Using Tweets For Rainfall Monitoring

[1,2]Luiz Eduardo Guarino de Vasconcelos, [2]Eder C. M. dos Santos, [1]Mário L. F. Neto, [1]Nelson Jesuz Ferreira, and [1, 2]Leandro Guarino de Vasconcelos

[1]Brazilian Institute for Space Research (INPE), Brazil
[2]FATEA, Lorena, Brazil

{du.guarino, sep4eder, le.guarino}@gmail.com

Abstract. In Brazil, the summer season is the wettest period in which many disasters can happen, such as landslides and floods. In recent years, even with the rainy season, the metropolitan region of São Paulo (Brazil) suffers a severe water crisis. Given this scenario, monitoring of rainfall is fundamental for taking preventive actions and planning in the various business branches. Thus, the use of computers to develop tools that assist the rainfall monitoring can help extend the coverage of the existing solutions. Moreover, it is known that, every day, the number of social media users is increasing, and consequently increases the amount of content published in these medias. The objective of this study is to analyze the contents of the Twitter social media, especially the tweets related to rainfall events in order to determine whether this information can contribute to the monitoring of rainfall events in Brazil. More than 1 million tweets published in Brazil related to rainfall were collected in a period of 30 days. Gathered tweets were analyzed and evaluated taking into account the data collected by automatic weather stations (AWS or EMA). The results were satisfactory and indicate a relationship between the geolocated tweets and data from AWS.

Keywords : twitter; hashtags; geolocation; social media analysis

1 Introduction

The number of people connected to the Internet has increased. According to Internet World Stats [1], 3.2 billion people in the world have connected the Internet. Among these people, more than 2 billion have profiles on social media and messaging services [2]. This number represents 27% of the world population, which is approximately 7.2 billion people in 2015. In Brazil, the population is approximately 204 million people [3] and 54% of the brazilian population access the Internet. Furthermore, there are 96 million active social networking accounts in the country.

Through content and data posted on social media, it is possible to collect, analyze and extract information and knowledge that can be useful to public institutions, companies and any citizen. In addition, the majority of people feel the desire and / or need to publish content on the Internet, both to show a moment of personal life as moments related to occupation.

Due to the success of social media in Brazil, especially Twitter, there was the hypothesis to monitor rainfall through tweets. The hypothesis arose from the need to improve rainfall estimation accuracy in the Brazilian current scenario, after all, this estimate is crucial in several business fields, such as water management, agriculture, disaster prevention, and others.

The most used tools in the rainfall monitoring are rainfall networks and remote sensing. Surface networks represent the intensity of rainfall on time with high reliability, but they have problems to represent its spatial variability [4, 5, 6]. Measures with rain gauges can be influenced by exposure to wind, which can cause up to 20% underestimation of the measurements [7].

Remote sensing is an alternative. Weather radars are able to represent the spatial structure of rainfall systems, but they have several sources of error inherent in the rainfall intensity [8]. The main problems are related to droplet spectrum, the presence of bright band, sampling problems for gate, refractive index of the atmosphere (anomalous propagation), etc. [9, 10, 11]. Also, the radars are costly and do not cover the whole national territory. Surface measures networks and radar rainfall estimates can be combined to reduce the magnitude errors [6].

Given the above, this work aims to analyze the contents of the Twitter social media, especially the tweets related to rainfall events in order to determine whether this information may contribute to the rainfall monitoring in Brazil. For this, an experimental application was developed by the Division of Satellites and Environmental Systems (DSA) of Weather Forecast and Climate Studies Center (CPTEC), under the Brazilian Institute for Space Research (INPE). The DSA meets the demand for information from weather satellites in Brazil. The application developed gathers Twitter data for specific hashtags and/or search terms. Tweets which have geolocation information are displayed on a map in order to facilitate visual analysis.

To validate the experiment, the rain gauge data are shown on the map, collected in near real time, provided by the National Institute of Meteorology (INMET) [12]. There was also a visual analysis with radar data provided by the Meteorological Network (REDEMET) of Aeronautics [13]. The results of initial experiments are satisfactory and are presented in Section III.

