FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

TwitterJam: Identification of mobility patterns in urban centers based on Tweets

Francisco José Moura de Bastos Rebelo

Mestrado Integrado em Engenharia Eletrotécnica e de Computadores

Supervisor: Carlos Manuel Milheiro de Oliveira Pinto Soares (PhD)

Co-Supervisor: Rosaldo José Fernandes Rossetti (PhD)

July 30, 2015

U.PORTO
FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

MIEEC - MESTRADO INTEGRADO EM ENGENHARIA ELETROTÉCNICA E DE COMPUTADORES | **2014/2015**

A Dissertação intitulada

"TwitterJam: Identification of Mobility Patterns in Urban Centers Based on Tweets"

foi aprovada em provas realizadas em 27-07-2015

o júri

Presidente Professor Doutor Luis Miguel Pinho de Almeida
Professor Associado do Departamento de Engenharia Eletrotécnica e de Computadores da Faculdade de Engenharia da Universidade do Porto

Professor Doutor Paulo Jorge de Sousa Azevedo
Professor Auxiliar do Departamento de Informática da Universidade do Minho

Professor Doutor Carlos Manuel Milheiro de Oliveira Pinto Soares
Professor Associado do Departamento de Engenharia Informática da Faculdade de Engenharia da Universidade do Porto

O autor declara que a presente dissertação (ou relatório de projeto) é da sua exclusiva autoria e foi escrita sem qualquer apoio externo não explicitamente autorizado. Os resultados, ideias, parágrafos, ou outros extratos tomados de ou inspirados em trabalhos de outros autores, e demais referências bibliográficas usadas, são corretamente citados.

Autor - Francisco José Moura de Bastos Rebelo

Faculdade de Engenharia da Universidade do Porto

Resumo

TwitterJam: Identificação de padrões de mobilidade em centros urbanos com base em Tweets

No início do século XXI, as redes sociais serviam apenas para darmos a conhecer ao mundo os nossos gostos, partilhar as nossas fotografias e partilhar alguns pensamentos. Volvida uma década, percebeu-se o grande potencial destes serviços. A quantidade de informação que as redes sociais contêm é enorme, foi então que as pessoas e empresas começaram a extrair informação das redes sociais. O TwitterJam é uma ferramenta que analisa o conteúdo da rede social Twitter para extrair eventos relacionados com o tráfego automobilístico. Para chegar a este objectivo, começou-se por analisar os tweets recolhidos para obter só os que são relacionados com trânsito rodoviário. Em seguida, recolheram-se tweets de uma fonte oficial, a conta do Centro de Operações da Prefeitura do Rio de Janeiro. Para saber se a informação é fidedigna usamos a correlação, uma sobre o número de tweets obtidos de cada tipo e a segunda uma correlação espacial entre os tweets normais e os oficiais. Também foram desenvolvidas duas hipóteses para correlacionar os tweets normais e oficiais nas duas dimensões, a de volume de dados e a espacial. Os resultados não são perfeitos mas têm uma magnitude aceitável. Também foram analisadas ferramentas para a visualização dos dados e decidimos qual a melhor abordagem. Foi desenvolvida, também, uma aplicação web que mostra e que permite a análise dos resultados.

Abstract

In the early twenty-first century, social networks served only to let the world know our tastes, share our photos and share some thoughts. A decade later, these services are filled with an enormous amount of information. Now, the industry and the academia are exploring this information, in order to extract implicit patterns. TwitterJam is a tool that analyses the contents of the social network Twitter to extract events related to road traffic. To reach this goal, we started by analysing tweets to know those which really contains road traffic information. The second step was to gather official information to confirm the extracted information. With these two types of information (official and general), we correlated them in order to verify the credibility of public tweets. The correlation between the two types of information was done separately in two ways: The first one concerns the amount of tweets in a certain time of day and the second one the localization of these tweets. Two hypothesis were also devised concerning these correlations. The results were not perfect but where reasonable enough. We also analysed tools suitable for the visualization of data to decide what is the best strategy to follow. At the end we developed a web application that shows the results, to help the analysis of results.

Agradecimentos

Agradeço em primeiro lugar ao meu orientador, Prof. Carlos Soares, pela oportunidade de começar este projecto e a sua ajuda essencial em ultrapassar todas as adversidades que foram surgindo. Agradeço também ao meu co-orientador, Prof. Rosaldo Rossetti, pois todos os seus pontos de vista fizeram melhorar todo este projecto.

Não sendo, oficialmente, nada nesta Dissertação, agradeço ao Doutor por todo apoio, ajuda e resolução de problemas e também a criação de problemas. Com estas três pessoas, consegui trazer o projecto a Bom Porto.

Agradeço aos aos meus pais, por me darem todas as oportunidades que se podem dar a um filho. Educação, valores e amor. Esta é a maior herança que um filho pode ter dos seus pais. Agradeço também à minha namorada, Sara, por estar sempre ao meu lado, estando bem ou mal, está para me apoiar e para me pôr nos trilhos do bom caminho. Agradeço aos meus Padrinhos por me proporcionarem sempre um ponto de vista diferente para me fazerem pensar e crescer de outra maneira bem como aos meus primos, por me incentivarem a experimentar tudo, desde tirar uma negativa a chumbar mas também por me terem ajudado a ser uma criança feliz junto deles.

Não menos importantes em todo este processo são os meus amigos. Os de sempre que mesmo estando longe, espalhados por aí, sei que estão comigo.

