Dengue surveillance based on a computational model of spatio-temporal locality of Twitter*

Janaína Gomide
Computer Science
UFMG - Brazil
janaina@dcc.ufmg.br

Adriano Veloso
Computer Science
UFMG - Brazil
adrianov@dcc.ufmg.br

Wagner Meira Jr.
Computer Science
UFMG - Brazil
meira@dcc.ufmg.br

Virgílio Almeida
Computer Science
UFMG - Brazil
virgilio@dcc.ufmg.br

Fabrício Benevenuto
Computer Science
UFOP - Brazil
benevenuto@gmail.com

Fernanda Ferraz and Mauro Teixeira
Biochemistry and Immunology
UFMG - Brazil
{ferrazicb,mmtex}@gmail.com

ABSTRACT

Twitter is a unique social media channel, in the sense that users discuss and talk about the most diverse topics, including their health conditions. In this paper we analyze how Dengue epidemic is reflected on Twitter and to what extent that information can be used for the sake of surveillance. Dengue is a mosquito-borne infectious disease that is a leading cause of illness and death in tropical and subtropical regions, including Brazil. We propose an active surveillance methodology that is based on four dimensions: volume, location, time and public perception. First we explore the public perception dimension by performing sentiment analysis. This analysis enables us to filter out content that is not relevant for the sake of Dengue surveillance. Then, we verify the high correlation between the number of cases reported by official statistics and the number of tweets posted during the same time period (i.e., $R^2 = 0.9578$). A clustering approach was used in order to exploit the spatio-temporal dimension, and the quality of the clusters obtained becomes evident when they are compared to official data (i.e., $RandIndex = 0.8914$). As an application, we propose a Dengue surveillance system that shows the evolution of the dengue situation reported in tweets, which is implemented in `www.observatorio.inweb.org.br/dengue/`.

Categories and Subject Descriptors

J.4. [**Computer Applications**]: Social and behavioral sciences Miscellaneous; H.3.5 [**Online Information Services**]: Web-based services

General Terms

Human Factors, Measurement, Web, Data Mining

*This work was partially supported by CNPq, CAPES, FAPEMIG, InWeb, and INCT on Dengue.

WebSci '11, June 14-17, 2011, Koblenz, Germany.

Keywords

Twitter, Dengue, Surveillance, Spatio-temporal data mining

1. INTRODUCTION

Twitter is amongst the new media channels that are challenging how we communicate. Users discuss and talk about the most diverse topics, including their health conditions. In this paper we analyze how dengue epidemics reflected on Twitter and to what extent that information can be used for surveillance. Dengue is a mosquito-borne infection that causes a severe flu-like illness, and sometimes a potentially lethal complication called dengue hemorrhagic fever. Indeed, despite massive media advertisement and expensive household control measures, governments have failed to decrease the prevalence of the mosquito and dengue epidemics [5].

Traditional disease surveillance comprises a set of epidemiological procedures that monitor the spread of a disease and determine how it is spreading. Surveillance reports quantify the number of cases for each location and time period. Rapid turnaround surveys are undoubtedly an essential tool for providing timely and sensitive information to public health agencies. However, for several diseases, including dengue, such surveillance results usually take weeks to be made public, when it is often too late for taking effective counter measures. Fortunately, social media channels, such as Twitter, offer a continuous source of epidemic information, arming public health agencies with the ability to perform real-time surveillance.

In this work we propose, for the sake of surveillance, to consider not only ***how much***, ***where*** and ***when*** dengue incidence happened, but also an additional dimension enabled by social media, which is ***how*** the population faces the epidemics. We then introduce an active surveillance framework that analyzes how social media reflects epidemics based on four dimensions: volume, location, time, and public perception. Further, we address a fundamental question that is to what extent tweets support spatio-temporal predictions of dengue cases and propose a methodology for such assessment, which is based on comparing different patterns and models for both actual disease reports and Twitter data. In particular, for predicting the volume we employ regres-

sion techniques, while the spatio-temporal predictions are based on clustering. Regarding public perception, we analyze how users refer to dengue in his/her tweets, again considering both time and location dimensions, and applying state of the art techniques for sentiment analysis. We apply the methodology to dengue-related data from Brazil, spreading from 2009 to 2011 and show that twitter is a powerful source of information. We also present and validate a simple surveillance system that is built on top of our framework.

In summary, this work has the following contributions: (i) proposal and application of a four-dimensional framework for assessing the use of social media data in active surveillance; (ii) proposal, implementation and evaluation of an active surveillance system; (iii) instantiation of both the framework and the surveillance system for dengue.