2 Theoretical

2.1 Social Networks

A social network is composed by people or organizations connected to a computer network [14].

Recuero [15] also reported that two elements define a social network: actors (individuals, institutions or groups) and their connections (interactions or social ties).

Social networking sites like Facebook, YouTube and Twitter have users or company profiles, which can be considered actors in a social network. It is known that there is a lot of content generated from the connections between the actors. The interactions can be exemplified by comments made by some users from posting or even by the actions linked to the post, like "retweet" and "favorite" of Twitter (which mean

respectively replicate a post from a user to a list of followers, and show that you liked the post content, leaving it in a list of posts with same criteria).

Social ties can be exemplified when there is interaction between two most users of any social media; when all the communication generated by these users becomes frequent and generates a connection between them.

According Recuero [15], social ties are classified into two types: associative and dialogical. Associative occurs when there is interaction between various authors on a social network, such as sharing a link to a news story by a tweet. Dialogic occurs when it does not depend on interactions, requiring only belong to a group or a community, for example decide to follow a user on Twitter.

As reported by Seron [16], the social networking sites are very important in disseminating information, where users have the opportunity both to send and to receive this information. In addition, you can reach thousands of users per second, only with a message posted on Twitter or Facebook.

2.2 Twitter

Every social media has a purpose. It can be said that Twitter is intended to convey information quickly, since each tweet is limited to 140 characters. Furthermore, Twitter allows ease of use of the information through its API (Application Program Interface), mainly through hashtags, resource created by own microblog.

As mentioned, there is a growth in the use of social media. On Twitter, for example, there are more than 316 million monthly active users and over 500 million tweets are sent per day [17].

Even with all the publicity, advertising and marketing of Twitter, beyond good and quick way to personal interaction among users, the microblogging is far from social media most used by Internet users in the world. With Facebook on the list of worldwide market share of social media, with just over 85%, Twitter is in third place with just over 3%, just behind the Pinterest, with nearly 4% (data updated in June 2015) [18].

2.3 Social Network Analysis

With the large volume of data and information provided in social media, it is possible to study patterns of interactions between the actors of social networks [15].

The term "Social Network Analysis" (SNA) has existed for a long time and their use has grown in recent years due to the widespread use of social media.

Researchers from various fields of knowledge are very important in SNA [19]. A large part of social media has powerful tools that analyze user profiles. For example, it can get information about the best days and times for content of publications, based on the estimation of content viewing; view graphics on the effect of any content posted; find out how much people viewing some content; among others.

2.4 Weather Information

For the acquisition of meteorological data in Brazil, automatic weather stations (AWS or EMA in portuguese) and meteorological radars are used.

An automatic weather station collects, minute by minute, weather information (temperature, humidity, atmospheric pressure, rainfall, wind speed and direction, solar radiation) representing the area where it is located. Every hour, these data are available for transmission, via satellite or mobile telephony, to the headquarters of INMET in Brasilia. The set of received data is evaluated by a quality control and stored in a database.

In addition, the data is freely available in real time via internet [12] for the development of weather forecast and the different meteorological products of interest to industry users and the public in general, and for a wide range of applications in research meteorology, hydrology and oceanography.

Another important feature in meteorology is the use of weather radar. The radars are used to perform the meteorological monitoring on vulnerable cities the occurrence of floods, flash floods and geological events such as landslides. Radars produce information necessary for the preparation of warnings about possible disasters related to rainfall.

Weather radar in Brazil are the responsibility of RedeMet [13] which aims to integrate the meteorological products focused on civil and military aviation, in order to make access to this information faster, efficient and safe.

Furthermore, in Brazil there is the monitoring of rainfall using satellite images. The satellite-based measurements establish relationship with meteorological variables estimated by existing devices on satellites. The accuracy of the information depends on the resolution of the images of the satellite. In this paper will not be covered comparative analyzes with satellite images.

3 Case Study

This section presents the solution developed by the Division of Satellites and Environmental Systems (DSA) for monitoring rainfall through the use of tweets. We present the architecture of the solution, the storage solution, data distribution and the data analysis.

