Traffic Observatory: a system to detect and locate traffic events and conditions using Twitter

Sílvio S. Ribeiro Jr.
Universidade Federal de Minas Gerais, Brazil
silviojr@dcc.ufmg.br

Diogo Rennó R. Oliveira
Universidade Federal de Minas Gerais, Brazil
renno@dcc.ufmg.br

Tatiana S. Gonçalves
Universidade Federal de Minas Gerais, Brazil
tati.sg@dcc.ufmg.br

Clodoveu A. Davis Jr.
Universidade Federal de Minas Gerais, Brazil
clodoveu@dcc.ufmg.br

Wagner Meira Jr.
Universidade Federal de Minas Gerais, Brazil
meira@dcc.ufmg.br

Gisele L. Pappa
Universidade Federal de Minas Gerais, Brazil
glpappa@dcc.ufmg.br

ABSTRACT

Twitter has become one of the most popular platforms for sharing user-generated content, which varies from ordinary conversations to information about recent events. Studies have already showed that the content of tweets has a high degree of correlation with what is going on in the real world. A type of event which is commonly talked about in Twitter is traffic. Aiming to help other drivers, many users tweet about current traffic conditions, and there are even user accounts specialized on the subject. With this in mind, this paper proposes a method to identify traffic events and conditions in Twitter, geocode them, and display them on the Web in real time. Preliminary results showed that the method is able to detect neighborhoods and thoroughfares with a precision that varies from 50 to 90%, depending on the number of places mentioned in the tweets.

Categories and Subject Descriptors

H.3 [**Information Systems**]: Information Storage and Retrieval

General Terms

Experimentation

Keywords

Twitter, Traffic, Geocoding

1. INTRODUCTION

Twitter has become a popular platform for content sharing, where users posts may vary from ordinary conversation to relevant information about events in real time. Previous works have shown that the content generated in Twitter has a high degree of correlation with the real world, and that has led to the development of applications that cover a wide range of events, from epidemics to elections [11, 9].

Traffic updates are very common in Twitter, since many users tweet to inform about problems they have when moving around in the city, reporting problems such as accidents, cars with mechanical problems, demonstrations, among others that may affect traffic. In this scenario, users who are voluntarily informing about traffic conditions can be viewed as sensors of a phenomenon that is evolving in real time. There are even Twitter accounts created specifically to inform about traffic conditions and events in big cities. Some of them are operated by official traffic departments, and are useful sources of information for drivers who follow them. In this scenario, there is a huge amount of unstructured information about traffic spread in different Twitter accounts, and a rising interest in the development of methodologies, techniques and tools that can collect, organize, integrate and publish this information consistently.

Twitter information can be used to complement whatever is generated by cameras and physical sensors, guiding the actions of public agents in promoting traffic improvements. It can also be used to support driver decisions on routing in the cities. The growing popularity of online social networks, especially Twitter, indicates that, in a short time, the information generated and published by the citizens themselves may become the main tool to evaluate traffic conditions in real time.

In this paper, we describe the initial phase of study and implementation of the Traffic Observatory (*Observatorio do Transito*, in Portuguese). Traffic Observatory is a text mining system that works on Twitter's stream, looking for relevant text patterns that indicate the traffic condition in specific locations. The information gathered is available in a web interface, freely accessed by users. The proposed methodology was evaluated using data collected about the city of Belo Horizonte, center of Brazil's third largest metropolitan area, with a population of about 2.5 million people. We achieved promising results on the detection of relevant traffic events and the places where they occur.

The remainder or this paper is organized as follows. Section 2 describes some related work, while Section 3 details the proposed method. Section 4 presents the datasets used in the evaluation process, including the Twitter dataset and the gazetteer used to support the process. Section 5 describes the experiments and their results and, finally, Section

ACM SIGSPATIAL LBSN '12, November 6, 2012. Redondo Beach, CA USA

6 presents conclusions and future work.

