Out of Sight, Not Out of Mind

On the Effect of Social and Physical Detachment on Information Need

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

The information needs of users and the documents which answer it 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 paper we investigate how the information need of users 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 three major events, we show that social and physical detachment levels of users are a major influence on their information needs, as manifested by 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%.

Categories and Subject Descriptors

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

General Terms

performance,experimentation

Keywords

social,physical,distance,information,need

SIGIR'11, July 24–28, 2011, Beijing, China.

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 seeking of information using web search engines. Their knowledge of the event and its different facets is shaped 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 that have friends or relatives which 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. As Chang *et al.* [10] showed, the distance of an event from the USA is one of the most predictive attributes for its media coverage. Furthermore, Wu [30] 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.

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 graphs[1] of news coverage, social media coverage, and information seeking for three news events in Figure 1. Although the trends, in general, correlate with each other for a given event, the information-seeking behavior of users is sufficiently different from that of media output to motivate a deeper study of information seeking in isolation.

In this paper, we investigate the information need of individuals in response to special events, both adverse and otherwise, as manifested by active information seeking on the web. We investigate this need as a function of users' relationship to the event, specifically their social detachment

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

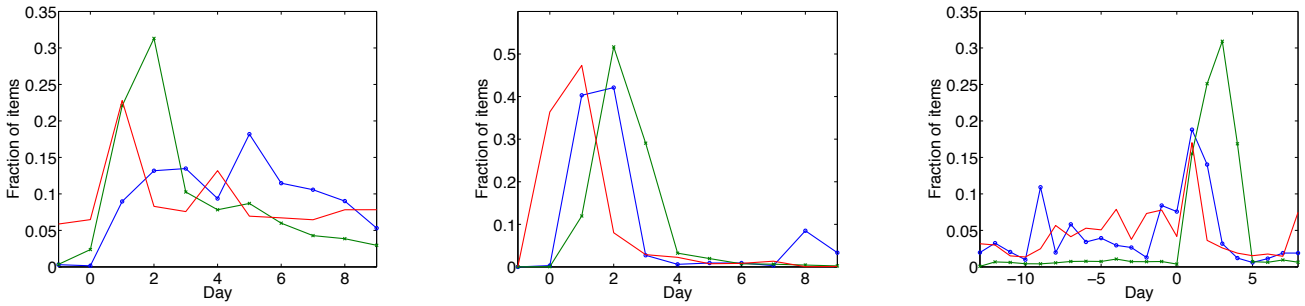

Figure 1: Fraction of mainstream media outlet items (blue, circles), social media items (red) and queries (green, crosses) posted over time in response to the San Bruno (left), New York (middle), and Alaska (right) events. These graphs are best printed in color.

from the event and their geographic distance from it. We show that information need is strongly dependent on social and geographical attachment. We demonstrate that knowledge of social and geographic information can improve retrieval by accurately identifying queries related to an event and consequently retrieving the appropriate documents.

The paper is structured as follows: In Section 2 we survey relevant related work. In Section 3 we enhance a previously-proposed model of information need to account for user characteristics. Section 4 introduces the data we analyzed. In Section 5 we present an analysis of our data and demonstrate the effect of social and geographic detachment levels on information need. Finally, Section 6 demonstrates two information-retrieval applications which benefit from the use of social and geographic detachment.

2. RELATED WORK

Information seeking 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 [27]. 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.*[1] developed a model for pinpointing events based on the location of people querying about them. In this way, both stationary terms (sports teams) and moving events (hurricanes) were pinpointed to within a few kilometers. In both [1] and [27] it was 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 [15]. Similarly, Hassan *et al.* also found that these geographic features improved general web search [17].

The social context of the user has received somewhat less treatment than geography. Carmel *et al.* [8] 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 homophily, rather than social connection, is the source of improvement. Teevan *et al.* found that personalizing at the group scale improved effectiveness for group-related queries [28]. The size of a population which can obtain knowledge of an event through their social network has been studied using both empirical data and theoretic models in [5] and [4]. For example, [5] estimated that between one in 17 and one 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 [24], 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.* [20] 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 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 discussed above).