Uma palavra de apreço a todos os que comigo vestiram Capa e Batina. São Capas, São Fitas, a Praxe Continua. Lavrador, Rute, Maia, Chuck, Autómato, Espaço, Dezanove e Ariel, com o vosso apoio, tudo é superável, nada é impossível. Para onde quer que vá levo-vos comigo. Dred, Inácio, Cardoso, Volume, Joaninha, Vazio, Timoneiro, Ana, Jarbas, Tocha, desejo que muitas tesouras pairem sobre o vosso cabelo. A vocês, um abraço sentido por tudo o que fizeram e fazem por mim. A todos os Unjosjoutros pelas quartas à noite na porta do EVC.

Agradeço também a onde tudo isto foi possível, a Faculdade de Engenharia da Universidade do Porto e às pessoas do Sapo Labs.

Ao Pub 1808 e todos os seus constituintes, agradeço todo o apoio líquido e comestível. Se há lugar onde se está bem é nesse vosso/nosso alpendre!

Agradeço, ao Cacete e ao Pedro por me terem acompanhado durante a estadia e em todas as festas de Madrid.

Aos companheiros de trabalho, as pessoas incríveis que me ajudarem dentro do Call Center de Campanhã: à Deusa Sabine, à Joana, ao César, à Paula, ao Rocha, ao Pinho e à Diana bem como à Tia Alice.

Francisco Moura Rebelo

"Eles comem tudo, eles comem tudo
Eles comem tudo e não deixam nada"

José Afonso

Contents

List of Figures

List of Tables

List of listings

Symbols and Abbreviations

API	Application Programming Interface
DB	Database
JSON	JavaScript Object Notation
NLP	Natural Language Processing
PCC	Pearson's Correlation Coefficient
POS	Part-of-speech
URL	Uniform Resource Locator
WWW	World Wide Web

Chapter 1

Introduction

The World Wide Web (WWW) is, since the beginning, a world in constant evolution. In the WWW we can access an huge amount of information, from news to job offers. In the last decade there was an evolution, a new paradigm. Now we have new ways to inform, be informed and work. This evolution took us to the creation of social media content websites. There is a variety of them, blogs or social networks, like Facebook[1] or Instagram[2]. Or we can find Wikis where we can work with people around the world. Micro-blogging is a phenomenon that was born in this paradigm of WWW. They can provide the user a new sharing method of information like, short messages or pictures. The most popular platforms are Twitter[3], Tumblr[4], and Plurk[5]. Data Mining researchers have been studying these in order to acquire useful information [10].

1.1 Context

Mobility patterns identification, in urban areas, is important for activities like transports planning and management, the adoption of urban planning, the adoption of new marketing strategies or to identify potential markets for new products. The main data collected is on national questionnaires. The increasing of information in the WWW, the growth of social networks, the sharing of preferences from users through web applications, are reasons to explore about mobility patterns, namely, about traffic. The detection of these patterns can be essential to urban planning by the responsible authorities.

1.2 Motivation

Twitter contains a high amount of information about politics or information provided by its users. These can contain information about road traffic events, like, accidents, jams or maintenance of

[1] https://www.facebook.com/ - accessed 10-01-2015
[2] http://instagram.com/ - accessed 10-01-2015
[3] https://about.twitter.com/company/ - accessed 10-01-2015
[4] https://www.tumblr.com/ - accessed 10-01-2015
[5] http://www.plurk.com/ - accessed 10-01-2015

roads. However, due to the amount of information in Twitter, this process is almost impossible to be done manually. The motivation for this dissertation lies in the construction of a Web platform that will display the results automatically, showing road traffic tweets on the map. The main contribution for this project is the creation of a correlation coefficient scheme that can provide credibility to the general tweets.

1.3 Objectives

This dissertation has three goals: 1) research techniques of Web Mining that are appropriate for the interpretation of events related to road events from Twitter, 2) investigate data visualization techniques to allow easy interpretation of the knowledge acquired on 1) and 3) implementation of a Web application that allows the visualization of the final result.

1.4 Planning: Initial vs Final

The work plan for the dissertation was divided in small tasks to divide problems so we can treat them all. The initial plan didn't suffer major changes through development. Some problems were encountered regarding correlation coefficients leading to some delays and to turn the completion of web application in a secondary objective, due to lack of time. Overall, the plan was met. In the Figure 1.1 the project's final Gantt Diagram can be seen.

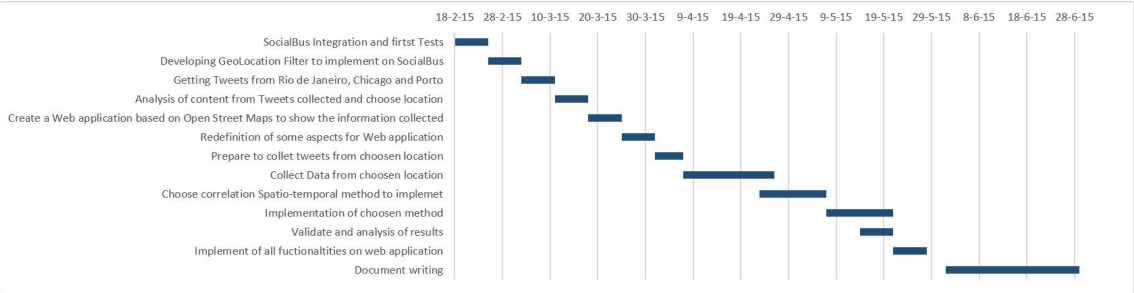


Figure 1.1: Final Gantt Diagram.

1.5 Structure

This dissertation is organized as follows: Chapter 2 contains the state of the art for the scientific topics related to this project, namely Text Classification and Visualization Support. Chapter 3 explains the approach used to solve the problem. Chapter 4 shows the primary analysis of the acquired data and Chapter 5 explains how the data was validated. Chapter 6 reveals the main insights and conclusions and Chapter 7 suggests some possible future work to be done.