2. BACKGROUND ON DENGUE

Dengue is caused by a virus - the Dengue virus - that belongs to the family Flaviviridae and present four known viral serotypes. Infection of humans causes disease of varying severity, from classical Dengue Fever (DF) to severe dengue infection and death [17]. The disease is very significant in tropical and subtropical regions of the planet and affects more than 100 countries in developing and underdeveloped countries. The World Health Organization estimates approximately 2 billion people are at risk of infection and about 50 million infections occur globally every year. In 2009, 18,000 deaths occurred according to the WHO, although exact numbers are not known due to underreporting [28]. The growth of the dengue epidemic is exponential, not only in terms of number and severity of cases but also in the number of new areas on to which the disease is now found. Epidemiological vigilance coupled to adequate control measures are the mainstay of the prevention of infectious disease outbreaks, especially for those in which vaccines are not available. Environmental and social features involved in disease dynamics on a certain population are important factors when considering the best epidemiological and control measures against an outbreak. In the context of dengue infection, spatial and temporal dynamics of the disease are pivotal factors that must be considered in surveillance actions. In most dengue endemic areas the transmission is seasonal and increases during the wet season. Outbreaks generally occur in inter-epidemic periods. Thus, surveillance systems must be capable of detecting these events in order to provide reliable indicators to orient control actions. The development of a population-based system designed to monitor spatial and temporal distribution of dengue and also capable of registering the clinical forms of dengue cases would have enormous public health importance for establishing the priority areas for intervention and the alert threshold point in outbreaks [23]. The strategy and the system proposed in this work aim to fulfill such demand in the context of dengue.

3. RELATED WORK

The Web has attracted lots of attention because of the immense publicly accessible data and its real-time character. Some examples of Web data are the user generated content such as the messages posted in online social medias or the queries submitted to a Web site that can be used to know what is talked or discussed, or what are the interests of the user.

Recently, some articles have demonstrated how social media content like chatter from Twitter can be used to predict real-world events. In [24] Twitter users are considered as sensors and are used to create a mechanism for the detection of earthquakes. They produced a probabilistic spatio-temporal model that can find the center of the earthquake and its trajectory. Their approach was able to send alerts faster than meteorological agencies. Similarly, Asur and Huberman [3] demonstrated how messages from Twitter can be used to forecast box-office revenues for movies. They propose a model based on the rate at which tweets are created about particular topics and show that their approach was able to outperform market-based predictors. In [25], Tumasjan *et al.* showed that opinions identified in tweets related to the German federal election are able to reflect offline political sentiment.

The Web is an important source of health information as people use it to search for information about specific diseases or medical information, especially about infectious diseases. As a consequence, the volume of Web search queries provides an alternative to disease surveillance. Although Eysenbach [12, 13] had no access to search engine query logs, he has collected the number of clicks on a "sponsored link" from Google Adsense, which appeared for Canadian searches only, who entered "flu" or "flu symptoms". He used this data to accurately anticipate the Flu reports collected by the Public Agency of Canada. More recently, Google developed a system to predict seasonal flu activity based on search queries that indicate influenza-like illness activity [14]. By applying a linear regression fit with official health reports, they achieved on average a linear correlation of 90%, predicting influenza rates 1-2 weeks in advance. Similarly, in [22] the frequency of influenza-related queries on Yahoo! search engine has been proven to be correlated with influenza and mortality rates in the United States. In [18] user queries on a Swedish medical website with influenza and influenza-like illness words have been used to learn ILI rates in Sweden. Web search queries were also used for other purposes. Goel *et al.* [15] showed that the subject associated with consumers searches on the Web can predict their future collective behavior with days or even weeks in advance. They used search query volume and linear regression models to forecast the box-office revenue, sales of video games, and the rank of songs.

The messages posted on Twitter with influenza-like words were recently used as a real-time indicator of Flu activity [21, 20]. They compared the volume of these tweets with data from the Health Protection Agency of United Kingdom to obtain a linear correlation greater than 95%. In references [7, 1] a framework, namely SNEFT, is introduced as a continuous data collection engine that combines the detection and prediction capability of social networks in discovering real world flu trends and use tweets to correlate to official statistics. The content of the messages in Twitter are analyzed in [8], where the authors monitor the use of terms related to Flu and validate Twitter as a real-time content, sentiment, and public attention trend-tracking tool.

Complementary to the above studies, our work aims at using the user generated content available on the Web to predict a real-world event. The event we are interested is the dengue epidemic, a neglected tropical disease that, to the best of our knowledge, has never been explored in the context of the Web. Additionally, differently from the above efforts, we propose a framework that not only considers the linear regression on the volume of data. Our dengue surveillance approach considers four dimensions: volume, location, time, and public perception. Next we describe our methodology.

4. METHODOLOGY

In this section we propose a methodology to perform active dengue surveillance based on four dimensions that are associated with Twitter data: volume, location, time, and public perception. Volume represents the amount of tweets mentioning the word "*dengue*", and it can be used to approximate dengue incidence rate. Location is the geographic information associated with tweets mentioning the word "*dengue*". Time refers to when tweets mentioning the word "*dengue*" were posted. The last dimension, content, is the overall population perception/sentiment about dengue epidemics. In order to implement our four-dimension surveillance technique, we consider Twitter as a constant stream of epidemiological data, which is freely reported by the online population.