The temporal analysis of news events also received some attention. For example, Leskovec *et al.* showed that media interest in an event peaks within two days and decays somewhat quicker than it arises [18]. 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 [11]. In the context of web ranking, researchers have found that

a general retrieval model fails for late-breaking news events and specialized query treatment was required to satisfy user intent [12]. This specialized ranker can be further improved by analyzing the discussion of the topic in social media [13].

3. A MODEL OF USER PREFERENCE FOR TOPIC INFORMATION

Define the primal information object as a topic, $T \in \mathcal{T}$, which is information pertinent to a defined subject. Each topic is comprised of two sets: the set of queries users may issue when seeking information about that topic, Q_T, and the set of documents which may satisfy users seeking information about that topic, R_T. Each topic may be partitioned into aspects $\{a_i\} \in T,\ i = 1, 2, \ldots N$, which describe different facets of the topic. Thus, a user may be interested in one or more of a topic's aspects.

Given a user, u, we claim that $P(T|u)$, i.e. 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, e.g. $P(a_i|u)$. These properties include a users' 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 (for example, time of day, exposure to different media, etc).

In this paper 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 as 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 paper we analyze three events, which are briefly described below:

1. **San Bruno event**: On September 9th, 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. The sound of the explosion led some local residents to initially believe that it was an earthquake or an airplane crash. 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 the 16th September 2010, hitting the boroughs of Queens, Brooklyn, and Staten Island. The storm reached tornado-level in Flushing Meadows (part of Queens), leaving one person dead and causing widespread damage to property.

3. **Alaska elections**: The elections to the USA Senate was held on the 2nd November 2010. The election in Alaska, although part of the electoral process in 33 other USA 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.

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. In the case of San Bruno, this was set to 5km, which is roughly the area directly disrupted by the explosion. In New York, we set this radius to 30km, to include all the neighborhoods hit by the storm. The radius for Alaska was set to 500km, to include most of Alaska's population, but not any other state.

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 the first two events, and 10 for the last), until 8 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 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. For example, the list of keywords for the New York event were:

1. **Descriptions**: Tornado, Storm

2. **Location**: New York, NY, NYC, Flushing, Queens, Brooklyn, Statue of Liberty, Staten Island.

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 his physical distance, where a larger distance indicates a larger separation, and a social separation, measured by the number of contacts local to the event, where 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

	Number of relevant queries	Number of unique relevant queries	Number of users posting relevant queries
San Bruno	194,184	45,843	79,134
New York	54,678	4,406	34,069
Alaska	281,851	23,921	183,258

Table 1: Query log statistics for the three events

through users' IP address, 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.

Finally, we used the list of contacts in the Yahoo Instant Messenger (IM) application as a proxy for users' social network. The number of contacts per person is power-law distributed ($\alpha = -0.99$, $R^2 = 0.93$), with a median of 6 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 are 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 4.2 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 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 $P(T|u)$.

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. 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. We then, for geographic distance, computed the correlation with $P(T|\mathcal{U}_{d_{\text{geo}}})$, the probability of group interest in T; similarly for social affinity. We define this precisely as,

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

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, and linearly so in the case of social affinity.

Table 2 shows the correlations between d_{geo} with $P(T|\mathcal{U}_{d_{\text{geo}}})$ for all three events; similarly for social affinity. Additionally, this table shows the polyserial correlation[14] between the log-transformed physical distance and the number of local contacts[4]. 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, [23]) 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 (The New York and Alaska events decay after approximately three days, as seen in Figure 1). 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.

5.2 Measuring the Effect of Detachment on Event Aspects

The fact that users post queries related to an event is, as we demonstrated above, 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 users' 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.

