The Effect of Social and Physical Detachment on Information Need

ELAD YOM-TOV and FERNANDO DIAZ, Yahoo Research New York

The information need of users and the documents which answer this need are frequently contingent on the different characteristics of users. This is especially evident during natural disasters, such as earthquakes and violent weather incidents, which create a strong transient information need. In this article, we investigate how the information need of users, as expressed by their queries, is affected by their physical detachment, as estimated by their physical location in relation to that of the event, and by their social detachment, as quantified by the number of their acquaintances who may be affected by the event. Drawing on large-scale data from ten major events, we show that social and physical detachment levels of users are a major influence on their search engine queries. We demonstrate how knowing social and physical detachment levels can assist in improving retrieval for two applications: identifying search queries related to events and ranking results in response to event-related queries. We find that the average precision in identifying relevant search queries improves by approximately 18%, and that the average precision of ranking that uses detachment information improves by 10%. Using both types of detachment achieved a larger gain in performance than each of them separately.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Search process*; H.4.3 [**Information Systems Applications**]: Communications Applications—*Information browsers*

General Terms: Performance, Experimentation

Additional Key Words and Phrases: Social, physical, distance, information, need

ACM Reference Format:
Yom-Tov, E. and Diaz, F. 2013. The effect of social and physical detachment on information need. *ACM Trans. Inf. Syst.* 31, 1, Article 4 (January 2013), 19 pages.
DOI: http://dx.doi.org/10.1145/2414782.2414786

1. INTRODUCTION

In today's world, people are ever more informed of news events that take place far from their homes very shortly after events occur. This is especially true of events which are limited in time and location. The information about these events comes from diverse sources, including media outlets (e.g., television, newspapers, media sites on the Web), social media (e.g., Twitter, Facebook), and active searching of information using Web search engines. Their knowledge of the event and its different aspects is shaped as

---

This article is a substantially enhanced version of a paper presented in *Proceedings of the 34th Annual International ACM SIGIR Conference.*
E. Yom-Tov is currently affiliated with Microsoft Research Israel and F. Diaz is currently affiliated with Microsoft Research New York.
Authors' addresses: E. Yom-Tov, Microsoft Research Israel, 13 Shenkar St., Herzelia 46733, Israel; email: elad@ieee.org; F. Diaz, Microsoft Research New York, 1290 Avenue of the Americas, New York, NY 10104.

DOI: http://dx.doi.org/10.1145/2414782.2414786

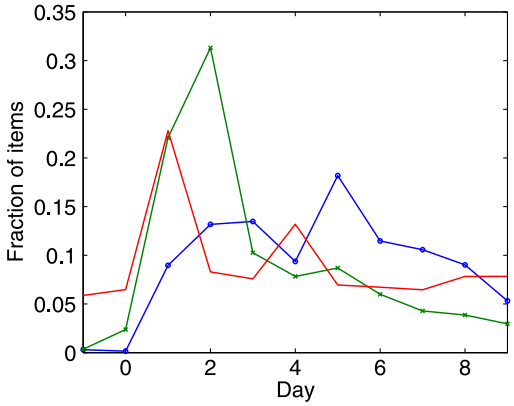

Fig. 1. Fraction of mainstream media outlet items (blue, circles), social media items (red), and queries (green, crosses) posted over time in response to the San Bruno event.

much by external forces (which items are reported by news channels they choose to watch) as much as by their preference for information.

As a motivating example, consider an earthquake event, its different information aspects, and how these might be consumed by users. People close to the event might be interested in emergency services and evacuation procedures. People far from the earthquake may want to know more general details of the earthquake and how they might help people affected by it. Finally, people not in the vicinity of the event but have friends or relatives who they think may be affected might seek the names of those injured by the earthquake.

News coverage, which may reflect people's interest in an event, is known to be influenced by the physical distance of consumers from the location of the event. Chang et al. [1987] showed that the distance of an event from the U.S is one of the most predictive attributes for its media coverage. Furthermore, Wu [1998] showed that the volume of media related to news events in Canada and Mexico by U.S-based news channels is partially explained by the distance of the events from the closest U.S border.

However, while news coverage measures information production with respect to an event, we are interested in more directly measuring people's information demand with respect to an event. Consider the graph[1] of news coverage, social media coverage, and information searching for an explosion in San Bruno, California, shown in Figure 1. Although the trends correlate with each other in general, the information-searching behavior of users is sufficiently different from that of media output to motivate a deeper study of information searching in isolation.

In this article, we investigate the information need (as defined in Carmel et al. [2006]) of individuals in response to special events—both adverse and otherwise. Because a user's true information needs are unobservable to a production Web search service, we approximate them by the user's queries, which are one partial facet of the information need. Although noisy and incomplete, queries can be harvested at scale and capture one aspect of a user's information searching. We investigate the propensity for users to query about an event as a function of their relationship to the event, specifically their social detachment from the event and their geographic distance from it. We show that a user's querying is strongly dependent on social and geographical attachment. We demonstrate that knowledge of social and geographic

[1]Details of the event and the data are given in Section 4.

information can improve retrieval by accurately identifying queries related to an event and consequently retrieving the appropriate documents.

2. RELATED WORK

Information searching with respect to news events is gaining increasing attention from the scientific community. Within this area, our work focuses on analyzing the influence of the geographic and social context of the user on querying behavior during an event.

The geographic context of the user has repeatedly been found to have an influence on querying behavior. For example, almost immediate detection of earthquakes in Japan can be performed with very high confidence by tracking the micro-blogging service Twitter [Sakaki et al. 2010]. In fact, a service based on Twitter sent alerts within 20 seconds of the events, compared to 6 minutes by the Japan Meteorological Agency. Earthquakes could also be localized by considering the propagation time of earthquakes and the reported location of users. Similarly, Backstrom et al. [2008] developed a model for pinpointing events based on the location of people querying them. In this way, both stationary terms (sports teams) and moving events (hurricanes) were pinpointed to within a few kilometers. Both Backstrom et al. [2008] and Sakaki et al. [2010] 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 the context of aggregated search, geographic features of the user as well as those mentioned in the query have been found to be very predictive of user interest in news [Hassan et al. 2009]. Similarly, Hassan et al. [2008] also found that these geographic features improved general Web search.

The social context of the user has received somewhat less treatment than geography. Carmel et al. [2009] demonstrated that in an intra-organizational setting, the fact that two people are familiar with each other can improve the suggestion of interesting content by serving one person the content which his colleague has marked as interesting. However, that study also concluded that suggesting similar content to similar people can obtain even better results, which may suggest that homophily rather than social connection is the source of improvement. Teevan et al. [2009] found that personalizing at the group scale improved effectiveness for group-related queries. The size of a population that can obtain knowledge of an event through its social network has been studied using both empirical data and theoretic models Bernard et al. [1991, 2001]. For example, Bernard et al. [2001] estimated that between 1 in 17 and 1 in 140 people knew a person who was directly influenced by the attacks against the World Trade Center on 9/11/2001. Information propagation in electronic social networks in response to a crisis event was studied in Palen et al. [2009], where it was shown how students utilized social media to form collective response decisions in response to a shooting incident at Virginia Tech University. We note that our study does not analyze the question of which information is passed via the social network. Instead, given a social network and the location of people, we study which information is of interest to them. In a recent paper, Mendoza et al. [2010] analyzed the information published and disseminated by users of Twitter during an earthquake in Chile. Their analysis demonstrated that while both rumors and facts related to the earthquake were disseminated, rumors tended to be questioned by users more than facts, which suggests that the community filters false information to some extent. Although it is a single event, information regarding the earthquake was disseminated for at least nine days, much longer than observed for many news stories in the public media (as just discussed).

