

TWinner: Understanding News Queries with Geo-content using Twitter

Satyen Abrol

Department of Computer Science
University of Texas at Dallas
Richardson, TX
satyen.abrol@student.utdallas.edu

Latifur Khan

Department of Computer Science
University of Texas at Dallas
Richardson, TX
lkhan@utdallas.edu

ABSTRACT

In the present world scenario, where the search engines wars are becoming fiercer than ever, it becomes necessary for each search engine to realize the intent of the user query to be able to provide him with more relevant search results. Amongst the various categories of search queries, a major portion is constituted by those having news intent. Seeing the tremendous growth of social media users, the spatial-temporal nature of the media can prove to be a very useful tool to improve the search quality. In our work we examine the development of such a tool that combines social media in improving the quality of web search and predicting whether the user is looking for news or not. We go one step beyond the previous research by mining *Twitter* messages, assigning weights to them and determining keywords that can be added to the search query to act as pointers to the existing search engine algorithms suggesting to it that the user is looking for news. We conduct a series of experiments and show the impact that TWinner has on the results.

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: Selection process

General Terms

Algorithms and Experimentation

Keywords

Geographic Information Retrieval, news queries, search engines

1. INTRODUCTION

It's 12th November 2009, John is a naïve user who wants to know the latest on the happenings related to the incident that occurred at the army base in Fort Hood. John opens his favorite search engine site and enters "Fort Hood", expecting to see the news. But unfortunately the search results that he sees are a little different from what he had expected. Firstly, he sees a lot of timeless information such as Fort Hood on maps, the Wikipedia article on Fort Hood, the Fort Hood homepage, etc. clearly indicating that the search engine has little clue as to what the user is looking for. Secondly, among the small news bulletins that get displayed on the screen, the content is not organized and the result is that he has

hard time finding the news for 12th November.

Companies like Google, Yahoo and Microsoft are battling to be the main gateway to the Internet. Since a typical way for internet users to find news is through search engines and a rather substantial portion of the search queries is news related where the user wants to know about the latest on the happenings at a particular geo-location, it thus becomes necessary for search engines to understand the intent of the user query, based on the limited user information available to it and also the current world scenario.

We hypothesize that the best way to know the user's opinion is by 'asking' the user. In order to do that, we analyze the content of a popular social networking site, *Twitter*, to understand which news topics are popular among the users. The system is based on the intuition that if an event has occurred at a location, then the frequency of *Twitter* messages mentioning this location increases tremendously. In addition to this, these messages can very efficiently summarize the event, and provide us with selected keywords that may prove useful to enhance the user query. As in the case for the Fort Hood shootings, *Twitter* can point out the occurrence of an event as well as provide us with suitable keywords like 'shooting', 'killing', 'suspect' etc. to enhance the query. TWinner collects the *Twitter* messages, assigns weights to individual keywords, measures the semantic similarity and chooses k optimum keywords.

The application of social media, including *Twitter*, presents several challenges. Foremost among these is the raw nature of the text, consisting of slang and incorrect grammar. Other challenges faced include the identification of spam messages, repetitive broadcasting of a single message by the same user.

TWinner makes two novel contributions to the field of Geographic information retrieval. First, it helps the search engine to identify the intent of the user query, whether he is interested in general information or the latest news. Second, TWinner adds additional keywords to the query so that the existing search engine algorithm understands the news intent and displays the news articles in a more meaningful way.

The research paper is organized as follows. Section 2 surveys the related work in this domain and points out the novelty in our approach. Section 3 discusses Twitter and its application as source for news. Section 4 describes our methodology in understanding the intent of the user query. Section 5 and 6 deal with assigning weights to the *Twitter* messages collected and inclusion of semantic similarity of words in selecting the keywords to enhance the search query. Section 7 discusses the experiments and the outcomes. Section 8 talks about time complexity of the method.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
GIR'10, 18-19th Feb. 2010, Zurich, Switzerland.

Copyright © 2010 ACM ISBN 978-1-60558-826-1/10/02... \$10.00.

We conclude in section 9, by giving a few pointers for the future work.

2. RELATED WORK

Geographic information retrieval is a well discussed topic in the past, where a lot of research has been done to establish a relationship between the location of the user, and the type of content that interests him. Researchers have analyzed the influence of user's location on the type of food he eats, the sports he follows, the clothes he wears, etc. But it is important to note here that most of the previous research does not take into account the influence of 'time' on the preferences of the user.

Previously, a lot of work has been done to identify and disambiguate the location of users by a lot of researchers. Most of the research can be broadly classified into two approaches. One, involving the concepts of Natural Language Processing and the other using data mining approach. Most of the work done using NLP techniques consists of input text that is structured and well-edited. Li et al. [14] combined these two methodologies and used typical 5-step approach, first short listing the keywords appearing in the gazetteer and then applying NLP techniques to remove non-geo terms. A precision of 93.8% was reported using their approach.

Significant work has been done in the past to establish the relationship between the location of the news and the news content. Mehler et al. [8] developed a model for estimating and evaluating spatial significance of entities using NLP techniques. Liu et al. [10] do a similar geo-analysis of the impact of the location of the source on the viewpoint presented in the news articles. Sheng et al. in [11] discussed the need for reordering the search results (like food, sports, etc.) based on user preferences obtained by analyzing user's location.

Other previous research attempts [7, 12] focused on establishing the relationship between the location obtained from IP address and the nature of the search query issued by the user. In our work, we do not include the location of the user into our consideration, since it may not be very accurate in predicting the intent of the user.

