep-2008	Ike	0.019	0.0003	63.3
25-Sep-2008 to 27-Sep-2008	debate	0.081	0.0012	67.5
28-Sep-2008 to 30-Sep-2008	economy	0.078	0.0145	5.4
2-Oct-2008 to 4-Oct-2008	palin	0.031	0.0035	8.9
2-Oct-2008 to 4-Oct-2008	biden	0.014	0.0014	10.0
5-Oct-2008 to 7-Oct-2008	election_express_yourself	0.017	0.0006	28.3
5-Oct-2008 to 7-Oct-2008	missouri	0.015	0.0004	37.5
5-Oct-2008 to 7-Oct-2008	Eey	0.014	0.0006	23.3
5-Oct-2008 to 7-Oct-2008	express	0.014	0.0005	28.0
8-Oct-2008 to 10-Oct-2008	cutting_costs	0.015	0.0006	25.0
22-Oct-2008 to 24-Oct-2008	mccain_palin	0.023	0.0029	7.9
27-Oct-2008 to 29-Oct-2008	obama	0.094	0.0176	5.3
27-Oct-2008 to 29-Oct-2008	situation_room	0.055	0.0042	13.1
3-Nov-2008 to 5-Nov-2008	election08	0.158	0.0343	4.6
4-Nov-2008 to 6-Nov-2008	Prep	0.039	0.0014	27.9
6-Nov-2008 to 8-Nov-2008	Gay	0.036	0.0012	30.0
6-Nov-2008 to 8-Nov-2008	marriage	0.036	0.0011	32.7
6-Nov-2008 to 8-Nov-2008	lesbian	0.029	0.0009	32.2
6-Nov-2008 to 8-Nov-2008	Pets	0.016	0.001	16.0
8-Nov-2008 to 10-Nov-2008	View	0.019	0.0009	21.1
10-Nov-2008 to 12-Nov-2008	sarah_palin	0.019	0.002	9.5
12-Nov-2008 to 14-Nov-2008	obama_administration	0.029	0.0014	20.7
15-Nov-2008 to 17-Nov-2008	wildfires	0.018	0.0015	12.0
18-Nov-2008 to 20-Nov-2008	bailout	0.042	0.0024	17.5
18-Nov-2008 to 20-Nov-2008	big_three	0.037	0.0004	92.5
14-Dec-2008 to 16-Dec-2008	holidays	0.015	0.0009	16.7
17-Dec-2008 to 19-Dec-2008	opinion	0.036	0.0086	4.2
17-Dec-2008 to 19-Dec-2008	sound_off	0.03	0.0066	4.5
17-Dec-2008 to 19-Dec-2008	comment	0.029	0.0066	4.4

18-Dec-2008 to 20-Dec-2008	wintry	0.04	0.0006	66.7
21-Dec-2008 to 23-Dec-2008	christmas	0.02	0.0008	25.0

