Transient News Crowds in Social Media

Janette Lehmann[1], **Carlos Castillo**[2], **Mounia Lalmas**[3], **Ethan Zuckerman**[4]
1. Universitat Pompeu Fabra, Barcelona, Spain; janette.lehmann@gmx.de.
2. Qatar Computing Research Institute, Doha, Qatar; chato@acm.org.
3. Yahoo! Labs, Barcelona, Spain; mounia@acm.org.
4. MIT Center for Civic Media, Cambridge, MA, USA; ethanz@media.mit.edu.

Abstract

Users increasingly inform themselves of the latest news through online news services. This is further accentuated by the increasingly seamless integration of social network platforms such as Twitter and Facebook into news websites, allowing easy content sharing and distribution. This makes online social network platforms of particular interest to news providers. For instance, online news producers use Twitter to disseminate articles published on their website, to assess the popularity of their contents, but also as an information source to be used on itself. In this paper, we focus on Twitter as a medium to help journalists and news editors rapidly detect follow-up stories to the articles they publish. We propose to do so by leveraging "transient news crowds", which are loosely-coupled groups that appear in Twitter around a particular news item, and where transient here reflects the fleeting nature of news. We define transient news crowds on Twitter, study their characteristics, and investigate how their characteristics can be used to discover related news. We validate our approach using Twitter data around news stories published by the BBC and Al Jazeera.

1 Introduction

The Web has totally changed the news landscape, causing a significant drop in newspaper and radio audiences, and becoming the second source of news in the US, after television (Pew Research Center 2012). Users are increasingly informing themselves through online news portals and social media platforms. With respect to news, we see social media as a transformative force that is not a replacement for traditional news sources, but a different media on its own. Indeed, online news providers, their journalists and news editors, use social media platforms such as Twitter and Facebook to spread news recently published on their websites, to assess the popularity of such news in different segments of their audience, but also to enrich the stories they publish on their websites.

Social media can be a powerful tool for journalists at multiple stages of the news production process: detection of newsworthy events, interpretation of them as meaningful developments, and investigation of their factual veracity (Fishman 1980). Although Twitter users tweet mainly about daily activities, they also share URLs related to news (Java et al. 2007). Indeed, 59% of them tweet or retweet news headlines (Pew Research Center 2012), which account for 85% of the trending topics on Twitter (Kwak et al. 2010).

In 2011, the fourth annual Digital Journalism Study[1] polled 478 journalists from 15 countries and found that 47 per cent of them used Twitter as a source. Twitter is used by journalists and news editors of mainstream media sites to enrich their articles (Diakopoulos, Choudhury, and Naaman 2012; Subasic and Berendt 2011). They do so by analyzing responses (post-read actions) to news articles (Stajner et al. 2013; Tsagkias, de Rijke, and Weerkamp 2011) as these can provide additional information about the topic of the news, contained in discussions and opinions (Stajner et al. 2013), but also in URLs of related news published by other news sites (Chen et al. 2010).

Many users in Twitter also devote a substantial amount of time and effort to *news curation*. Digital news curation is an emerging trend where users carefully select and filter news stories that are highly relevant for a specific audience.[2] News curators can reach a high level of engagement and specialization, becoming a sort of *distant witnesses* (Carvin 2013) of news stories of global significance. Twitter users can benefit from news curation by using the "lists" feature, that allows them to organize the people they follow into arbitrary topics; these lists are routinely used by many organizations including media companies to collect content around developing news stories. However, users still have to create such lists by hand and update them when deemed necessary.

In this work, we propose a radically new approach: aggregating all the users who tweet a particular news item: a *transient news crowd*, and then use an automatic method to detect contents posted by them that are *related* to the original news item. The advantages are many-fold: (1) the crowd is created automatically and available immediately, (2) we can account for the fleeting nature of news, (3) there is no need to maintain a list or follow a number of experts or curators on Twitter. In addition, by extracting relevant contents from transient news crowds, journalists can cover news *beats* incorporating the shifts of interest of the audiences that follow

[1] http://www.oriellaprnetwork.com/

[2] http://www.pbs.org/idealab/2010/04/our-friends-become-curators-of-twitter-based-news092.html

those beats. This represents an important step: given that journalists can be seen as members of an interpretive community (Zelizer 1993) who come together to make sense of events and translate their importance, transient news crowds might represent individual news users demanding to be part of that interpretive community.

Even a casual examination of the data can show the potential of news crowds. For instance, on January 6, 2013, an article with title "Delhi police dispute India gang-rape account" was posted in the Al Jazeera English website and attracted 147 users who tweeted its link in the first six hours after its publication. Two days later, 27 of those users (18%) tweeted a link to a Huffington Post article with title "Father of Indian gang-rape victim speaks out". If we were able to detect such articles automatically, we could generate a timely alert for the author of the first article pointing to the related article found by the crowd. Of course, we do not assume that *every* subsequent posted article will be related. Instead, we show that such related articles exist and that it is possible to detect them automatically.

Our goals are therefore three-fold: (1) define the notion of "transient news crowds" on Twitter, (2) study their characteristics, and (3) investigate how these can be exploited to discover related news posted on Twitter.

2 Related Work

Recommender systems. Twitter has been used as a source of news recommendations, typically by exploiting Twitter-specific features extracted from post-read social responses (Agarwal, Chen, and Wang 2012; Morales, Gionis, and Lucchese 2012; Phelan et al. 2011), tweets content (hashtags, topics, entities), users followees and followers, public timelines and retweeting behavior. However these works aim at building personalized recommender systems, suggesting news articles based on the inferred topical interests of a user.

Our objective is entirely different, as we want to follow specific stories over time and offer related news articles to the authors of such stories. We want to provide journalists and editors a tool to discover new content that can complement or extend the one that they have produced.

Community detection. Many studies aiming at detecting Twitter communities around topics exists (Greene, O'Callaghan, and Cunningham 2012; Michelson and Macskassy 2010). The methods used rely on the extraction and incorporation of numerous features, such as user tweets (Zhang, Wu, and Yang 2012; Gupta, Joshi, and Kumaraguru 2012), but also user profiles and link similarity: how often two users retweeted, mention or reply to a common third person tweets (Gupta, Joshi, and Kumaraguru 2012). The similarity of the tweet text, URLs, and hashtags have also been considered in the creation of such communities (Zhang, Wu, and Yang 2012), as well as user mentions (addressing/replying), retweets, follower networks, and user lists (Michelson and Macskassy 2010).