The temporal analysis of news events also received some attention. For example, Leskovec et al. [2009] showed that media interest in an event peaks within two

days and decays somewhat quicker than it rises. Social media outlets have a similar attention timespan and peak, on average, 2.5 hours later than mainstream media. In the context of aggregated news search, Diaz found (somewhat unsurprisingly) that temporally local variables, such as query volume, were important in detecting the newsworthiness of a query [Diaz 2009]. In the context of Web ranking, researchers have found that a general retrieval model fails for late-breaking news events and that specialized query treatment was required to satisfy user intent [Dong et al. 2010]. This specialized ranker can be further improved by analyzing the discussion of the topic in social media [Dong et al. 2010].

There are parallels between our work and that of understanding information seeking behavior in Web users (reviewed in Martzoukou [2005]). For example, Kim [2009] described three tasks involving trip planning, finding medical information, and learning about health issues. These studies usually provide a deeper analysis into the behavior of a small number of users as they find information. However, much of the work in this area has been performed on tasks which are static in nature, that is, they do not change significantly whether they are performed today or in the future. Moreover, this article is focused on the information need as expressed in their queries and the documents they viewed, regardless of the information seeking process. Future research is required in order to use methods from information seeking research for an analysis of information needs during transient events.

3. A MODEL OF USER PREFERENCE FOR TOPIC INFORMATION

A user's information need is comprised of the knowledge and news a user seeks for a specific subject [Carmel et al. 2006]. Define the primal information object as a topic, $T \in \mathcal{T}$, which is information pertinent to a defined subject. Each topic may be partitioned into aspects $\{a_i\} \in T,\ i = 1, 2, \ldots N$, which describe different subtopics of the topic. For example, information about an earthquake event may be subdivided into information about emergency services, evacuation procedures, magnitude, and so forth. A user may be interested in one or more of a topic's aspects. Note that as this article is concerned with information needs during transient events, we refer to topics and events synonymously, though obviously this identity does not hold true in general.

Each topic is associated with two types of data: the set of queries users may issue when searching for information about that topic, Q_T, and the set of documents which may satisfy users searching for information about that topic, R_T [Carmel et al. 2006].

Given a user, u, we claim that $P(T|u)$, that is, the probability of the user's interest in T, and the specific aspects a_i of T which will be of interest to u depend heavily on different properties of u, for example, $P(a_i|u)$. These properties include a user's demographic information (age, gender, etc), social parameters (circle of friends, acquaintances, and work associates), geographic parameters (location vis-a-vis topic location), as well as other parameters (e.g., time of day, exposure to different media, etc).

In this article we are concerned with two main attributes of users: their location and social parameters. Specifically, the first property we investigate is $d_{\text{geo}}(u, T)$, which is the geographic distance between the user and epicenter of the event. We term this property the geographic detachment of a user from an event. We hypothesize that $P(T|u)$ is negatively correlated with $d_{\text{geo}}(u, T)$.

The second property we investigate is $a_{\text{soc}}(u, T)$, which is the social affinity between the user and those users geographically close to the epicenter of an event. We term this property social attachment. A simple way of thinking about social affinity is the number of friends or family close to the event epicenter. We hypothesize that $P(T|u)$ is positively correlated with $a_{\text{soc}}(u, T)$.

Note that both geographic distance and social affinity cannot be defined for every topic. However, for any event which has a clear geographic epicenter, these detachment levels can be defined and measured.

4. EXPERIMENTAL SETUP

4.1. Case Studies

In this article we analyze ten events, seven of which were man made and are briefly described next.

(1) *San Bruno event*. On September 9, 2010, at 6:11PM local time, a large pipe carrying natural gas exploded in the city of San Bruno (near San Francisco), California. The explosion was registered as a 1.1 magnitude earthquake on the Richter scale. Eight people died as a result of the explosion and 38 houses were destroyed.
(2) *New York storm*. A violent storm passed through New York City on September 16, 2010, hitting the boroughs of Queens, Brooklyn, and Staten Island. The storm reached tornado-level status in Flushing Meadows (part of Queens), leaving one person dead and causing widespread damage to property.
(3) *Alaska elections*. The elections to the U.S Senate were held on November 2, 2010. The election in Alaska, although part of the electoral process in 33 other U.S states, drew significant attention because of a three-way race, which included one candidate who lost the primary elections of her party but decided to participate in the elections as an independent candidate. She went on to win the elections.
(4) *Chicago blackouts*. Chicago suffered from widespread blackouts following a storm which hit on June 22, 2011.
(5) *Last shuttle landing*. The space shuttle Atlantis landed for the last time in Merritt Island, Florida, on July 21, 2011, after which the space shuttle was shut down.
(6) *Fatal shooting in Texas*. A gunman shot and killed six people during a birthday party in Grand Prairie, Texas, during a birthday party at a skating rink.
(7) *Indiana stage collapse*. The stage at the Indiana State Fair collapsed during a performance on August 15, 2011 when the above-stage fly system was hit by a high-velocity wind gust in front of a severe thunderstorm, killing 7 people and injuring 43.
(8) *Virginia earthquake*. A 5.8 magnitude earthquake hit Louisa County, Virginia, on August 23, 2011. This earthquake caused damages to property as far away as Washington D.C. and New York.
(9) *Austin fires*. Austin, Texas, suffered from major forest fires beginning on June 6, 2011. These fires resulted in 700 houses being burnt.
(10) *Fatal shooting in Nevada*. A gunman shot and killed four people at a restaurant in Carson City, Nevada, on June 6, 2011.

These events were chosen because they were physically localized in their scope and thus have a clear epicenter. Therefore, social and physical detachment 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 determined a radius of influence for each event based on the radius of the community affected. For example, in the case of San Bruno, this was set to 5 km, which is roughly the area directly disrupted by the explosion. In New York, we set this radius to 30 km to include all the neighborhoods hit by the storm. The radius for Alaska was set to 500 km to include most of Alaska's population but not any other state. See Section 5.3 on automatically setting this radius.

4.2. Data

We used two types of data in our study. First, we extracted query-log data of the Yahoo search engine from several days before the event (one day for all events except the Alaska elections event for which we extracted ten days of data prior to the event), until eight days after the event. For each query, we extracted its text, time, a unique identifier of the user who posted it, the results displayed to the user, and the pages he or she selected to view as a result of the query.