To collect the tweets is necessary to consume the data from the Twitter API, which provides a defined set of request and response messages. This API is well documented and enables the development of various applications such as websites, widgets and other projects that can interact with Twitter. The API also allows you to use almost all the features that are available on the site. For example, the user information, the timeline, their friends and followers, tweets and retweets, search for tweets, among others. It isn't possible, for example, consume information from a profile that has no relationship with the application that collects the data. Only public information can be collected through the APIs, that is, information that users without access restriction. The application communicates with the API over HTTP requests and receives data in JSON format (JavaScript Object Notation). To query information is used the GET

method and to send information is used the POST method. Several error messages can be obtained of the API to verify if is possible to get information of the tweets [20].

The Twitter API allows extract various information of the social network, such as: friendly relations among users, who users are following, friends, the user profile and posted public tweets. The disadvantage of this API is the limit of requests, which is 180 requests per 15 minute intervals depending on the resource being requested. In addition, each request returns up to 100 tweets in one request.

For access the API's information, the application should be authenticated through the OAuth protocol [21]. There are two types of requests: refers to public information and user data modification. To manipulate data of a user, as well as authorization to access the API, it is also required user authorization. To develop an application that accesses data public must obtain a token that is provided on the Twitter site. Thus, the application can get the JSON responses of the API.

The Twitter API provides various information related to a particular tweet, such as date and time of creation, location, among others. Information relating to the user's location are disabled by default in Twitter. If user enable this option, the application can get the information about tweet location. This information includes the latitude and longitude of the user.

3.1 Application Architecture

The Twitter's data stream is continuous, however the API sets limits for data consuming, and it was necessary to develop a tool that collects and store the tweet's information and the user who made the publication. The application architecture can be seen in Figure 1. Briefly, the information of the tweets are collected through the Twitter API (a) and rainfall information are collected from AWS (b). The information from these sources are stored in a database (c) and are consumed through a web application (d).

Fig. 1. Application Architecture

In the application were used:

1. The Java programming language, along with the Twitter4J, which is an abstraction of the Twitter API.
2. Database PostgreSQL to the storage of all data collected by the application;
3. Web service in Java, hosted on Apache Tomcat 7 Server;
4. Web application that consumes and uses the web service to view the data on a map.

The application uses the Twitter information, by Twitter4J, using a request with some parameters such as: amount of tweets to be collected, default language of the posted tweets and tweets from a date. These parameters are processed by Twitter4J, being sent to the Twitter API and then are returned in JSON files.

All tweets collected by Twitter4J are returned in JSON files, received by the application and then are entered into the database.

In addition, another application receives the data from the AWSs, stores them in the database and makes the rainfall data are available in JSON files.

The web application consumes the web services that return information, for example, what are the tweets of a particular hashtag, what are the tweets of a date; what are AWSs with millimeters of rainfall greater than zero; what are AWSs with millimeters of rainfall greater equal zero. The information returned are shown on a map using latitude and longitude of each information (tweet or AWS).

3.2 Experiments and Results

Initially, we evaluated several words that users can used to represent rainfall. For this, we used the site TopSys [http://topsy.com/analytics] that shows the approximate amount of tweets for a period and word. The words used in the search are in Table 1. This search occured between 6 August 2015 to 06 September 2015.

Table 1. Words used to rainfall Monitoring

Word (in Portuguese)	Word (in English)	Amount of tweets	%
chuva	rain	793,132	71,15
chuvisco	drizzle	786	0,07
trovoada	thunderstorm	10,544	0,95
nevoeiro	fog	5,310	0,05
tempestade	storm	76,231	6,84
raio	lightning	60,087	5,39
garoa	drizzle	4,934	0,44
trovão	Thunder	19,012	1,71
temporal	temporal	112,505	10,09
relâmpago	lightning	26,158	2,35
dilúvio	flood	23,790	0,21
chuvarada	rain	4,219	0,38
precipitação	rainfall	1,456	0,38

After the Table 1 analysis, it was defined #chuva (rain) hashtag and the search term "chuva." We collected 1,328,221 tweets (posted by anybody) between days 08 September 2015-08 October 2015. The collection was done through the application.