Topic engagement (e.g. whether a user will join a discussion) has also been predicted (Purohit et al. 2011; Welch et al. 2011). The content of tweets has been found to be a significant feature for this task, and retweeting the tweets of a user has been found to be a stronger indicator of topical interest than following a user.

Our approach is a natural complement to these works, which carefully craft a topically-focused community around a topic, and then assume that all the content produced by that community is on-topic. Instead, we put together a set of users that have a priori only one element in common (they tweeted a URL), and carefully filter the tweets they produce in order to find relevant on-topic content. Of course, both approaches can be combined.

User lists on Twitter have been used to detect communities (Greene, O'Callaghan, and Cunningham 2012). Recent studies are concerned with recommending new tweets to a list (Duh et al. 2012), understanding the nature of curators, e.g. member and subscriber (García-Silva et al. 2012), and investigating users interests (Kim et al. 2010).

Our work can be viewed as a means to automatically build such lists, which we recall are built manually, but accounting for the fleeting and volatile nature of news and with the aim to discover and recommend related news around a topic.

Expert finding. One way to remain informed about a particular topic is to follow Twitter users that are expert on that topic (for example to add them in a user list build around the topic). It has been shown that experts are often able to identify interesting news very early (Hsieh et al. 2013). Various means to automatically detect expert users have been proposed. As for community detection and Twitter-based news recommendation, typical approaches are based on features such as tweets content, follower network (Weng et al. 2010), and retweet-networks (Kong and Feng 2011). More sophisticated features experimented with distinguished topic-related tweets by retweets, conversational and normal tweets (Pal and Counts 2011). Overall, expert detection in Twitter is a difficult task. Studies show that tweets content provides less useful information than contextual data (profile, user list, etc.) (Liao et al. 2012; Ghosh et al. 2012). On the other hand, manual expert detection revealed that decisions are influenced by shallow factors such as author names (Pal and Counts 2011).

In this work, we therefore do not attempt to identify the specific users who are experts on some topic; instead, we consider the crowd of all the users who tweeted an article, and extract from the crowd certain characteristics (many of them referred to in this section) that can be carefully combined to discover news related to the original article.

3 Data and processing

We describe the data used in our work, how it was created, and various processing performed.

3.1 Data Extraction

We collected news articles published in early 2013 on two major online news websites, BBC and Al Jazeera English (AJE). The news websites represent large media organizations, seeking adoption of their content in a wide range of international markets. From the BBC, we collected separately articles from World Service (BBC-WORLD) and BBC

UK (BBC-UK), each forming a different dataset. We downloaded periodically the homepage of each website, from which we sampled at random a subset of news articles. We focused on the headline news: opinions, magazine and sport news were not included. The sampled articles cover a variety of stories such as Obama's inauguration, the conflict in Mali, the pollution in Beijing, and road accidents in the UK.

For each of the sampled articles, we started a process that used Twitter's API[3] to periodically find tweets including that article's URL. The earliest tweets followed almost immediately the publication of the article, as each of these news organizations disseminate their content via their own twitter account(s) (e.g. *@BBCWorld*, *@AJEnglish*). We define the *crowd* of a news article as the set of users that tweeted the article within the first 6 hours after the first tweet on that article. We selected this time period because it encompasses about 90% of the tweets an article receives (87% for BBC-WORLD, 91% for BBC-UK, and 91% for AJE). We followed users on each crowd during one week, recording every public tweet they posted during this period.

3.2 Data Filtering

In Twitter there is a substantial amount of spam. Spammers routinely abuse Twitter to promote products and services. Successful spammers attract followers and bypass filtering mechanisms by posting a mixture of reputable tweets and advertising (Benevenuto et al. 2010). Spam can negatively affect our results, and given that the Twitter API has strict rate limitations, it can also reduce the coverage of our dataset by forcing us to waste our quota downloading useless tweets. Hence, it is important to define some criteria to filter out at least the most active spammers.

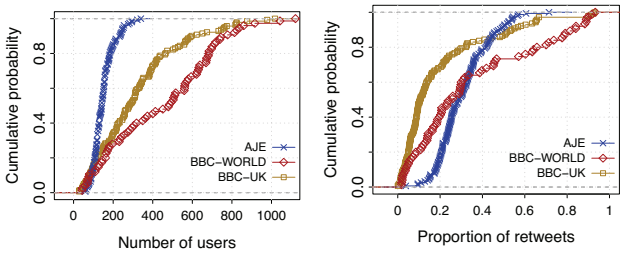
Informed by previous works (Wang 2010; Benevenuto et al. 2010), as an heuristic to remove spammers, we removed users with an abnormally high tweeting activity (98 tweets per day), whereby most of the tweets were retweets (90% retweets) or tweets containing URLs (92% URL-tweets). We also examined manually the most prolific accounts and defined a blacklist of high-throughput automatic accounts that do not focus on any particular region, topic, or news provider. We removed only accounts having clearly anomalous behavior, and tried to keep our manual intervention to a minimum, discarding less than 5% of the users in total.

Finally, news articles with very small crowds (lower 15% of the distribution) or very large ones (upper 5% of the distribution) were excluded. We kept articles with 50–150 users for BBC-WORLD and BBC-UK news articles and 70–360 users for AJE. The resulting number of articles, the sizes of the crowds, and the number of tweets collected for each dataset are summarized in Table 1. As shown in Figure 1(a), these distributions are very skewed and there are crowds that are substantially larger than the average, as well as users that are substantially more prolific than the average. We observe that the crowds around articles in AJE are smaller than the ones of BBC-WORLD and BBC-UK, a reflection of the different sizes of their audiences.

[3]http://dev.twitter.com/

Table 1: General characteristics of our datasets: number of articles, users, and tweets.

		Users		Tweets	
Dataset	Articles	Total	Per crowd	Total	Per crowd
BBC World Service	75	13.3K	177	35.5K	201
BBC News UK	141	13.1K	92	47.8K	339
Al Jazeera English	155	8.3K	53	24.0K	154



(a) Users per crowd (b) Retweets at creation

Figure 1: Distributions of the number of users per crowd. Proportion of retweets during each crowd's creation.