The query log was parsed to identify those queries which were likely relevant (i.e., describing an information need pertaining) to the event using a term-matching scheme. We did this by manually generating a list of keywords for each event. The keywords were drawn from several categories: 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 to remove irrelevant queries. 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. The list of keywords for each of the events is given in Table I. We encode data at the granularity of a day in order to remove diurnal effects.

Each user was represented by two measures of separation from the event: a physical separation, measured by their physical distance, where a larger distance indicates a larger separation, and a social separation, measured by the number of contacts local to the event. For the latter, consider an event influencing a radius R around its epicenter L. The social separation for the ith user with a set of N_i contacts $\{U_j\}_{j=1}^{N_i}$ is defined as $\left|U_j \ s.t. \ d\left(U_j, L\right) \leq R\right|$, where $d\left(U_j, L\right)$ is the physical distance between user U_j and the event epicenter. Thus, a larger number indicates smaller separation.

We used the zip code given by users at the time of registration with Yahoo to identify their approximate location. An alternative way to determine location could have been through users' IP addresses, but as some users use proxy servers, it is not clear that this would be a superior way of measuring location. We computed the physical distance of each user to the event using the Haversine formula [Gellert and Hellwich 1989].

Finally, we used the list of contacts in the Yahoo Instant Messenger (IM) application as a proxy for users' social networks. The number of contacts per person is power-law distributed $\left(\alpha = -0.99, R^2 = 0.93\right)$, with a median of six connections per person and an average of 22.7 connections.[2] The list of contacts was used to determine the number of contacts each user had in the area of the event, which is defined as the number of contacts inside the radius of influence of each event. Thus, each query is described by a tuple of its text, time, and date, and the physical and social separation of the user who posted it from the event. Table II provides statistics on the number of queries related to each event which were posted during the days we observed the query log, the number of unique relevant queries, and the number of users which posted these queries.

We quantified the daily media volume of news outlets related to an event by counting the number of document found on Yahoo News[3] every day, which contained all the words used in any of the 50 most popular queries identified using the procedures

[2]Note that users who were not using the IM application had, by definition, zero connections. Therefore, the average and median number of connections is higher than what would have been obtained using the entire population

[3]www.news.yahoo.com.

Table I. Event Keywords

Event	Category 1	Category 2	Category 3	Category 4	Exclude
San Bruno explosion	bruno, california, sf francisco	fire, explosion, crash, gas, gasoline, pg&e, pge			mar, mtv, lyric, lady polyethylen, toni, buddi
New York storm	tornado, storm	ny, nyc, flushing, newyork york, liberty, jersey staten, brooklyn, queens			
Alaska elections	murkowski, miller, senator, mcadams	alaska, poll, campaign, caucus, caucus, lisa, palin, race, winning, democrats, tea, voting, republican			all other states
Indiana fair	indiana, fair, fairgrounds	collapse	bareilles, stage sugarland		
Shuttle landing	shuttle, sts-135, atlantis	nasa, kennedy	landing, final	ferguson, hurley, walheim, magnus	
Chicago blackouts	chicago, downers grove, prospect	coned, commonwealth, edison	storm, tornado	electricity, blackout, brownout, power	
Virginia earthquake	earthquake, aftershock	virginia, delaware, jersey, york, washington, pennsy, richmond, louisa			
Texas shooting	roller, prairie, texas	shooting, gunman, killing			
Austin fires	forest, firefighter fire, wildfire,	austin, texas, bastrop, montgomery, grimes travis, steiner			
Carson City shooting	gunman	carson, nevada, ihop pancakes, reno	shot, kill	guard	

outlined in Section 4. Similarly, the volume of social media was measured via the number of Twitter messages which contained the words used in these queries.

5. EMPIRICAL ANALYSIS

5.1. Measuring the Effect of Detachment on Interest in an Event

We begin our analysis by showing how the probability of a user's interest in topics related to the event of interest depend on his geographic and social affinity, that is, we measure $P(T|u)$.

Table II. Query Log Statistics for the Analyzed Events

	Number of relevant queries	Number of unique relevant queries	Number of users posting relevant queries
San Bruno explosion	194,184	45,843	79,134
New York storm	54,678	4,406	34,069
Alaska elections	281,851	23,921	183,258
Indiana fair	510,064	5,479	63,116
Shuttle landing	15,418	2,665	6,845
Chicago blackouts	19,320	3,637	5,427
Virginia earthquake	309,274	11,772	74,964
Texas shooting	8,143	1,898	4,964
Austin fires	316,378	23,423	100,263
Carson City shooting	1,561	823	875

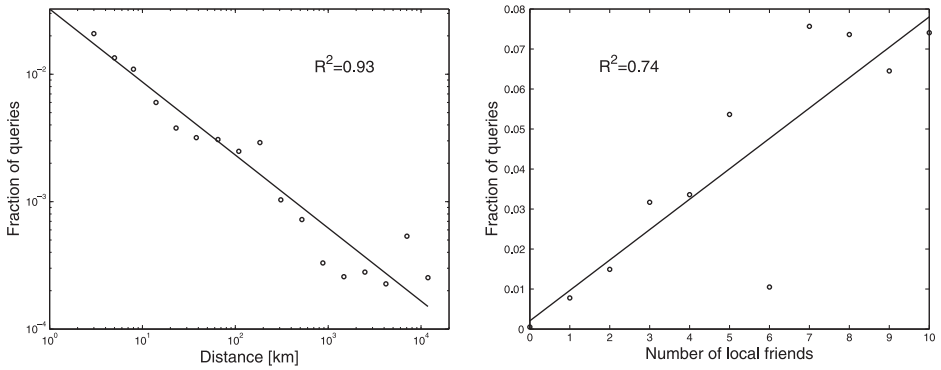

Fig. 2. Fraction of queries as a function of the physical distance (left) and the number of local contacts (right). The regression line on the left is an exponential fit with $R^2 = 0.93$ and a linear fit on the right with $R^2 = 0.74$. Note the logarithmic axes of the left figure.

Recall that we are interested in examining the relationship between $P(T|u)$ and our detachment measures. In our analysis, we binned users into logarithmically spaced bins according to their geographic detachment from the event and used the raw social detachment values (e.g., the number of contacts local to the event). Let $\mathcal{U}_{d_{\text{geo}}}$ be the set of users within a distance $d_{\text{geo}} \pm \delta$ of the event; similarly for social affinity. Thus Q_u denotes the queries of users within that affinity range. We then, for geographic distance, computed the correlation with $P\left(T|\mathcal{U}_{d_{\text{geo}}}\right)$, that is, the probability of group interest in T; similarly for social affinity. We define this precisely as the following.

$$P\left(T|\mathcal{U}_{d_{\text{geo}}}\right) = \frac{\sum_{u \in \mathcal{U}_{d_{\text{geo}}}} |Q_u \cap Q_T|}{\sum_{u \in \mathcal{U}_{d_{\text{geo}}}} |Q_u|}. \tag{1}$$

Figure 2 shows $P(T|\mathcal{U}_{d_{\text{geo}}})$ related to the San Bruno event as a function of d_{geo} and as a function of a_{soc}. These figures show that interest is correlated with physical and social affinity. Indeed, the fraction of event-related queries decays exponentially as physical detachment weakens; linearly so in the case of social affinity.