Table 5. Top 50 tags which saw a burst in a five-day window

Date	Tag	Min r(d,5)	Ave r(d)	Min r(d, 5) /Ave r(d)
18-Jun-2008 to 22-Jun-2008	flooding	0.027	0.006	4.5
17-Jul-2008 to 21-Jul-2008	movies	0.006	0.0013	4.6
24-Jul-2008 to 28-Jul-2008	black_in_america	0.102	0.0092	11.1
02-Aug-2008 to 06-Aug-2008	fan_zone	0.011	0.002	5.5
09-Aug-2008 to 13-Aug-2008	sport	0.012	0.0017	7.1
09-Aug-2008 to 13-Aug-2008	affair	0.007	0.0014	5.0
12-Aug-2008 to 16-Aug-2008	georgia	0.062	0.0032	19.4
12-Aug-2008 to 16-Aug-2008	russia	0.059	0.0031	19.0
14-Aug-2008 to 18-Aug-2008	ossetia	0.015	0.0008	18.8
18-Aug-2008 to 22-Aug-2008	hurricane	0.011	0.0017	6.5
18-Aug-2008 to 22-Aug-2008	severe_weather	0.011	0.0014	7.9
28-Aug-2008 to 01-Sep-2008	gustav	0.006	0.0006	10.0
29-Aug-2008 to 02-Sep-2008	Vp	0.041	0.0044	9.3
30-Aug-2008 to 03-Sep-2008	sarah	0.008	0.0004	20.0
01-Sep-2008 to 05-Sep-2008	mccain	0.07	0.0109	6.4
01-Sep-2008 to 05-Sep-2008	2008_election	0.063	0.0093	6.8
02-Sep-2008 to 06-Sep-2008	Rnc	0.007	0.0006	11.7
15-Sep-2008 to 19-Sep-2008	Ike	0.008	0.0005	16.0
17-Sep-2008 to 21-Sep-2008	republican	0.008	0.0018	4.4
20-Sep-2008 to 24-Sep-2008	biden	0.008	0.0009	8.9
24-Sep-2008 to 28-Sep-2008	debate	0.02	0.0018	11.1
28-Sep-2008 to 02-Oct-2008	economy	0.062	0.0153	4.1
28-Sep-2008 to 02-Oct-2008	bailout	0.026	0.0012	21.7
02-Oct-2008 to 06-Oct-2008	palin	0.029	0.0039	7.4
08-Oct-2008 to 12-Oct-2008	cutting_costs	0.009	0.0007	12.9
15-Oct-2008 to 19-Oct-2008	election_express_yourself	0.007	0.0013	5.4
16-Oct-2008 to 20-Oct-2008	Joe	0.007	0.0005	14.0
21-Oct-2008 to 25-Oct-2008	mccain_palin	0.021	0.003	7.0
22-Oct-2008 to 26-Oct-2008	obama_biden	0.018	0.0044	4.1
23-Oct-2008 to 27-Oct-2008	election_2008	0.005	0.0007	7.1
27-Oct-2008 to 31-Oct-2008	obama	0.094	0.0183	5.1
27-Oct-2008 to 31-Oct-2008	situation_room	0.055	0.0047	11.7
27-Oct-2008 to 31-Oct-2008	battleground	0.005	0.0005	10.0
29-Oct-2008 to 02-Nov-2008	undecided	0.009	0.0021	4.3
29-Oct-2008 to 02-Nov-2008	halloween	0.005	0.0006	8.3
31-Oct-2008 to 04-Nov-2008	early_voting	0.011	0.0008	13.8
03-Nov-2008 to 07-Nov-2008	Prep	0.008	0.0015	5.3
06-Nov-2008 to 10-Nov-2008	Gay	0.022	0.0016	13.8
06-Nov-2008 to 10-Nov-2008	lesbian	0.019	0.0012	15.8
06-Nov-2008 to 10-Nov-2008	marriage	0.019	0.0015	12.7
10-Nov-2008 to 14-Nov-2008	obama_administration	0.012	0.0014	8.6
10-Nov-2008 to 14-Nov-2008	message_for_obama	0.007	0.0006	11.7
11-Nov-2008 to 15-Nov-2008	Pets	0.006	0.0015	4.0
14-Nov-2008 to 18-Nov-2008	wildfires	0.013	0.0016	8.1
18-Nov-2008 to 22-Nov-2008	big_three	0.012	0.0007	17.1
14-Dec-2008 to 18-Dec-2008	holidays	0.015	0.001	15.0
18-Dec-2008 to 22-Dec-2008	weather	0.046	0.0095	4.8
18-Dec-2008 to 22-Dec-2008	wintry	0.04	0.0009	44.4

18-Dec-2008 to 22-Dec-2008	snow	0.012	0.0009	13.3
18-Dec-2008 to 22-Dec-2008	christmas	0.008	0.0007	11.4