[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]`news.yahoo.com`

[4]Polyserial correlation is required because physical distance is a continuous variable, whereas number of local contacts is an ordinal variable

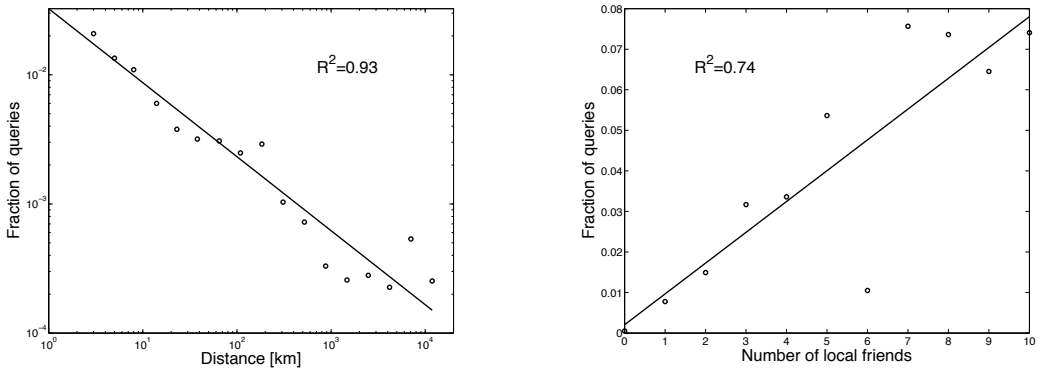

Figure 2: Fraction of queries as a function of the physical distance (left) and the number of local contacts (right). The regression line on the left is an exponential fit with $R^2 = 0.93$ and a linear fit on the right with $R^2 = 0.74$. Note the logarithmic axes of the left figure.

	Polyserial 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	-0.61	0.93	0.74	-0.57	0.008
New York	-0.47	0.71	0.18	-0.33	0.012
Alaska	-0.78	0.60	0.26	-0.22	$8 \cdot 10^{-5}$

Table 2: Correlation between physical and social detachment, and the correlations between binned physical detachment and social affinity with the fraction of queries for all three events. Linear regression was used for the physical detachment, and exponential regression for social affinity. All regression results are statistically significant at $p < 0.01$.

Clusters of queries have been shown in the past to represent different information needs [6, 7]. Therefore, we clustered the queries to identify the different information needs 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 4 clusters for the San Bruno and New York events, and three clusters for the Alaska event. We note that we did not use measures of detachment during clustering.

If the clustering has identified different aspects of the topic, we expect users to have a tendency to post queries which will be part of the same clusters. We measured this hypothesis using the Pearson goodness-of-fit test. This measures, for each user, if the distribution of queries to clusters is significantly different from the same distribution averaged for all users.

We limited this test to users who posted 5 or more queries related to an event so as to obtain good estimates of the distribution. Significance level was determined using a False Detection Rate [2] level of 0.1%.

We found that for the 27% of users of the San Bruno event, 18% of users of the New York event, and 7% of users of the Alaska event had a distribution statistically significantly different from the prior. As a comparison, running this test with randomly permuted data resulted, on average, in less than 0.1% of users who had significantly different distributions. This is a demonstration that the clusters typify different aspects of the event.

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[29], $p < 10^{-5}$) 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. 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 pat-

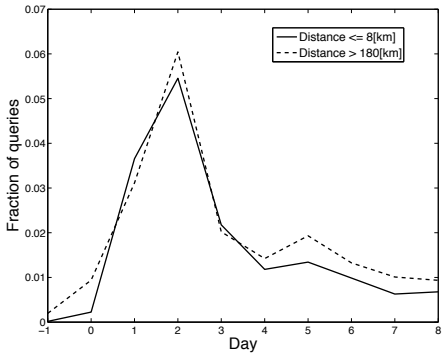

Figure 3: Fraction of event-related queries as a function of time in the San Bruno event, partitioned by the physical distance of users from the event. The fraction of queries for users far from the event was multiplied 50 times, to be on the same scale as users close to the event. Day zero denotes the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

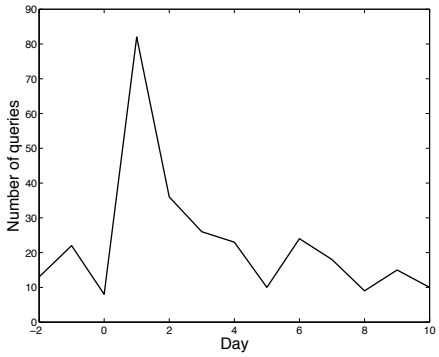

Figure 5: Number of times the query "PG&E" appeared during the San Bruno event.

terns. 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.

