# 'What are you Tweeting about?': A survey of Trending Topics within Twitter

Marc Cheong

Clayton School of Information Technology,
Monash University
Victoria, 3800
Australia

marc.cheong@infotech.monash.edu.au

**Abstract.** Twitter allows users to observe the top ten popular terms or topics of discussion at any given moment through its 'Trending Topics' feature. In this paper, we monitor the Trending Topics chart to survey the topics frequently discussed within the Twitter community (i.e. the 'Twitterverse'). This allows us to gain better insight into the collective viewpoint and *zeitgeist* exhibited by the Twitter ecosystem as a whole; and also to learn more about the subjects of interest of the typical Twitter user by looking at the 'big picture' context of the common types of Twitter messages that are promoted to Trending Topic status.

## 1 Introduction

Twitter [13] has evolved from being a simple microblogging service with the sole objective of answering the trivial question "what are you doing?" to an Web 2.0 phenomenon that has a large user base with exponential growth rates [18]. It is now being used globally by people from all walks of life, including celebrities, politicians, and organizations [2].

One of the interesting features of Twitter is the existence of Trending Topics, which is a list of top ten most tweeted topics ranked by Twitter's proprietary algorithm ('Tweeting' is a term for writing Twitter messages). This list is readily available at `http://twitter.com/` as can be seen in Figure 1.

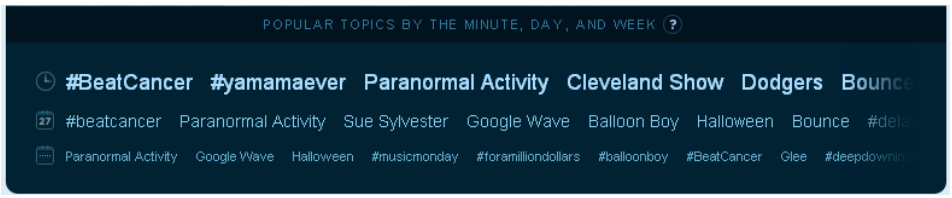


**Fig. 1.** A screenshot of Twitter Trending Topics accessible via `http://twitter.com/`.

Trending Topics are a useful tool to discover the topics of interest to the 'Twitterverse' – a term used by Twitter users to describe the Twitter community as a whole, similar to the usage of the term *blogosphere* for blogs – as a whole. In other words, it is a simple way of measuring the collective and emergent behavior exhibited by Twitter users worldwide at any given moment.

The motivation behind this paper is to survey the list of keywords present on Trending Topics, study the significance these topics have on the Twitter user base, and observe how these topics evolve over time. This ideally gives us a 'big picture overview' to better understand the nature of information that can be mined from the Twitterverse, and also the nature of the latent demographics behind the Twitterverse itself.

## 2 Related Work

As academic research on Twitter is generally a new field, there exists a limited amount of research into the Twitter ecosystem.

Specifically, discussion on Twitter Trends is limited to Cheong & Lee [2] who analyzed the demographics behind a sampling of both trending and non-trending topics by discovering the inherent user base and manually identifying properties of the users as discovered via the Twitter API [13]. They have found that by retrieving the messages discussing about a particular trend, the user demographics and Twitter usage patterns exhibited will mirror the subject's real-world properties. As an example, they determined that the Trending Topic *Grey's Anatomy* (a US television drama) has a Twitter user demographic that accurately resembles the real-world viewership of the TV show [2].

The key difference between this paper and [2] is that Cheong & Lee have performed a 'top-down' analysis detailing the specifics of the user base (demographics, usage habits) of a particular trend. In this paper, however, we perform an overview of the Trending Topics themselves – a high-level, 'big picture' overview – without assuming anything about the underlying user base.