3.3 Shortened URL Handling

The limitation of number of characters in tweets is viewed as one of the key elements of the success of Twitter as a sharing platform. However, it also imposes constraints for people who want to post URLs, which are usually long strings. Hence, a number of *URL shortening* services have appeared in recent years, providing on-demand URL alias such as "http://tinyurl.com/2g774x". URL shortening services typically generate different shortened URLs for different users, given the same input URL. Expanding shortened URLs requires at least one network access, thus creating a bottleneck for many applications that should be avoided when possible.

We therefore expanded only a subset of the URLs appearing in our dataset. To determine this subset we rely on the text of the tweets containing the URL. That text is stemmed, stopwords are removed, and word bigrams are extracted; the latter are used as the tweet representation. Two URLs are considered equal if they appear in tweets having a Jaccard similarity of at least θ under this representation. The threshold is determined by using the outcome of a crowdsourcing task in which 300 random pairs of tweets from our collection were annotated by humans (details of the crowdsourcing service used are given in Section 5.2). We set $\theta = 0.25$, which has a precision and recall of 0.84 on this test set.

A shortened URL, without the need to be expanded, is thus represented as a cluster (computed by a greedy clustering algorithm) of equal URLs as calculated above. Only one of the URLs in each cluster needs to be expanded, and the *frequency* of a URL is the number of tweets in its cluster. This definition of frequency is used in the remainder of this paper, particularly in Section 5.3 in the task of discovering popular related stories.

4 Characterizing Transient News Crowds

To the best of our knowledge this type of transient crowd in the news domain has not been studied in depth. We summarize key observations about the characteristics of these crowds in terms of their creation, members and dynamics.

4.1 Crowd Creation

There are two main mechanisms by which a user can tweet about a news article and hence become a member of a crowd: direct tweets and re-tweets. *Direct tweets* can be done by the user by clicking on a "tweet" button provided by the news website, or by using a bookmarklet, or by copying and pasting the URL in a Twitter client. *Re-tweets* are created by users in a Twitter client or directly on the Twitter website, and correspond to re-posting what other users have shared.

Figure 1(b) depicts the proportion of retweets for our three datasets. This proportion is basically below 0.4. This indicates that a large proportion of the activity around a news URL on Twitter can be traced back directly to the news website, and not to word-of-mouth/propagations effects in Twitter. However, in AJE we observe a stronger word-of-mouth effect than in the other two sites, which is consistent with previous observations (Lotan, Gaffney, and Meyer 2011).

4.2 Crowd Members

We look at the behavior of the users belonging to news crowds during the one-week period following their creation (starting with the first tweet about a news article). In Figure 2(a) we plot the distribution of the average number of tweets per day of crowd members, which peaks at around 40 tweets/day for AJE and 60 tweets/day for BBC-WORLD and BBC-UK. In any case, these are unusually high numbers, given that the overall average is around 2.5 tweets/day.[4]

Indeed, after our spam filtering heuristics (Section 3.2), crowds still include many Twitter accounts that post tweets automatically but are not spammers. These include corporate accounts operated by the news networks themselves, such as *@BBCWorld* and *@AJELive* (from Al Jazeera). They also include services provided by third parties, such as *@bbcnews_ticker* that tweets all the news in the BBC news ticker, *@AmmanHashtags* that automatically re-tweets news mentioning the capital of Jordan, and *@TwittyAlgeria* that tweets news containing the word "Algeria" extracted from a variety of sources.

At the same time, there are several accounts that do not seem to be operated automatically and that are examples of good news curators. For instance, *@thomas_wiegold* has a few thousand followers and manually curates a set of conflict stories around the globe and adds commentary aimed at a German-speaking audience.

Crowds can also be described by the characteristics of the tweets posted by their members. The fraction of tweets that are re-tweets, shown in Figure 2(b), is consistent with Figure 1(b), showing a large proportion of re-tweets in AJE.

[4]By the end of 2012, CNET reported that the number of tweets per day was close to 500 million (http://cnet.co/U3hOUW), while the number of active users, according to The Guardian, was around 200 million (http://gu.com/p/3cjvf/tw).

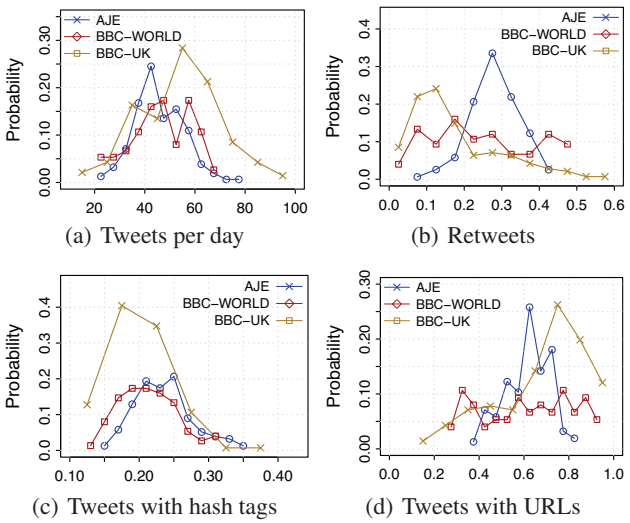


Figure 2: Distributions of number of tweets per day and different type of tweets.

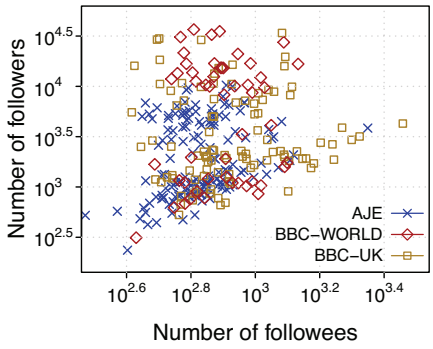


Figure 3: Average number of followers and followees of users per crowd. Each data point corresponds to one crowd.

The fraction of tweets containing hashtags (Figure 2(c)), or URLs (Figure 2(d)) indicates that in comparison with the other datasets, users of BBC-UK use slightly less hashtags (peak at 0.2 vs peak at 0.25) and have a higher proportion of tweets with URLs (peak at 0.8 vs peak at 0.6).