Chapter 2

State of the Art

A review of the state of the art of the techniques and technologies related to this project was done. First, it was necessary to understand the micro-blogging world (see Section 2.1). Then a tool used to retrieve data from Twitter was studied (see Section 2.2). Afterwards, a brief overview of a field called Natural Language Processing (NLP) and how it was leveraged in order to retrieve road traffic information was done (see Section 2.3). Finally, some of the most relevant techniques used in visualization systems focused towards social media and other relevant systems were studied (see Section 2.4).

2.1 Micro-blogging

2.1.1 Twitter

For Micro-blogging, Twitter is one of the most used tools [11]. Created in 2006 by Evan Williams, Biz Stone and Jack Dorsey, Twitter was designed to allow users to give and receive information in the fastest way, without any kind of disturbances. As a matter of fact, using Twitter is extremely easy: it provides each user a maximum of 140 characters to publish a single message, this message can be anything, a weather information or a random comment about the queue at the gas station. These messages are known as "tweets". Twitter is can be considered a social network between users and it has different types os relations. One of these facilities is the possibility of each user having followers, which means that those followers will receive a notification everytime user publishes a tweet. Another of these facilities is the fact that one can respond directly to another exclusively using what are called the "tags" or participate in an already existing discussion using "hashtags". Tags are a representation of the user's Twitter and it is marked with an at sign "@" before the user's name. Also, "hashtag" is generally a word that identifies the discussion topic. It is inserted in the user's tweet and it is marked with a number sign "#" before the topic related word. Figure 2.1 illustrate all the concepts present earlier based on this tweet. Twitter has been registering a growing popularity over the years as the number of users has increased considerably since it was created in 2006[1]. This growing pattern was studied by Smith and Brenner[12] who estimated that around 8% of the adults that are acquainted with the Internet use Twitter. As far as

Figure 2.1: Example of a tweet.

social networks are concerned, over the years, this tool has been gaining importance because of its usefulness on giving data about a user or expressing opinions about certain matters. Despite being mostly used as a social media, it is now considered a useful tool for communication. As an example of Twitter's popularity, in Iran, during the 2009 elections, when Mahmoud Ahmadinejad won, a huge amount of opponents expressed their disfavor with the outcome in Twitter. After this media coverage, the department of the American State created a new account on Twitter to publish opinions on international facts [13] [14]. Regarding the necessity of innovation, Hughed and Palen [15] studied the possibility of adopting this social network as a media tool or as a way of publishing an emergency. The main result of their study is that important news are correlated by a big number of "tweets" and new users. Carvalho et al. [16] [17] made work to identify road traffic in Twitter messages. Passos et al. [18] made work in simulation traffic analysis in various approaches.

2.1.2 Twitter API

Twitter's data can be accessed by using two Application Programming Interfaces (API): REST API [19] and Streaming API [20]. The REST API provides access to read and write data, allows access to information about user, friends and followers. It is based on requests, whose responses are returned in Javascript Object Notation (JSON) and its use requires an OAuth authentication. The latter uses events to provide information and requires oAuth or HTTP basic authentication. However, this REST API limits the number of requests allowed per user. The schema of the connection with this API can be seen in Figure 2.2. The Streaming API uses real-time data,

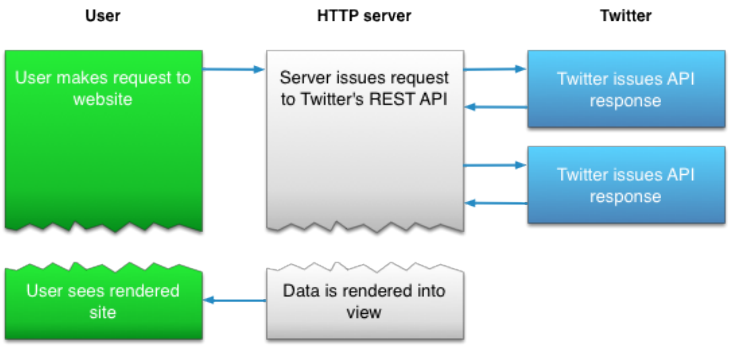


Figure 2.2: Connection using REST API [1].

limiting the amount of data available according to the session's beginning timestamp. There are

three types of streams: *Public*, *User* and *Site*. The first one streams all public data flowing through this social network. User and Site streams are filtered to only collect data from one user or multiple users. Figure 2.3 shows the schema of a Streaming API connection.

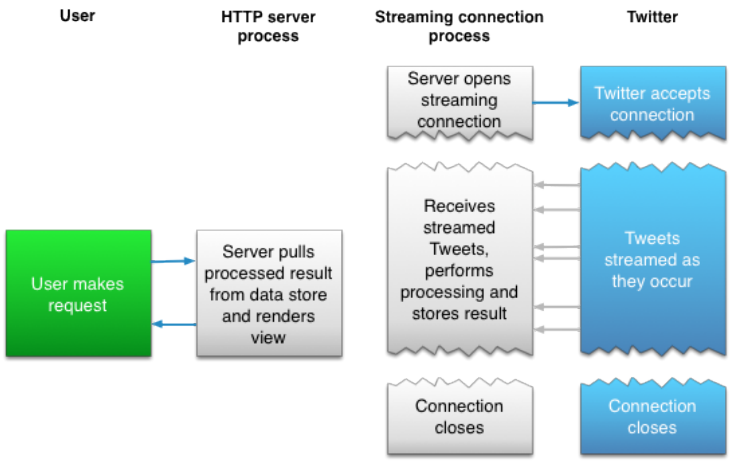


Figure 2.3: Connection using Streaming API [1].