On a broader scope, literature on various aspects of Twitter in academic research has been growing within the past two years. Krishnamurthy et al. [12] has performed research on the growth rates, geographic spread, and other statistical properties on the Twitter user base. Java et al. [11] and Huberman et al. [9] performed similar research, but with more emphasis on the social network perspective of Twitter. To a certain extent, Java et al. [11] has worked on analyzing popular keywords on Twitter by focusing on phrase rankings restricted to a particular Twitter user network (composed of a subset of users interlinked with each other via a *follow* connection on Twitter).

A focused usage of Twitter on emergency response and mass convergence – Twitter usage habits during hurricanes, and US presidential campaigns, respectively – has been studied by Hughes & Palen [10]. A prototype framework of using Twitter to enable authorities to better respond to acts of terror has been proposed by Cheong & Lee [3]. Such uses of Twitter are given consideration in our review of prior work, as crisis topics and topics of mass convergence (as clas-

sified by [10]) are expected to be promoted to a Trending Topic if given enough attention by the Twitterverse. A case in point would be the activism campaigns in the Iran elections of 2009 [7] (hashtag `#IranElection`) and the Moldovan "Twitter Revolution" [16] (hashtag `#Moldova`) which have been promoted to Trending Topic status. In the case of `#IranElection`, it has been promoted to first place and has been persistently first for a brief period of time due to worldwide awareness and conversation.

In the non-academic domain, coverage on Twitter Trends in mainstream and popular media has been steadily increasing. A search on Google News for the key phrase `Twitter Trends` returned approximately 1700 news stories. [1] One such example of the media putting Twitter Trends in the spotlight is of *The Independent* – a UK-based newspaper – which has a weekly online feature dissecting each week's popular trends [19].

## 3 Methodology

### 3.1 Our 'listening post'

The first part to our textual survey of the Trending Topics on Twitter is the development of a simple Java program based on the Twitter API [13]. It polls the Twitter API to retrieve a list of ten Trending Topic strings after a specified time interval. The interval of five minutes is chosen to avoid wasting resources and clog server traffic (abiding by the API call limit restriction, so as to not abuse whitelisting permission granted by Twitter for our research), while at the same time providing sufficient granularity in observing the trends' movements.

Our 'listening post' records every trend as reported verbatim by the API, then marks its minimum and maximum position on the Trending list ($1-10$ inclusive). To track its permanence on the Trending list, a simple counter mechanism is implemented, where every 5-minute poll of the API increments the counter; and if the trend is seen for the first time, it is the timestamped.

We let our 'listening post' run for ten days, beginning the $11^{\text{th}}$ of November until the $21^{\text{st}}$ of November, and the results obtained at the end of the observation period are dumped into a CSV file for ease of processing and analysis.

### 3.2 Sanitization and annotation

The data is then sanitized to merge duplicates caused by irregular casing. An important point to note here is that in our observation, Twitter's proprietary trending algorithm exhibits case-sensitivity with respect to the strings found. Some of the Trending Topic strings have differing case in between different API polling results – e.g. from `lowercase` to `CamelCase`.

The noun, concept, or meme behind each trend string is then interpreted using *What The Trend?* [20], a collaborative website designed for users to describe and explain the meaning behind each trend string (Figure 2).

[1] Search performed on Google News on $3^{\text{rd}}$ December via `http://news.google.com/news?q=twitter+trends` returned 1661 unique news entries.


**Fig. 2.** *What The Trend?*, a crowdsourced trend description page, available at `http://www.whatthetrend.com/`.

The justification behind using *What The Trend?* to provide crowdsourced interpretations of the trend strings, as opposed to other methods, is that the official Twitter site itself [13] utilizes *What The Trend?* to explain the trends on their website. For trends that have no explanation, Google (specifically its Translate and News services) is consulted to provide interpretations to trends in other languages or trends that are not defined on the *What The Trend?* website – e.g. by translating the string from a different language into English, searching for proper context in newswire services.

Based on the explanation by *What The Trend?*, each entry is tagged with a category, which will be discussed shortly ahead. A country code is also assigned to each trend, as some trends are localized and/or location information is provided from the crowdsourced explanations.