4.1 Content analysis

Content analysis may provide important clues about public perception, that is, the attitude associated with tweets mentioning dengue. Classification techniques may be used to estimate (or score) sentiments expressed in tweets. First, we create sentiment categories, and then we use a selective sampling approach [19] to create a representative training dataset. Finally, a classifier uses this dataset in order to estimate the likelihood for each category.

We employed the same taxonomy proposed in [8] in order to assess the population's aggregate perception about dengue epidemics. The taxonomy is composed of 5 sentiment categories:

1. Personal experience: tweets expressing dengue cases (i.e., "You know I have had dengue?").
2. Ironic/sarcastic tweets.
3. Opinion: tweets expressing the opinion about some fact related to dengue (i.e., "Very cool the campaign against dengue. The small gestures that end up preserving the lives of many people").
4. Resource: informative tweets (i.e., "Dengue virus type 4 in circulation").
5. Marketing: tweets repercuting public campaigns (i.e., "All against dengue. Brazil relies on you").

A selective sampling strategy [19] was carried in order to build a small, but representative training dataset. Then, an associative classifier [27], which is shown to produce accurate models even when only few training examples are made available, produces a sentiment model, which is extracted from the training dataset. The classifier employs association rules mapping textual-patterns to specific sentiment categories. These rules have the form $\{x \rightarrow y\}$, where x is a textual-pattern and y is any of the categories in our taxonomy. Rules are first extracted from the training dataset, and then each rule $\{x \rightarrow y\}$ is interpreted as a vote given by x for y. These votes have different weights, depending on the confidence [2] of the corresponding rules. The weighted votes for sentiment y_i are summed, giving the score for sentiment y_i with regard to a specific tweet m, as shown in Equation 1:

$$s(m, y_i) = \sum \theta(x \rightarrow y_i) \quad (1)$$

Finally, the scores are normalized, as expressed by the scoring function $\hat{p}(y_i|m)$, shown in Equation 2. It is important to notice that the same tweet may be simultaneously associated with different sentiment categories, and thus, instead of predicting a single category, we employ a scoring function (shown in Equation 2) which estimates the likelihood of sentiment y_i being the implicit attitude of tweet m.

$$\hat{p}(y_i|m) = \frac{s(m, y_i)}{\sum_{j=0} s(m, y_j)} \quad (2)$$

4.2 Correlation analysis

Now we will consider the volume dimension. Basically, we want to fit a regression model that may approximate dengue incidence rate. We fit linear models using the amount of tweets mentioning the word "*dengue*", in order to predict the official amount of cases reported by the Health Ministry. Regression models are further enhanced by considering only tweets expressing high levels of personal experience. The intuition is to improve our model by considering only tweets related with real dengue cases. Our models are based on one of the following variables:

1. the volume of tweets related to dengue, posted by Brazilian users
2. the proportion of tweets expressing personal experience, posted by Brazilian users, given as:

$$PTPE = \frac{\sum \hat{p}(\text{Personal Experience}|m)}{\#tweets}$$

The first regression model simply uses the volume of tweets in order to predict the amount of dengue cases, in the same time period, as shown in Equation 3:

$$\#cases_t = \beta_0 + \beta_1 \times \#tweets_t + \beta_2 \times \#tweets_{t-1} + \epsilon \quad (3)$$

where variable $\#tweets_t$ gives the number of tweets mentioning the word "*dengue*" posted on time period t. The second model refines the first one by considering only tweets

expressing personal experience (i.e., PTPE), as shown in Equation 4:

$$\#cases_t = \beta_0 + \beta_1 \times PTPE_t + \beta_2 \times PTPE_{t-1} + \epsilon \quad (4)$$

4.3 Spatio-temporal analysis

When the epidemics location is early detected, it becomes possible to enhance health assistance, by properly intensifying disease control measures. In order to enable government agencies to concentrate efforts on critical locations in the right time, we must discover groups of cities that are near each other, and having similar dengue incidence rates at a given point of time.

In this sense, location and time dimensions are used to perform spatio-temporal predictions by means of clustering. Specifically, we want to find close cities with similar dengue incidence rates at the same time. Therefore, groups of cities are created by taking into account the spatial distance, as well as the difference in dengue incidence at a given point of time. To this end, we used a state-of-the-art spatial clustering technique, namely ST-DBSCAN [4], which does not require apriori specification of the number of clusters. This algorithm is based on DBSCAN [11] and it has the ability of discovering clusters with arbitrary shape. Further, it determines clusters according to non-spatial, spatial and temporal information. In our context, non-spatial information consists of dengue incidence rates for each observed city (these rates are approximated by the proportion of tweets expressing personal experience − PTPE value). Spatial information consists of the city location, given by its latitude and longitude. Finally, temporal information consists of the month in which the corresponding incidence rate was observed.