Figure 3 depicts each crowd from the perspective of the average number of followers and followees of its members. We observe that crowds in BBC-WORLD and BBC-UK have on average a larger number of followers than those in AJE. Overall, these values are relatively high considering that a majority of Twitter users have less than 50 followers.[5]

The average is dominated by crowd members having an extremely large number of followers, such as some of the accounts we have mentioned. For instance, *@TwittyAlgeria*, despite being evidently generated automatically, has over 240,000 followers (as of March 2013). Even if some of these followers were in turn automatically-created accounts, these large numbers indicate that people perceive their tweets

[5]http://www.beevolve.com/twitter-statistics/

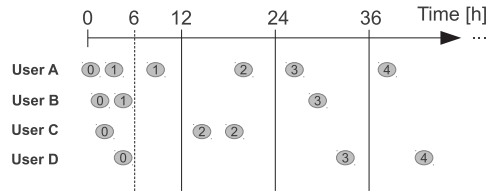


Figure 4: Depiction of our assignment of slices to tweets in the data. Each row corresponds to a user and each oval to a tweet, numbered with the time slice it belongs to. All the tweets containing the original URL are assigned to slice t_0, and must be posted within 6 hours to grant crowd membership to its user. Subsequent tweets are assigned to other slices as per the diagram.

as valuable, as otherwise they would have ceased to follow them ("unfollowing" an account in Twitter is as easy as following it). In other words, automatically generating/aggregating content does not seem to be perceived a priori as negative by Twitter users. Therefore, we do not attempt to remove automatic users from our crowds, but we do weight their influence carefully (as we show in Section 5.3).

Recurring crowd members. Crowd members on a given website often overlap. About 74% (sd=0.13) of people who tweet an article in AJE tweet at least one other article in AJE during our observation period. Something similar happens with BBC-WORLD and BBC-UK, where respectively 61% (sd=0.24) and 75% (sd=0.22) of crowd members tweet more than one article. Again, these recurring crowd members include a mixture of automatic accounts but also manually-operated ones that choose to tweet from the same sources repeatedly. This reinforces the need to weight their influence carefully in the news discovery task.

4.3 Crowd Dynamics

To study the dynamics of crowds over time we discretize time into slices of a fixed size. We illustrate this in Figure 4 for an example set of four users. The tweets that create a crowd are assigned to the slice t_0 and are posted within 6 hours of each other. The remaining tweets from these users are assigned to a time slice according to the time passed since that first tweet. We perform this assignment independently in each of the crowds of our three datasets.

Time granularity. The choice of the appropriate time granularity for time slices depends on the application. In our case, we are interested in the news discovery problem described in Section 5, and hence, this problem informs our choice of a time granularity.

We focus on the phenomenon of *topic drift*, by virtue of which each crowd "disperses" in terms of the content of their tweets. We can quantify this dispersion by first measuring the expected similarity between tweets in a time slice, and then observing if this expected similarity changes over time. The similarity of two tweets is measured using the Jaccard coefficient of their representations as bags of words after stemming and stopword removal (see Section 3.3).

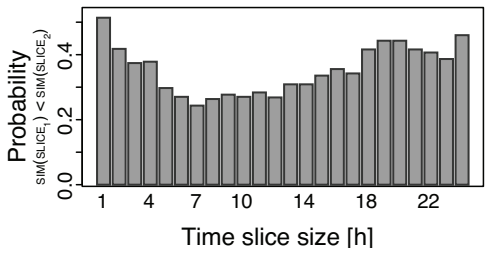


Figure 5: Probability that a crowd's tweets become more similar on the second time slice (compared to the first time slice) for different choices of time granularity.

Over time, we expect that the average similarity becomes smaller. In particular, we expect that given our choice of time granularity, tweets on the first time slice of a crowd are more similar to each other than tweets on the second time slice of the same crowd. With this in mind, we study different time granularities ranging from 1 hour to 24 hours, and measure the probability that in a crowd the average similarity on the second slice is (contrary to our expectations) higher than the average similarity on the first slice.

Figure 5 shows the results for this test. For small granularities (e.g. 1 hour) the probability is close to 0.5, indicating that using that time granularity crowds can get either farther apart or closer together. This can be viewed as random, and any values above or below can be considered as a signal. For granularities between 7 and 12 hours a minimum of less than 0.3 is attained, indicating that crowds have at least a 70% chance of becoming more dispersed in the slice t_2 with respect to slice t_1. We chose a time granularity of 12 hours in the remainder of the paper, as it is easy to interpret. In the Figure we observe that for larger time granularities, we return slowly to random behavior, reaching 0.5 at granularities of 18-24 hours.

Correlation test. Next we must determine if at least part of the activities of a crowd are related to the article that created each crowd. In order to do this, we conduct a randomized test. We consider random pairs of crowds whose first slice overlaps (i.e. the original articles are posted within 12 hours of each other). First, we look at the similarity of the original articles, by measuring the average similarity of tweets containing the original URL (belonging to slice t_0 in both crowds). Second, we perform the same measure in the slice t_3 of both crowds. This test attempt to answer the following question: if two articles posted today are similar to each other, will people who tweeted about those articles tweet about similar things tomorrow?

The correlation obtained in general between both similarities is $r^2 \approx 0.4$. Figure 6 depicts the distribution of similarities in slice t_3 for different buckets of similarity at slice t_0. We can see a clear trend in which articles that are not similar to each other rarely have crowds that tweet about the same topics in the future, while this often happens in crowds originating from articles that are similar to each other. This clearly shows that crowds are not formed randomly. Next, we use them for a news discovery task.

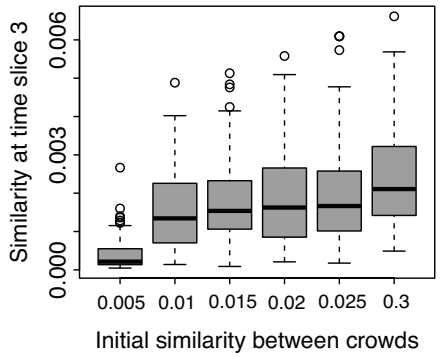


Figure 6: Distribution of similarity at slice t_3 for pairs of crowds at different levels of similarity at slice t_0.

5 Crowd-based News Discovery

This study is motivated by the intention of discovering news items with the help of news crowds. In this section, we describe a method for performing such discovery. We formulate the discovery task as follows: *given a news crowd and a time slice, find URLs in that time slice that are related to the article that created the crowd.*

A number of steps are followed. First, we extract from each slice the most frequently posted URLs, as described in Section 5.1. Next, we use a supervised learning approach, which we explain in Section 5.2. We employ 18 features, inspired from related works (see Section 2), and also from the observations reported in Section 4 on the characteristics of transient news crowds. Three types of features were employed, frequency-based, text-based and user-based, described in Sections 5.3, 5.4 and 5.5, respectively. Our results are discussed in Section 5.6, and we also suggest an application to crowd visualization over time task in Section 5.7.