### 3.3 Category selection criteria

There is no conclusive set of category labels that can be directly used for our trend data, as categorization of trends on Twitter have not been given much emphasis as per our review of prior work. On the other hand, there exists a (non-exhaustive) listing of categories used by studies on social awareness (e.g. Naaman et al. [17]). However, these are more focused on the information needs of a particular individual in the domain of individual computer-human interaction – which is unsuitable for our categorization task.

Therefore, we have decided to create a list of categories tailored to classify our obtained trends. Information sources and prior research used to develop this list include:

1. Tags available on *What The Trend?* [20]
2. The general categories of topics surveyed by Cheong & Lee [2] in their original Twitter Trends research paper.
3. Habits of information sharing, c.f. Dearman et al. [5].
4. The habits of live reporting, c.f. O'Reilly & Milstein [18] and Ebner & Schiefner [6].
5. Internet memes and viral information sharing, c.f. Arbesman [1], Wasik [21], and Hodge [8].

Our final list of categories used to annotate the tags are as per Table 1.

| **Category** | **Trend refers to** |
|---|---|
| activism | usage of Twitter for activism (e.g. to spread awareness about a charity) |
| conference | users 'conference-Tweeting' about an ongoing conference |
| culture | popular culture |
| entertainment | entertainment, e.g. music, television, movies and celebrities |
| general | common phrases and proper nouns, but without sufficient context to frame the trend |
| meme | Internet and Twitter-based memes |
| meme+entertainment | memes that originate from popular entertainment (e.g. started by a celebrity) |
| news | current affairs and local/global news (includes crisis events) |
| science | scientific news (excluding IT/technology-related subjects) |
| spam | spam, phishing attempts, malicious activity |
| sport | sporting events and sports news |
| tech | specifically IT/technology-related subjects (e.g. games, gadgets, software) |
| Twitter | official changes to the Twitter service introduced by Twitter Inc. |
| viral marketing | use of Twitter to virally promote a product (without malicious intent) |

**Table 1.** Categories used for annotating the harvested trend keywords.

### 3.4 Data aggregation and interpretation

Once the list of trend strings have been sanitized and annotated, they are then imported into a spreadsheet for study. As this is a high-level exploratory survey on the trends themselves, no user demographic information will be obtained on the Twitter user base (in contrast with research such as [2], [11], and [9]). Rather, information on trends themselves such as statistics and trend rankings are interpreted, to identify the common interests, tweeting habits and collective behavior of the Twitterverse.

## 4 Findings, interpretation, and discussion

### 4.1 Obtained data

Over the observation period ($11^{\text{th}}$ November – $21^{\text{st}}$ November), we have observed 677 Trending Topics, which are case-sensitive.

*A note on the continuity of data.* The data collection was briefly interrupted for at most a few hours during the course of data collection due to network issues and Twitter scheduled maintenance; the 'listening post' program was resumed immediately after connectivity was established.

The only preprocessing performed on the strings are merging all differing variations in case (as described in Section 3.2). As a result of the sanitization process, 466 unique topic strings are obtained.

Note that similar strings (distinguished based on their spelling and phrasing) are *not* collated together, as the purpose of this study is to survey the obtained trends without performing additional assumptions e.g. disambiguating terms or collating them to a common subject. Following from this, several points can be observed about the behavior of Twitter's Trending algorithm.

1. It does not automatically group together a string and its related hashtag [prefixed with a hash (`#`) symbol].
   **Example**: '`#oprah`' and '`oprah`' are treated as separate Trending Topics, not as one single topic.
2. Different variations in phrasing the same subject or typographical differences create two or more Trending Topic strings.
   **Example**: "`chrome os`", "`google chrome os`", "`with chrome os`", "`google chrome os to`" [2] are four separate topics.
3. Due to the above, certain trend strings have no particular meaning until it is rephrased in the context of the original Twitter messages.
   **Example**: the keyword '`lax`' observed does not refer directly to the LAX Airport as a subject, nor the English adjective 'lax' [i.e. not strict]; it actually refers to news of Mike Tyson arrested at LAX airport.