Table 6. Top 50 tags which saw a burst in a nine-day window

Date	Tag	Min r(d,9)	Ave r(d)	Min r(d, 9) /Ave r(d)
24-May-2008 to 01-Jun-2008	ireport_for_cnn	0.02	0.0039	5.1
31-Jul-2008 to 08-Aug-2008	energy_fix	0.001	0.0003	3.3
02-Aug-2008 to 10-Aug-2008	fan_zone	0.009	0.0024	3.8
07-Aug-2008 to 15-Aug-2008	sport	0.009	0.0018	5.0
10-Aug-2008 to 18-Aug-2008	georgia	0.035	0.0045	7.8
10-Aug-2008 to 18-Aug-2008	russia	0.035	0.0043	8.1
12-Aug-2008 to 20-Aug-2008	ossetia	0.004	0.001	4.0
12-Aug-2008 to 20-Aug-2008	putin	0.002	0.0002	10.0
26-Aug-2008 to 03-Sep-2008	2008_election	0.044	0.008	5.5
28-Aug-2008 to 05-Sep-2008	Usa	0.003	0.0009	3.3
29-Aug-2008 to 06-Sep-2008	mccain	0.048	0.0113	4.2
29-Aug-2008 to 06-Sep-2008	sarah	0.005	0.0005	10.0
15-Sep-2008 to 23-Sep-2008	biden	0.006	0.0009	6.7
15-Sep-2008 to 23-Sep-2008	republican	0.006	0.0019	3.2
15-Sep-2008 to 23-Sep-2008	Ike	0.005	0.0007	7.1
15-Sep-2008 to 23-Sep-2008	sarah_palin	0.005	0.0011	4.5
15-Sep-2008 to 23-Sep-2008	alaska	0.002	0.0004	5.0
15-Sep-2008 to 23-Sep-2008	independent_convention	0.002	0.0003	6.7
15-Sep-2008 to 23-Sep-2008	lou_dobbs	0.002	0.0003	6.7
15-Sep-2008 to 23-Sep-2008	middle_class	0.002	0.0003	6.7
15-Sep-2008 to 23-Sep-2008	united	0.002	0.0006	3.3
15-Sep-2008 to 23-Sep-2008	Vice	0.002	0.0004	5.0
15-Sep-2008 to 23-Sep-2008	Eu	0.001	0.0002	5.0
15-Sep-2008 to 23-Sep-2008	Nato	0.001	0.0002	5.0
16-Sep-2008 to 24-Sep-2008	corruption	0.002	0.0004	5.0
19-Sep-2008 to 27-Sep-2008	cnn_money	0.003	0.0009	3.3
24-Sep-2008 to 02-Oct-2008	economy	0.057	0.0153	3.7
25-Sep-2008 to 03-Oct-2008	bailout	0.015	0.0013	11.5
15-Oct-2008 to 23-Oct-2008	election_2008	0.003	0.0006	5.0
16-Oct-2008 to 24-Oct-2008	Joe	0.003	0.0006	5.0
17-Oct-2008 to 25-Oct-2008	mccain_palin	0.018	0.003	6.0
19-Oct-2008 to 27-Oct-2008	worldview	0.003	0.0005	6.0
21-Oct-2008 to 29-Oct-2008	Palin	0.023	0.0063	3.7
21-Oct-2008 to 29-Oct-2008	obama_biden	0.016	0.0046	3.5
21-Oct-2008 to 29-Oct-2008	early_voting	0.006	0.0004	15.0
23-Oct-2008 to 31-Oct-2008	halloween	0.003	0.0004	7.5
27-Oct-2008 to 04-Nov-2008	battleground	0.005	0.0007	7.1
27-Oct-2008 to 04-Nov-2008	Vote	0.004	0.0011	3.6
03-Nov-2008 to 11-Nov-2008	Prep	0.006	0.0017	3.5
06-Nov-2008 to 14-Nov-2008	Gay	0.011	0.0018	6.1
06-Nov-2008 to 14-Nov-2008	marriage	0.009	0.0017	5.3
06-Nov-2008 to 14-Nov-2008	Lesbian	0.008	0.0014	5.7
06-Nov-2008 to 14-Nov-2008	message_for_obama	0.004	0.0006	6.7
10-Nov-2008 to 18-Nov-2008	obama_administration	0.009	0.0017	5.3
10-Nov-2008 to 18-Nov-2008	issue_1	0.003	0.0006	5.0
16-Nov-2008 to 24-Nov-2008	cnn_heroes	0.002	0.0006	3.3
18-Nov-2008 to 26-Nov-2008	big_three	0.005	0.0009	5.6
13-Dec-2008 to 21-Dec-2008	holidays	0.008	0.0012	6.7