2.2 Socialbus

The Socialbus project [21] is a research platform for extracting, storing and analysing the Portuguese Twittosphere for R&D and journalistic purposes. Its system architecture is presented in figure 2.4. Socialbus, previously called TwitterEcho, collects data in real-time using Twitter's Streaming API. These tweets are sent to a message broker and processed on two components: stream processing and pre-processing. The resulting data is stored in MongoDB[1]. Data is stored in the form of JSON objects.

[1] https://www.mongodb.org/ - online accessed on 15/04/2015

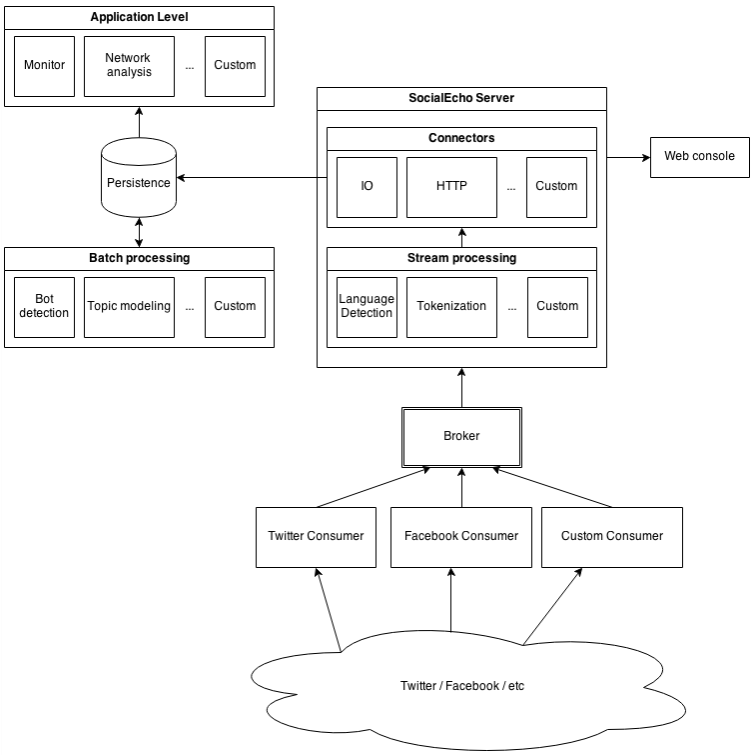


Figure 2.4: Socialbus high-level architecture[a].

[a]Image obtained from http://reaction.fe.up.pt/socialbus/

2.3 Natural Language Processing

Natural Language Processing began in the fifties as a intersection of artificial intelligence and linguistics [22]. NLP is used to characterize how software or hardware analyze spoken or written language [23]. NLP explores how computers are used to understand or manipulate natural language text or speech to do useful things. Applications of NLP include a number of fields of study, such as machine translation, natural language text processing and summarization [24]. Sentence delimiters, tokenizers and part of speech (POS) taggers are important tools for every NLP process [23]. Tokenizers can do segmentation on a stream of characters into units called tokens, a simplistic view, a token is any sequence of characters separated by spaces [23]. For example, in the sentence "I will be back" we have 4 tokens. POS taggers are constructed upon tokenizers and sentence delimiters, each word in a sentence is tagged [23]. Can be a noun, a verb, an adjective, etc. Another example, "My name is Francisco" we have a pronoun, a noun, a verb and a noun in the end. Wanichayapong et al. [25] used NLP for tokenize tweets and then apply a filter to extract road traffic information. Ribeiro et al. [26] treat tweets using NLP for removing hastags, mentions or links from twitter and then search for keywords on tweet context. Endarnoto et al. [2] created an application for visualization of traffic condition information on Android. This system extract information from Twitter account of Traffic Management Center Jakarta Metropolitan Police. They

use NLP for tokenize tweet context and then use POS tagger for a sentence analysis. This android application display the traffic condition in a map form, with 3 different colors for different traffic conditions, green for normal, yellow for crowded, red for jammed [2]. In Figure 2.5 it is shown user inteface of this application.

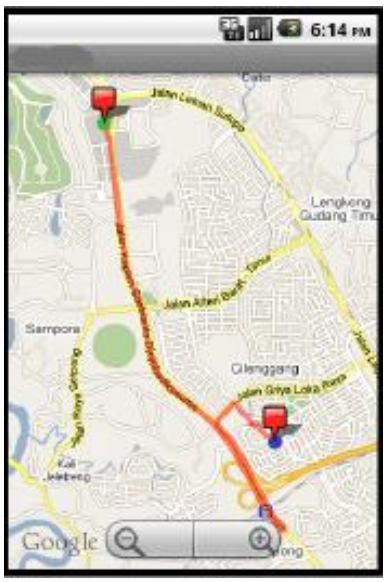


Figure 2.5: Graphic Interface of Android Application [2].

2.4 Visualization Support

In Data Mining, one of the most important steps, it's the interpretation of results. The major properties of the these tools must follow three important characteristics: 1) the appearance of data, 2) the temporal behavior and 3) properties of entire scene [27]. Visualization tools requires a pragmatic and critical review of the ways visualization can be used to represent and to analyze data [28]. In this section we will review some techniques to represent data collected, in this case, from Twitter.

2.4.1 Twitter Data Visualization

In section 2.1.2 we referred what data we can extract from Twitter. Based on this information we can collect, multiple visualization techniques can be used. The following list presents some of those techniques.

