Location and Timeliness of Information Sources During News Events

Elad Yom-Tov
Yahoo Research
111 W 41st st.
New York, NY 10018, USA
eladyt@yahoo-inc.com

Fernando Diaz
Yahoo Research
111 W 41st st.
New York, NY 10018, USA
diazf@yahoo-inc.com

ABSTRACT

People nowadays can obtain information on current news events through media outlets, social media, and by actively seeking information using search engines. In this paper we investigate the temporal relationship between news coverage by media outlets, social media, and query logs and show that social media frequently precedes other information sources. Additionally, we demonstrate that there is strong negative correlation between the probability for reporting of an event and the distance of the information source from the event.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search process; H.4.3 [**Communications Applications**]: Information browsers

General Terms

Algorithms

Keywords

social,physical,distance,information,need

1. INTRODUCTION

The information people can receive of news events comes from diverse sources, including media outlets (TV, newspapers, media sites on the web), social media (Twitter, Facebook, etc), and active seeking of information using search engines on the web. These news sources assign importance to events (as manifested by the likelihood and extensiveness they will report on them) based on attributes of the event and on features of the news source itself.

Coverage of news is known to be influenced by the physical distance of consumers from the location of the event. As [2] showed, the distance of an event from the USA is one of the most predictive attributes for its media coverage. Furthermore, Wu [5] showed that the volume of media related to news events in Canada and Mexico by USA-based news channels is partially explained by the distance of the events from the closest USA border.

Historically, news outlets were the source of many people's information of world events. More recently, social media has emerged as a timely source of information. For example, almost immediate detection of earthquakes in Japan can be performed with very high confidence by tracking the microblogging service Twitter [4]. Active information seeking by users, through search engine queries, is another important way by which users can acquire information. This is exemplified by the work of Backstrom et al. [1], who developed a model for pinpointing events based on the location of people querying about them. Both [1] and [4] found that the closer people are to an event, the more likely they are to seek information about the event or to report about it.

In this paper, we investigate the relationship between the volume of information as a function of location and time during unusual news events. We show that social media frequently precedes other information sources, and that the probability to report an event has a high negative correlation with the location of the information source.

2. DATA

In this paper we analyze three events, which are briefly described below (for more details, see [6]):

1. **San Bruno event**: On September 9th, 2010 a large pipe carrying natural gas exploded in the city of San Bruno (near San Francisco), California. The explosion killed eight people and destroyed 38 houses.
2. **New York storm**: A violent storm passed through New York city on the 16th September 2010, killing one person and causing widespread damage to property.
3. **Alaska elections**: The elections for representatives from Alaska to the USA Senate on the 2nd November 2010 drew significant attention because of a three-way race, which was won by an independent candidate.

These events were chosen because they were physically localized in their scope and thus have a clear epicenter. Therefore, the physical distance of users from the event are clearly defined for these events. Furthermore, because they are temporally limited, they act as an impulse to the system, and thus create high levels of interest for a relatively short and clearly-defined period.

We used three types of data in our study: First, we extracted query-log data (text, time, and an identification of the user who posted them) of the Yahoo search engine from several days before the event (one day for the first two events, and 10 for the last), until 8 days after the event.

The query log was parsed to identify those queries which were likely relevant to the event, using a term-matching scheme. This was done by manually generating a list of

SIGIR'11, July 24–28, 2011, Beijing, China.
ACM 978-1-4503-0757-4/11/07.