Hassan et al. in [9] focus their work on establishing a relationship between the geographic information of the user and the query issued. They examine millions of web search queries to predict the news intent of the user, taking into account the query location confidence, location type of the geo-reference in the query and the population density of the user location. But they do not consider influence of time at which the user issued the query, which can negatively affect the search results for news intent. For example, a query for 'Fort Hood' 5 months back would have less news intent and more information intent than a query made in second week of November (after the Ft. Hood shootings took place).

Twitter acts as a popular social medium for internet users to express their opinions and share information on diverse topics ranging from food to politics. A lot of these messages are irrelevant from an information perspective and are either spam or pointless babble. Another concern while dealing with such data is, that it consists of a lot of informal text including words such as 'gimme', 'wassup', etc. and need to be processed before traditional NLP techniques can be applied to them.

Nagarajan et al. [16] explore the application of restricted relationship graphs or resource description framework (RDF) and statistical NLP techniques to improve named entity annotation in

challenging informal English domains of social networking sites such as *MySpace*.

It is vital to understand the contribution of *TWinner* in establishing a relationship between the search query and the social media content. In order to do so we suggest *Twitter*, a popular social networking site to predict the news intent of the user search queries.

3. TWITTER AS NEWS-WIRE

Twitter is a free social networking and micro-blogging service that enables users to send and read messages known as *tweets*. Tweets are text posts of up to 140 characters displayed on the author's profile page and delivered to the author's subscribers who are known as *followers*.

San Antonio based market research firm Pear Analytics [24] analyzed 2,000 tweets (originating from the US and in English) over a two week period from 11:00am to 5:00pm (CST) and categorized them as:

- News
- Spam
- Self-promotion
- Pointless babble
- Conversational
- Pass-along value

Tweets with news from mainstream media publications accounted for 72 tweets or 3.60 percent of the total number [19]. Realizing the importance of *Twitter* as a medium for news updates, the company emphasized on news and information networking strategy in November 2009 by changing the question it asks users for status updates from "What are you doing?" to "What's happening?".

The growth of *Twitter* attributed to the fact that it is free, highly mobile, very personal and very quick. It's also built to spread, and spread fast. *Twitterers* like to append notes called hash tags — #theylooklikethis — to their tweets, so that they can be grouped and searched for by topic; especially interesting or urgent tweets tend to get picked up and retransmitted by other users, a practice known as re-tweeting, or RT. And *Twitter* is promiscuous by nature: tweets go out over two networks, the Internet and SMS, the network that cell phones use for text messages, and they can be received and read on practically anything with a screen and a network connection. Each message is associated with a time stamp and additional information such as user location and details pertaining to his social network can be easily derived.

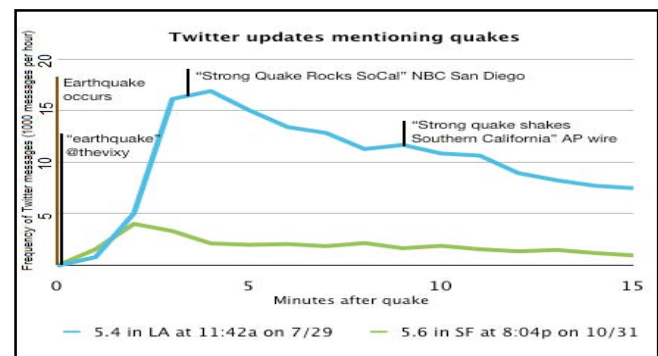


Fig 1 shows the *Twitter* message graph after the Southern California earthquakes. [2]

The impact of *Twitter* on news can be understood further by its coverage of two very crucial recent events, the 29th July earthquake in southern California and the turbulent aftermath of Iran's Elections in June'09.

This chart illustrates the beginning of this morning's earthquake followed seconds later by the first *Twitter* update from Los Angeles. About four minutes later, official news began to emerge about the quake. By then, "Earthquake" was trending on *Twitter* Search with thousands of updates and more on the way. Many news agencies get their feed from a news wire service such as the Associated Press. "Strong quake shakes Southern California" was pushed out by AP about 9 minutes after people began *Twittering* primary accounts from their homes, businesses, doctor's appointments, or wherever they were when the quake struck. [2]

The second example would be that of the elections in Iran this year. Time in partnership with CNN discuss the impact of *Twitter* on the coverage of developments after Iran elections [3]. On June 12th, Iran held its presidential elections between incumbent Ahmadinejad and rival Mousavi. The result, a landslide for Ahmadinejad, has led to violent riots across Iran, charges of voting fraud, and protests worldwide. Even as the government of that country was evidently restricting access to opposition websites and text-messaging, but on *Twitter*, a separate uprising took place, as tweets marked with the hash tag *#cnnfail* began tearing into the cable-news network for devoting too few resources to the controversy in Iran. US State Department officials reached out to *Twitter* and asked them to delay a network upgrade that was scheduled for Monday (June 15th) night. This was done to protect the interests of Iranians using the service to protest the presidential election that took place on June 12.

4. DETERMINING NEWS INTENT

In this section we give a detailed description of the process that we undertake to understand the intent of the user query.

Fig 2 shows the architecture of TWinner. In the first step the user enters his search consisting of a location. In the next step, the location is uniquely identified, and the Frequency Population Ratio (FPR) is calculated at that instant using *Twitter*. If the FPR is significantly higher than 1, it indicates that the topic is popular on *Twitter* and the query is tagged as a news intent query and further steps are taken to enhance the query to yield better search results. For this, the system collects all the messages posted on *Twitter* in the last 24 hours. Next, we assign weights to them keeping in mind their likelihood of containing news content. Then, we reassign the weights taking into account the semantic similarity between the keywords. Finally we extract the top k keywords that are semantically dissimilar but contribute the maximum weight.