- Flow Map: displays movements of objects or subjects from one place to another by means of lines or arrows [28]. The data used for these maps have got different initial and final geographical locations, for example, migration patterns between regions. Flow maps are mostly static, however they can be dynamic using time sequence animation [28]. Figure 2.6 is an example of a flow map;

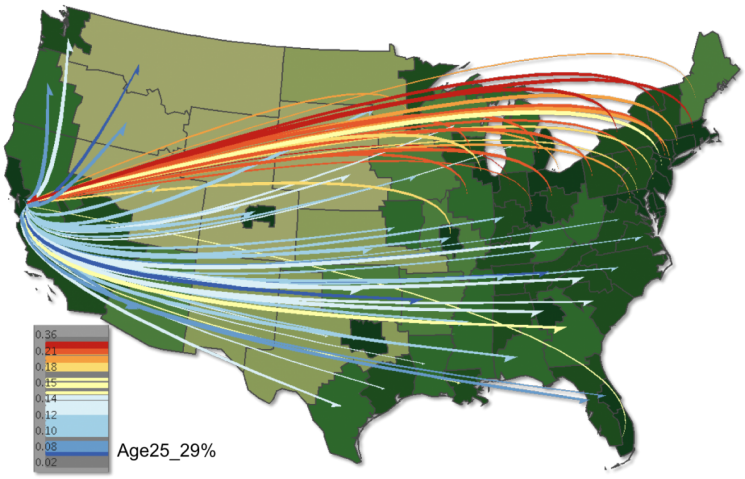


Figure 2.6: Flow Map example [3].

- Choropleth Maps: It represent aggregated measures of pre-defined regions with crisp boundaries. Choropleth mapping is particularly useful in providing comparative summaries over specified geographies [28]. Figure 2.7 is an example of a Choropleth Map.

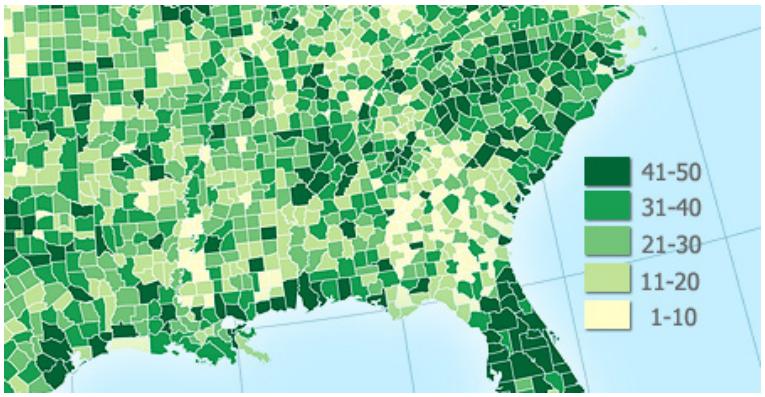


Figure 2.7: Choropleth Map example [4].

2.4.2 Existing Systems

In this section will be presented some visualization tools, how they work and see how can they positively affect systems to take advantage of social media data, like Twitter.

2.4.2.1 CompactMap

CompactMap [5] is an on-line visualization application that packs text clusters, with stable generation of layouts to maintain user's mental map. It achieves spatio-temporally coherent layouts by dynamically matching clusters across time, and removing cluster overlaps according to spatial proximity and constraints. CompactMap allows:

- A dynamic visualization technique that displays clusters in text streams as stable, space-efficient layouts;

- A real-time visual search engine that supports arbitrary keyword search combined with semantic analysis of topics;

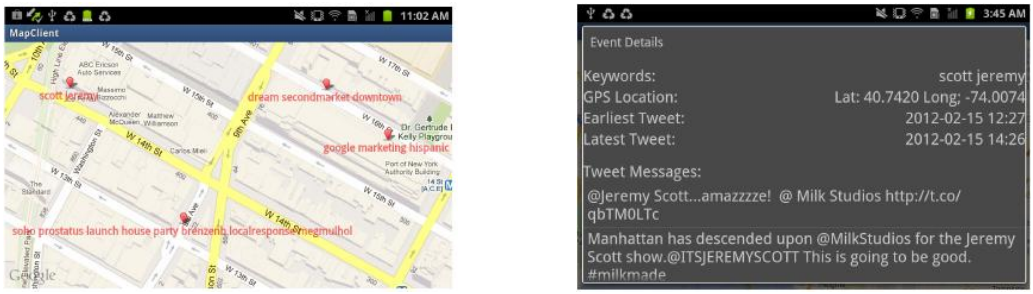
- An enhanced real-time visual analysis system that enables users to explore and compare discussions and topics of interest.

In Figure 2.8 we can see an example of CompactMap.

Figure 2.8: Clustering Visualization of the tweets for the search term "technology" on May 16, 2013 [5].

2.4.2.2 EventRadar

EventRadar [6] is a local event detection method with data collected form Twitter. The detection of events use two phases, first is selected a subset of keywords from a bigger set of words, second phase is try to identify potential events in a set of recent tweets for each subsets of words. Uses maps visualization techniques to display the position o tweets with respect to events. In Figure 2.9 we have an example of EventRadar.



(a) User Interface on EventRadar.

(b) Event Details on EventRadar.

Figure 2.9: User Interfaces on EventRadar [6].

2.4.2.3 TweetDrops

TweetDrops [7] is a visualization designed for people who have not paid attention to sustainability. It is an opportunity to learn about energy conservation. It has two main visual components: one is the background rain drops, which represent the accumulation of energy related tweets collected from Twitter and the second clickable foreground tweets with detailed content [7]. In Figure 2.10 is shown the interface of TweetDrops.

Figure 2.10: TweetDrops interface with background drops and foreground tweets [7].