From the collected data, it was found that only 0.1% of tweets had the latitude and longitude in the tweets. To increase the amount of tweets with geolocation, the city defined in the user profile was used which is available in each tweet. When the user city is used, a attribute in database is setted to indicate that latitude and longitude have been adjusted by the application. For this, they were stored all geolocation information (latitude and longitude) of Brazilian cities. These are available in the Brazilian Geodesic System (SGB), of the Brazilian Institute of Geography and Statistics (IBGE) [22]. For each tweet without geolocation information, were assigned for tweet the latitude and longitude of the user city provided by SGB. This allowed 23.78% of tweets were geolocation information.

From this information, the web application was developed for viewing the location of the tweets on the map. This application makes use of the Google Maps API. Figure 2 shows the map of Brazil with the collected tweets (marker in red color), the AWS who had a change in rainfall (indicating that there was rainfall on site - Umbrella icon), and the AWS who had no change the rainfall (in green marker). Each marker on the map represents a tweet with their respective coordinates (latitude, longitude).

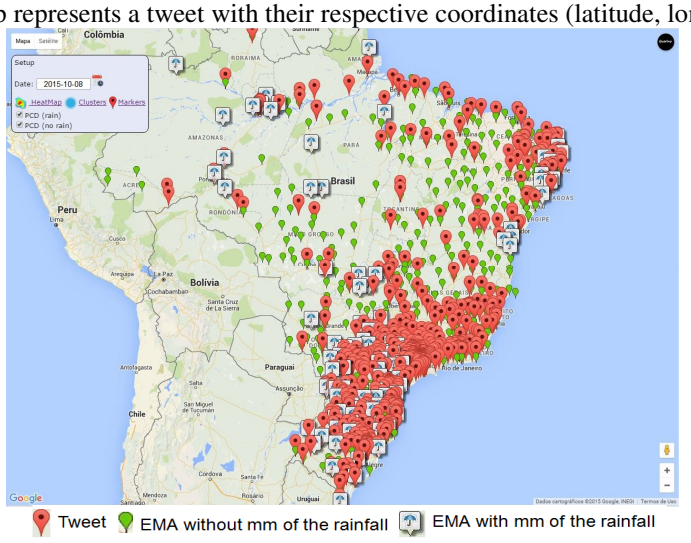

Fig. 2. Chart with markers - October 8, 2015

In Figure 3 is shown a heat map, which allows viewing of the places with more tweets. Red marker indicates the region that had more published on the subject, while that the marker light blue represents the region with less tweets.

The information in Figures 2 and 3 are of the day 8 October 2015, which accounted 48.500 tweets about rainfall. In addition, 102 AWS received mm of rainfall and 439 AWS had no rainfall that day.

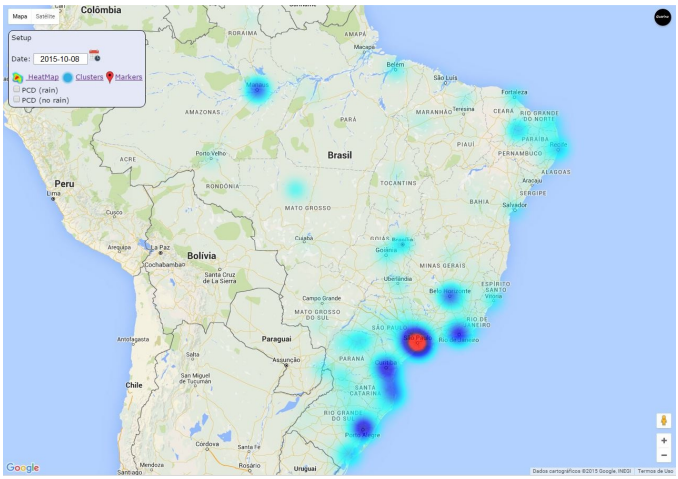

Fig. 3. Table 2. Heat Map chart - October 8, 2015

In Figure 4 is shown a region of the map of Brazil. In addition, markers that are close to each other are grouped, forming clusters. This makes it easy to display on the map. If the user zooms in application, clusters are dismembered. Cluster in red color indicates greater number of counters in the same area and the blue color indicates a smaller amount. Within each cluster is shown the amount of the markers that were grouped. In this figure, it also is possible to see the distance between the tweets and AWS. In Figure 5 is shown the same region of Figure 4, but without clusters of tweets. Red markers represent the tweets. It can view a good concentration of tweets close to the AWSs.