2. RELATED WORK

Usually associated to the georeferencing of postal addresses, geocoding is now understood as the process of obtaining locations from descriptions of places [8]. Since such descriptions often include place names, gazetteers, or toponymic dictionaries, constitute important sources of information. Most gazetteers do not include urban detail, i.e., information on street names, neighborhoods or urban landmarks [10], and thus geocoding is harder in urban scales, unless structured postal addresses are used. Even when a proper gazetteer is available, geocoding is a challenging task, because of common problems such as ambiguity: many places can have the same name, and a place name can also be used to reference other entities [5]. Furthermore, the use of abbreviations and simplifications, which are very common in tweets due to space limitations, also complicates the recognition of place names.

Some techniques have been proposed in order to recognize and interpret place names contained in text. Twaroch et al. [12] presents the detection of place names in the Web using expressions related to the context of geographic location, along with a gazetteer. Amitay and Har-El [2] propose a method to solve the ambiguity between place names and with names of other entities. Their approach uses a worldwide hierarchical gazetteer in order to determine a single place within a certain level of confidence, using evidence such as the correlation between an ambiguous name and others that have been previously identified. Delboni et al. [7] propose the recognition of relevant place names in a text by looking at the vicinity of positioning expressions such as "close to" or "at walking distance from". The characterization of synonyms for each of these expressions is also proposed, in an attempt to improve the precision of the response. Cardoso et al. [3] present a prototype of a geographic information retrieval system that aims at capturing implicit location evidence, such as company or building names, and use them along with explicit references to places as a way to improve retrieval results.

Cheng et al. [4] propose an approach to locate Twitter users based on tweet contents. Using only the text of the messages, they developed a probabilistic framework to estimate the location of a Twitter user at the city level. A classifier is used to automatically identify words within the tweets that are strongly related to a local geographic scope, and then user locations can be estimated using a smoothing model that searches for the identified words in the messages. Alencar et al.[1] present a method to extract such location-related words and expressions from Wikipedia. Davis Jr. et al. [6] introduce a method to infer the location of Twitter users at the city level, based on following-follower relationships involving users that authorize the publication of their location.

For this work, it is insufficient to geocode users or tweets at the city level. We need to obtain information as to the specific point within the city to which the message refers. A small share of the tweets originates from mobile devices that, if properly authorized by the user, can associate spatial coordinates to the message, within the locational accuracy of the hardware – and assuming that the tweet has been posted while the user was in the vicinity of the event. In many cases, therefore, the location must be obtained by looking at the text of the messages, extracting references to urban places and landmarks, such as street names and points of interest. The use of a gazetteer, in this situation, is only viable if its contents include fine-grained urban detail, which is unusual [10].

3. DETECTING TRAFFIC CONDITIONS AND EVENTS IN TWITTER

As previously explained, we propose a method to detect and locate traffic events and conditions from tweets. The method comprises four phases: (i) preprocessing of the messages' content, (ii) traffic event identification/detection, (iii) detection of locations using exact string matching, (iv) enhancement of the location information using approximate string matching.

Preprocessing includes removing accent marks (since the tweets being processed are written in Portuguese), links and mentions to other Twitter accounts (e.g., *@BHTrans*). In this stage, messages referring to other cities are identified and removed from the pipeline. This is necessary because one of the accounts used in our experiment (*@WayTaxi*) reports traffic conditions for many Brazilian state capitals, but the name of the city is always explicitly mentioned in each tweet.

In the second step, we identify traffic-related events. In this work, we manually listed the most frequent types of events and terms used to express traffic situations. This static list ensures that we are only considering tweets about traffic conditions. The set of events and expressions used are described in Table 2. In the future, we intend to use machine learning techniques to dynamically detect events.

We classified traffic information into two main categories: condition and event. Conditions refer to the status at a given location at a given moment (e.g., "slow"), while events correspond to situations that may change the traffic status (e.g., "accident"). We tried to cover as many events and conditions as possible, but we are aware that our list is not exhaustive.