4.1 Identification of Location

In the first step we attempt to geo-tag the query to a location with certain confidence. We use the system described in [20]. The system builds a city language model, which is a probabilistic representation of the language surrounding the mention of a city in web queries. We use several features derived from these language models to identify the words in the query that refer to a location. For example a query like "Disney world" implies that the city is Orlando, Florida. The system has over 90% precision and more than 74% accuracy for the task of detecting users' implicit city level geo intent. For further details please refer to [20].

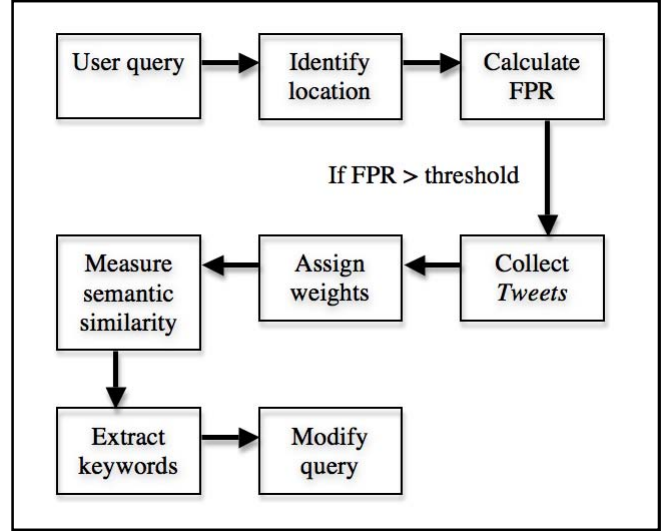


Fig 2 shows the architecture of news intent system TWinner

4.2 Frequency – Population Ratio

Once the location mentioned in the query has been identified explicitly, the next step is to assign a news intent confidence to the query.

Coming back to the Fort Hood query, once we are able to identify Fort Hood as a unique location, our next task is to identify the intent of the user. Intuition tells us that if something has happened at a place, the likelihood of people talking about it on *Twitter* will increase manifolds.

To understand this concept, we define an index called the Frequency - Population Ratio (FPR) which in the absence of a news making event will always remain constant irrespective of the location.

It is essential to note here that to make FPR to be a constant value, irrespective of the location type; we shall have to take certain factors into consideration. To determine the effect of geo-type and the population density on the news content of the messages we perform a series of experiments. We collect a dataset of 10 thousand twitter messages, all having a reference to a location in them. For the first experiment, we divide the messages on the basis of population density. Figure 3 shows how the frequency of *Twitter* messages is affected by the population density of the geo-location mentioned in it. The horizontal axis represents the population density in number of persons per sq. miles and the y axis represents the number of tweets per hour.

It can be clearly observed that the higher the population density of the location, the higher is the frequency of tweets. In other words, there will be more *Twitter* messages on New York than on a small town like Bryan, Texas.

The second experiment we perform is to determine the relation between the message frequency and the geo-type of the reference location. For the same dataset, we classified each message on the basis of geo-reference. Figure 4 shows the outcome of the experiment. We determine that state or country names are more likely to appear in *Twitter* messages rather than the county names.

Now we are ready to define the FPR as a function of these two factors, population density and geo-type, represented by constants

α and β respectively. We further state that the values of α and β are such chosen that, the FPR for all locations is a constant value equal to 1 irrespective of any factor. In other words, the value of α is inversely proportional to the population density and β depends on the geo-type as shown in Fig 5.

Thus,

$$FPR = (\alpha + \beta) * N_t$$

Where α is the population density factor, N_t is the number of tweets per minute at that instant and β is the location type constant.

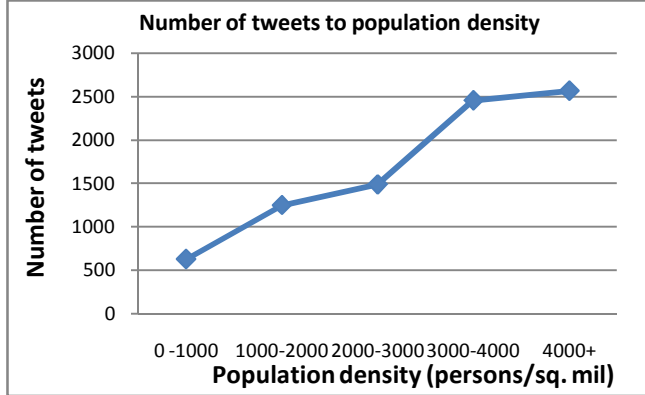


Fig 3 shows the percentage of news messages versus the population density of user's location in persons per square miles

Table 1 shows some sample geo-locations, the chosen values of alpha and beta and the resulting FPR ratio for weekdays based on a one week time period. It is very important to note here that FPR is a constant (equal to 1) on regular days when the geo-location is not in news or is not a popular topic on *Twitter*. But in events such as the Fort Hood shootout incident, the new FPR calculated using the pre-determined values of α and β increases by manifolds. We make use of this feature to determine whether a geo-location is in news or not.

For example, we calculated the average FPR for 'Fort Hood' during the week of 5th to 12th November using the values of α and β and found it to be 1820.7610 which is seemingly higher than 1, indicating that people were talking about Fort Hood on *Twitter*. And we take that as a pointer that the place is in news.