For a cluster to be formed, it is necessary a minimum number of cities ($MinPts$) that are near each other (the distance between cities must be $< Eps1$) and with the similar dengue incidence rates (difference between dengue rates associated with two cities must be $< Eps2$) in the same month. In this way, $MinPts$, $Eps1$, $Eps2$ are the three input parameters needed by the clustering algorithm.

4.4 Surveillance

Surveillance systems collect and monitor data for disease trends in order to detect outbreaks as fast as possible, and also to guide immediate actions, prioritizing the allocation of health resources and enhancing both the timeliness and quality of information provided [5, 23]. We propose a surveillance approach using Twitter data that considers the four dimensions already discussed. Basically, we aim to analyze the proportion of tweets expressing personal experience (i.e., PTPE value) in a weekly basis, and for all cities in Brazil. The intuition is that, an abrupt increase on PTPE may indicate outbreaks in the corresponding cities.

A simple graphic model based on heat maps is able to capture variations on PTPE. In this case, weeks associated with colors tending to red may indicate dengue outbreaks occurring on the corresponding city.

5. DATASETS

In this paper, we employ data collected from two different sources. One is the official dengue reports made available by the Brazilian Health Ministry. The other is composed of Twitter messages mentioning the word "*dengue*". Next, we describe in details both datasets and also discuss some of their limitations.

5.1 Official dengue reports

To measure the occurrence of dengue cases, we used reports from the Brazilian Health Ministry related to the dengue epidemics. This data provide regional statistics for Brazil, based on the notified cases of dengue. Particularly, we used the publicly available historical data about dengue [9]. This dataset contains the number of dengue cases per city, notified between 2007 and 2010. As this dataset is not frequently updated, it was not possible to obtain data from 2011.

5.2 Twitter

This dataset is composed by Twitter messages (tweets) related to dengue posted in two distinct periods: from 2006 (when Twitter was created) to July 2009 and from December 2010 to April 2011. The first set of tweets was obtained from a previous measurement study that collected the complete history of tweets posted by all users as of July 2009 [6]. The second part consists of tweets we collected through Twitter Streaming API [26]. We define the tweet as related to dengue if it contains in its text the word "*dengue*".

The first dataset has 27,658 tweets related to dengue, out of which 90.27% are from 2009. As there are not enough messages before 2009, we decided to consider only tweets from January 2009 until May 2009, which is the dengue season in Brazil. The second dataset contains 465,444 tweets, and some of them were posted during the dengue 2011 season in Brazil. The number of tweets and users from both datasets are shown in Table 1.

From these datasets we are particularly interested in three pieces of information. The first is the text tweet, used to analyze the public perception about dengue. Also, the exactly time the tweets were posted and the user location is crucial to know exactly when and where people commented about dengue. Text and time were obtained directly from the Twitter API, whereas the user location requires one last step to be gathered properly.

The location information written in user profiles of Twitter is in free text form and often contains invalid location like "Mars" or "everywhere", making it difficult to automate the process. We filtered out invalid locations and inferred plausible locations of users by using the Google Geocoding API [16], which converts addresses or city names written in free text form into geographic coordinates of latitude and longitude. As our study is restricted to users who are posting from Brazil, we only consider tweets that are posted from Brazilian users. Table 1 shows the information of both datasets and the fraction of tweets with user location identified.

5.3 Data limitations

Although both datasets described above give us a unique opportunity to use the Web as a surveillance tool for dengue,

Table 1: Tweets with location details

	Jan2009 - Mai2009	Dez2010 - Apr 2011
Tweets	12256	465444
without location	2310 (18.85%)	107354(23.06%)
from Brazil	3242(26.45%)	280770(60.32%)
with city	2447(19.97%)	181845(39.07%)
Users	11643	458045
without location	2206(18.95%)	105926(23.13%)
from Brazil	3100(26.63%)	275318(60.11%)
with city	2338(20.08%)	178045(38.87%)

these datasets have a few limitations. First, the dataset containing official dengue reports that is made available by Brazilian Healthy Ministry might not contain all dengue cases occurred in Brazil, but only the registered by the doctors and reported by the government. Second, many tweets are discarded due to the lack of user location, leading us to discard from our analysis the cities with very few tweets. Finally, user location was not obtained for each individual tweet, but for each user. Thus, our analysis ignores user mobility.

6. EXPERIMENTAL EVALUATION

In this section, we present the evaluation and experimental results for the four-dimension methodology proposed in Section 4, which includes content, correlation and spatio-temporal analysis. Further, we present surveillance results.