In the third stage we perform an exact string matching using a gazetteer to find street and neighborhood names, as described in Section 3.1. Finally, in the fourth step we expand on the results obtained in the previous step using approximate string matching this time, as shown in Section 3.2. We employed exact matching first for performance reasons, since approximate matching is more costly, and we also took advantage of this two-step matching to verify how often exact matching succeeds.

3.1 Detecting locations from tweets

One of the main tasks required by our system is the identification of urban references in tweets. Street names, neighborhood names and other landmarks must be found in the message, and related to geographic locations so that the accumulated results can be presented to users in a consistent way. We used a subset of places from a gazetteer [6], selecting elements contained within Belo Horizonte. This subset included 9,514 thoroughfare names and 40,749 street crossings and their related thoroughfares, along with their geographic representations. We also generated a set of 47,211 thoroughfare segments, dividing each street's geometry at their intersection with neighborhood boundaries, so that we can locate more precisely references of the type "street X at

neighborhood Y".

Even with this gazetteer data, recognizing references to urban locations in tweets is hard. Since tweets are limited to 140 characters, messages usually employ shortened place names, using abbreviations and omitting parts. Furthermore, many typos occur, and sometimes there are references to variations, such as historical versions or popular nicknames, of the names of streets, neighborhoods and landmarks. Because of that, we also used gazetteer data on alternative names, and created a dictionary of common abbreviations for thoroughfare types, such as "Av." for "Avenue". Putting all those resources together, we generated a final set of place names, which we call GEODICT from this point on.

The proposed method initially searches GEODICT using exact string matching. In this stage, we search for substrings from the tweet that can be found in GEODICT. For thoroughfare names, we used two variations: the official name with the thoroughfare type (e.g. "Fleming Avenue") and the name alone (e.g. "Fleming"). Notice that the first variation also includes abbreviated forms of the thoroughfare type (e.g. "Fleming Av."). Both variations were tested, and the results were compared. While the first variation gains in precision, the latter gains in recall.

3.2 Enhancing location data

In this step, we search for street and neighborhood names that are related to the places identified in the previous step, now using approximate string matching. In this case, two situations may happen. If the previously identified place is a street, we try to find names of other related streets (crossings) and neighborhoods (crossed by the street), using information from GEODICT. Likewise, if the previously identified place is a neighborhood, we try to find the names of related streets from GEODICT. With this, we intend to narrow down the position to which the tweet refers, since street crossings are represented by points and street segments within neighborhoods are usually much shorter than the entire throughfare. Furthermore, finding related geographic references reduces the chance of misplaced results.

The string matching technique used in this step is the fuzzy string searching. While the previous technique looks up for substrings that are identical to thoroughfare and neighborhood names listed in GEODICT, fuzzy matching returns a score that varies from 0 (completely different strings) to 100 (completely identical strings), according the to the similarity between the substrings and the place names. Matching is achieved if the score is higher than a predetermined threshold.

Notice that, in this stage, if two streets that have a common crossing are cited in the same tweet, the traffic condition or event is geocoded to the location of that crossing. If the cited streets have more than one common crossing, geocoding is incomplete, and therefore abandoned. Although we have precise information about the geometry of each street, we can only geocode the event or condition when other references help us point to a location with some certainty. Messages such as "Avenue X is really slow today" may be useful to the user only if, by experience, she knows the spots along Avenue X that are usually problematic, but this sort of message does not have enough information that would allow us to pinpoint a traffic event. A more precise location can be obtained from messages such as "Accident on Avenue X at Y".

4. THE TWITTER DATASET

The dataset used to evaluate our method was collected from Twitter through a controlled process. We collected tweets from ten profiles whose main purpose is to inform traffic conditions in Belo Horizonte and other cities in Brazil. These accounts include *@TransitoBH* (a profile that collects volunteered information), *@Transito98FM* (managed by a radio station) and *@waytaxi* (managed by a taxi company). Furthermore, for comparison, tweets from *@OficialBHTrans*, the official account of the city's traffic department, were collected to be considered as ground truth. In the future we are going to contrast these datasets with those obtained by generic collections, from accounts whose purpose is not to inform about traffic. Using information from such accounts will probably be more inclusive, but more problems for the identification of relevant tweets will certainly arise.