5.1 Candidate Generation

We recall that given a URL (article) around which a crowd has been formed, the aim is to discover articles (their URLs) related to the original article. The first step is to extract a pool of candidate URLs from all the URLs tweeted by the crowd. In each time slice, we generate the top URLs having the largest frequencies, where the URL frequencies are computed using the method described in Section 3.3. We remove all URLs having frequency less than 3. This still yields a substantial number of candidates, 41.2 (sd=23.8) per time slice on average for BBC-WORLD, 54.8 (sd=23.8) for BBC-UK, and 15.7 (sd=4.7) for AJE.

Many of these candidate URLs are not related to the original article. We illustrate this with an example using two articles published on AJE on January 13th, 2013. Both articles correspond to ongoing armed conflicts in the Middle East ("Syria allows UN to step up food aid") and Africa ("French troops launch ground combat in Mali"). We identify the crowds of each story and follow them during 14 time slices of 12 hours each, i.e. one week. Next, we manually assess whether each candidate is related to the original story or not. The result of this manual process is shown in Table 2.

Table 2: Example of candidates found for two stories published on January 13, 2013. A candidate is a URL posted by 3 users or more during each of the time slices ($t_1 \ldots t_{14}$). We include the number of candidates related to the original story, and the number and topics of those that are not related.

	Syria allows UN to step up food aid			French troops launch ground combat in Mali		
	Rel.	Not related		Rel.	Not related	
t_1	7	0		1	3	Zero Dark Thirty (film), Algeria×2
t_2	7	0		1	1	Spain
t_3	9	0		0	0	
t_4	5	1	Algeria	3	1	Algeria
t_5	5	1	Algeria	1	0	
t_6	5	2	Crabtree (football), Iran	0	2	Manti Te'o (football), Algeria
t_7	8	1	Algeria	1	1	Chardy (tennis)
t_8	9	4	Mali, Obama, Davos, Batman (film)	1	4	Algeria×2, Soccer, Israel
t_9	8	0		1	1	Algeria
t_{10}	13	2	Iraq, Federer (tennis)	0	1	Flanders
t_{11}	10	1	Obama	1	3	Algeria×2, MLK
t_{12}	10	0		0	1	Algeria
t_{13}	5	2	Lady Gaga (artist), Algeria	1	2	Djokovic (tennis), Jordan
t_{14}	13	2	Beyonce (artist), Palestine	1	0	
	114	16	Total	12	20	Total

For the crowd of the story on Syria, we can see that the number of candidates that are related to the original story consistently exceeds the number of candidates that are not related. For instance, in time slice t_5 we have five related candidates (including stories about the Taftanaz Military Airport, the Kilis refugee camp, etc.) and one unrelated candidate about a hostage crisis in Algeria. For the crowd of the story on Mali, there are actually more unrelated candidates than related ones. Still, some time slices such as t_4 have three related candidates (two about the movements of troops in Mali and one about a related statement by French President Hollande) and one unrelated candidate, again about the hostage crisis in Algeria.[6]

There can be many reasons for the differences, one being that the conflict in Mali is more recent than the one in Syria, hence the latter has many more stories, and a more cohesive crowd of people following the news related to it. It is however clear that relying solely on frequency information (URLs in our case) will often not be sufficient to identify related stories. Other features are important and need to be incorporated, as described next, using a learning approach.

[6]How we define "relatedness" may have an effect on the results. Indeed, with a less restrictive definition than adopted here, the news on the Algerian hostage crisis could be considered as related to the news on the French troops in Mali, because the main demand of the kidnappers was the end of the French military operations in Mali.

5.2 Learning Framework

Learning scheme. We experimented with several learning schemes on our data and obtained the best results using a random forest classifier as implemented in Weka.[7] Given the large class imbalance, we applied asymmetric misclassification costs. Specifically, false negatives (classifying a relevant article as non relevant) were considered five times more costly than false positives; values close to this number did not change substantially the obtained results. For consistency and simplicity, we use the same cost across the three datasets, even if their priors are different.

Evaluation metrics. We use standard evaluation metrics including precision, recall, and AUC, all measured after ten-fold cross validation. Given that the targeted users of this system (journalists) do not expect nor need to have perfect results, we decide to aim for a level of precision close to two-thirds, as we considered it would be satisfactory for them to see twice as many related stories than unrelated stories. Hence, a key metric in our evaluation is the *recall at two-thirds precision*, which measures the probability that a related story is found in our system, if we allow it to generate at most one-third of unrelated stories in its output.

Training data. We collected about 22,500 labels for about 7,500 training examples through Crowdflower,[8] a crowdsourcing provider that provides an interface to a variety of crowdsourcing platforms. We sampled uniformly at random 160 crowds: 80 from AJE, 40 from BBC-WORLD, and 40 from BBC-UK. For each crowd, we selected 5 slices at random, and up to 10 random candidates (URLs having a frequency of 3 or more) from each selected slice.

For each candidate, we showed to three different crowdsourcing workers a sample tweet from the original story and a sample tweet from the candidate URL, and asked them to label the pair using the following instructions:

> *You will be presented with two Twitter messages ("tweets") on current news stories. Please indicate how these two stories are related:*
>
> - Strongly related*: Same ongoing news story (e.g. two articles about nuclear inspections in Iran).*
> - Weakly related*: Not same story, but same location, person, or topic (e.g. two articles about nuclear proliferation).*
> - Not related*.*
>
> *Having "Al Jazeera", "BBC", etc. in both tweets does NOT automatically mean they are related.*

We merged the first two classes as simply *related* for the purpose of our experiments. We ignored the pairs for which the confidence (based on the agreement of the workers) was less than 0.8 and the label was *not related*, as these were almost always borderline cases and are not useful for training or evaluation purposes. Overall, the assessors determined that for BBC-WORLD 4.9% of the candidates were related, for BBC-UK 8.2% and for AJE 9.3%.

[7] http://www.cs.waikato.ac.nz/ml/weka/

[8] http://www.crowdflower.com/

The ratio of weakly related candidates per strongly related candidate varies greatly across datasets: 1.6:1 for BBC-WORLD, 8.5:1 for BBC-UK, and 0.9:1 for AJE. In other words, while in AJE the assessors found candidates that were strongly or weakly related in roughly similar proportions, in the case of BBC-UK there are more than eight weakly related candidates for each strongly related one. This in fact has an effect on the performance obtained for BBC-UK, as described in Section 5.6.