6.1 Content analysis

For content analysis, we considered Twitter datasets discussed in Section 5.2. First, we want to verify the sentiment distribution for both datasets over time. Figure 1 shows, for each month, the percentage associated with each sentiment. Content analysis is extremely useful, either to capture public perception about dengue, and also to reduce noisy for the sake of surveillance, since we may consider more heavily tweets expressing personal experience (instead of considering all tweets indiscriminately). This selection is mandatory for dengue surveillance using Twitter, because it enables us to focus only on tweets that are related to dengue cases.

Next, we want to verify the discrepancy in the proportion of tweets expressing personal experience – PTPE, coming from representative cities that actually showed different dengue incidence rates (according to official reports). In the beginning of 2011, the city of Manaus suffered with a sudden outbreak, which resulted in many deaths. During the same period, the city of Belo Horizonte reported a decrease in dengue cases (when compared to the last year). Figure 2 shows the variation in PTPE values for Manaus and Belo Horizonte from December 2010 to April 2011. As it can be seen, for all months considered, Manaus has larger PTPE values when compared to PTPE values associated with Belo Horizonte.

6.2 Correlation analysis

In order to understand whether activity on Twitter indeed reflects dengue incidence in Brazil, we examine linear regression models discussed in Section 4.2. The first model was fit simply using the volume of tweets mentioning the word

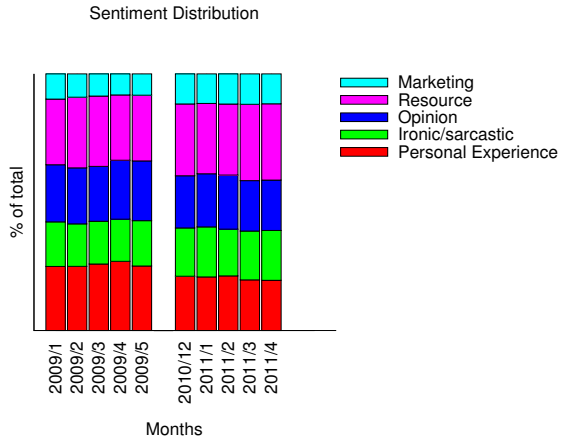

Figure 1: Sentiment distribution over 2009 dengue season and from Dec-2010 until Apr-2011.

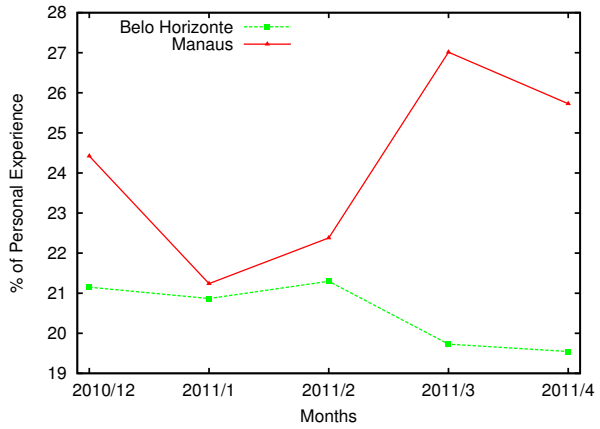

Figure 2: Sentiment distribution from Dec-2010 to Apr-2011

"*dengue*" each month. The second model was fit using only PTPE values, thus reducing the impact of tweets that are not relevant for the sake of dengue surveillance.

For this experiment, we used only the Twitter dataset collected from Jan-2009 to May-2009, since there is no official data reported by the Brazilian Health Ministry for 2011. Results are shown in Table 2. As it can be seen, the model that was produced using all tweets mentioning the word "*dengue*" is well correlated with official reports, but the model becomes much better when we use only PTPE values. In conclusion, we can accurately estimate the current level of monthly dengue activity using tweets expressing personal experience.

Table 2: Regression-analysis results

Model	R^2
Tweets	0.7829
Personal experience	0.9578

6.3 Spatio-temporal analysis

Geographic and time dimensions were, again, analyzed using Twitter data collected from Jan-2009 to May-2009, since there is no official data reported by the Brazilian Health Ministry for 2011. For this experiment, we first selected, for each month, only the cities that appear both in the Twitter and in the official data. This selection was performed in order to guarantee that both datasets have the same cities: 332 cities remained after this selection. Then, we calculate the incidence rate associated with each of the 332 cities, by dividing the official number of dengue cases by the corresponding population. This way, we have, for each city, the proportion of the population that are infected by dengue. Similarly, the number of tweets associated with each city was also divided by the population of that city. After that, a normalization was made to have both rates of dengue with values between 0 and 1. Again, we also use PTPE values (associated with each city) in order to consider only the relevant content for the sake of dengue surveillance.