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

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 which 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 seemingly queries 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' information needs, improving retrieval [12], or aggregating news content [11].

As noted above, each query is represented by its text, as well as 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 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 1M queries appeared at least 100 times for each of the three 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 five-fold 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 [25] 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 3 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

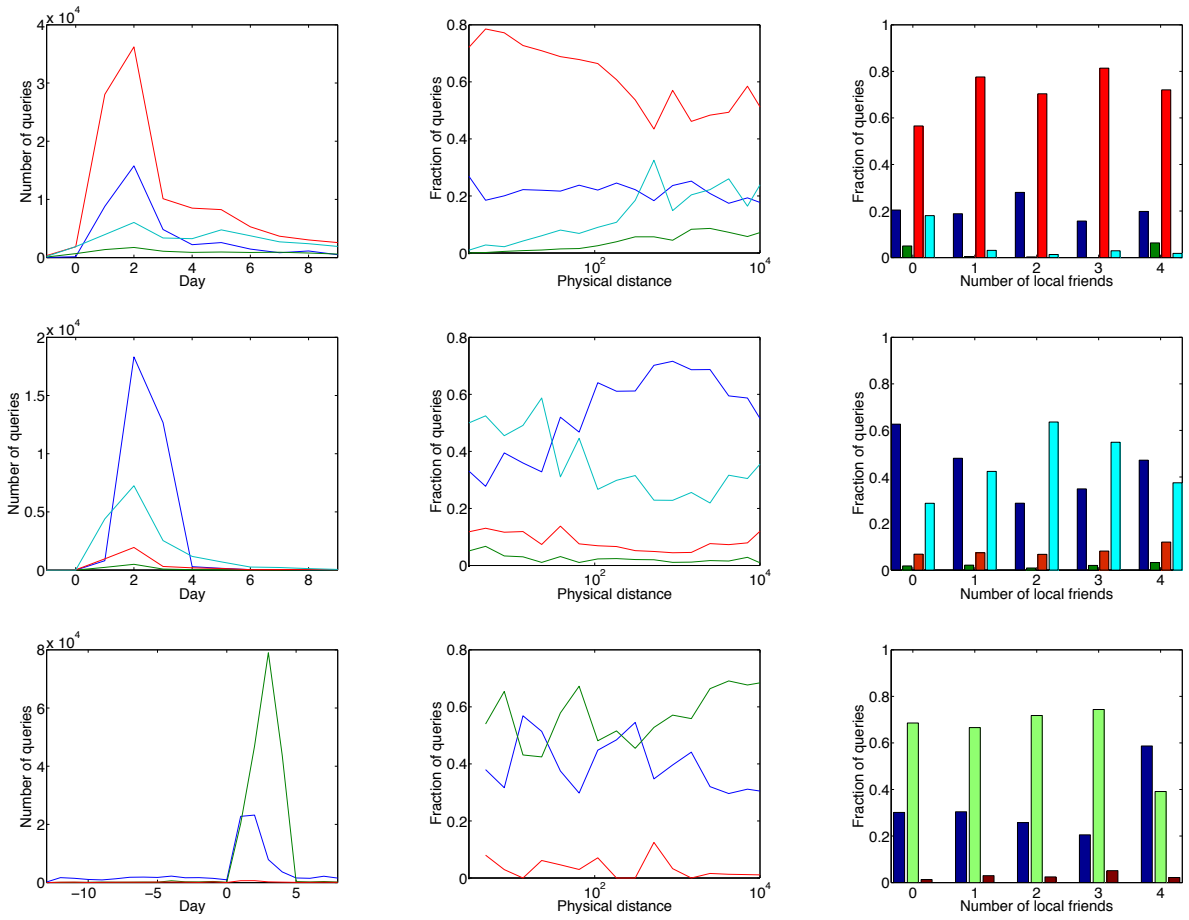

Figure 4: The number of queries from each cluster as a function of time, and the fraction of queries from each cluster as a function of physical and social affinity. Colors denote different clusters: Cluster 1 (blue), Cluster 2 (green), Cluster 3 (red), and Cluster 4 (cyan). Each row presents the data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

three times more event-related queries in this set of queries than in a 0.01-fraction random sample of the queries. Lift measures the precision at a given threshold [9], albeit scaled such that it can be larger than 1.