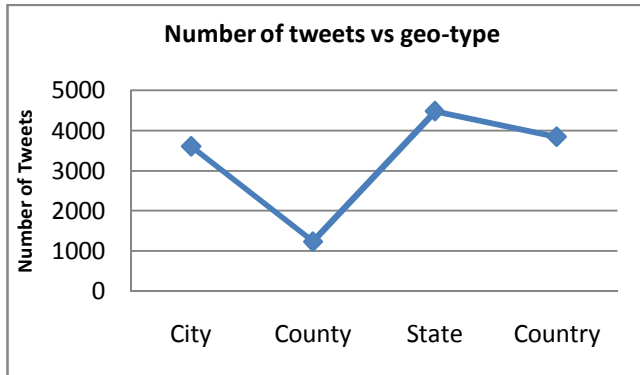


Fig 4 shows the percentage of tweets corresponding to type of geo-reference them.

An evident drawback of this approach is that while considering the FPR, we are not taking into account the geographical relatedness of features. For example, if the user enters Tehran and is looking for Iran elections, while calculating the FPR, in addition to the *Twitter* messages for 'Tehran' we need to consider messages containing keywords 'Iran' and 'Asia' as well. Therefore, we modify our earlier definition for FPR to

$$FPR = \sum \delta_i (\alpha_i + \beta_i) * N_t$$

The constant δ_i accounts for the fact that each geo-location related to the primary search query contributes differently. That is, the contribution of *Twitter* messages with 'Fort Hood' (primary search location) will be more than that of messages with 'Texas' or 'United States of America'.

Table 1 shows some sample locations, and the corresponding estimated alpha and beta values and the Frequency Population Ratio (FPR)

Example of Location	Value of Alpha	Value of Beta	Frequency Population Ratio (FPR)
Fort Hood (City)	1.21	0.1096	1.025
Los Angeles (City)	0.02	0.1096	1.100
Collin (County)	0.749	0.3396	1.084
Texas (State)	0.0045	0.09459	0.997
North Dakota (State)	0.233	0.09459	1.129
Australia (Country)	0.104	0.1073	0.988

5. ASSIGNING WEIGHTS TO TWEETS

Once we have determined to a certain confidence level the news intent of the user query, the next step is to add certain keywords to the query which act as pointers to the current search engine algorithm telling it that the user is looking for news.

To begin with we collect all *Twitter* messages posted in the last 24 hours containing a reference to either the geo-location (e.g. Fort Hood) or the concepts that subsume it (e.g. Texas, United States of America, etc.). We then assign weights to each *Twitter* message based on the likelihood of its accuracy in conveying the news. In the following subsections we describe the various factors that might affect the possibility of a *Twitter* message having news content. The weight carried by each message is the function of spam message content, user location weight and hyperlink weight.

5.1 Detecting Spam Messages

On close observation of the *Twitter* messages for popular topics, it was noticed that some of the *Twitter* messages are actually spam messages, where the spammer has just used the popular keywords so that his message reaches out to the people who are looking at this trending topic. In other words a significant percentage of the *Twitter* messages are actually spam and carry little or no relevant information. It is thus important to recognize such messages and

give lower weight to them. In this section we briefly describe our method of identifying whether a message is spam or not.

To determine whether a message is spam or not, it is important to understand the nature of spam messages. For this, we collect a set of 1000 messages manually annotated as spam. The methodology we use is based on analyzing the social network of the user posting the message. The social network of a user on *Twitter* is defined by two factors, one, the people he is *following* and the other people *following* him. We observed that the ratio of the number of followers to the number of people a spammer is following is very small. The second observation is that a spammer rarely addresses his messages to some specific people, that is, he will rarely reply to messages, re-tweet other messages, etc. Fig 5 shows the profile of a typical spammer. Note that he is following 752 people and is just being followed by 7 people.



Fig 5 shows profile of a typical spammer

Based on these two hypotheses, we come up with a formula that tags to a certain level of confidence whether the message is spam or not. The spam confidence Z_i is defined as

$$Z_i = \frac{1}{(N_p/N_q) + \mu * N_r}$$

Where N_p and N_q are the number of followers and number of people the user is following respectively. μ is an arbitrary constant and N_r is the ratio of number of tweets containing a reply to the total number of tweets.

It is important to note here the higher the value of the spam confidence, Z_i , the greater is the probability of the message being spam and therefore its contribution to the total weight is lowered.

5.2 On Basis of User Location

In this section we describe the experiments that we conducted to understand the relationship between *Twitter* news messages and the location of the user. We performed experiments on two different samples of data each comprising of 10 thousand tweets, and calculated the distance between the location of the user and the location mentioned in the message. The results of the experiment are shown in figure 6.

It can be interpreted from the findings that people located in the same state, same country and also neighboring countries are more

likely to talk about a news event as compared to the people located immediately next to the location (within a ten mile radius) or very far from it (different continent). We use these experiments as the baseline and use the inferences to assign weights for messages on future topics.