15-Dec-2008 to 23-Dec-2008	Snow	0.009	0.001	9.0
15-Dec-2008 to 23-Dec-2008	christmas	0.007	0.0008	8.8

Similar to tags with spikes, tags with bursts also appear to dovetail with the occurrence of significant news events identified in Table 1. For example, tags such as 'china' and 'torch' saw bursts from April 8 to 10 which coincided with the Olympic torch relay in the U.S. leg on April 9. However, given that they did not see a burst in a five- nor nine-day window, the interest of iReport contributors on this event did not appear prolonged.

The tag 'earthquake', which saw a burst in a three-day window from May 19 to 21, can be attributed to the Sichuan disaster reported by the mainstream media on May 12. The tag 'gas_prices' saw a burst in a three-day window from May 27 to 29 as well as in a five-day window from May 27 to 31 in response to the increasing oil prices. Related to the Russia-Georgia war are tags including 'georgia', 'russia' and 'ossetia' which saw bursts consistently over a three-, five- and nine-day window. These tags unanimously saws bursts from August 15 to 17; the tags 'georgia' and 'russia' saw bursts from Aug 12 to 16, as well as from Aug 10 to 18 while the tag 'ossentia' saw a burst from Aug 14 to 18 and August 12 to 20. This suggests that the conflict, which occurred on August 7, had captured the attention of iReport contributors over a nine-day period.

Tags related to Obama's victory include those surrounding the U.S. presidential election such as 'obama_biden', 'mccain', 'sarah', 'palin', 'biden', 'debate', 'election08' and 'obama_administration'. The tag 'obama_biden' saw a burst over a three-day window from August 23 to 25 following the mainstream news on August 23 that Obama had announced Biden as his running mate, while the tags 'mccain' and 'sarah' saw bursts over a nine-day window from August 29 to September 6, in response to McCain's introduction of Palin as his running mate on August 29. The tags 'palin' and 'biden' unanimously saw bursts from October 2 to 4. The tag 'palin' persisted in a burst further over a five-day window from October 2 to 6 while tag 'debate' saw a burst over a nine-day window from October 1 to 9. These dates were in conjunction with a televised debate held on October 2 in which both vice presidential candidates pitted against each other on a range of foreign and domestic policies. The tag 'election08' saw a burst over a three-day window from November 3 to 5, as well as a five-day window from November 1 to 5, reflecting heightened interest among iReport contributors on the 2008 U.S. presidential election held on November 4. The tag 'obama_administration' saw a burst over a three-, five and nine-day window from November 12 to 14, November 10 to 14 and November 10 to 18, following Obama's euphoric victory.

4.5. Comparing spikes and bursts with ground truth events

Table 7 summarizes the analyses of the tags' spikes and bursts in relation to the ground truth events. Of the 10 events, seven could be detected from spikes while six could be detected from bursts. Events undetected from either spikes or bursts were Cyclone Nargis, Betancourt's rescue and the Lehman Brothers' bankruptcy. Further inspection of the iReport postings revealed that tags associated with Cyclone Nargis and Betancourt's rescue were too cryptic (e.g. 'news_to_me', 'situation_room', 'ireport_for_cnn') to offer any discernible cues to the events. Tags related to the Lehman's Brothers' bankruptcy were scant vis-à-vis those related to the government's decisions to bailout AIG. Further inspection of the iReport postings showed that the latter news story, which was not one of the ground truth events but had been reported in mainstream media on September 16, could have eclipsed the news on Lehman's Brothers.