2.4.2.4 SensePlace2

SensePlace2 [8] is a geovisual analytics support for situation awareness for crisis events using data collected from Twitter. It focuses on leveraging explicit and implicit geographical information for tweets and on providing visual interface methods to enable understanding of place, time, and theme components of evolving situations [8]. Is user-centered, using scenario-based design methods that include formal scenarios to guide design and validate implementation as well as a systematic claims analysis to justify design choices and provide a framework for future testing [8]. The SensePlace2 interface includes 5 core components: query panel, timeline display/control, tweet list, tweet map, and history view. This components can be seen in Figure 2.11.

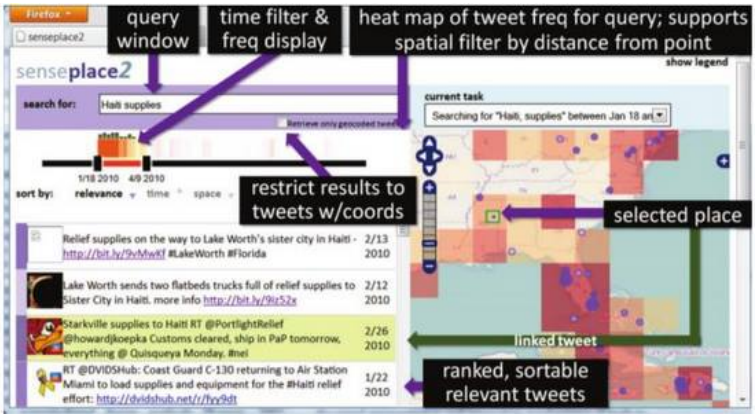


Figure 2.11: SensePlace2 interface: time constraint on query for "Haiti supplies" plus spatial selection of relevant results [8].

Chapter 3

Methodology

In this section will be presented the methodology of the work for this dissertation. We have three sections one regarded to the data , other for data modelling and the last regards to the web tool. In Data section 3.1 we will approach how we collect data and how we process it. In Section 3.2 we will explain how we select the region and the correlation methods. On the Web tool Section 3.3 it is presented the system architecture and the functionalities.

3.1 Data

3.1.1 Data Collection

To collect data from Twitter, we used SocialBus. SocialBus, as it was explained on Section 2.2, is a tool that collects data from Twitter and saves it on a Database (DB). In SocialBus we have access to several filters, including filter tweets by users or by keywords in text. However, for our problem we were required to collect tweets from a certain area. This filter was not yet created on SocialBus, so we developed it and added this functionality to the software package. The new filter created, which we dubbed as GeoLocation filter, filters data by two pairs of longitude and latitude points that create a Bounding Box. One point refers to the South-west point and the other to the North-east point. Afterwards, the Streaming API creates the Bounding Box and filters all tweets whose coordinates match with the bounding box. Alternatively, if the tweet does not present coordinate data, we match its location with the field place to check if the tweet is on the area that is specified. In Figures 3.1 and 3.2 is presented the work flow for general and official tweets data collecting, respectively.

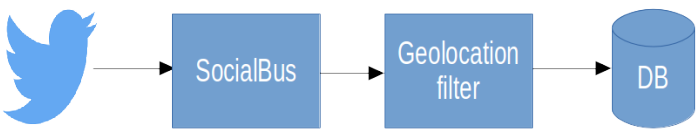


Figure 3.1: Work flow for collecting general tweets.

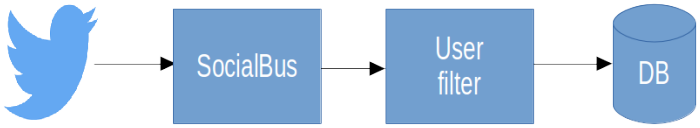


Figure 3.2: Work flow for collecting official tweets.

3.1.2 Data Processing

To process data we have two methods, one for general tweets and the other for the official tweets. General tweets are collected from common users and official tweets are concern to a official account from "Centro de Operações da Prefeitura do Rio de Janeiro" where they text about weather conditions and, mostly, about road traffic in Rio de Janeiro, Brazil.

The work flow of this process is showed in Figure 3.3

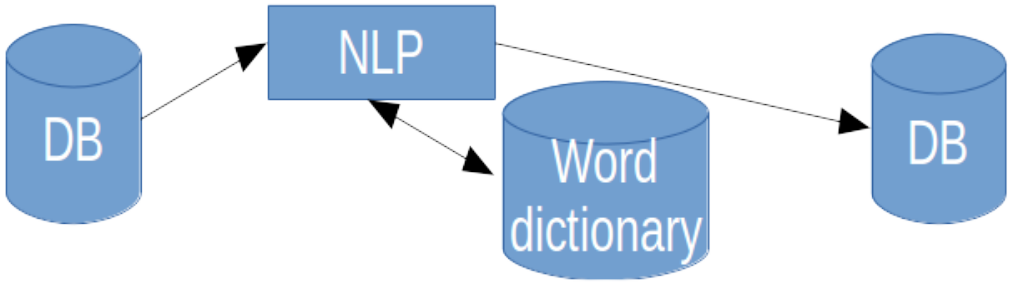


Figure 3.3: Work flow for collecting general tweets.

For the general tweets, we have to analyse the context to know if it contains information regarding road traffic. We created a dictionary of Portuguese words that are related to road traffic. This filtering was built using NLP. The first step on NLP involved text cleaning tasks, such as URL links, accents and punctuation marks. Then, we convert all tweets into lower-case and tokenize them. With these tokens we managed to see if the tweet contains road traffic information or not. If it contains information, we store it on a DB with the original text, latitude and longitude coordinates and temporal data. The listings 3.1 and 3.2 are examples of general tweets extracted directly by filter created in SocialBus and general tweets stored after retrived road traffic information, respectivelly.