Table III shows the correlations between d_{geo} with $P\left(T|\mathcal{U}_{d_{\text{geo}}}\right)$ for all events and similarly for social affinity. We tested the variables for normality using the

Table III. Correlation between Physical and Social Detachment between Binned Physical Detachment and Social Affinity with the Fraction of Queries for All Events

	Spearman correlation between the physical detachment and social affinity	Correlation (R^2) between the fraction of relevant queries and:		Slope of the correlation between the fraction of relevant queries and:	
		physical detachment	social detachment	physical detachment	social detachment
San Bruno	–0.13	0.93	0.74	–0.57	0.008
New York	–0.18	0.71	0.18	–0.33	0.012
Alaska	–0.06	0.60	0.26	–0.22	$8 \cdot 10^{-5}$
Indiana fair	–0.01	0.54	0.79	–0.23	0.001
Shuttle landing	–0.05	0.16	0.94	–0.19	$2 \cdot 10^{-4}$
Chicago blackouts	–0.27	0.23	0.60	–0.21	$3 \cdot 10^{-5}$
Virginia earthquake	–0.20	0.44	0.80	–0.36	$5 \cdot 10^{-5}$
Texas shooting	–0.18	0.51	0.86	–0.42	$4 \cdot 10^{-4}$
Austin fires	–0.18	0.33	0.82	–0.26	0.002
Carson City shooting	–0.11	0.51	0.80	–0.52	0.002

Exponential regression was used for the physical detachment, and linear regression for social affinity. All regression results are statistically significant at $p < 0.01$.

Jarque-Bera test (with Bonferonni correction) and found that the null hypothesis (data are normally distributed) cannot be rejected at $p < 0.05$. Additionally, we performed a Chi-square lackness-of-fit test (with Bonferroni correction) on the model residuals and found that the null hypothesis (the model fits well) cannot be rejected at $p < 0.05$. Additionally, this table shows the Spearman correlation [Drasgow 1986] between the log-transformed physical distance and the number of local contacts. The correlation of the two detachment methods is significant but not very high. This is to be expected, since people closer to an event can be expected to have more local relationships. While both physical detachment and social affinity are significantly correlated with the fraction of event-related queries, physical detachment seems to be a stronger indicator for interest compared to social affinity. This is in line with several studies (for example, Nitzan and Libai [2010]) who have found that social attributes are secondary in importance to individual traits. Thus, we conclude that the probability that a user will be interested in an event-related topic is highly correlated with social and geographic detachment.

The events we analyzed are characterized by a strong temporally-limited interest of users. Figure 3 shows the fraction of queries related to the San Bruno event as a function of time, partitioned according to the physical distance of users from the event. Day zero marks the day of the explosion. Since the explosion happened in the early hours of the evening, significant interest in the event starts on Day 1 and peaks only on Day 2. Interest in the event decays rapidly and could have been faster if not for a spike on Day 5, which is likely related to new findings about the event being reported in the media. Indeed, the volume of queries decreases to below 25% of its maximum after 3.1 days, averaged across the ten events. (See details for each event in Table IV). Strikingly, as the figure shows, interest in the San Bruno event decays in an identical manner for users, independent of whether they were physically close to the event or far from it ($R^2 = 0.95$ between the two time series), though a logical hypothesis would have been that closer users would retain interest in the event for a longer time period. A similar image appears when considering users with no local contacts versus those with many such contacts.

Table IV. Number of Days it Took for the Volume of Queries to Decrease Below 25% of its Maximum

Event	Decay time [days]
San Bruno explosion	3
New York storm	4
Alaska elections	5
Indiana fair	4
Shuttle landing	3
Chicago blackouts	2
Virginia earthquake	1
Texas shooting	4
Austin fires	3
Carson City shooting	2

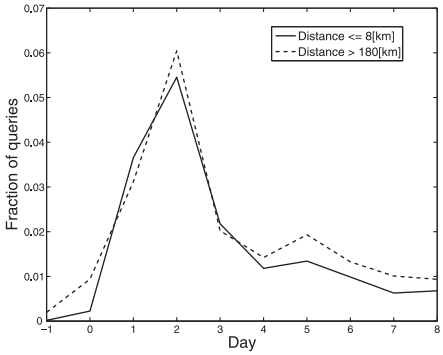

Fig. 3. Fraction of event-related queries as a function of time in the San Bruno event, partitioned by the physical distance of users from the event. The fraction of queries for users far from the event was multiplied 50 times to be on the same scale as users close to the event. Day zero denotes the day of the explosion.

5.2. Measuring the Effect of Detachment on Event Aspects

The fact that users post queries related to an event is, as we previously demonstrated, significantly tied to their detachment from that event. However, queries may have been due to different aspects of the topic. Indeed, our hypothesis is that the distribution of aspects is dependent on a user's relationship to the event. In this section, we identify different aspects of the event-related topics and show how differences in the degree of detachment influence the need for information.

In order to identify which words most typify how users describe the event based on the social and physical attachment, we defined those users living within under 10 km as physically close to the event and those users living over 500 km from the event as physically far. Similarly, we defined those users who had two or more contacts in the area of the event as being socially close, whereas those users with no contacts in the area of the event as socially far. We created a model for the users of each combination of social and physical distance. This model is a count of the words used by queries of each population. Therefore, for each word w_i, we denote the model as $C_j(i) = \|w_i\| \;\; j = 1, .., 4$.

For each model, we found the closest of the three other models using cosine similarity and computed which words occur in it with a probability of 20% or more than in the closest model. For example, if model j was the closest model to model i, we identified those words where

$$\frac{C_i / \sum (C_i)}{C_j / \sum (C_j)} > 0.2. \tag{2}$$

Table V. Words Used to Describe the San Bruno Event, Depending on the Level of Attachment

		Physical distance	
		Close	Far
Social Distance	Close	News, line, San, map	Video, natural, pipeline, San, map, news, fire, line, CA
	Far	San, news, fire, CA, map, Bruno	Gas, explosion, California

Table V shows these words for each of the four combinations in the San Bruno event. Evidently, users who were physically close to the event used similar words to search for it. However, users who were physically far but socially close tended to look for news and images from the event, whereas those who were far from the event in both measures used only the general terms to describe it.

Similar phenomena were observed in the other events; for example, in the New York event, users physically close used the words "Queens," "Picture," "Storm," "City," and "NYC" to search for it. Users who were physically far but socially close used the words "Brooklyn," "NY," "2010," "New," and "York," and users far by both measures used the words "City," "NYC," "Statue," "Liberty," and "Storm," evidently because they were interested in images of the Statue of Liberty during the storm.

The conclusion of this analysis is that users close to the event use similar words to express their information need, while those farther away can be differentiated in their information need according to their social attachment to people in the area of the event.