In the next three sub-sections, we described the three sets of features employed in our learning algorithm.

5.3 Frequency-Based Features

For each candidate URL we compute its relative frequency, i.e. the frequency of its URL divided by the frequency of the most frequent URL in the slice (we name this feature *CandidateNormalizedFrequency*).

As we described in Section 5.1, even candidates having a high frequency are often not related to the original news item. Often breaking news about events of global significance appear in many crowds at the same time. To alleviate this problem, we incorporate a feature, analogous to the inverse document frequency in Information Retrieval, that measures how specific is a candidate with respect to a given crowd. If there are n crowds that have time slices that overlap with a candidate appearing in n_c of them, then *CandidateInverseDocFrequency* $= \log(n/n_c)$.

We also observe that repeatedly the top URLs on a given slice can be traced back to prolific users (such as those mentioned in Section 4.2) that post hundreds of tweets per day. These observations inform the design of the features described in Section 5.5.

5.4 Text-Based Features

To remove candidates not related to the original story, we employ a text-similarity approach. We use the same representation of a URL that we used to compute its frequency: a cluster of tweets that contain variants of a URL. Given this representation, we compute the similarity between two URLs by concatenating all the tweets in each cluster in a document, and compute the Jaccard similarity between such documents. Since this approach do not require the web page content of the original news article and the candidate URLs, we are able to access non-textual candidates such as videos, pictures or podcasts. Moreover our approach is computational more efficient as we deal with less content.

First, we measure how similar are the tweets containing the candidate URL to the ones containing the article that created each crowd. We compute four features based on different representations of the content: word unigrams (*SimOriginalUnigrams*), word bigrams (*SimOriginalBigrams*), hash tags (*SimOriginalHashtags*) and capitalized terms (*SimOriginalCapitalized*). The latter is equal to word unigrams except that only words starting with a capital letter are considered – an heuristic that is effective in our dataset given the news headlines writing style.

Second, we measure how similar are the tweets containing the candidate URL to other tweets that appear in candidates from other crowds. We consider only the slices of

the other crowds that overlap with the candidate's slice and use text similarity measures to compute how unique is a candidate with respect to a given crowd. Again, we computed four features based on unigrams, bigrams, hashtags and capitalized terms, but determined through experimentation that only one of them was significant: *SimOthersHashtags*. In total, we used 5 text-based features.

5.5 User-Based Features

Based on the analysis of Section 4, in particular the presence of prolific automatic accounts, it was deemed important to consider features related to the users that contributed each candidate. We therefore incorporated weighted frequency features, in which each user that posts the URL of a candidate contributes a "vote" to that candidate that is weighted by a user-dependent feature. The purpose of these weights is to increase the influence of users that are focused in a single topic, and conversely reduce the influence of users who post tweets about many different topics. Additionally, we want to increase the influence of users who have attracted a significant number of followers.

Specifically, we consider that a user voted according to (i) its ratio of followers to followees, *WeightedSumFollowerFollowees*, (ii) the inverse of the number of crowds s/he belongs to, *WeightedSumInverseCrowds*, and (iii) the inverse of the number of distinct sections of the crowds s/he belongs to, *WeightedSumInverseSections*. For the latter, *sections* correspond to different topics/regions in the news websites we work with, and we associate crowds to a section by looking at the prefix of the path of the article originating each crowd. For instance, articles under http://www.bbc.co.uk/news/wales/ correspond to the section "Wales" of BBC-UK. In websites organized in a different manner, other ways of defining sections may be necessary.

Additionally, we characterize the activity of users contributing to a candidate by averaging the following quantities in each crowd: their overall volume of tweets per day (*UserTweetsDaily*), their number of followers and followees (*UserFollowers* and *UserFollowees*), and how many tweets they have favorited (*UserFavorites*). We also obtained statistics from their tweets by computing the fraction of their tweets that contains a re-tweet mark "RT", a URL, a user mention or a hashtag (respectively *UserFracRetweets*, *UserFracURL*, *UserFracMention*, and *UserFracHashtag*).

5.6 Results

The performance of our automatic method for discovering related stories is shown in Table 3. This method was applied over the three most frequent URLs on each slice. This was found to be much faster than considering all candidates and, in addition, it led to a similar accuracy than considering them all – this means that this set usually contains the related articles that matter.

We include results with the 2 frequency-based features (*CandidateNormalizedFrequency* and *CandidateInverseDocFrequency*), the 5 text-based features, the 11 user-based features, and combinations of them. We observe that as expected the combination of these features yields the best performance. User-based features are valuable, even if they cause a decrease of 3 points of recall (at the desired level of precision) for BBC-WORLD; they bring a substantial increase of 10 points for AJE and BBC-UK.

Table 3: Evaluation of the discovery of related articles, in terms of area under the ROC curve (AUC) and recall at 2/3 precisions (R@2/3). Each row corresponds to a set of features. Empty cells indicate that a set of features is unable to attain 2/3 precision.

		AJE		BBC-WORLD		BBC-UK	
Features	#	AUC	R@2/3	AUC	R@2/3	AUC	R@2/3
Frequency	2	0.65	-	0.64	-	0.54	-
Text-based	5	0.87	0.40	0.85	0.44	0.66	-
User-based	11	0.81	0.30	0.70	-	0.64	-
Freq+Text	7	0.89	0.62	**0.88**	**0.52**	0.66	0.04
Freq+User	13	0.79	0.32	0.72	-	0.64	0.11
Text+User	16	0.92	0.66	0.80	0.43	0.73	0.14
All	18	**0.92**	**0.72**	0.85	0.49	**0.71**	**0.14**

In the case of BBC-UK we discover 14% of the related stories using our method. In the cases of AJE and BBC-WORLD we can discover half or more of the related articles in each crowd at the same level of precision. The difference in performance can be partially explained by the high proportion of weakly-related stories in BBC-UK (see end of Section 5.2), e.g. road accidents that are related to other road accidents but often do not belong to long-standing issues such as the ones covered by BBC-WORLD and AJE.

Our features largely complement each other, as several feature selection methods failed to produce consistent gains in terms of these metrics. We can apply a feature selection method to BBC-WORLD to make it perform better, but if we use the same feature selection method in the other datasets we decrease the effectiveness of our models. In a real-world deployment of such a system, it will therefore be important to identify the particular combination of features that lead to the best performance on a specific dataset.