Table 3 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 (especially in the San Bruno and Alaska cases), though there is no clear preference for use of geographic versus social information. Interestingly, 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 [11].

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 (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.

	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

Table 3: 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. The column denoted by "All" represents results when using appearance counts, physical detachment and social detachment information.

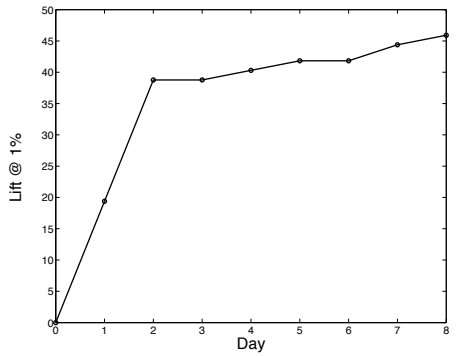

Figure 6: Lift at 1% obtained by using the data from days 0 to N, for the San Bruno event

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 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 her specific detachment information.

One way to use detachment information is to augment queries with attachment data and use a ranking algorithm [22] 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 far physically 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 can be trained for those users who are 10km or less from the event, and another model can be trained for those users who are located 10km 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 three 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 [16]. 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, e.g., 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 web site, and are especially useful when distinguishing news and non-news web sites.

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 and for the combination of the detachment levels. As a baseline, we trained a single model for the entire data (without 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 relevant documents was a decision tree classifier. We used five-fold cross validation in our experimentation, and measured the mean average precision (MAP) and the average precision at the top 5 and top 10 documents (denoted as P@5 and P@10, respectively).

Table 4 shows the results of our experiments. The models which use detachment information obtained a statistically significant improvement over the baseline. Separately, using social detachment gave a larger improvement over the baseline than physical detachment. The improvement in MAP using social detachment information over the baseline was, on average 10%. Improvement in P@5 was 5% and in P@10 2%. Interestingly, using both detachment parameters gave only marginal improvements over social detachment 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.

	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

Table 4: Improvement in retrieval results when using detachment information. 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.

7. DISCUSSION

In this paper we investigated the effect of social and geographic detachment of users of information need. 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 [19, 21, 26]. 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 non-zero 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 seeking, 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 election is almost solely related to the results of the election and not to candidates' work or people's need for information about voting procedures, etc, prior to the election. This might hint that query log volume is driven to some extent by media interest, but further experimentation is required to validate this link.

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 which is 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 [1]. 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 frequency of queries does not follow a simple exponential.

Bernard et al. [3] discussed the question of estimating the number of people who died in an earthquake from an entire population based on sampling a population and determining how many people they knew, first-hand, who had died, as well as the average number of acquaintances each person has (e.g., their immediate social network). 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. We plan to verify the utility of this method and, as it is considered a lower bound on the number of affected users, measure how close its estimates are to the ones identified by our heuristic.

8. 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] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate - a new and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57:289–300, 1995.

[3] H. Russell Bernard, Eugene C. Johnsen, Peter D. Killworth, and Scott Robinson. Estimating the size of an average personal network and of an event subpopulation. In M. Kochen, editor, *The small world*, pages 159–175. 1989.

[4] H. Russell Bernard, Eugene C. Johnsen, Peter D. Killworth, and Scott Robinson. Estimating the size of an average personal network and of an event subpopulation: Some empirical results. *Social science research*, 20:109–121, 1991.

[5] H. Russell Bernard, Peter D. Killworth, Eugene C. Johnsen, Gene A. Shelley, and Christopher McCarty.