Collection took place throughout a period of three months, from April to June 2012, with a total of 10,005 tweets from the selected profiles in 91 days. From those, 1,137 were retweets of messages already collected. Retweets between the selected profiles were removed to avoid repeated texts, thus leaving 8,868 unique tweets. Table 1 presents a comparison between the behavior of the official city's department account (*@OficialBHTrans*) and the other 10 unofficial accounts that also report traffic conditions in the city. It shows the number of tweets in the period, the number of days when there was at least one tweet, the average number of tweets considering the whole period, and the average number of tweets considering active days (days with at least one tweet) from the official and unofficial accounts. It is noticeable that, although the official profile provided an average number of tweets significantly larger than the average number provided by the other profiles, it generated messages in only 47 of the 91 days analyzed. The other profiles covered 88 of the 91 days. This confirms our hypothesis that there is information about traffic spread throughout several profiles, and that consistently gathering all this content will generate better information.

Figure 1 shows the volume of tweets for the analyzed period grouped by hour of the day. We can observe that, as expected, the number of tweets about traffic is higher in the morning, around 9 AM, and in the evening, around 6 PM. Through this observation it is possible to notice a correlation between the frequency of traffic-related tweets and traffic events in the real world, since those times of the day coincide with rush hours.

5. EXPERIMENTS AND RESULTS

Based on the proposed method and the tweets database, we created the Traffic Observatory [1], which shows the current traffic situation in Belo Horizonte. In order to evaluate the accuracy in locating traffic events and conditions in the Observatory, we manually annotated 505 tweets, identifying, for each of them, the names of the thoroughfares and neighborhoods and their associated traffic conditions or events. Furthermore, the tweets have been geocoded and displayed using a kernel density map. The next section discusses these topics.

[1] http://inweb-dev.speed.dcc.ufmg.br/transitobh/

Table 1: Profiles specialized in traffic conditions

Profiles	# of tweets	# of active days	Tweets per day (average)	Tweets per active day (average)
OficialBHTRANS	1,543	47	16.9	32.8
Other profiles	8,462	88	7.7 (per profile)	8.0 (per profile)
Total	10,005	91	109.9	-

Table 2: Most frequent traffic events and conditions found in the dataset

Event/Condition	Number of Tweets	Event/Condition	Number of Tweets
slow	2000	**stopped**	209
accident	582	**free**	198
stuck	499	**jammed**	100
regular	373	**demonstration**	86
intense	305	**blocked**	48
pay attention	277	**complicated**	31

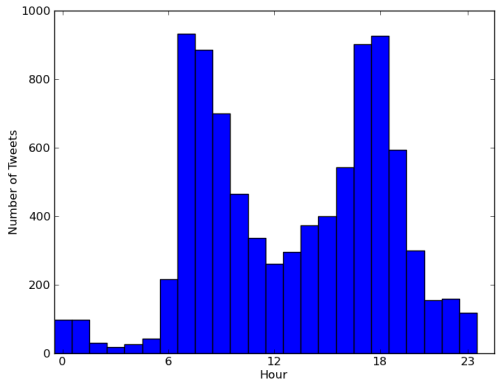

Figure 1: Number of Tweets per Hour

5.1 Identifying Streets and Neighborhoods

Two experiments were performed considering the manually annotated tweets. The first considers all the 505 labeled tweets (this set is shown as *All Tweets* in the related tables), including those for which our method did not find any local. The second considers only the *Classified Tweets*, i.e., the tweets from which our method extracted information about at least one thoroughfare or neighborhood.

For each scenario, we calculated the precision considering hits and partial hits. We consider a *hit* if the set of thoroughfares and neighborhoods found in a tweet is identical to that annotated by humans. A *partial hit*, in contrast, occurs when the set of locations found by the proposed method is a subset of the annotated locations. The evaluation of precision is performed considering three cases: identification of thoroughfares, identification of neighborhoods and identification of the previous two together.