We observe that across datasets some features are always valuable, while others contribute only in some cases. Table 4 shows the features sorted by decreasing order of importance, using an aggregation (Borda count) of their rankings by chi-squared tests in each dataset. The most important features include the similarity to the original story, as well as measures of how unique is the association of the candidate URL and its contributing users to the specific story's crowd. This interesting result is well aligned with previous works (tweet contents as an important feature) but also with the characteristics of the transient news crowds we reported in Section 4.

5.7 Application to Crowd Summarization

The discovery of related news stories can help summarizing the evolution of a crowd over time, as we illustrate briefly in this section. We use as example the article "Central African rebels advance on capital", posted in AJE on 28 December, 2012. We considered a baseline that selected up to 3 candidates, posted by at least 3 users each, based on their frequency. This is the method employed to produce Table 2. We compared this against our method that classified each of

Table 4: Aggregated ranking of features by importance (most important first) across the three datasets.

1	*SimOriginalBigrams*	10	*UserFavorites*
2	*SimOriginalCapitalized*	11	*WeightedSumFollowerFollowees*
3	*WeightedSumInverseCrowds*	12	*CandidateNormalizedFrequency*
4	*SimOriginalUnigrams*	13	*UserFracHashtag*
5	*CandidateInverseDocFrequency*	14	*UserFracMention*
6	*UserTweetsDaily*	15	*UserFracURL*
7	*SimOthersHashtags*	16	*UserFollowees*
8	*WeightedSumInverseSections*	17	*UserFracRetweets*
9	*UserFollowers*	18	*SimOriginalHashtags*

these candidates as relevant or not. We took the output of both systems and used frequent words used in the tweets containing each URL to create word clouds for the time slices t_1, t_8 and t_{14} of this crowd, as show in in Figure 7. As usual, font sizes are proportional to word frequencies.

The word clouds show that the candidates filtered by our method belong to follow-up articles of the original one. Four days after the news article was published (t_8), several members of the crowd tweeted an article about the fact that the rebels were considering a coalition offer. Seven days after the news article was published (t_{14}), crowd members posted that rebels had stopped advancing towards Bangui, the capital of the Central African Republic. If we do not filter the candidates (using our method) we find articles on a wide range of topics that are popular, but weakly related or not related at all to the original news article. The method we use to discover related articles can yield a method for representing the evolution of a news story over time.

6 Conclusions

In this paper, we are interested in Twitter as a medium to help journalists and news editors of mainstream media outlets in rapidly acquiring follow-up stories and information about the articles they publish. We propose to do so by leveraging "transient news crowds", which are loosely-coupled groups that appear in Twitter around a particular news item. That a person posts a news article to a microblogging site may seem a trivial action, and indeed in terms of individuals, not much can be read from this event. However, when we consider people in aggregate, we uncover a noisy yet usable signal that can lead to new insights and to the discovery of new related content. This may be of particular interest in the context of news, as studied in this paper.

We have observed that after they tweet a news article, people's subsequent tweets are correlated during a brief period of time. We have shown that such correlation is weak but significant, in terms of e.g. reflecting the similarity between the articles that originate a crowd. We have also showed that just as the majority of crowds simply disperse over time, parts of some crowds come together again around new newsworthy events. As an application of this observation, we have designed and validated experimentally a method for uncovering related contents to a news article. This method can be used to build a practical system in which a journalist can be presented with a selection of new stories that are often related to the one s/he originally authored.

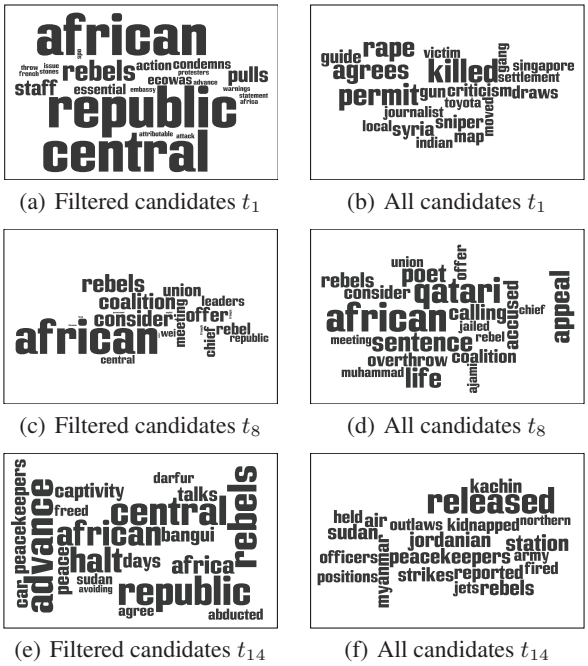


Figure 7: Word clouds generated for the crowd on the AJE story "Central African rebels advance on capital", by considering the 20 most frequent terms appearing in stories filtered by our method (left) and on the top-3 candidates by frequency (right).

A fundamental concern for journalists is how to find reliable information sources on a given topic. These sources are usually either (i) primary sources that can provide authoritative information by themselves, such as eye witness of a developing crisis, or (ii) secondary sources that are proficient at aggregating information from primary sources, such as specialized news curators. The study of transient news crowds can help in both tasks.

Future work. Any content producer wants to learn as much as possible about their engaged audience. In the case of online news writers and editors, this includes a number of questions whose answer can be obtained from a careful analysis of news crowds, including information about their demographics and ideological/political leaning. Whereas in the past the news rooms were responsible for setting agendas (McCombs and Shaw 1972), nowadays social media users have the power to influence news rooms.[9] The question to address next is how influential news crowds are to news provider, for instance, in triggering them to create more content related to popular or salient news.

One important and so far not considered aspect is the nature of news curators that belong to a crowd, for instance, the ways in which they end up curating and influencing the information their peers have about news and other top-

[9] http://www.niemanlab.org/2012/03/mohamed-nanabhay-on-al-jazeeras-online-growth-and-the-future-of-news-distribution/

ics (Katz and Lazarsfeld 1955). We have observed that news crowds contain a mixture of automatic and manually-operated accounts that are highly specialized on a particular news topic or region. We have experimented with a set of user-centric features to automatically determine news curators for a given story (Lehmann et al. 2013). Our next step is to identify types of news curators, such as opinion leaders, who are news curators that tweet directly from the news website (using the 'share' buttons and not retweeting). Opinion leaders are considered to mediate between mass media and the public in the so called two-step flow of information (Wu et al. 2011).