Third, before running ST-DBSCAN, it is necessary to determine its input parameters. A simple heuristic is presented in [4], which determines parameters Eps and $MinPts$. The heuristic suggests $MinPts \approx \ln(n)$ where n is the size of the dataset, and Eps must be chosen depending on the value of $MinPts$. The first step of the heuristic method is to determine the distances to the *k-nearest neighbors* for each point (i.e., city), where k is equal to $MinPts$. Then this *k-distance values* should be sorted in descending order. Finally, we should determine the threshold point which is the first "valley" of the sorted curve. Also, we should select Eps to be less than the distance defined by the first valley.

In our context, the value for the parameter $MinPts$ is set to 2 because we want to find clusters with at least two cities with similar incidence rates. The value of the maximum distance between the cities, $Eps1$, and the maximum difference between dengue incidence, $Eps2$, were determined by the heuristic described above. The minimum number of points and the distance between the cities are the same for both datasets and as incidence rates were normalized (i.e., values are between 0 and 1), it has the same value too. The input parameters assigned as $Eps2 = 0.025$ when using the entire Twitter data, and $Eps2 = 0.037$ when using only PTPE values. Parameters $Eps1 = 2.5$, and $MinPts = 2$ are equal for both datasets.

The clusters found using the entire Twitter data are very correlated with the clusters found using the official data. The mean value for Rand Index is 0.8506 (min= 0.7887 and max= 0.9310). By considering only tweets expressing personal experience, the clusters found are strongly correlated with the clusters found using the official data. Specifically, the mean value for Rand Index grows to 0.891437 (min= 0.8327 and max= 0.9284). Again, we were able to obtain much better results by considering only PTPE values. The clusters formed with the official dengue statistics at monthly basis, as well as with PTPE values data are shown in Figure 3.

The number of clusters obtained in each month and the mean number of cities per cluster are shown in Table 3. As can be seen, the number of clusters obtained from both Twitter and official data are similar as well as the average of cities per cluster.

Table 3: Number of clusters and average of cities per cluster and number

Months	Twitter dataset		Official dataset	
	#Cluster	Average of Cities	#Cluster	Average of Cities
1	7	3	4	4
2	6	3	5	4.8
3	11	13	10	4.5
4	18	4.89	19	4.63
5	19	5.58	15	7.2

6.4 Surveillance

In our experiments for evaluating the surveillance ability using Twitter, we consider the top 22 cities in terms of number of tweets. These tweets correspond to roughly 60% of the total of tweets in our dataset (i.e., Twitter data from Dec-2010 to Apr-2011), as shown in Figure 4. Our surveillance approach can be viewed as a heat map that shows weekly PTPE values for each city. We show the results obtained using Twitter data in Figure 5.

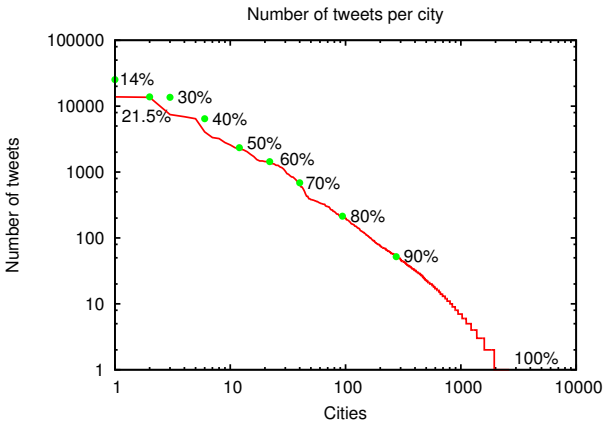

Figure 4: Number of tweets per cities in logscale

Although there is no official dengue reports for 2011, we can use a document provided by Brazilian Health Ministry that summarizes dengue situation in Brazil from January to March 2011 [10], to analyze the effectiveness of the surveillance approach proposed.

Brazil is divided into 26 states, which are grouped into five regions: North (N), Northeast (NE), Middlewest (MW), Southeast (DE) and South (S). About 68% of dengue cases notified concentrates in 7 states which are all represented in Figure 5 by one or more cities. Some examples that we can highlight by contrasting Figure 5 and the epidemiological report summary are described in the following. For instance, in the North region, the two cities with highest dengue incidence were Manaus and Rio Branco. Both cities show high PTPE values during most of the weeks. In the Northeast region, the city that had the highest incidence was Fortaleza,