Table 3 presents the precision considering hits and partial hits for the two scenarios described above. It also shows the recall obtained. The recall is calculated as the ratio between the number of tweets with at least one location extracted and the number of tweets that refer to any location in the city (according to human annotators). Results have been obtained using three different methods, as explained below.

The first method, used as a baseline, consists of looking up for exact string matches with the thoroughfare and neighborhood names extracted from the gazetteer, without considering the variations added to GEODICT. The second and third methods are variations of the strategies mentioned in Section 3. Initially, the following experiments were performed: first we used the thoroughfares full name (e.g.,"Afonso Pena Avenue") or its name alone ("Afonso Pena"). Experiments showed that using the full names obtained better results than considering only the name, and for the sake of simplicity, only the successful case is presented here. Knowing that it is better to consider the thoroughfare's full name, we did the same type of experiment for neighborhood. In Brazil, it is common that the name of a neighborhood comes after the word *Bairro*. Hence, we performed one experiment using the word (called *Full Names*, as both thoroughfares and neighborhoods have their full names considered) and another ignoring the neighborhood full name (*No Neighborhood Word*).

Table 3 shows the precision obtained when identifying only thoroughfares (T), only neighborhoods (N) and both neighborhoods and thoroughfares simultaneously (NT). For example, if in a tweet with one street and one neighborhood the method identifies only the street, it counts as a hit for **T**, and a miss for both **N** and **NT**.

The baseline method presents high precision for the partial hits, specially when dealing with thoroughfares and neighborhoods independently. For the hits, however, a low precision is obtained, specially for streets. Besides, its recall is very low, making its use impracticable.

Recall increased significantly when applying the two variations of the method proposed in Section 3. While the percentage of tweets from which we identify at least one location increased from 12% to 86%, the highest precision loss (in partial hits) was around 10%. Furthermore, the precision gain for hits was higher than 47% for *All Tweets*. One of the negative impacts of the method over the exact match was a loss in precision for hits in neighborhoods in the *Classified Tweets*. The main reason of this side effect is the erroneous identification of some neighborhoods due to its common names ("Hills", "Airport"). Finally, the relaxation of the use of the definitions for the identification of neighborhoods and streets provided an increase of 9% in recall over the original method, and a precision gain of 27% for hits considering neighborhoods. This is due to the fact that neighborhoods are often mentioned without the word "Bairro". The precision for hits considering both thoroughfares and neighborhoods increased from 57% to 74%.

Using the third method variation (*No Neighborhood Word*), which presented the best results overall, we analyzed all the tweets – annotated or not –, aiming to characterize their content according to the type of location found. The

Table 3: Identification precision for Neighborhood and Thoroughfares (NT), Thoroughfares (T) and Neighborhood (N) for each method

	Baseline			Full Names			No Neighborhood Word		
	NT	**T**	**N**	**NT**	**T**	**NP**	**NT**	**T**	**N**
All Tweets/Hits	0.29	0.32	0.73	0.57	0.79	0.73	0.74	0.79	0.90
All Tweets/Partial Hits	0.98	0.99	0.99	0.88	0.90	0.98	0.82	0.90	0.92
Classified Tweets/Hits	0.50	0.52	0.93	0.50	0.80	0.66	0.69	0.75	0.87
Classified Tweets/Partial Hits	0.83	0.87	0.96	0.81	0.84	0.87	0.76	0.87	0.89
Recall	0.12			0.86			0.95		