In order to analyze the distribution of aspects, $P(a_i|u)$, we need to first enumerate the set of aspects for each topic. Clusters of queries have been shown in the past to represent different information needs [Carmel et al. 2006; 2008]. Therefore, we clustered the queries to identify the different aspects in our data (separately for each event). After stemming and stop-word removal, each query was represented by its TF-IDF model. The queries were then clustered using the k-means algorithm with a cosine similarity measure. The number of clusters was determined by running the algorithm five times using an increasing number of clusters and using the largest number of clusters which did not generate singleton clusters during any of the five runs. This resulted in four clusters for the San Bruno, New York, Carson City, and Indiana fair events and three clusters for all other events. We note that we did not use measures of detachment during clustering.

Although the clusters were generated using only the text of the query, we found that they created a partition which is statistically significant (Kruskal-Wallis [Upton and Cook 2002], $p < 10^{-3}$ for all events) for both social affinity and physical detachment. This suggests that the textual description of an event (or indeed, the aspects of an event sought by different users) is correlated with the detachment level of users.

Figure 4 shows the number of queries from each cluster posted on each day and the relationship of social and physical detachment on the fraction of queries from each cluster for each of the events. This figure shows that some clusters have a high preferential physical detachment, for example, Cluster 3 in the San Bruno event (Speaman correlation, $\rho = -0.88$), while others, such as Cluster 4 of the San Bruno event, have a more pronounced social affinity (Speaman correlation, $\rho = 0.92$). Some of the clusters exhibit weak correlations with detachment levels. These may be a manifestation of an aspect which is commonly queried across geographic or social radii. Interestingly, different clusters have different temporal patterns. For example, whereas Cluster 1 of the Alaska event has a peak lasting for two days, Cluster 2 has a spike lasting only one day.

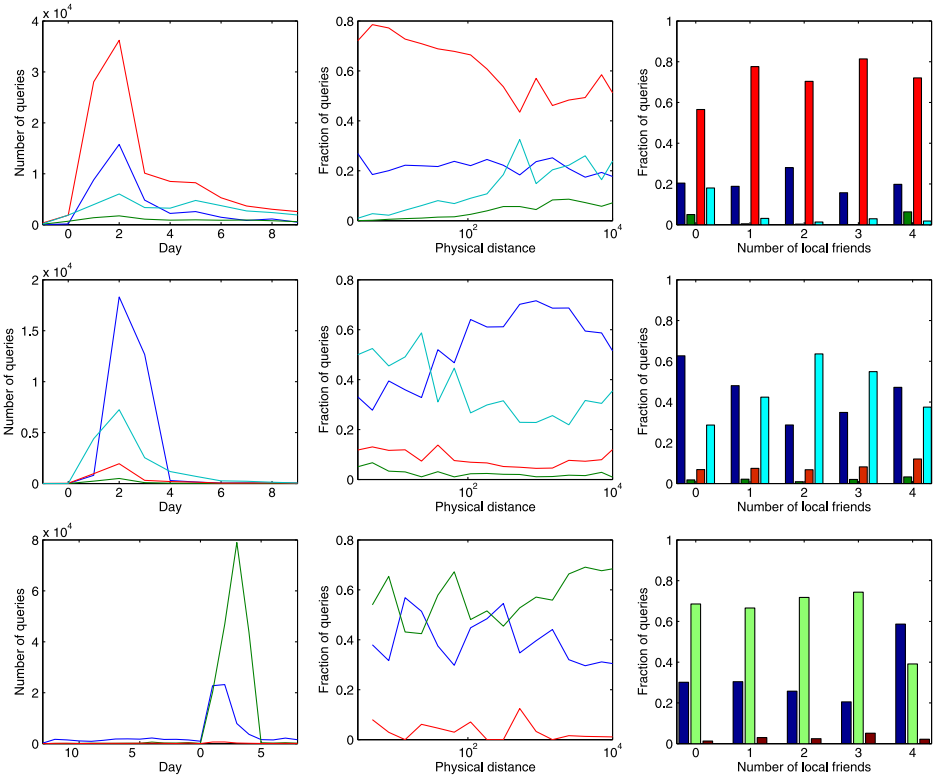

Fig. 4. The number of queries from each cluster as a function of time and the fraction of queries from each cluster as a function of physical and social affinity. Colors denote different clusters: Cluster 1 (blue), Cluster 2 (green), Cluster 3 (red), and Cluster 4 (cyan). Each row presents the data from one event, which are, from top to bottom, San Bruno, New York, and Alaska.

Therefore, our results indicate that users seek different aspects of information regarding events and that their interests strongly correlate with their social and geographic detachment. These aspects have different temporal profiles and textual manifestations.

5.3. Automatically Setting the Radius of Influence

Setting the number of local users for finding those who may be influenced by an event was performed in this study using a simple heuristic, that of finding all the people in a radius enclosing the urban population which may have been affected by the event. However, if we aim to automate the improvement in search result, the finding of this radius needs to be automated.

One way to estimate the radius of influence is derived in Backstrom et al. [2008]. However, in that study, only geographic location was taken into account as influencing the radius of influence. Furthermore, the frequency of queries was posited to decay exponentially as a function of distance. However, our observations are that because of the location of population centers in relation to the location of an event, the decay in the frequency of queries does not follow a simple exponential.

Bernard et al. [1989] discussed the question of estimating the number of people who died in an earthquake (e) from an entire population (t) based on sampling a population and determining how many people they knew first-hand who had died (p), as well as

the average number of acquaintances each person has (c). This was shown to be lower-bounded by the following.

$$n \approx t \cdot \left(1 - (1-p)^{1/c}\right) \tag{3}$$

Applying this method to our work, the radius of influence can be set to contain the estimated number of affected users given the other parameters, that is setting n to be the number of local people affected by an event. Similarly, t is the entire (U.S.) population, p is the fraction of users who posted queries related to the event, and c is the average number of connections in the social network of people (which is approximately 22.7, as reported in Section 4). Unfortunately, p is only a very rough estimate, since some people will not search for information regarding an event even if they have acquaintances in the area, either because they are not interested in it or because they use other means to obtain relevant information.

The U.S. population in 2010 was 307.2 million people [Bureau 2009]. Yahoo IM users comprised of approximately 40% of the IM usage population.[4] Therefore, we set $t = 123M$.

Given this information, we can set the radius of influence to contain approximately n users. Alternatively, we now proceed to computing the expected number of local users e in the events analyzed and compare this number to the one used by the heuristically set radii.

We computed the parameter p by counting the fraction of users who posted information about the event, divided by the total number of active users during the time period examined.

Our findings indicate that the computed number is always smaller than the actual number of local users. The correlation between the estimated and actual local users is of medium quality ($\rho = 0.21$, Spearman). This likely is due both to the relatively rough approximation method, as well as to the fact that many local users we identified are not active users, that is, they include users who registered but are no longer using the service.