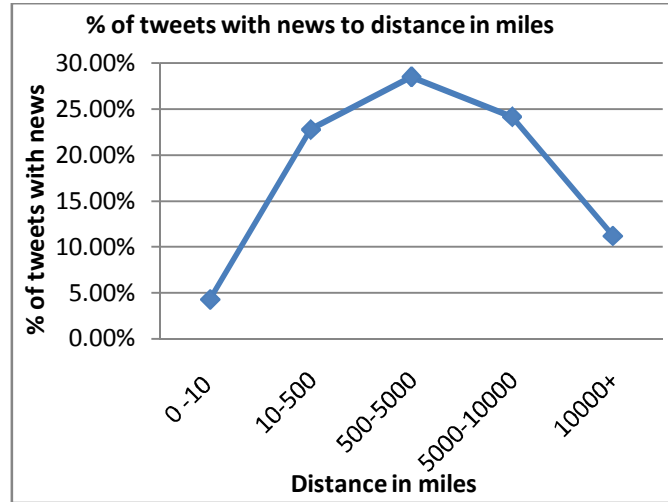


Fig 6 displays the relation between number of tweets to the distance between the *Twitter* user and query location

5.3 Using Hyperlinks Mentioned in Tweets

An interesting observation that we make from our experiments is that 30-50% of the general *Twitter* messages contain a hyperlink to an external website and for news *Twitter* messages this percentage increases to 70-80%. Closer analysis indicate that firstly, a lot of popular news websites tweet regularly and secondly, mostly people follow a fixed template of writing a short message followed by a link to the actual news article.

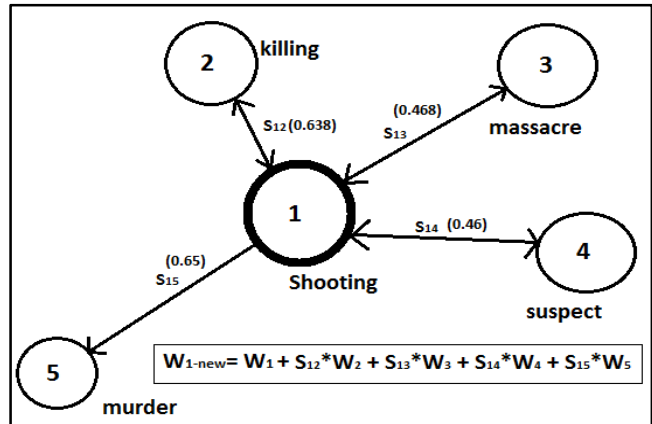


Fig 7 shows the revised calculation of weights taking semantic similarity into account.

So we make use of this pointer in the *Twitter* message for extra information and crawl the links to analyze the content. Hence in addition to the previously mentioned two strategies, the weight for the message is also affected by the content of the website mentioned in the message. A weight, which is a function of the factors such as the type of site (news, spam, blog etc.), the currency of the site, etc., is assigned to each message.

6. SEMANTIC SIMILARITY

Now that we have assigned the weights to each *Twitter* message (section 5), it becomes essential for us to summarize them into a couple of most meaningful keywords. We now break down each message into keywords and each keyword in a message is assigned the same weight as carried by the message. Finally we to obtain the total weight of keyword, we sum up the weights of the messages that contain it.

Various approaches can be employed to choose k keywords that enhance the user query. A naïve approach would be to rank the keywords in descending order of weight and pick the first k keywords. But one evident disadvantage of this approach would be that it would not take into account the semantic similarity of the keywords involved. E.g. ‘shooting’ and ‘killing’ are treated as two separate keywords, in spite of their semantic proximity. In this section we describe a process that in the first step reassigns weights to the keywords on the basis of semantic relatedness and in the second step picks k keywords that are semantically dissimilar but have maximum combined weight.

As mentioned earlier any two words are rarely independent and are semantically related to each other. E.g. ‘shooting’, ‘killing’ and ‘murder’ are semantically very similar words. Since our work is based on news intent, to calculate the similarity, we use the New York Times corpus that contains 1.8 million articles. The semantic similarity, S_{xy} , of two words x and y is defined as

$$S_{xy} = \frac{\log M - \min \{ \log f(x), \log f(y) \}}{\max \{ \log f(x), \log f(y) \} - \log f(x,y)}$$

where M is the total number of articles searched in New York Times Corpus; $f(x)$ and $f(y)$ are the number of articles for search terms x and y , respectively; and $f(x, y)$ is the number of articles on which both x and y occur.

Now we reassign the weight of all keywords on the basis of the following formula:

$$W_i^* = W_i + \sum S_{ij} * W_j$$

Where W_i^* is the new weight of the keyword i , W_i is the weight without semantic similarity, S_{ij} is the semantic similarity derived from semantic formula and W_j is the initial weight of the other words being considered as shown in Figure 7 where the weight of the keyword ‘shooting’ is boosted by the presence of similar words such as ‘killing’, ‘murder’, ‘suspect’ and ‘massacre’.

After all the n keywords are reassigned a weight, we go to our next step that aims at identifying k keywords that are semantically dissimilar but contribute together contribute maximum weight. In other words choose words $W_1 \dots W_k$ such that

- (1) $S_{pq} < S_{\text{Threshold}}$, the similarity between any two word(p) and word(q) belonging to the set k is less than a threshold, and
- (2) $W_1 + W_2 + \dots + W_k$ is maximum for all groups satisfying (1).

It can be easily shown that the complexity of the above described method is exponential in n . We briefly describe three techniques that are lower in complexity to approximately come up with the k keywords.

6.1 Greedy Approach

First we applied the greedy algorithm approach. For this, we arrange all the words in decreasing order of their weights. We start with the keyword with the maximum weight that is W_1 , put it in the basket and starting traversing the array of words. Next, we define an objective function defined by

$$\Theta_i = (1/E_i) * W_i$$

Where E_i is the sum of semantic similarity of the word with all the words in the basket and W_i is its own weight. Hence at each step of the algorithm we choose a word that maximizes the objective function (Θ).

6.2 Hill Climbing Approach

The second approach is hill climbing approach. We choose a set of k random words that satisfy the condition. Next, we randomly select a word, check if it satisfies the condition of semantic similarity threshold with all the k words. If its weight is more than the weight of the lightest word in the random list, we replace the two. We keep repeating the process until the random word selected does not satisfy the condition.