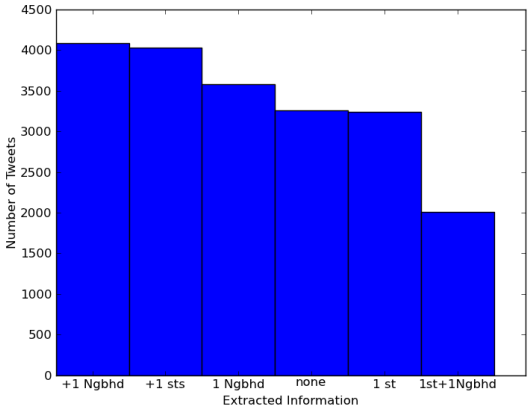

Figure 2: Extracted information

results are shown in Figure 2, where each bar represents a group of tweets. The first group contains tweets where our method found one or more neighborhoods, the second stands for one or more thoroughfares, the third is for exactly one neighborhood, the fourth is for tweets where the method did not find any location, the fifth is for exactly one thoroughfare and the last is for exactly one neighborhood and one thoroughfare. Notice that tweets that reference more than one street support the identification of crossings, which enable us to pinpoint the event's location.

We manually annotated a sample of 3% from each group. In Table 4, we present the precisions according to hits and partial hits for each group (bar) of tweets shown in 2. Notice that our method has a higher accuracy for the group in which we found exactly one street and one neighborhood. In 39% of the tweets containing one or more locations, our method was unable to determine it; therefore, the recall was 0.61.

From the results in Table 4, we observe that at least one street was identified in 81% of the classified tweets, and at least one neighborhood in 72%. Notice that the precision is higher for partial hits when we find more than one street or neighborhood. The best precision is achieved when we find exactly one neighborhood and one thoroughfare.

5.2 Geocoding

Using the crossings and the thoroughfare-neighborhood pairs found by our method, we geocoded the approximate latitude and longitude of the events and traffic conditions. In order to visualize the regions the users most tweet about, we used the tweets in our dataset to generate Figures 3(a) and 3(b). The areas in red represent regions with the highest number of tweets collected. Conversely, the areas in green contain the lowest volume. Comparing Figures 3(a) and 3(b), and considering that most tweets are related to bad traffic conditions (see data in Table 2), there are trends in different regions of the city according to the period of the day. In many regions, these figures are visually similar to maps published online by the traffic authorities, based on sensors embedded in the pavement. However, since these sensors do not cover all streets, a direct comparison is not possible at this point.

5.3 Towards Identifying References and Directions

We used regular expressions to look up for the most frequent expressions that indicate that someone is using a known point of interest (POI) – such as a shopping mall, bar or even nearby streets – in the tweets as a spatial reference. We found that at least 7% of the tweets in our dataset contained references to nearby locations. We validated this result sampling and manually labeling 3% of these tweets. The results showed that 70% of tweets reference a nearby place, while 30% reference a nearby street or neighborhood. We also sampled and labeled 3% of the tweets for which we did not obtain any match using our regular expressions. Results showed that 10% of tweets actually contained some reference to known places, but our method could not find them because there was a ill-formed text or uncommon expression indicating a reference. In future work we intend to incorporate additional data containing the geolocation of POIs, and geocode tweets also taking these references into account. In this way, we expect to increase the number of geocoded tweets.

Another interesting concern in the case of bidirectional thoroughfares is on the recognition of the direction of flow in which there is a traffic problem. We know, for example, that during the morning more problems (e.g., "slow traffic") occur towards downtown, while the opposite is true in the end of the afternoon. We characterized the existence of this information in our dataset. We found that at 29% of the tweets in our dataset contain expressions that indicate the direction in which the problems are occurring. We manually labeled 3% of the tweets in which our system did not find any indication of direction, and found it missed approximately 8% of tweets that actually contained this type of information. Again, most of these errors were due to ill-formed text. In a future work, we plan to use this information to distinguish between the status of both ways of a thoroughfare when direction information is available.

Table 4: Precision for Hits and Partial Hits according to the number of locations our method found in tweets

	Hit (%)	Partial Hit (%)
Nothing	61	61
Only one neighborhood	60	78
Only one thoroughfare	74	76
One thoroughfare + one neighborhood	90	90
At least one neighborhood	48	72
At least one thoroughfare	68	81

(a) Morning

(b) Afternoon

Figure 3: Density of geocoded tweets in two different periods of the day

6. CONCLUSION AND FUTURE WORK

This work proposed a new method for identifying traffic events in current traffic in real time using Twitter. We created the Traffic Observatory, and performed a set of experiments to evaluate the effectiveness of the method and possible future work directions. The method works uses an enhanced gazetteer, which contains urban detail, takes into account popular names of streets and neighborhoods, and performs both exact and fuzzy matching.