### 4.2 Generalizations on Trending Topics

We categorize our list of Trending Topic strings firstly by the absence or presence of hashtags. The presence of hashtags indicate the use of social tagging to categorize posts to allow ease of communication and searching for related posts (further discussion on hashtags can be found in [18], [2], and [4]).

| **Hashtag presence** | **Percentage** |
|---|---|
| Hashtag present | 28.76% (134 out of 466) |
| No hashtag | 71.24% (332 out of 466) |

**Table 2.** Presence versus absence of hashtags

[2] The last phrase was originally '*Google Chrome OS To Launch Within A Week*', quoted verbatim from TechCrunch, an influential technology blog. The process of quoting the phrase verbatim (a 'retweet' or RT in Twitter) caused it to be a separate Trending Topic. The original URL written by Michael Arrington is at `http://www.techcrunch.com/2009/11/13/google-chrome-os-to-launch-within-a-week/`.

Our results in Table 2 indicate that almost 30% of Trending Topics in our survey contain hashtags. In our interpretation of this, social tagging and self-organizing behavior is present among the user base contributing to the discussion of a Trending Topic to a certain extent. As discussed in [18], hashtags are a user-created "ad hoc solution... to categorize a message", where there is no tagging system in Twitter to begin with. We now focus our attention to the categories of our Trending Topic strings. Based on the annotated categories as per Table 1, we obtain the following distribution over our sample of Trending strings in Figure 3.

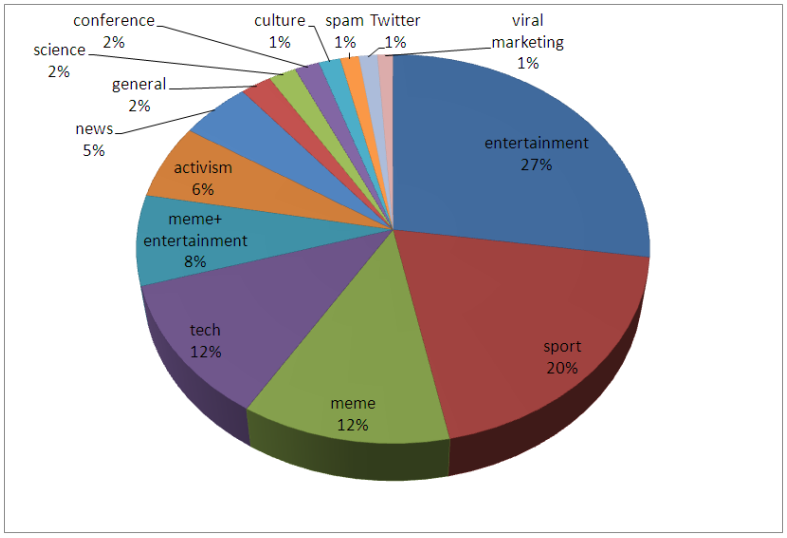


**Fig. 3.** Breakdown of Trending strings by category in our surveyed dataset (rounded to the nearest percentage point).

Majority of the trends surveyed consist of entertainment-related topics (27.25%, or 127 trends), followed by sports news (19.53%, or 91 trends), then by Internet memes and technology news (approximately 12% each; with 56 and 54 trends respectively). Memes that have their origins in entertainment (e.g. started by celebrities, or about a particular artist) account for 35 trends (7.51%); if we merge both memes and memes in entertainment as one combined category, it would have the same percentage of trends as sports news.

An interesting entry is the presence of changes to the Twitter official web service and/or API, which accounted for 6 trends (1% of the total) – this indicates that users discuss about the Twitterverse itself in the context of everyday Twitter messages (which may be beneficial to research on social presence and similar fields: examples of related research include [15] and [14]). To follow up on Cheong & Lee's observation [2], we calculated the amount of spam topics in

our survey, which turns out to be just about 1% (6 trend strings) of all total strings.