Estimating the ripple effect of a disaster. *Connections*, 24(2):18–22, 2001.

[6] David Carmel, Elad Yom-Tov, Adam Darlow, and Dan Pelleg. What makes a query difficult? In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 390–397. ACM, 2006.

[7] David Carmel, Elad Yom-Tov, and Haggai Roitman. Enhancing digital libraries using missing content analysis. In *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*, JCDL '08, pages 1–10. ACM, 2008.

[8] David Carmel, Naama Zwerdling, Ido Guy, Shila Ofek-Koifman, Nadav Har'el, Inbal Ronen, Erel Uziel, Sivan Yogev, and Sergey Chernov. Personalized social search based on the user's social network. In *Proceeding of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 1227–1236, 2009.

[9] Rich Caruana and Alexandru Niculescu-Mizil. Data mining in metric space: an empirical analysis of supervised learning performance criteria. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, pages 69–78. ACM, 2004.

[10] 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.

[11] Fernando Diaz. Integration of news content into web results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, 2009.

[12] Anlei Dong, Yi Chang, Zhaohui Zheng, Gilad Mishne, Jing Bai, Ruiqiang Zhang, Karolina Buchner, Ciya Liao, and Fernando Diaz. Towards recency ranking in web search. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 11–20. ACM, 2010.

[13] Anlei Dong, Ruiqiang Zhang, Pranam Kolari, Jing Bai, Fernando Diaz, Yi Chang, Zhaohui Zheng, and Hongyuan Zha. Time is of the essence: improving recency ranking using Twitter data. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 331–340. ACM, 2010.

[14] Fritz Drasgow. Polychoric and polyserial correlations. In S. Kotz and N. Johnson, editors, *The Encyclopedia of Statistics, Volume 7*, pages 68–74. Wiley, 1986.

[15] Ahmed Hassan, Rosie Jones, and Fernando Diaz. 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, pages 33–41. ACM, 2009.

[16] Thorsten Joachims. Unbiased evaluation of retrieval quality using clickthrough data. In *SIGIR Workshop on Mathematical/Formal Methods in Information Retrieval*, 2002.

[17] Rosie Jones, Ahmed Hassan, and Fernando Diaz. Geographic features in web search retrieval. In *Proceeding of the 2nd international workshop on Geographic information retrieval*, GIR '08, pages 57–58. ACM, 2008.

[18] 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.

[19] Xiao Li, Ye-Yi Wang, and Alex Acero. Learning query intent from regularized click graphs. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 339–346. ACM, 2008.

[20] Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. Twitter under crisis: Can we trust what we RT? In *ACM SIGKDD 2010 Workshop on Social Media Analytics (SOMA)*, 2010.

[21] Donald Metzler, Susan T. Dumais, and Christopher Meek. Similarity measures for short segments of text. In *ECIR*, pages 16–27, 2007.

[22] Ramesh Nallapati. Discriminative models for information retrieval. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 64–71, 2004.

[23] Irit Nitzan and Barak Libai. Social effects on customer retention. *Marketing Science Institute working paper*, pages 10–107, 2010.

[24] Leyseia Palen, Sarah Vieweg, Sophia B. Liu, and Amanda Lee Hughes. Crisis in a networked world: Features of computer-mediated communication in the april 16, 2007, Virginia Tech Event. *Social Science Computer Review*, 2009.

[25] Yossi Richter, Elad Yom-Tov, and Noam Slonim. Predicting customer churn in mobile networks through analysis of social groups. In *Proceedings of the SIAM International Conference on Data Mining, SDM 2010*, pages 732–741, 2010.

[26] Mehran Sahami and Timothy D. Heilman. A web-based kernel function for measuring the similarity of short text snippets. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 377–386. ACM, 2006.

[27] 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.

[28] Jaime Teevan, Meredith Ringel Morris, and Steve Bush. Discovering and using groups to improve personalized search. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 15–24. ACM, 2009.

[29] Graham Upton and Ian Cook. *Oxford dictionary of statistics*. Oxford University Press, 2002.

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