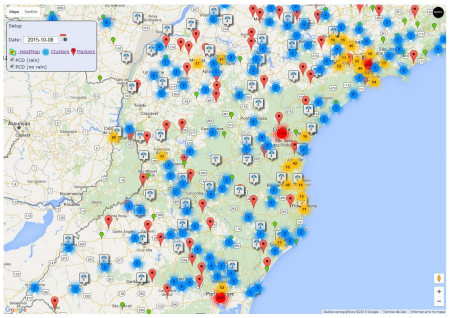

Fig. 4. Chart with tweets (clusters) and AWS's of the Southern Brazil region – October 8, 2015

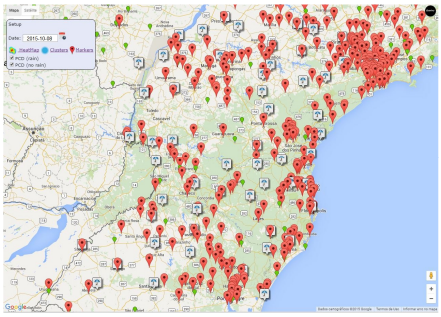

Fig. 5. Chart with tweets (markers) and AWS's of the Southern Brazil region – October 8, 2015

In figure 6 its is possible to visualize the relationship between the amount of tweets and the amount of mm of the rainfall per day. It can be seen that there is a relationship between these two variables. Blue line refers to the tweets and the red line refers to amount of mm of the rainfall.

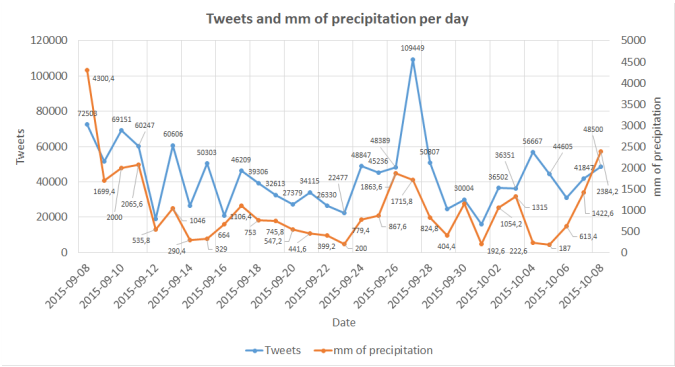

Fig. 6. Amount of tweets and mm of the rainfall per day

4 Conclusion

In this work, we performed social network analysis of Twitter, in particular, the tweets related to rainfall events in order to determine whether this information could contribute to the monitoring of precipitation events in Brazil. It can be seen that a close relationship exists between the amount of tweets and the amount of mm precipitation detected by AWS.

The solution used in this study is a tool that can assist in monitoring rainfall events. Through it, there is the possibility, for example, to provide warnings about certain weather conditions and especially supplement the information provided by other monitoring means (eg AWS, weather radar, satellite imagery).

It is also concluded that the study and analysis of social networks and their media is very important, not only to the field of technology, but also for other fields of knowledge.

Considering possible future work, it is possible to evaluate the possibility of dismembering the information of tweets to analyze content, making sure the content is positive or negative; analyzing feelings of tweets by checking the behavior of users through the content posted; and expand the analysis considering evaluate other social media, such as Instagram, which has the use of interactions through hashtags, similar to those of Twitter, with the difference being only through photos or videos.

Moreover, given the success of the experimental application, CPTEC should make the release of a specific hashtag so that people can inform precipitation events. The suggested hashtag is #CPTECCHUVA.