Finally, we take a look at the country and regional statistics for the trend strings gathered in our survey. In our survey methodology, trends which are of global concern or where the country or region is not explicitly specified are not annotated with a specific country. Out of the total, 253 trends have been associated with a particular country (with the remaining 213 not considered for analysis). Some trends are associated with two or more countries, hence they are tagged with 2 country labels. Absolute percentages are not used due to this potential overlap.

The table in Figure 4 contains the breakdown by country, while the bar graph illustrates the trend statistics aggregated by region (regions are adapted from the grouping proposed in [11] and [12]).

| **Country** | **Percentage** |
|---|---|
| US | 149 |
| UK | 49 |
| Brazil | 10 |
| Indonesia | 10 |
| Phillipines | 8 |
| France | 6 |
| Ireland | 6 |
| Canada | 5 |
| Australia | 3 |
| Korea | 3 |
| New Zealand | 3 |
| Japan | 2 |
| Belgium | 1 |
| Chile | 1 |
| Mexico | 1 |
| Netherlands | 1 |
| Puerto Rico | 1 |
| Singapore | 1 |

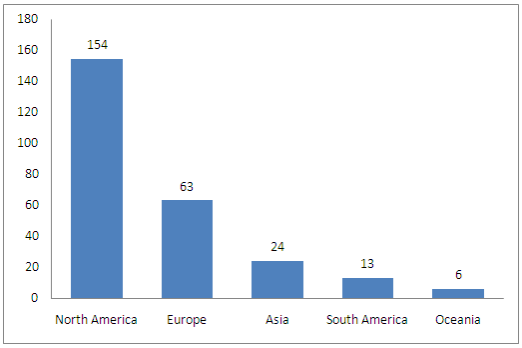


**Fig. 4.** Table (on the left) shows the breakdown of trends associated with with a particular country. The bar graph (on the right) shows the aggregated trends by region.

Looking at the aggregated results of trends by geographic region, the distribution of trends approximate the distribution of global Twitter users, as shown in studies by Java et al. [11] and Krishnamurthy et al. [12], albeit with minor differences of the last two regions' rankings.

From our survey, we have come up with an observation based on locality and categories of trend strings. Twitter's user base is predominantly focused in the United States where Twitter had its origins, which explains why certain trends are highly specific (localized) to the US. The majority of sporting events

surveyed involve American-based sports – such as National Football League, the National Basketball Association tournament and Major League Baseball – all of which are highly popular in the US. Entertainment news such as those regarding American celebrities and artists are also commonly found in our studied dataset.

### 4.3 Popular topics

From our dataset, we then analyze the topics based on their popularity, and duration of time they remain on the top ten Trending list. As explained in our methodology, we have a counter to track how many intervals a particular trend stays on the list of top ten trends, and also a record of the highest possible rank a topic has on that list.

By generating a scatterplot of the trend's highest rank versus the number of time intervals recorded, we find that there exists a logarithmic correlation between the two with an approximate R-squared value of 0.60, as Figure 5 illustrates. Several findings can be drawn from our analysis.