Listing 3.1: JSON of general tweets obtained from SocialBus

```
{
"_id":ObjectId("55239db044ae61308b601c2d"),
...
"lang":"pt",
```

```
...
"place":{
"id" : "97bcdfca1a2dca59","bounding_box" : {"type" : "Polygon",
"coordinates":[[[-43.795449,-23.08302],[-43.795449,-22.7398234],
[-43.0877068,-22.7398234],[-43.0877068,-23.08302]]]},
"place_type":"city",
"name":"Rio de Janeiro",
...},
"coordinates":{..."coordinates" : [-43.189162,-22.796322]},
"text":"Ainda ta noiteeeeee .. SOCORROOOO kkkkkk'",
...
"created_at":ISODate("2015-04-07T09:04:47.000Z"),
"user":{..."name" : "@JaiSouza",...}
}
```

Listing 3.2: JSON of general tweets after treatment

```
{
"_id":ObjectId("5551c5fc44aecba97e19aeed"),
"lat":-22.92319,
"LONG":-43.567822,
"texto":"transito , transito e+ transito",
"tipo":"coordinates",
"dia":"07/04/2015",
"hora":"11:44:31"
}
```

For official tweets the method is distinct because they don't have active geolocation. We identified the ten busiest avenues on Rio de Janeiro, Brazil and used NLP to ascertain whether the tweet text has specified one of the avenues. If the tweet contains these data, then we move to another step to discover location. This step uses Google geocoder[1] that provides latitude and longitude points from a given address. The final step is to store the tweet with original text, latitude, longitude, day and time and the location retrieved by geocoder on a DB. The listings 3.3 and 3.4 are examples of official tweets extracted directly using user filter created in SocialBus and official tweets stored after retrived road traffic information, respectivelly. The main difference between this two examples is the location, in listing 3.3 location is null in three different fields, but in listing 3.4, after the geocoding process, we already have location.

Listing 3.3: JSON of official tweets obtained from SocialBus

```
{
"_id" : ObjectId("5523b0b544ae0ceb9bbc47d0"),
"filter_level" : "low",
"retweeted" : false,
```

[1] https://developers.google.com/maps/documentation/geocoding/

```
...
"lang" : "pt",
...
"place" : null,
"coordinates" : null,
"text" : "Acidente interdita parcialmente o Vdto Eng. Noronha, sentido Btfogo,
logo apos a saida do Tn. Sta Barbara. Retencao http://t.co/9Tdb78A5Qx",
...
"geo" : null,
...
"screen_name" : "OperacoesRio",
...
"created_at" : ISODate("2010-12-14T02:53:26.000Z")
}
```

Listing 3.4: JSON of official tweets after treatment

```
{
"_id" : ObjectId("555b102d44aef6a124165c31"),
"lat" : -22.86123085021973,
"LONG" : -43.45218658447266,
"dia" : "07/04/2015",
"hora" : "21:29:05",
"localizacao" : "Avenida Brasil, Rio de Janeiro - RJ, Brazil",
"texto" : "@operacoesrio av brasil sentido zona
oeste altura da fio cruz parada"
}
```

3.2 Data Modelling

3.2.1 Region Selection

For this work we had some troubles in defining the region to be analysed. We tried some cities like Porto and Lisbon from Portugal, Boston and San Francisco, U.S.A and Rio de Janeiro, Brazil. In Portugal, Twitter is a social media with low relevance, only 9% of internet users have account on Twitter [29] unlike the U.S.A. and Brazil two of the countries with the largest number of Twitter accounts [30]. According to Leetaru et al. [30] Rio de Janeiro have 1.39% of georeferenced tweets. We have created a collection of tweets in the cited cities and in Rio de Janeiro the amount of tweets were much higher than in the others. We collected, in average 200k tweets per day (these number reflect less than 1% of total tweets provided by Twitter Streaming API). A study made by TomTom [31] confirms Rio de Janeiro, Brazil as the third city where people spent more hours on road traffic. After the selection of Rio de Janeiro, Brazil, we collected tweets from this region from April 7 to 26, 2015. Another point in favor of Rio de Janeiro: is the existence of an official account

of "Centro de Operações da Prefeitura do Rio de Janeiro", that tweets about weather conditions and above all road traffic information. This account was use as ground truth for data analysis and for correlation method.

3.2.2 Avenues Selection

After the decision of collecting tweets located in Rio de Janeiro, Brazil, we had another issue because official tweets are not geographically located, i.e. they don't have latitude and longitude coordinates. We selected a list of busy avenues: Avenida Brasil, Avenida Francisco Bicalho, Avenida Niemeyer, Avenida Lúcio Costa, Linha Amarela, Linha Vermelha, Avenida das Américas, Avenida Ernâni Cardoso, Avenida Presidente Vargas, Avenida General Dantas.

3.2.3 Correlation

Correlation between official tweets and general tweets consists in two coefficients: normal correlation coefficient and spatial correlation coefficient. With the data collected, we process, as explained on Section 3.1.2 and create four different matrices:

- Number of tweets by day and hour - general tweets;
- Number of tweets by day and hour - official tweets;
- Location midpoint by day and hour - general tweets;
- Location midpoint by day and hour - official tweets;

The matrices were organized by hours in rows and days in columns. The first two matrices show how many tweets were collected, the last two show the location midpoint. A example of Matrix is shown on Table **??** .