Nevertheless, our results indicate that it is possible to set the radius automatically using the previously mentioned computation, though further studies are required in order to find more accurate approximation methods.

6. APPLICATIONS

6.1. Detecting Queries Related to an Event

Results in previous sections suggest that $P(T|u)$ is strongly related to $d_{\text{geo}}(u, T)$ and $a_{\text{soc}}(u, T)$. Therefore, in this section we use this finding to identify event-related queries. In Section 4, we identified queries relevant to the event using a term-matching method. However, while this method is easy to use in practice, it is likely to miss some queries that used phrases which a human annotator did not consider when generating the list of keywords for term matching. In this section, we show how profiles created using detachment information can be used to identify queries seemingly relevant to the event but which are not in our seed set.

As an example for seemingly unrelated queries, consider the query "PG&E" (Pacific Gas and Electric), which is the name of a local gas company in San Bruno. This query is submitted on regular days, but during the San Bruno event, it was posted with much higher frequency, as shown in Figure 5. It also exhibited a very different distance profile. Detecting these types of queries can be useful for understanding users'

[4]http://www.businessinsider.com/chart-of-the-day-instant-messenger-services-2010-8

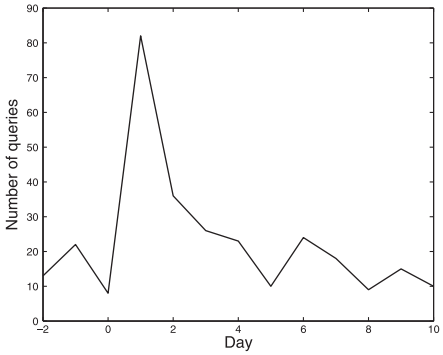

Fig. 5. Number of times the query "PG&E" appeared during the San Bruno event.

information needs, improving retrieval [Dong et al. 2010], or aggregating news content [Diaz 2009].

As previously noted, each query is represented by its text as well as by the time it was submitted, and the physical and social detachment of the user who submitted the query. For popular queries, we can represent each query as the distribution of these three parameters (time, physical attachment, and social attachment). Since in some case there may not be sufficient data to estimate the entire distribution with enough confidence, it may be necessary to resort to sufficient statistics.

Our analysis focused on popular queries, which we defined as those queries which appeared at least 100 times during the event interval. We represented each such query by the number of times it appeared on each day and by the average distance of users who posted this query on that day. This use of averages rather than a distribution allowed us to use queries which appeared relatively few times in the data, while maintaining high confidence in the estimation of the query distribution. Approximately 1 million queries appeared at least 100 times for each of the events.

We compared the representation of the popular queries to that of the clusters of queries identified by term matching (see Section 5.2 for details of the clustering process). The appearance count was compared using linear correlation, and the average distances using Euclidian distance. A linear predictor was then trained to decide on the weight of each parameter (correlation and distance) in the final scoring of each popular query. Positive (event-related) queries were replicated 100 times to mitigate the effect of their sparseness in the data (approximately 0.5% of queries were event related).

In our experiments, we used fivefold cross validation to reduce the chance of overfitting. After finding the appropriate weight for each parameter, we measured how well the term-matched queries were identified by the method. The use of term-matched queries as the target queries means our results should be considered an underestimate of the true result.

Term-matched queries are rare in the most popular queries (in the range of 0.1%–0.3%). We therefore report two measures of success in classification: average precision and lift. For any given fraction of the queries $0 < f < 1$, lift [Richter et al. 2010] is defined as the ratio between the number of event-related queries among the fraction of T queries that are ranked highest by the proposed system, and the expected number of event-related queries in a random sample from the general query pool of equal size. For example, a lift of three at a fraction T = 0.01 means that if we scan the 1% of queries ranked highest by the proposed system, we expect to see three times more event-related queries in this set of queries than in a 0.01-fraction random

Table VI. Average Precision and Maximum Lift Obtained in Identifying Event-Related Queries Using Appearance Counts Per Day (Count) and When Additionally Using Physical and Social Detachment Profiles

	Average Precision				Lift			
	Count	Physical	Social	All	Count	Physical	Social	All
San Bruno	0.0185	0.0238	0.0119	0.0143	33.7	43.9	50.0	48.0
New York	0.0170	0.0255	0.0117	0.0154	24.5	25.2	25.2	25.2
Alaska	0.0058	0.0051	0.0054	0.0052	11.1	10.4	11.3	10.4
Indiana fair	0.0099	0.0115	0.0150	0.0262	25.0	21.9	21.9	34.4
Shuttle landing	0.0014	0.0015	0.0007	0.0008	15.0	15.0	8.0	10.0
Chicago blackouts	0.0003	0.0001	0.0001	0.0001	5.3	2.9	2.9	2.9
Virginia earthquake	0.0156	0.0174	0.0172	0.0179	14.7	21.8	21.0	21.5
Texas shooting	0.0002	0.0004	0.0009	0.0017	2.2	4.4	13.3	20.0
Austin fires	0.0370	0.1024	0.0527	0.1978	21.7	40.2	34.9	50.7
Average improvement over count profile		59%	2%	9%		23%	10%	9%
Average rank	2.5	2.9	2.2	2.3	1.8	2.3	2.9	2.9

Note: The column denoted by "All" represents results when using appearance counts, physical detachment, and social detachment information. The Carson City shooting event is excluded because no relevant queries appeared with sufficient frequencies.

sample of the queries. Lift measures the precision at a given threshold [Caruana and Niculescu-Mizil 2004], albeit scaled such that it can be larger than one.

Table VI shows the average precision and lift obtained by the system when attempting to identify term-matched queries in the popular queries. The lift is given for 1% of the queries, with and without the use of the physical and social detachment measures. The proposed method accurately identifies the term-matched relevant queries and improves lift and average precision by 18% over count profiles. This is very useful both for improving query routing and for human annotators to be able to understand the kind of phrases users make use of when searching for information about the event.

The detachment profiles improve results. On average, the physical detachment information improved retrieval the most (59% for average precision and 23% in lift) over query counts alone, followed by social detachment and both profiles. However, we could not demonstrate that these differences are statistically significant (Friedman test [Demsar 2006]). Using both social and geographic detachment reduced precision. We attribute this to the sparsity of the positive examples, which accounted for around 0.5% of the data.

Anecdotally, when not using the distance profile, the queries ranked high by the predictor, and which are unrelated to the event, are mostly associated with other news events which peaked in interest around the same time as the event of interest. This is consistent with the use of such features when detecting newsworthy queries in general rather than for a specific topic [Diaz 2009].

In the case of the San Bruno event, we scanned the highest-ranked queries and found that (excluding the term-matched queries) these included mostly names of local news channels and event-related queries which contained partially entered words.

Identifying relevant queries in retrospect (i.e., using the query profile over the entire time span of the event) is useful for some applications. However, in most cases, one would like to identify event-related queries soon after the event begins. Figure 6 shows the lift obtained by our method when considering the data as it is collected. In this experiment, feature vectors for day n were constructed using the data from days –1 to day n (inclusive). As this figure shows, even after only one day of the event, reasonable lift is obtained by our method. This indicates that our method can assist in pinpointing event-related queries soon after they begin.