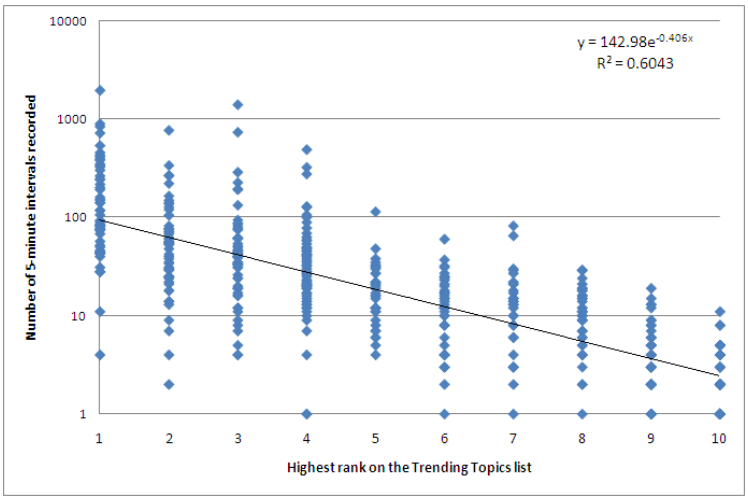


**Fig. 5.** Trends' highest rank versus the number of time intervals recorded.

1. Trends which are persistently in the top ten in terms of ranking spend a long time in Trending Topics, indicating a continuous mention of the topic by the Twitterverse.
2. Most of the trends which only persist for a brief period of time tend to drop below the top ten rankings immediately after it 'Trended'.
3. There are some trends that suddenly peaked to the top half of the trends list but quickly 'died off' (fell off the trends list) due to lack of attention. This may indicate the usage of alternate strings (such as hashtags or different

rephrasings of the topic) by users to carry on the conversation, or due to spam epidemics that quickly fell off due to user awareness and preventative measures by Twitter (cf. observation in [2]).

4. The top 26 trends (the first 5%, out of a possible 466) recorded over 50% of the cumulative time recorded by all trends on the Trending list (see Figure 6).

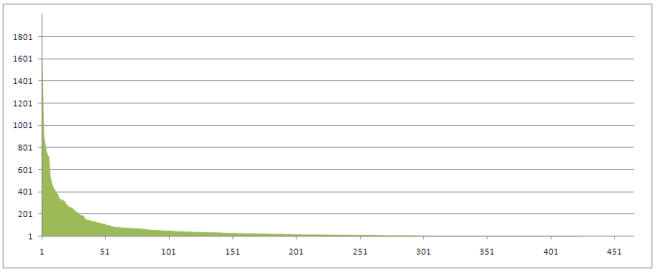


**Fig. 6.** Distribution of trends in order of time spent on the Trending list.

Our final study on this dataset is observing the composition of the top 5% of the trends, sorted in descending order by (a) time on the Trending list, and (b) highest rank on the Trending list. It is visualized and color-coded according to category according to Figure 7.

Most of the top trends are on memes originating from Twitter or other parts of the Internet (excluding those started by celebrities). Technology-related news and entertainment news are also visible in the list. However, news on activism and sport rarely reach the top ten Trending list.

## 5 Conclusion and further work

This paper has shown that an exploratory text survey of the top trend strings appearing on the Twitter Trends list can reveal much information about the interests and topics of discussion among the Twitterverse. To a certain extent, latent emergent behavior can be observed simply by assigning meaning and context to the trends themselves, as have been done here. By also tracking the movement in rank of trends over time, we are able to have an overview of how topics populate the Trending Topics chart with respect to its popularity; with applications in studying memetic behavior and viral information spread.

From this study, several areas can be further explored in the future. Firstly, by combining a 'big-picture' analysis of trends with top-down demographic mining, we can study the user base contributing to such topics in much detail and see how the examination of the user base's demographics and Twitter usage habits is reflected in the meaning and contexts of the trends. Memetic behavior can also