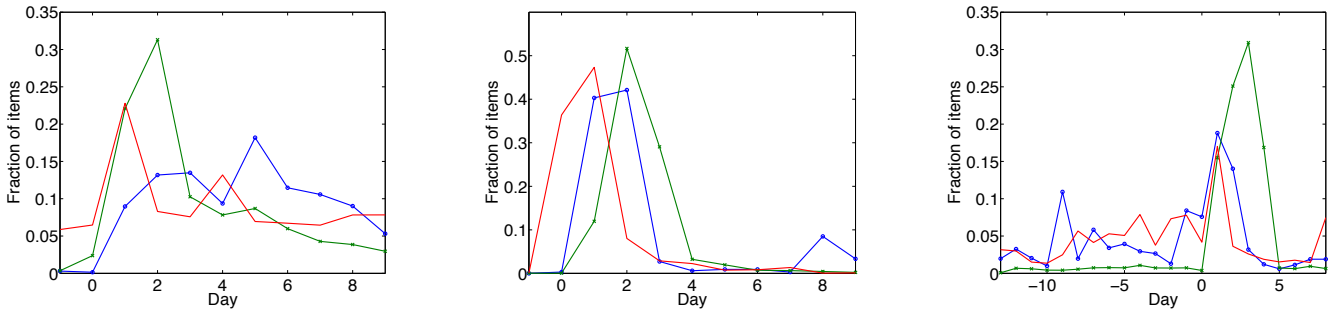

Figure 1: Fraction of media outlet items (blue, circles), social media items (red) and queries (green, crosses) posted in response to the San Bruno (left), New York (middle), and Alaska (right) events over time.

keywords for each event. The keywords were drawn from several categories such as where the event took place, who was involved in the event, and what happened at the event. We also generated a list of excluded words. Queries and keywords were stemmed using a Porter stemmer. A query was considered relevant to the event if keywords from at least two categories were used in the query, and none of the excluded words appearing in it.

We encode data at the granularity of a day in order to remove diurnal effects. The zip-code given by users at the time of registration with Yahoo was used to identify their approximate location. We computed the physical distance of each user to the event using the Haversine formula.

Daily media volume of news outlets related to an event was quantified by counting the number of document found on Yahoo News (`news.yahoo.com`) every day, which contained all the words used in any of the 50 most popular queries. Similarly, the volume of social media was measured via the number of Twitter messages which contained the words used in these queries.

3. RESULTS

Figure 1 shows the media and query volume related to the three events. Note the similarities in shapes between the different information sources, i.e. the spike on day 5 of the San Bruno event which is observed in both the query volume and general media volume. In general, the peak in social media volume precedes both general news media and query volume, and decays first. In the New York and Alaska events, query volume peaked and decayed later than media volume. San Bruno was unusual in that general media attention decayed more slowly than query volume. The cross-correlation of media volume with the number of queries per day, shown in Table 1, quantifies these findings and shows that news media and social media (Twitter) have a relatively strong correlation, as was also found by [3].

We binned the information sources according to their physical distance from the event, on a logarithmical-spaced axis, and computed the fraction of event-related items from the total items generated at this distance. Table 2 shows the Spearman correlation between the physical distance and the fraction of event-related items. The correlation is extremely high, validating the findings of [5] and extending them to social media and query volume.

Event	News-Queries	Twitter-Queries	Twitter-News
San Bruno	0.81 (-1)	0.87 (0)	0.87 (+1)
New York	0.93 (+1)	0.97 (+2)	0.98 (+1)
Alaska	0.84 (+2)	0.69 (+2)	0.82 (0)

Table 1: Maximum cross-correlation between media volume and the number of queries per day, for each of the three events. The number in parenthesis is the delay in days at which the maximum is reached, where a positive sign indicates that the first time series precedes the second time series.

Event	Queries	News	Twitter
San Bruno	-0.97	-0.97	-0.97
New York	-0.86	-0.95	-0.90
Alaska	-0.69	-0.91	-0.52

Table 2: Spearman correlation between physical distance and the fraction of media items and relevant queries, for each of the three events. All correlations are statistically significant at $p < 0.05$.

4. REFERENCES

[1] Lars Backstrom, Jon Kleinberg, Ravi Kumar, and Jasmine Novak. Spatial variation in search engine queries. In *Proceeding of the 17th international conference on World Wide Web*, WWW '08, pages 357–366. ACM, 2008.

[2] Tsan-Kuo Chang, Pamela J. Shoemaker, and Nancy Brendlinger. Determinants of international news coverage in the U.S. media. *Communications research*, 14(4):396–414, 1987.

[3] Jure Leskovec, Lars Backstrom, and Jon Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2009.

[4] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 851–860. ACM, 2010.

[5] Haoming Denis Wu. Geographic distance and US newspaper coverage of Canada and Mexico. *International Communication Gazette*, 60(3):253–263, 1998.

[6] Elad Yom-Tov and Fernando Diaz. Out of sight, not out of mind: On the effect of social and physical detachment on information need. In *Proceedings of the 34th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '11. ACM, 2011.