6.3 Simulated Annealing Approach

And our final method is that of simulated annealing. The advantage of simulated annealing as compared to hill climbing is that, it does not get stuck on local minima. Unlike hill climbing, even if the weight of the next randomly chosen word is less than the weight of the *lightest* word, it is still taken but with a probability that decreases each time a word that lowers the total weight is chosen as shown in Fig 8.

```
function SIMULATED-ANNEALING( problem, schedule ) returns a solution state
  inputs: problem, a problem to choose  $k$  keywords that maximize the weight
         schedule, a mapping from time to "temperature"
  local variables: current, current weight of the  $k$ -words
                  next, the weight after taking including the next keyword
                   $T$ , a "temperature" controlling prob. of downward steps

  current ← MAKE-NODE( INITIAL-STATE[problem] )
  for  $t \leftarrow 1$  to  $\infty$  do
     $T \leftarrow \text{schedule}[t]$ 
    if  $T = 0$  then return current
    next ← a randomly selected successor weight
     $\Delta E \leftarrow \text{VALUE}[\text{next}] - \text{VALUE}[\text{current}]$ 
    if  $\Delta E > 0$  then current ← next
    else current ← next only with probability  $e^{\Delta E/T}$ 
```

Fig 8 shows the modified simulated annealing algorithm for choosing k semantically dissimilar words with maximum weights

Hence in each step the randomly chosen word is included with a probability which decreases with each *light* weight word. And it can be easily shown that after not too many wrong choices the probability becomes 0. Amongst the three methods described above, simulated annealing produces the most accurate results, but on the other hand is slower than the other two. However, the running time of these methods heavily depends on the value of k . And since for our approach k is usually a very small number (for example 2), we can safely adopt simulated annealing to obtain the bag of k words.

These k keywords derived from reassigning the weights after taking semantic similarity into account are treated as special

words that act as pointers making the news intent of the query evident to the current search engine algorithm.

7. EXPERIMENTS AND RESULTS

To see the validity of our hypothesis, we performed experiments to determine two keywords ($k=2$) to enhance three queries which returned confidence values of ~ 1 indicating news intent. For the first experiment, we come back to our initial problem scenario, where John, a naïve user, is looking for the latest on the happenings in context to the Fort Hood incident. He enters a query with news intent on 12th November 2009. Now, TWinner determines the value of FPR, using the predetermined values of α and β as 1.21 and 0.1096 (Fig 5), to be 1820.7610. Since this is significantly greater than the standard value of 1, it suggests that Fort Hood is a popular topic on *Twitter*. We assign a confidence value of 0.999 on the news intent of the query based on the formula discussed in section 4.

The next part of TWinner is to determine k keywords that enhance the search query. For this, we collect 10 thousand *Twitter* messages for 12th November having the keywords “Fort Hood” in them and assign weights to them based on the criteria mentioned in section 5. Amongst these, 789 were determined to be spam and were hence assigned low weights. After this, keywords ‘murder’ and ‘suspect’ were selected by the algorithm to have the maximum cumulative weights. We added these keywords to the search query and observed the impact they had on the results returned by the search engine. The difference on the results is shown in Fig 9.

Similarly we conducted an experiment to enhance the query for ‘Russia’ on 5th December 2009. This is done keeping in mind people who wanted to know about the explosion in a night club in Russia that killed more than 100 people. The new FPR calculated using predetermined values of α and β was calculated to be 783.9702, indicating that Russia was a hot topic on *Twitter*. Next to determine the extra words to enhance the query, we chose all the *Twitter* messages containing the keyword ‘Russia’ and applied the algorithm to them. The algorithm returned the two words ‘night’ and ‘explosion’, but it was interesting to note here that two other sets of words ‘club’ and ‘explosion’ also had very similar collective weight. In such a scenario, the algorithm chooses all three words ‘night’, ‘club’ and ‘explosion’ to enhance the query.

The third experiment was performed to check the response of the system for keyword “Haiti” keeping in mind the user looking for news on 7.0 M_w earthquake that hit the country on 12th January 2010. For the experiment conducted on 18th January 2010, the system returned the keywords “earthquake” and “help”, fairly summing up the current happenings.

Fig 10 compares the accuracy of a naïve approach with TWinner. It is important to note here that for Fort Hood and Haiti, the number of news results (indicated by accuracy) varies between 60% and 76% for Fort Hood and between 46% and 60% for Haiti. But on the other hand for Russia, the number of news results is between 10% and 20%, indicating the severe problem faced by the search engine in understanding the news intent for popular queries which are not always in the news. This is where TWinner proves most useful, and provides better accuracy in determining the intent of the query.

It can be observed that without using TWinner, the search engine is not sure about the news intent of the user. As a result it displays

results that constitute a short news caption, the homepage of Fort Hood, the maps version, Wikipedia articles, etc. On the right side in figure 9 is an enhanced version of the query obtained after TWinner extracted the keywords ‘murder’ and ‘suspect’. The impact on the content of results is clearly visible. Majority of the search results are news oriented and are also in accordance with the date the user issued the query, that is, 12th November (The *Twitter* dataset was also collected for 12th November).

9. CONCLUSION AND FUTURE WORK

We present a system that takes into account location mentioned and time of query into consideration while assigning a confidence score to each user query containing reference to geo-location, predicting whether the query will receive a news click or not. We make use of the social networking site *Twitter* to understand the relationship between the geographic information and the likelihood of the news intent of the query. Our second contribution is the identification of keywords from the *Twitter* text that enhance the query and act as pointers to the search engine algorithm, telling it about the news intent of the search queries.