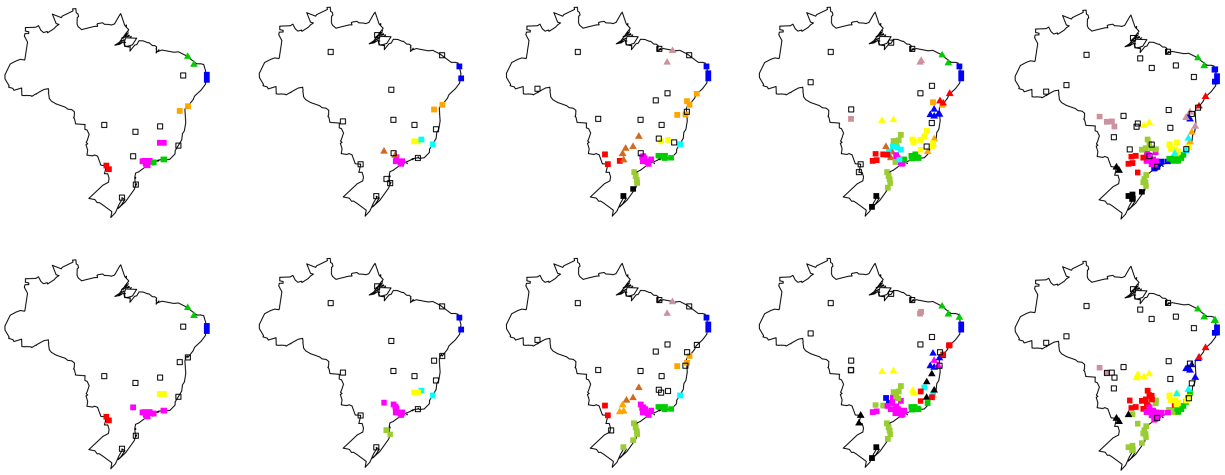

Figure 3: Clusters of cities with similar rates of dengue in the period from dengue season of 2009. The first line are the clusters that were obtained considering the percentage of personal experience tweets and the second line with data from official dengue reports and each map from left to right are the months from Jan-2009 to May-2009.

which, according to our surveillance approach, has a great number of weeks with high PTPE values. Another interesting example concerns the cities of Campinas and Ribeirão Preto, both cities are in the state of São Paulo, and they concentrate the highest dengue incidence in this state.

These techniques are implemented in `www.observatorio.inweb.org.br/dengue/`

7. CONCLUSIONS

In this paper we show the potential of Twitter data for the sake of dengue surveillance. We proposed a methodology that is based on four dimensions: volume, location, time and content. Specifically, we speculate how users refer to dengue in Twitter with sentiment analysis and use the result to focus only on tweets that somehow express personal experience about dengue. Then we constructed a highly correlated linear regression model for predicting the number of dengue cases using the proportion of tweets expressing personal experience. We showed that Twitter can be used to predict, spatially and temporally, dengue epidemics by means of clustering.

Finally, we propose a dengue surveillance approach, that is a weekly overview of what is happening in each city compared with the weeks before. While in this study we focused on the dengue disease in Brazil for the sake of having the data from Brazilian Health Ministry, this method may be extended to other countries and other diseases.

8. REFERENCES

[1] H. Achrekar, A. Gandhe, R. Lazarus, S. Yu, and B. Liu. Predicting flu trends using twitter data. In *International Workshop on Cyber-Physical Networking Systems*, 2011.

[2] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *ACM SIGMOD International Conference on Management of Data*, pages 207–216. ACM, 1993.

[3] S. Asur and B. A. Huberman. Predicting the future with social media. *CoRR*, abs/1003.5699, 2010.

[4] D. Birant and A. Kut. St-dbscan: An algorithm for clustering spatial-temporal data. *Data Knowl. Eng.*, 60:208–221, January 2007.

[5] Centers for Disease Control and Prevention. `http://www.cdc.gov/dengue/`.

[6] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *International AAAI Conference on Weblogs and Social Media*. AAAI Press, May 2010.

[7] L. Chen, H. Achrekar, B. Liu, and R. Lazarus. Vision: towards real time epidemic vigilance through online social networks. In *ACM Workshop on Mobile Cloud Computing Services: Social Networks and Beyond*, pages 1–5. ACM, 2010.

[8] C. Chew and G. Eysenbach. Pandemics in the age of twitter: Content analysis of tweets during the 2009 h1n1 outbreak. *PLoS ONE*, 5(11):e14118, 11 2010.

[9] DATASUS Dengue. `http://bit.ly/dGtFst`.

[10] Epidemiological report summary on Dengue. `http://portal.saude.gov.br/portal/arquivos/pdf/informe_dengue_2011_janeiro_e_marco_13_04.pdf`.

[11] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *International Conference on Knowledge Discovery and Data Mining*, pages 226–231. AAAI Press, 1996.

[12] G. Eysenbach. Infodemiology:tracking flu-related searches on the web for syndromic surveillance. In *AMIA Annu Symp Proc.*, pages 244–248, 2006.