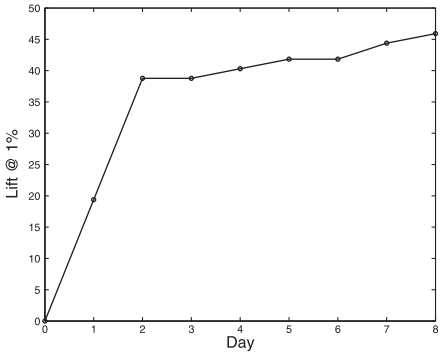

Fig. 6. Lift at 1% obtained by using the data from days 0 to N for the San Bruno event.

6.2. Retrieval for Queries Related to Events

In Section 5.2, we demonstrated that detachment affects the aspect of the topic which is sought by users. Therefore, we hypothesize that by using the detachment information of users, it should be possible to improve retrieval. Our goal is to identify which pages are likely to be of interest to a user given his specific detachment information.

One way to use detachment information is to augment queries with attachment data and use a ranking algorithm [Nallapati 2004] to find an improved page ranking. While this solution is possible, it should be noted that the distribution of user detachments is far from uniform (there are many more users who are physically far from any event than those close to it). Therefore, a large number of data points as well as care is required when training such a model to provide sufficient weight to examples of sparsely represented detachment values.

An alternative solution is to partition the data according to detachment metrics and train a different ranking model for each detachment partition. For example, one model could be trained for those users who are 10 km or less from the event, and another model could be trained for those users who are located 10 km or more from the event. In the following, we take this approach and build models according to the physical detachment level and/or the social detachment level of users.

We randomly chose approximately 16,000 queries for each of the events. For each of the queries, we extracted the pages presented to users as the search engine results, as well as whether the user clicked each page. A page was considered of interest if it was clicked by the user [Joachims 2002]. We note that our analysis, which was performed after the event, suffers from a certain selection bias by users, because the initial ranking of documents was generated by the search engine over which we had no control. Each page was represented by its rank on the results page (which is a proxy for its perceived importance as estimated by the search engine), and the second-level URL information, for example, the address `http://news.yahoo.com`, would be represented as *news*. This information is important because second-level domains are usually indicative of the type of information on the website and are especially useful when distinguishing news and non-news websites.

Users were partitioned into five groups according to their physical distance from the event using logarithmical-spaced bins or to one of three social attachment groups: no local contacts, one local contact, and more than one such contact. We trained one set of models for each detachment type separately as well as one for the combination of the detachment levels. As a baseline, we use the original order of the documents as presented to the user. This baseline is highly optimized for ranking documents for late-breaking Web search results. We note, though, that the features used in

Table VII. Improvement in Retrieval Results When Using Detachment Information

	MAP				P@5				P@10			
	Both	Physical	Social	No	Both	Physical	Social	No	Both	Physical	Social	No
San Bruno	0.693	0.692	0.716	0.607	0.182	0.181	0.185	0.164	0.234	0.234	0.236	0.228
New York	0.804	0.794	0.788	0.764	0.128	0.128	0.128	0.128	0.148	0.148	0.148	0.145
Alaska	0.889	0.886	0.898	0.831	0.146	0.146	0.146	0.144	0.167	0.167	0.168	0.167
Indiana fair	0.389	0.350	0.341	0.294	0.131	0.143	0.123	0.094	0.261	0.274	0.240	0.193
Shuttle landing	0.372	0.380	0.399	0.291	0.120	0.120	0.128	0.082	0.238	0.219	0.229	0.169
Chicago blackouts	0.567	0.527	0.579	0.324	0.072	0.161	0.179	0.090	0.1 00	0.232	0.253	0.184
Virginia earthquake	0.479	0.519	0.534	0.304	0.116	0.125	0.124	0.072	0.164	0.175	0.177	0.143
Texas shooting	0.507	0.486	0.509	0.335	0.158	0.142	0.144	0.091	0.228	0.208	0.207	0.177
Austin fires	0.338	0.375	0.355	0.302	0.095	0.105	0.095	0.080	0.150	0.179	0.168	0.154
Carson City shooting	0.509	0.410	0.452	0.316	0.118	0.097	0.117	0.078	0.165	0.138	0.170	0.146
Average improvement over no detachment	34%	31%	35%		29%	38%	40%		9%	15%	17%	
Average rank	4.0	1.9	1.8	2.3	3.3	2.3	2.2	2.3	2.7	2.8	2.1	2.2

Note: Columns denoted by "both" mark models which used both physical and social detachment information. Columns marked by "physical" and "social" denote cases where physical or social detachment information was used, respectively. In all cases and metrics, the use of detachment information improves retrieval in a statistically significant manner ($p < 10^{-2}$, sign test) over the baseline. The differences in MAP when using social, physical, and both measures are pairwise statistically significant ($p < 10^{-2}$, sign test) except for the use of physical versus both physical and social information. The differences in P@5 and P@10 when using social, physical, and both measures were not pairwise statistically significant.

this model do not capture detachment information. All else being equal, since this model was trained using more data, it could be expected to perform better than the partitioned model. The model used for identifying documents of interest was a decision tree classifier. We used fivefold cross validation in our experimentation and measured the mean average precision (MAP) and the average precision at the top five and top ten documents (denoted as P@5 and P@10, respectively).

Table VII shows the results of our experiments. The models which use detachment information obtained a statistically significant improvement over the baseline (Friedman test, $p < 0.01$ [Demsar 2006]). Separately, using physical detachment gave a larger average improvement over the baseline than social detachment for MAP, but not for precision. Interestingly, using both detachment parameters resulted in inferior improvements over either detachment levels alone. We attribute this to the correlation between the social and physical information, as well as to sparseness of the training data for some models.

We did not find a trend which suggests that the proposed method improves (or decreases) retrieval performance as a function of detachment levels. Rather, gain is distributed across all detachment levels.

Unsurprisingly, the most indicative attributes of models built for users closer to the event featured local news outlets (newspapers and TV channels), whereas for users far from the event, these included mostly national news channels and Web aggregators, such as Yahoo News.

We conclude that the use of detachment information is beneficial in improving retrieval but that specific measures of interest should be optimized using different detachment measures.

7. DISCUSSION

In this article, we investigated the effect of social and geographic detachment of users on querying behavior. We found that the level of detachment has a significant effect on both the level of interest and the type of interest that users have in an event.

The aspects we identified were limited by the fact that only query text was used as their input. In future work, we intend to use more robust clustering and similarity measures [Li et al. 2008; Metzler et al. 2007; Sahami and Heilman 2006]. This will allow us to better understand the semantics of aspects, especially as they are understood by the users.

The improvement attained by using social detachment in both applications was markedly better than that of geographic detachment when measured by high-precision metrics (lift, P@5, and P@10). Geographic detachment was better in terms of average precision. We speculate that this is due to the relative sparsity of nonzero social detachment compared to geographic data as well as to the importance of social connections.