Our current approach makes little attempt at understanding the content of the social media messages and the sentiment conveyed by them. Another pointer to the future work would be enhancing the accuracy of the system by combining our work with traditional methods to predict intent like location of user, population density of region etc.

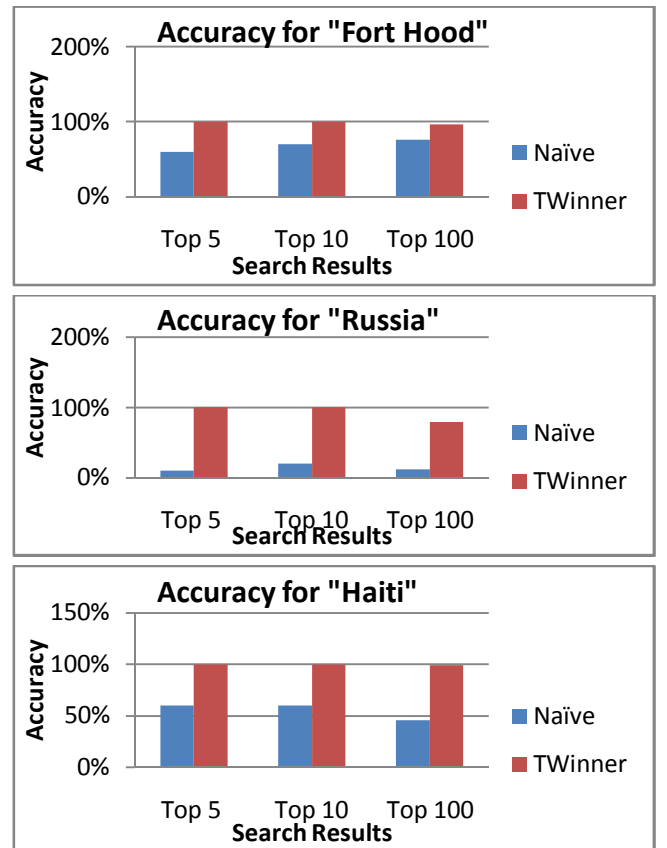



Fig 10 shows difference in accuracy for the search results without and with TWinner for “Fort Hood”, “Russia” and “Haiti” respectively.




[News results for fort hood](#)
[Imam's e-mails to Fort Hood suspect Hasan tame compared to online ...](#) - 1 day ago
[After accused Fort Hood shooter Nidal Malik Hasan started e-mailing in December, it increased the pace of his fundamentalist rhetoric on the Web, ...](#)
[Dallas Morning News - 87 related articles »](#)
[Kathy Ward: Fort Hood soldiers should be Texans of the Year -](#)
[Dallas Morning News - 925 related articles »](#)
[Will heads roll in Pentagon probe of Fort Hood shootings? -](#)
[Christian Science Monitor - 48 related articles »](#)

[The Fort Hood Homepage](#)
News, images, and links to other Fort Hood, Local Community, Army and Department of Defense sites.
[www.hood.army.mil/ - Cached -](#)

[Gunman kills 12, wounds 31 at Fort Hood - Crime & courts- msnbc.com](#)
Nov 5, 2009 ... An Army psychiatrist who opened fire at Fort Hood, Texas, killing 12 people and wounding 31 others, was shot but captured alive, ...
[www.msnbc.msn.com/id/33678801/ - Cached -](#)

[MSNBC Exclusive: Fort Hood Never Happened! - HUMAN EVENTS](#)
Nov 25, 2009 ... Olbermann has argued, on several occasions, that it is impossible, literally impossible, to commit mass murder at a military base.
[www.humanevents.com/article.php?id=34594 -](#)



Fort Hood
[pao.hood.army.mil](#)
1937 Bldg
Killeen, TX 76544
(254) 532-5229
[Get directions](#) - [Is this accurate?](#)
1 review - [Write a review](#)
[More information »](#)

©2009 Google Map data ©2009 Google

[Blog posts about fort hood](#)
[Obama Urges Congress To Put Off Fort Hood Probe, Warns Against ...](#) -
[Barack Obama on The Huffington Post - Nov 14, 2009](#)
[Brian Levin, J.D.: Fort Hood Tragedy Is Being Exploited To Bolster ...](#) -
[Terrorism on The Huffington Post - Nov 12, 2009](#)
[RealClearPolitics - An Officer's Outrage Over Fort Hood - RealClearPolitics - Articles - Nov 12, 2009](#)

[Fort Hood - Wikipedia, the free encyclopedia](#)
As originally constructed, Fort Hood had an area of 158706 Acres, with Billeting for 6007 Officers and 82610 Enlisted Personnel ...

[Fort Hood suspect charged with murder - CNN.com](#)
Nov 12, 2009 ... Maj. Nidal Malik Hasan has been charged with 13 preliminary counts of premeditated murder stemming from last week's shooting at Fort Hood ...
[www.cnn.com/2009/CRIME/11/12/fort.hood.../index.html - Cached -](#)

[Fort Hood suspect charged with murder - Tragedy at Fort Hood...](#)
Nov 12, 2009 ... Military prosecutors charge Fort Hood shooting suspect Maj. Nidal Malik Hasan of 13 counts of premeditated murder, officials say.
[www.msnbc.msn.com/id/33886322/.../us_news-tragedy_at_fort_hood/ - Cached -](#)