Other suggested work are (i) examine minimum and maximum distance of tweets regarding the AWS with and without precipitation statement; (ii) monitor hourly tweets relating them to rain gauges and radars; (iii) visually compare the tweets map with satellite images processed by DSA/CPTEC/INPE.

References

1. Internet World Stats. Available in <http://www.internetworldstats.com>
2. We Are Social. Digital, Social & Mobile Worldwide in 2015. Available in: <http://wearesocial.net/blog/2015/01/digital-social-mobile-worldwide-2015/>.
3. IBGE. Brazilian Institute of Geography and Statistics. Projeção da população do Brasil e das Unidades de Federação. Available in: <http://www.ibge.gov.br/apps/populacao/projecao/>.
4. Rocha Filho K. L. Modelagem hidrológica da bacia do Rio Pirajuçara com TopModel, telemetria e radar meteorológico. 138 f. Dissertação (Mestrado em Meteorologia) – Instituto de Astronomia, Geofísica e Ciências Atmosféricas, Universidade de São Paulo, São Paulo, 2010.
5. Silva, F. D. S. Análise objetiva estatística da precipitação estimada com radar e medida por uma rede telemétrica. 2006. 101 f. Tese (Mestrado em Meteorologia) – Instituto de Astronomia, Geofísica e Ciências Atmosféricas, Universidade de São Paulo, São Paulo, 2006.
6. Pereira Filho, A. J.: Crawford, K. C.; Hartzell, C. Improving WSR-88D hourly rainfall estimates. Weather and Forecasting, v. 13, n.4, p. 1016-1028, 1998.
7. Legate, D. R.; Deliberty, T. L. Measurement biases in the United States rain gauge network. Water Resource Bulletin. v. 29, p. 855-861, 1993.
8. Calvetti, L.; Beneti, C.; Pereira Filho, A. J. Integração do radar meteorológico doppler do SIMEPAR e uma rede pluviométrica para a estimativa da precipitação in: Simpósio Brasileiro de Sensoriamento Remoto, 2003, Belo Horizonte. Anais do Simpósio de Brasileiro de Sensoriamento Remoto, 2003. CD-ROM
9. Austin, P. M. Relation between measured radar reflectivity and surface rainfall. Monthly Weather Review, v. 115, p. 1053-1070, 1987.
10. Batlan, L. J. Radar Observations of the Atmosphere. Chicago: The University of Chicago Press, 324p, 1973.
11. Doviak, R. J.; Zrnic, D. S. Doppler radar and weather observations. Dover Publications, 1993.
12. INMET. Brazilian National Weather Institute. Available in: <http://www.inmet.gov.br/portal/index.php?r=estacoes/mapaEstacoes>
13. Redemet. Available in: <http://www.redemet.aer.mil.br/>
14. Garton, Laura; Haythornthwaite, Caroline; Wellman, Barry. Studying Online Social Networks. In: Journal of Computer-Mediated Communication. Available in: <http://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.1997.tb00062.x/full>.
15. Recuero, R. Redes sociais na internet. Sulina, 2009.
16. Seron, Wilson F. M. de S. Análise de Redes Sociais - Um Estudo do Twitter. São Paulo, 2015. Dissertação. Instituto de Ciência e Tecnologia. Universidade Federal de São Paulo.
17. Twitter. Uso do Twitter / Fatos Sobre a Empresa. Available in: <https://about.twitter.com/pt/company>.
18. AREPPIM. Available in: <http://stats.areppim.com/stats/stats_socmediaxsnapshot.htm>.
19. Matheus, Renato F; Silva, Antonio B. de O. Análise de redes sociais como método para a Ciência da Informação. Revista da Ciência da Informação, v. 7, n. 2, 2006. Available in: <http://www.dgz.org.br/abr06/Art_03.htm>.
20. Twitter. Error Codes & Responses. Available in: <https://dev.twitter.com/overview/api/response-codes>.
21. OAuth. Disponível em: <http://oauth.net/>.
22. IBGE. Brazilian Institute of Geography and Statistics. Available in: <http://www.ibge.gov.br/home/geociencias/geodesia/bdgpesq_googlemaps.php>