[13] G. Eysenbach. Infodemiology and infoveillance: Framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the internet. *J Med*

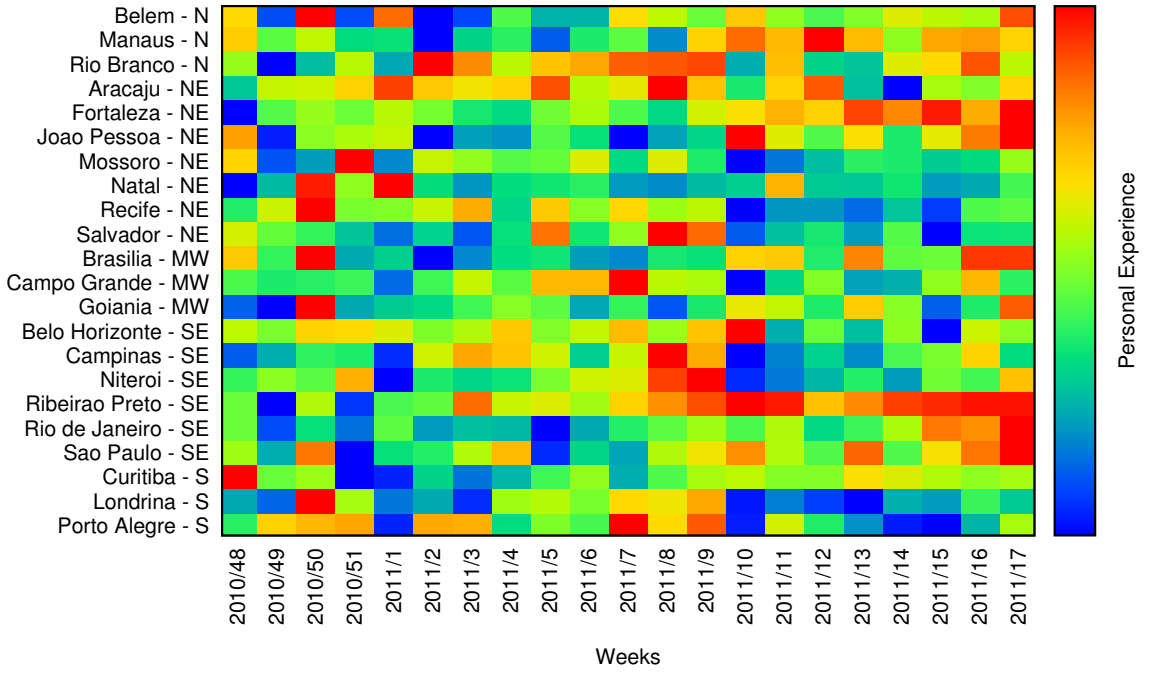

Figure 5: Surveillance approach: heat map with PTPE values for 22 cities during the weeks of Dec-2010 until Apr-2011

Internet Res., 11:e11, 2009.

[14] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457:1012–4, 2009.

[15] S. Goel, J. M. Hofman, S. Lahaie, D. M. Pennock, and D. J. Watts. Predicting consumer behavior with web search. *Proceedings of the National Academy of Sciences*, 107(41):17486–17490, October 2010.

[16] Google Geocoding API. `http://code.google.com/intl/en/apis/maps/documentation/geocoding/`.

[17] S. B. Halstead. Dengue. In *Lancet*, pages 1644–1652, 2007.

[18] A. Hulth, G. Rydevik, and A. Linde. Web queries as a source for syndromic surveillance. *PLoS ONE*, 4(2):e4378, 02 2009.

[19] J. Kivinen and H. Mannila. The power of sampling in knowledge discovery. In *ACM SIGACT- SIGMOD-SIGART Symposium on Principles of Database Systems (PODS)*, pages 77–85. ACM, 1994.

[20] V. Lampos and N. Cristianini. Tracking the flu pandemic by monitoring the social web. In *Workshop on Cognitive Information Processing (CIP 2010)*, pages 411–416. IEEE Press, 2010.

[21] V. Lampos, T. De Bie, and N. Cristianini. Flu detector: tracking epidemics on twitter. In *European conference on Machine learning and knowledge discovery in databases*, pages 599–602. Springer-Verlag, 2010.

[22] P. M. Polgreen, Y. Chen, D. M. Pennock, F. D. Nelson, and R. A. Weinstein. Using internet searches for influenza surveillance. *Clinical Infectious Diseases*, 47:1443–1448, 2008.

[23] S. Runge-Ranzinger, O. Horstick, M. Marx, and A. Kroeger. What does dengue disease surveillance contribute to predicting and detecting outbreaks and describing trends? *Tropical Medicine International Health*, 13(8):1022–1041, 2008.

[24] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *International conference on World wide web*, pages 851–860. ACM, 2010.

[25] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *International AAAI Conference on Weblogs and Social Media*. AAAI Press, 2010.

[26] Twitter Streaming API. `http://apiwiki.twitter.com/`.

[27] A. Veloso, W. Meira Jr., and M. J. Zaki. Lazy associative classification. In *International Conference on Data Mining*, pages 645–654. IEEE Computer Society, 2006.

[28] World Health Organization. `http://www.who.int/tdr/diseases/default.htm`.