Although our sample is limited, our results indicate that highly localized events, such as the San Bruno event, tend to be explained better by attachments than, events which take place over wider areas, such as the other two events. Further research of additional events is required in order to generalize these findings.

An interesting question which arises from our work is the relationship between media (social and commercial) and users' information searching, as manifested by the query log. The temporal behavior is highly correlated and includes the appearance of idiosyncratic peaks, such as those on day 4 of the San Bruno event. Furthermore, as Figure 1 shows, interest in the Alaska elections is almost entirely limited to dates after the elections themselves. This is surprising because it indicates that interest in these elections is almost solely related to the results of the elections and not to candidates' work or people's need for information about voting procedures, etc, prior to the elections. This may hint that query log volume is driven, to some extent, by media interest, but further experimentation is required to validate this link.

REFERENCES

Backstrom, L., Kleinberg, J., Kumar, R., and Novak, J. 2008. Spatial variation in search engine queries. In *Proceeding of the 17th World Wide Web Conference (WWW'08)*. ACM, 357–366.

Bernard, H. R., Johnsen, E. C., Killworth, P. D., and Robinson, S. 1989. Estimating the size of an average personal network and of an event subpopulation. In *The Small world*, M. Kochen, Ed. 159–175.

Bernard, H. R., Johnsen, E. C., Killworth, P. D., and Robinson, S. 1991. Estimating the size of an average personal network and of an event subpopulation: Some empirical results. *Soc. Sci. Res. 20*, 109–121.

Bernard, H. R., Killworth, P. D., Johnsen, E. C., Shelley, G. A., and McCarty, C. 2001. Estimating the ripple effect of a disaster. *Connections 24,* 2, 18–22.

Carmel, D., Yom-Tov, E., Darlow, A., and Pelleg, D. 2006. What makes a query difficult? In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'06)*. ACM, 390–397.

Carmel, D., Yom-Tov, E., and Roitman, H. 2008. Enhancing digital libraries using missing content analysis. In *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'08)*. ACM, 1–10.

Carmel, D., Zwerdling, N., Guy, I., Ofek-Koifman, S., Har'el, N., Ronen, I., Uziel, E., Yogev, S., and Chernov, S. 2009. Personalized social search based on the user's social network. In *Proceeding of the 18th ACM Conference on Information and Knowledge Management (CIKM'09)*. 1227–1236.

Caruana, R. and Niculescu-Mizil, A. 2004. Data mining in metric space: An empirical analysis of supervised learning performance criteria. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'04)*. 69–78.

Chang, T.-K., Shoemaker, P. J., and Brendlinger, N. 1987. Determinants of international news coverage in the U.S. media. *Commun. Res. 14,* 4, 396–414.

Demsar, J. 2006. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res. 7*, 1–30.

Diaz, F. 2009. Integration of news content into web results. In *Proceedings of the 2nd ACM International Conference on Web Search and Data Mining*.

Dong, A., Chang, Y., Zheng, Z., Mishne, G., Bai, J., Zhang, R., Buchner, K., Liao, C., and Diaz, F. 2010. Towards recency ranking in web search. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM'10)*. 11–20.

Dong, A., Zhang, R., Kolari, P., Bai, J., Diaz, F., Chang, Y., Zheng, Z., and Zha, H. 2010. Time is of the essence: improving recency ranking using Twitter data. In *Proceedings of the 19th World Wide Web Conference (WWW'10)*. 331–340.

Drasgow, F. 1986. Polychoric and polyserial correlations. In *The Encyclopedia of Statistics, Volume 7*, S. Kotz and N. Johnson Eds. Wiley, Hoboken, NJ. 68–74.

Gellert, G. and Hellwich, K. 1989. *The VNR Concise Encyclopedia of Mathematics,* 2nd Ed. Van Nostrand Reinhold.

Hassan, A., Jones, R., and Diaz, F. 2009. A case study of using geographic cues to predict query news intent. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS'09)*. 33–41.

Joachims, T. 2002. Unbiased evaluation of retrieval quality using clickthrough data. In *Proceedings of the SIGIR Workshop on Mathematical/Formal Methods in Information Retrieval*.

Jones, R., Hassan, A., and Diaz, F. 2008. Geographic features in web search retrieval. In *Proceeding of the 2nd International Workshop on Geographic Information Retrieval (GIR'08)*. 57–58.

Kim, J. 2009. Describing and predicting information-seeking behavior on the web. *J. Amer. Soc. Inf. Sci. Technol. 60,* 4, 679–693.

Leskovec, J., Backstrom, L., and Kleinberg, J. 2009. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Li, X., Wang, Y.-Y., and Acero, A. 2008. Learning query intent from regularized click graphs. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'08)*. ACM, 339–346.

Martzoukou, K. 2005. A review of web information seeking research: Considerations of method and foci of interest. *Inf. Res. 10,* 2.

Mendoza, M., Poblete, B., and Castillo, C. 2010. Twitter under crisis: Can we trust what we RT? In *Proceedings of the ACM SIGKDD Workshop on Social Media Analytics (SOMA)*.

Metzler, D., Dumais, S. T., and Meek, C. 2007. Similarity measures for short segments of text. In *Proceedings of the European Conference on Information Retrieval*. 16–27.

Nallapati, R. 2004. Discriminative models for information retrieval. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'04)*. 64–71.

Nitzan, I. and Libai, B. 2010. Social effects on customer retention. *Marketing Science Institute Working Paper*, 10–107.

Palen, L., Vieweg, S., Liu, S. B., and Hughes, A. L. 2009. Crisis in a networked world: Features of computer-mediated communication in the april 16, 2007, Virginia Tech Event. *Soc. Sci. Comput. Rev*.

Richter, Y., Yom-Tov, E., and Slonim, N. 2010. Predicting customer churn in mobile networks through analysis of social groups. In *Proceedings of the SIAM International Conference on Data Mining (SDM'10)*. 732–741.

Sahami, M. and Heilman, T. D. 2006. A web-based kernel function for measuring the similarity of short text snippets. In *Proceedings of the 15th World Wide Web Conference (WWW'06)*. 377–386.

Sakaki, T., Okazaki, M., and Matsuo, Y. 2010. Earthquake shakes Twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th World Wide Web Conference (WWW'10)*. 851–860.

Teevan, J., Morris, M. R., and Bush, S. 2009. Discovering and using groups to improve personalized search. In *Proceedings of the 2nd ACM International Conference on Web Search and Data Mining (WSDM'09)*. 15–24.

Upton, G. and Cook, I. 2002. *Oxford Dictionary of Statistics*. Oxford University Press.

U.S. Census Bureau 2009. 2009 population estimate. `http://quickfacts.census.gov/qfd/states/00000.html`.

Wu, H. D. 1998. Geographic distance and US newspaper coverage of Canada and Mexico. *Int. Commun. Gazette 60,* 3, 253–263.

Received April 2012; revised July, October 2012; accepted November 2012