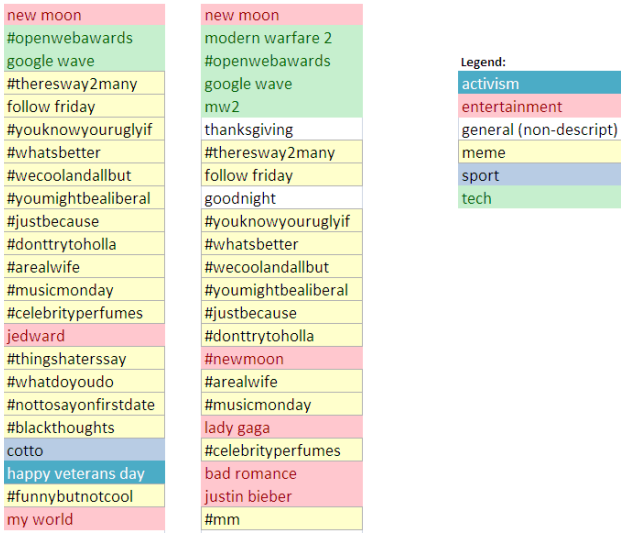


**Fig. 7.** Top 26 strings, in order of (a) time on the Trending list, and (b) highest rank on the Trending list.

be tracked on Twitter by coupling this research with methodologies of surveying social information spread. Finally, research in opinion mining and sentiment analysis can be performed by combining automated trend analysis, user base exploration, and clustering methods.

## References


1. S. Arbesman. The Memespread Project: An initial analysis of the contagious nature of information in social networks. Available from `http://www.arbesman.net/memespread.pdf`, 2004.
2. M. Cheong and V. Lee. Integrating web-based intelligence retrieval and decision-making from the Twitter Trends knowledge base. In *Proc. CIKM 2009 Co-Located Workshops: SWSM 2009*, pages 1–8, 2009.
3. M. Cheong and V. Lee. A microblogging-based approach to terrorism informatics: exploration and chronicling civilian sentiment and response to terrorism events via Twitter. 2009.
4. d. boyd, S. Golder, and G. Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *Proc. HICCS-43*, 2010.
5. D. Dearman, M. Kellar, and K. N. Truong. An examination of daily information needs and sharing opportunities. In *Proc. CSCW 2008*, pages 679–688, 2008.
6. M. Ebner and M. Schiefner. Microblogging - more than fun? In *Proc. IADIS Mobile Learning Conference 2008*, pages 155–159, 2008.
7. J. Fleishman. Mideast hanging on every text and tweet from Iran. *Los Angeles Times*, June 17 2009.
8. K. Hodge. It's all in the memes. *The Guardian*, August 10 2000.

9. B. A. Huberman, D. M. Romero, and F. Wu. Social networks that matter: Twitter under the microscope. Available from `http://ssrn.com/abstract=1313405`, 2008.
10. A. L. Hughes and L. Palen. Twitter adoption and use in mass convergence and emergency events. In *Proc. 6th International ISCRAM Conference*, 2009.
11. A. Java, X. Song, T. Finin, and B. Tsen. Why we Twitter: An analysis of a microblogging community. In *Proc. 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 118–138. Springer-Verlag, 2009.
12. B. Krishnamurthy, P. Gill, and M. Arlitt. A few chirps about twitter. In *Proc. WOSN 2008*, pages 19–24, 2008.
13. M. Mayer. *What The Trend?* Available from `http://www.whatthetrend.com`, 2009.
14. B. J. McNely. Backchannel persistence and collaborative meaning-making. In *Proc. SIGDOC'09*, 2009.
15. E. Mischaud. Twitter: Expressions of the whole self. Master's thesis, London School of Economics and Political Science, 2007.
16. A. Mungiu-Pippidi and I. Munteanu. Moldova's "Twitter Revolution". *Journal of Democracy*, 20(3):136–142, 2009.
17. M. Naaman, J. Boase, and C. Lai. Is it Really About Me? message content in social awareness streams. In *Proc. CSCW 2010*, 2010.
18. T. O'Reilly and S. Milstein. *The Twitter Book*. O'Reilly Media, Inc., Sebastopol, CA, 2009.
19. Relax News. Current Twitter trends: Google Wave, 'A real wife'. Available from `http://www.independent.co.uk/news/media/current-twitter-trends-google-wave-a-real-wife-1820222.html`, November 13 2009.
20. Twitter Inc. *Twitter*. Available from `http://www.twitter.com`, 2009.
21. B. Wasik. *And Then There's This: How Stories Live and Die in Viral Culture*. Penguin Group (USA), New York, NY, 2009.