Table 7. Summary of tags' spikes and bursts in relation to the 10 ground truth events

Event	Date Reported	Spike			Burst		
		Sample tags	Date	Delay (days)	Sample tags	Date*	Delay (days)
Chaos on Olympic torch's global route	Apr 6	tibet protest	Apr 9	3	china torch	Apr 8 – 10	2
Cyclone Nargis in Myanmar	May 2	-	-	-	-	-	-
Earthquake in China's Sichuan Province	May 12	earthquake	May 19	7	earthquake	May 19 – 21	7
Betancourt's rescue from Colombian guerrilla fighters	Jul 2	-	-	-	-	-	-
Oil price boost	Jul 11	oil	July 16	5	oil	Jul 15-17	4
Russia-Georgia war	Aug 7	russia georgia	Aug 17	10	russia georgia	Aug 10 – 18	3
Lehman Brothers' bankruptcy	Sep 15	-	-	-	-	-	-
Obama's presidential victory	Nov 4	obama_ administration	Nov 6	2	obama_ administration	Nov 10 – 14	6
Terrorist attacks in Mumbai, India	Nov 26	india	Dec 3	7	india	Dec 3 - 7	7
Israel air strike on Gaza Strip	Dec 27	israel	Dec 31	4	-	-	-

5. Discussion

Based on the results, four main findings emerge. First, the performance of using spikes and bursts as defined in this study to detect news events were found to be comparable. This was consistent with an earlier work [23]. Spikes are indicative of the intensity of attention trained on a news event while bursts represent the duration within which interest on the news is sustained. In terms of coverage, almost all news events detected via spikes were similarly detected via bursts. The exception was the Israel air strike on Gaza whose occurrence towards the end of the study period could have obscured the development of a possible burst. In terms of immediacy, bursts fared marginally better than spikes. This could be attributable to the different thresholds used to define spikes and bursts.

Second, compared to the dates reported by mainstream media on ground truth events, news events detected via spikes and bursts were found to lag ranging from a few days to more than a week. Even as the Internet has often been acknowledged for its efficacy in spreading news [28], this study uncovers a lag time before a breaking news event triggers a perceptible level of interest from citizen journalists. It seemed that iReport contributors took the cues from professional journalists insofar as the ground truth events were concerned. For instance, the Sichuan earthquake which made headlines in mainstream media internationally saw a delay of seven days before it could be detected in iReport. In the case of the Russia-Georgia war, a burst over a nine-day window was detected three days after the news was carried in mainstream media. Spikes in the associated tags, namely, 'russia' and 'georgia', were detected only after a 10-day delay, suggesting a gradual increase in interest in the news event.

Third, news events deemed to be significant by professional journalists did not always attract a high level of interest from iReport contributors. As gatekeepers of information, professional journalists determine the worthiness of a news event, establish the boundary of what to be made known, and deliver the news in a professional and impartial manner [29]. In contrast, citizen journalists contribute news which is most aligned to their interests and concerns. At times, being "in the wrong place at the right time" [13], citizen journalists frame a news event by relying on their personal observations and experiences. Thus, for example, even though the Lehman's Brothers' bankruptcy was regarded highly newsworthy, it failed to generate any vigorous reactions as the government's bailout for AIG did.

Fourth, even though citizen journalism transcends national boundaries via the Internet, news posted to iReport seemed to show proclivity towards local context. The iReport site, which is part of the U.S-based CNN, attracted contributors who appeared also to be primarily from the U.S. Being hyperlocal in focus [30], iReport contributors enjoyed geographical and emotional propinquity among themselves. This sense of connectedness in turn fuelled participation. For instance, while Cyclone Nargis which devastated Myanmar on an unprecedented scale could not be detected, declining car sales amid the financial crisis in the U.S. aroused intense discussions from iReport contributors. Furthermore, local news tended to attract more diverse and descriptive tags vis-à-vis international news. For example, local news events such as the U.S. presidential election that took place during the study period elicited a wide variety of tags including 'vote', '2008_election', 'republican', 'democrats', 'obama' and 'mccain'. In contrast, Betancourt's rescue in Colombia could not be detected; the terrorist attack in Mumbai and Israel air strike on Gaza drew only a single tag each.

6. Conclusion

When a major news event occurs, citizen journalists tend to contribute postings on the Internet with tags which are identical or semantically similar to those already used by others. This causes a sudden surge in the relative frequencies of tags which opens the possibility for news events to be detected. Riding on the increasing popularity of citizen journalism, this study seeks to use spikes and bursts seen in the relative frequencies of tags in iReport to detect major news events. In terms of coverage, seven out of 10 ground truth events could be detected from spikes while six could be detected from bursts. In terms of immediacy, news events detected via spikes and bursts were delayed for between two and 10 days.