[Fort Hood Suspect Maj. Nidal Hasan Faces Murder Charges](#)
Nov 12, 2009 ... Maj. Nidal Hasan is charged with 13 counts of premeditated murder for last week's deadly shooting rampage at Fort Hood, according to ...
[news.aol.com/article/fort-hood-suspect-maj-nidal-hasan.../765023 - Cached -](#)

[Army: Fort Hood suspect charged with murder - Yahoo! News](#)
Nov 12, 2009 ... Army: Fort Hood suspect charged with murder ... Hood, Texas, has been charged in a military court with 13 counts of premeditated murder. ...
[news.yahoo.com/s/ap/.../ap.../us_fort_hood_shooting_charges - Cached -](#)

[Army: Fort Hood Suspect Charged With Murder - ABC News](#)
Nov 12, 2009 ... Army investigators: Fort Hood shooting suspect charged with 13 counts of premeditated murder. By LOLITA C. BALDOR Associated Press Writer ...
[abcnews.go.com/Politics/wireStory?id=9065419 -](#)

[Army: Fort Hood suspect charged with murder](#)
Army: Fort Hood suspect charged with murder. Nov 12 01:33 PM US/Eastern By ANGELA K. BROWN and LOLITA C. BALDOR Associated Press Writers ...
[www.breitbart.com/article.php?id=D9BU59V80&show_article... - Cached -](#)

[Fort Hood suspect to be charged with murder - CNN.com](#)
Nov 12, 2009 ... Maj. Nidal Hasan will face 13 preliminary charges of premeditated murder stemming from last week's shooting at Fort Hood Army Post in Texas, ...
[sidebar.cnn.com/2009/CRIME/11/12/fort.hood.../index.html - Cached -](#)

[Fort Hood shooting suspect charged with 13 murders | Reuters](#)
Nov 12, 2009 ... If convicted of premeditated murder by a military court he could face the death penalty. ... Fort Hood suspect to be tried in military court ...
[www.reuters.com/article/topNews/idUSTRE5AB43E20091112 - Cached -](#)

[Army: Fort Hood suspect charged with murder - CHICAGO SUN-TIMES...](#)
Nov 12, 2009 ... Army: Fort Hood suspect charged with murder Premeditated murder charges make him eligible for death penalty if convicted - CHICAGO ...
[www.suntimes.com/.../11880927.fort-hood-shooting-charges-111209.article -](#)

[Fort Hood suspect to be charged with murder - Mixx](#)
Maj. Nidal Hasan will face 13 preliminary charges of premeditated murder, US military sources say.
[www.msnbc.msn.com/id/33886322/.../us_news-tragedy_at_fort_hood/ - Cached -](#)

Fig 9 shows the contrast in search results produced by using original query and after adding keywords obtained by TWinner.

10. REFERENCES

- <http://en.wikipedia.org/wiki/Twitter>.
- Biz Stone, co-founder of Twitter on Twitter blog, <http://blog.Twitter.com/2008/07/Twitter-as-news-wire.html>
- Time in Partnership with CNN, <http://www.time.com/time/world/article/0,8599,1905125,00.html>.
- Juan, Y.-F. & Chang, C.-C., in "An analysis of search engine switching behavior using click streams", Proc. WWW, 2005.
- M. D. Lieberman, H. Samet, J. Sankaranarayanan, and J. Sperling. Steward: architecture of a spatio-textual search engine in *GIS*, page 25. ACM, 2007.
- F. Diaz. Integration of news content into web results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, 2009.
- Z. Zhuang, C. Brunk, and C. L. Giles. Modeling and visualizing geo-sensitive queries based on user clicks. In *LocWeb*, pages 73--76, 2008.
- A. Mehler, Y. Bao, X. Li, Y. Wang, and S. Skiena. Spatial analysis of news sources. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):765--772, 2006.
- Hassan, A., Jones, R., Diaz, F. in "A case study of using geographic cues to predict query news intent", ACM GIS'09, November 4-6, 2009. Seattle, WA.
- J. Liu and L. Birnbaum. Localsavvy: aggregating local points of view about news issues. In *LocWeb*, pages 33--40, 2008.
- C. Sheng, W. Hsu, and M.-L. Lee. Discovering geographical-specific interests from web click data. In *LocWeb*, pages 41--48, 2008.
- L. Backstrom, J. M. Kleinberg, R. Kumar, and J. Novak. Spatial variation in search engine queries. In *WWW*, pages 357--366, 2008.
- Smith, D.A. and Crane G. in "Disambiguating geographic names in a historical digital library", 5th European Conference on Research and Advanced Technology for Digital Libraries (ECDL'01), September 2001.
- Li, H., Sihari, R.K., Niu, C., and Li, W. in "Location Normalization for Information Extraction", 19th International Conference on Computational Linguistics, Aug. 2002, Taipei, Taiwan.
- Meenakshi Nagarajan, Kamal Baid, Amit P. Sheth, and Shaojun Wang, *Monetizing User Activity on Social Networks - Challenges and Experiences*, 2009 IEEE/WIC/ACM International Conference on Web Intelligence, Sep 15-18 2009, Milan, Italy.
- Pear Analytics in "Twitter Study August 2009": <http://www.pearanalytics.com/wp-content/uploads/2009/08/Twitter-Study-August-2009.pdf>.
- M. Sanderson and J. Kohler. Analyzing geographic queries. In *Proceedings of Workshop on Geographic Information Retrieval SIGIR*, New York, NY, USA, 2004. ACM Press.
- Brown, L. D., Hua, H., and Gao, C. 2003. A widget framework for augmented interaction in SCAPE. In *Proceedings of the 16th Annual ACM Symposium on User Interface Software and Technology* (Vancouver, Canada, November 02 - 05, 2003)