Data availability. The dataset used in this study is available upon request for research purposes.

7 Acknowledgments

This work was partially funded by Grant TIN2009-14560-C03-01 of the Ministry of Science and Innovation of Spain. This work was carried out as part of Janette Lehmann internship at QCRI. The authors would like to thank Mohammed El-Haddad and Nasir Khan from Al Jazeera for insightful discussions and comments.

References

Agarwal, D.; Chen, B.-C.; and Wang, X. 2012. Multi-faceted ranking of news articles using post-read actions. *CoRR* abs/1205.0591.

Benevenuto, F.; Magno, G.; Rodrigues, T.; and Almeida, V. 2010. Detecting spammers on twitter. In *CEAS*.

Carvin, A. 2013. *Distant Witness*. CUNY Journalism Press.

Chen, J.; Nairn, R.; Nelson, L.; Bernstein, M. S.; and Chi, E. H. 2010. Short and tweet: experiments on recommending content from information streams. In *CHI*, 1185–1194.

Diakopoulos, N.; Choudhury, M. D.; and Naaman, M. 2012. Finding and assessing social media information sources in the context of journalism. In *CHI*, 2451–2460.

Duh, K.; Hirao, T.; Kimura, A.; Ishiguro, K.; Iwata, T.; and man Au Yeung, C. 2012. Creating stories: Social curation of twitter messages. In *ICWSM*.

Fishman, M. 1980. *Manufacturing the News*. University of Texas Press.

García-Silva, A.; Kang, J.-H.; Lerman, K.; and Corcho, Ó. 2012. Characterising emergent semantics in twitter lists. In *ESWC*.

Ghosh, S.; Sharma, N. K.; Benevenuto, F.; Ganguly, N.; and Gummadi, P. K. 2012. Cognos: crowdsourcing search for topic experts in microblogs. In *SIGIR*, 575–590.

Greene, D.; O'Callaghan, D.; and Cunningham, P. 2012. Identifying topical twitter communities via user list aggregation. *CoRR* abs/1207.0017.

Gupta, A.; Joshi, A.; and Kumaraguru, P. 2012. Identifying and characterizing user communities on twitter during crisis events. In *DUBMMSM*.

Hsieh, C.; Moghbel, C.; Fang, J.; and Cho, J. 2013. Experts vs the crowd: Examining popular news prediction perfomance on twitter. In *WSDM, 2013*.

Java, A.; Song, X.; Finin, T.; and Tseng, B. L. 2007. Why we twitter: An analysis of a microblogging community. In *WebKDD/SNA-KDD*, 118–138.

Katz, E., and Lazarsfeld, P. 1955. Personal influence. New York: The Free Press.

Kim, D.; Jo, Y.; Moon, I.; and Oh, A. 2010. Analysis of twitter lists as a potential source for discovering latent characteristics of users. *ACM CHI Workshop on Microblogging*.

Kong, S., and Feng, L. 2011. A tweet-centric approach for topic-specific author ranking in micro-blog. In *ADMA (1)*, 138–151.

Kwak, H.; Lee, C.; Park, H.; and Moon, S. B. 2010. What is twitter, a social network or a news media? In *WWW*, 591–600.

Lehmann, J.; Castillo, C.; Lalmas, M.; and Zuckerman, E. 2013. Finding news curators in twitter. In *WWW/SNOW workshop*.

Liao, Q. V.; Wagner, C.; Pirolli, P.; and Fu, W.-T. 2012. Understanding experts' and novices' expertise judgment of twitter users. In *CHI*, 2461–2464.

Lotan, G.; Gaffney, D.; and Meyer, C. 2011. Audience analysis of major news accounts on twitter. Technical report, SocialFlow.

McCombs, M. E., and Shaw, D. L. 1972. The agenda-setting function of mass media. *Public Opinion Quarterly* 36:176–187.

Michelson, M., and Macskassy, S. A. 2010. Discovering users' topics of interest on twitter: a first look. In *AND*, 73–80.

Morales, G. D. F.; Gionis, A.; and Lucchese, C. 2012. From chatter to headlines: harnessing the real-time web for personalized news recommendation. In *WSDM*, 153–162.

Pal, A., and Counts, S. 2011. Identifying topical authorities in microblogs. In *WSDM*, 45–54.

Pew Research Center. 2012. In changing news landscape, even television is vulnerable. trends in news consumption: 1991-2012. http://www.people-press.org/2012/09/27/in-changing-news-landscape-even-television-is-vulnerable/.

Phelan, O.; McCarthy, K.; Bennett, M.; and Smyth, B. 2011. Terms of a feather: Content-based news recommendation and discovery using twitter. In *ECIR*, 448–459.

Purohit, H.; Ruan, Y.; Joshi, A.; Parthasarathy, S.; and Sheth, A. 2011. Understanding user-community engagement by multi-faceted features: A case study on twitter. In *SoME Workshop/WWW*.

Stajner, T.; Thomee, B.; Popescu, A.; and Jaimes, A. 2013. Automatic selection of social media responses to news. In *ACM WSDM*.

Subasic, I., and Berendt, B. 2011. Peddling or creating? investigating the role of twitter in news reporting. In *ECIR*, 207–213.

Tsagkias, M.; de Rijke, M.; and Weerkamp, W. 2011. Linking online news and social media. In *WSDM*, 565–574.

Wang, A. H. 2010. Don't follow me: Spam detection in twitter. In *SECRYPT*.

Welch, M. J.; Schonfeld, U.; He, D.; and Cho, J. 2011. Topical semantics of twitter links. In *WSDM*, 327–336.

Weng, J.; Lim, E.-P.; Jiang, J.; and He, Q. 2010. Twitterrank: finding topic-sensitive influential twitterers. In *WSDM*, 261–270.

Wu, S.; Hofman, J. M.; Mason, W. A.; and Watts, D. J. 2011. Who says what to whom on twitter. In *WWW*, 705–714.

Zelizer, B. 1993. Journalists as interpretive communities. *Critical Studies in Mass Communication* 10(3):219–237.

Zhang, Y.; Wu, Y.; and Yang, Q. 2012. Community discovery in twitter based on user interests. *Journal of Comp. Inf. Systems*.