Two limitations inherent in this paper must be acknowledged. One, the computation of the relative frequencies did not take into account whether a given tag had been reused multiple times by the same contributor on a given day. Should a contributor submit a large number of postings with the same tag repeatedly, a spike or a burst might have been created. Two, the dataset used in this study was drawn from iReport where most of its contributors were from the U.S. Expanding the dataset to include citizen journalism sites in other countries is likely to offer a more international and generalizable perspective.

With massive amounts of new content emerging from the Internet daily, news consumers are invariably inundated with information. It would be difficult, for example, to find and track news events which are of interest to them. News event detection through tags certainly offers sufficient depth and breath for further investigation. Thus, one direction for future research would be to replicate this study and compare results obtained among other citizen journalism Websites such as www.citizenside.com, (affiliated to The France-Presse Agency) www.jasminenews.com (from Sri Lanka) and www.merinews.com (from India). A second area of investigation could center around the types of topics that are contributed in citizen journalism sites since we found some differences between iReport contributors and mainstream news. Doing so would better illuminate the interests and concerns of citizen journalists, which could lead to more effective discovery and delivery of personalized news stories. A final suggestion for future research is to augment the event detection techniques used in this paper. Possible data points admitted for analysis could include the tagging patterns of individual contributors, the growth of tag vocabularies [31] and the rate of tag reuse [17]. These analyses could shed light in the tagging behaviour of citizen journalists which will facilitate in the automatic detection of events. As events are often evolving to other related events, tags could be harnessed in conjunction with techniques which examine the development of the topics within a given news event [32].

Acknowledgement

This work is partly funded by A*STAR grant 062 130 0057.

Chua, A., Razikin, K.B. & Goh, D. H. (2011). Social tags as news event detectors. Journal of Information Science, 37(1): 3-18

References

[1] D. Gillmor, We the media: The rise of citizen journalists, *National Civic Review* 93(3) (2004) 58-63.

[2] D, Ahlers, News consumption and the new electronic media, *Harvard International Journal of Press-Politics* 11(1) (2006) 29-52.

[3] N. Thurman, Forums for citizen journalists? Adoption of user generated content initiatives by online news media, *New Media & Society* 10(1) (2008) 139 – 157.

[4] D.H. Goh, R.P. Ang, A.Y.K. Chua, and C.S. Lee, Why we share: A study of motivations for mobile media sharing. *In Proceedings of the International Conference on Active Media Technology AMT 2009*, Lecture Notes in Computer Science 5820, 195-206 (2009).

[5] K. Good, The Rise of the Citizen Journalist, *Feliciter* 52(2) (2006) 69-71.

[6] D.H. Goh, A.Y.K. Chua, C.S. Lee, and K. Razikin, Resource discovery through social tagging: A classification and content analytic approach, *Online Information Review* 33(3) (2009) 568-583.

[7] H. Wu, M. Zubair. And K. Maly, Harvesting social knowledge from folksonomies. *In Proceedings of the 17th conference on Hypertext and hypermedia*, 111 – 114 (2006).

[8] H. Becker, M. Naaman, and L. Gravano, Learning similarity metrics for event identification in social media. *In Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*, 291-300 (2010).

[9] J. Curran, and J. Seaton, Power without Responsibility: The Press, Broadcasting and New Media in Britain. (Routledge, London, 2003).

[10] J. Y. M. Nip, Exploring the second phase of public journalism, *Journalism Studies* 7(2) (2006) 212-236.

[11] E. Tilley, J. Cokley, Deconstructing the discourse of citizen journalism: Who says what and why it matters, *Preview Pacific Journalism Review* 14(1) (2008) 94-114.

[12] C.P. Wei and Y.H. Lee, Event detection from online news documents for supporting environmental scanning, *Decision Support Systems*, 36(4) (2004) 385-401.

[13] J. Allan, R. Papka, and V. Lavrenko, On-line new event detection and tracking. *In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 37–45 (1998).