One of the first things we intend to do now is to identify the types of events and conditions dynamically, instead of just using a static list. Another interesting direction is to identify tweets that are talking about the same traffic events and conditions in different ways. In order to do that, we first need to characterize the traffic events and conditions. This characterization will also allow us to correlate traffic events to the conditions they may cause.

Furthermore, in this work we analysed the traffic conditions individually. A more challenging research direction is to consider how an event in a given location will impact the region surrounding it, and how long it will take for it to happen. This might be useful to offer users an alternative route selection service, based on traffic simulations and on the seasonality of traffic issues.

Although we can currently geocode many traffic events and conditions, many of them are left uncoded because we do not know exactly in which position in a given street or neighborhood the event happened. A possible way to increase the number of geocoded traffic conditions is to use references to points of interest, such as shopping malls, bars, squares, etc., data on which are to be included in the gazetteer.

Finally, it would be interesting to study the differences of information provided by ordinary users and those specialized in traffic information. It will be interesting to determine how much information we can gain if we also consider tweets from ordinary users. The answer to this question is also the subject of future work.

7. REFERENCES

[1] R. O. Alencar and C. A. Davis Jr. Geotagging aided by topic detection with wikipedia. In S. Geertman, W. Reinhardt, F. Toppen, W. Cartwright, G. Gartner, L. Meng, and M. P. Peterson, editors, *Advancing Geoinformation Science for a Changing World*, volume 1 of *Lecture Notes in Geoinformation and Cartography*, pages 461–477. Springer Berlin Heidelberg, 2011.

[2] E. Amitay, N. Har'El, R. Sivan, and A. Soffer. Web-a-where: geotagging web content. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 273–280, New York, NY, USA, 2004. ACM.

[3] N. Cardoso, M. J. Silva, and D. Santos. Handling implicit geographic evidence for geographic ir. In *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 1383–1384, New York, NY, USA, 2008. ACM.

[4] Z. Cheng, J. Caverlee, and K. Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 759–768, New York, NY, USA, 2010. ACM.

[5] B. Daniel. *Handbook of Research on Methods and Techniques for Studying Virtual Communities: Paradigms and Phenomena*. Number vol. 1. Igi Global, 2010.

[6] C. A. Davis Jr., G. L. Pappa, D. R. R. de Oliveira, and F. L. de Arcanjo. Inferring the location of twitter messages based on user relationships. *T. GIS*, 15(6):735–751, 2011.

[7] T. M. Delboni, K. A. V. Borges, and A. H. F. Laender. Geographic web search based on positioning expressions. In *Proceedings of the 2005 workshop on Geographic information retrieval*, GIR '05, pages 61–64, New York, NY, USA, 2005. ACM.

[8] D. Goldberg, J. Wilson, and C. Knoblock. From text to geographic coordinates: the current state of geocoding. *URISA Journal*, 19(1):33–47, 2007.

[9] J. Gomide, A. Veloso, W. Meira Jr., F. Benevenuto, V. Almeida, F. Ferraz, and M. Teixeira. Dengue surveillance based on a computational model of spatio-temporal locality of twitter. In *Proc. of the 3rd International Conference on Web Science*, pages 1–8, 2011.

[10] I. M. Machado, R. O. de Alencar, R. de Oliveira Campos Jr., and C. A. Davis Jr. An ontological gazetteer and its application for place name disambiguation in text. *J. Braz. Comp. Soc.*, 17(4):267–279, 2011.

[11] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *ICWSM'10*, pages 178–185, 2010.

[12] F. A. Twaroch, P. D. Smart, and C. B. Jones. Mining the web to detect place names. In *Proceedings of the 2nd international workshop on Geographic information retrieval*, GIR '08, pages 43–44, New York, NY, USA, 2008. ACM.