[14] C.C. Chen, M.C. Chen and M.S. Chen, An adaptive threshold framework for event detection using HMM-based life profiles. *ACM Transactions on Information Systems*, 27(2) (2009) Article 9.

[15] F. Can, S. Kocberber, O. Baglioglu, S. Kardas, H.C. Ocalan and E. Uyar, New event detection and topic tracking in Turkish, *Journal of the American Society for Information Science and Technology* 61(4) (2010) 802 – 819.

[16] Z. Li, B. Wang, M. Li and W.Y. Ma, A probabilistic model for retrospective news event detection. *In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 106 – 113 (2005).

[17] S. Sen, S.K. Lam, A.M. Rashid, D. Cosley, D. Frankowski, J. Osterhouse, F.M. Harper and J. Riedl, Tagging communities, vocabulary, evolution. *In Proceedings of the Conference on Computer Supported Cooperative Work*, 181-190 (2006).

[18] S.Y. Neo, Y. Ran, H.K. Goh, Y. Zheng, T.S. Chua and J. Li, The use of topic evolution to help users browse and find answers in news video corpus. *In Proceedings of the 15th International Conference on Multimedia*. 198 – 207 (2007).

[19] S.A. Golder and B.A. Huberman, Usage patterns of collaborative tagging systems, *Journal of Information Science* 32(2) (2006) 198-208.

[20] G. Geisler and S. Burns, Tagging video: conventions and strategies of the YouTube community. *In Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, 480 – 480 (2007).

[21] G.W. Furnas, C. Fake, L. von Ahn, J. Schachter, S. Golder, K. Fox, M. Davis, C. Marlow, C., and M. Naaman, Why do tagging systems work, In CHI '06 Extended Abstracts on Human Factors in Computing. 36-39 (2006).

[22] T. Sakaki, M. Okazaki, and Y. Matsuo, Earthquake shakes Twitter users: real-time event detection by social sensors, *In Proceedings of the 19th International Conference on World Wide Web*, 851-860 (2010).

[23] M. Thelwall, R. Prabowo and R. Fairclough, Are raw RSS feeds suitable for broad issue scanning? A science concern case study, *Journal of the American Society for Information Science and Technology*, 57(12) (2006) 1644 - 1654.

[24] A. Y. K. Chua, D.H. Goh, and K. Razikin, Detecting news event from a citizen journalism Website using tags. *In the Proceedings of the 5th International Conference on Active Media Technology*, 478-489 (2009).

[25] D. Gruhl, R. Guha, D. Liben-Nowell and A. Tomkins, *Information diffusion through Blogspace* (2004). Paper presented at the WWW2004, New York. Available at http://www.www2004.org/proceedings/docs/1p491.pdf. (accessed 1 August 2010)

[26] R. B. Cialdini, Influence: science and practice (Allyn and Bacon, Boston, 2001).

[27] J. Kleinberg, Temporal dynamics of on-line information streams. In M. Garofalakis, J. Gehrke & R. Rastogi (eds.), *Data stream management: Processing high-speed data streams*. (Berlin: Springer, 2005).

[28] J. Dimmick, Y. Chen and Z. Li, Competition Between the Internet and Traditional News Media: The Gratification-Opportunities Niche Dimension, *Journal of Media Economics* 17(1) (2004) 19 – 33.

[29] D. Domingo, T. Quandt, A. Heinonen, A. Paulussen, J. B. Singer and M. Vujnovic, Participatory journalism practices in the media and beyond, *Journalism Practice* 2(3) (2008) 326-342.

[30] Z. Reich, How citizens create news stories: The "news access" problem reversed, *Journalism Studies* 9(5) (2008) 739-758.

[31] U. Farooq, T.G. Kannampallil, Y. Song, C.H. Ganoe, J.M. Carroll and L. Giles, Evaluating tagging behavior in social bookmarking systems: metrics and design heuristics. *In Proceedings of the 2007 International ACM Conference on Supporting Group Work*, 351 – 360 (2007).

[32] C. C. Yang, X. Shi, and C.P. Wei, Discovering event evolution graphs from news corpora, *IEEE Transactions of Systems, Man, and Cybernetics – Part A: Systems and Humans* 39(4) (2009) 850-863.