X	7/04/2015	8/04/2015	...	26/04/2015
0h	α	β	...	δ
1h	α_1	β_1	...	δ_1
2h	α_2	β_2	..	δ_2
...	...	...	...	...
23h	α_{23}	β_{23}	..	δ_{23}

Table 3.1: Matrix example

3.2.3.1 Normal Correlation

To calculate the correlation, we used Pearson's Correlation Coefficient (PCC). In Equation 3.1, x_{ij} is the number of general tweets to each hour and day, y_{ij} is the number of official tweets to each hour and day, $\bar{x}$ is the average number of general tweets and $\bar{y}$ is the average number of official

tweets. This correlation coefficient relates the amount of tweets from the official with tweets from the general sources

$$\rho_1 = \frac{\sum_{i=1}^{20}(\sum_{j=0}^{23}((x_{ij}-\bar{x})\cdot(y_{ij}-\bar{y})))}{\sqrt{\sum_{i=1}^{20}\sum_{j=0}^{23}(x_{ij}-\bar{x})^2}\cdot\sqrt{\sum_{i=1}^{20}\sum_{j=0}^{23}(y_{ij}-\bar{y})^2}} \tag{3.1}$$

3.2.3.2 Spatial Correlation

To calculate spatial correlation we substitute, in PCC, $(x_i-\bar{x})$ for the haversine distance calculation between two points, $haversine(x_{ij},\bar{x})$. We chose haversine formula instead of Spherical Law of Cosines and Vincenty formula. The execution time was a good argument to exclude Vincenty formula that was the slowest one [32]. In computational language haversine formula is more accurate since it Spherical Law of Cosines [33]. Haversine formula estimates the distance, in this case, in kilometers, between two points characterized by their latitude and longitude. This replacement gives the difference, a number equal or greater than zero. In the Equations 3.2, 3.3 and 3.4 is explained hot to calculate haversine distance between two points. In 3.2 φ is latitude and λ is longitude, in 3.4 R is earth's radius (mean radius = 6372.8 kilometers).

$$a = \sin^2(\frac{\Delta\varphi}{2}) + \cos\varphi_1 \cdot \cos\varphi_2 \cdot \sin^2(\frac{\Delta\lambda}{2}) \tag{3.2}$$

$$c = 2\cdot\arctan 2(\sqrt{a},\sqrt{(1-a)}) \tag{3.3}$$

$$d = R\cdot c \tag{3.4}$$

The previously mentioned Equations 3.2, 3.3 and 3.4 are computed in the following code excerpt (see listing 3.5).

Listing 3.5: Haversine distance calculation in Java

```
{
public static double haversine(
double lat1, double lon1, double lat2, double lon2) {
public static final double R = 6372.8; // In kilometers
double dLat = Math.toRadians(lat2 - lat1);
double dLon = Math.toRadians(lon2 - lon1);
lat1 = Math.toRadians(lat1);
lat2 = Math.toRadians(lat2);

double a = Math.sin(dLat / 2) * Math.sin(dLat / 2) + Math.sin(dLon / 2)
* Math.sin(dLon / 2) * Math.cos(lat1) * Math.cos(lat2);
double c = 2 * Math.atan2(Math.sqrt(a), Math.sqrt(1-a));
return R * c;
}
```

```
}
```

This computation yields the distance between every midpoint on the matrix and the average of that matrix. Equation 3.5 explains how we calculate spatial correlation coefficient, x_{ij} is a pair of latitude and longitude coordinates, in average, for each hour and day of general tweets, $\bar{x}$ represents the latitude and longitude average of general tweets. For official tweets, y_{ij} is a pair of latitude and longitude coordinates, in average, for each hour and day of official tweets and $\bar{y}$ is the latitude and longitude average of official tweets. This correlation coefficient only treats spatial dimension.

$$\rho_2 = \frac{\sum_{i=1}^{20}\sum_{j=0}^{23}(haversine(x_{ij},\bar{x}) \cdot haversine(y_{ij},\bar{y}))}{\sqrt{\sum_{i=1}^{20}\sum_{j=0}^{23} haversine(x_{ij},\bar{x})^2} \cdot \sqrt{\sum_{i=1}^{20}\sum_{j=0}^{23} haversine(y_{ij},\bar{y})^2}} \tag{3.5}$$

3.2.3.3 Space extended correlation coefficient

For this correlation we had some issues and we tried two different hypothesis. The first one, expressed on Equation 3.6, is a bond between Equation 3.1 and the Haversine distance. The elements $(x_{ij} - \bar{x})$ and $(y_{ij} - \bar{y})$ are the same of Section 3.2.3.1 and the element $haversine(G_{ij}, O_{ij})$ is the distance between every general tweets midpoint and official tweets midpoint. This approach was to join in the same equation the two dimensions, quantity and space.

$$\rho_3 = \frac{\sum_{i=1}^{20}(\sum_{j=0}^{23}((x_{ij} - \bar{x}) \cdot (y_{ij} - \bar{y}) \cdot haversine(G_{ij}, O_{ij})))}{\sqrt{\sum_{i=1}^{20}\sum_{j=0}^{23}(x_{ij} - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{20}\sum_{j=0}^{23}(y_{ij} - \bar{y})^2} \cdot \sqrt{\sum_{i=1}^{20}\sum_{j=0}^{23} haversine(G_{ij}, O_{ij})^2}} \tag{3.6}$$

The second option was joining Equations 3.1 and 3.5 obtaining the Equation 3.7. This hypothesis is the geometrical average between the two correlation coefficients and joins the normal and spatial correlation, yielding a two dimensional correlation (quantity and space).

$$\rho_4 = \sqrt{\rho_1 \cdot \rho_2} \tag{3.7}$$

3.3 Web Application

This section explains the web application we developed to interact with the data and results of our work. In here, we discuss the System Architecture and the functionalities.

3.3.1 System Architecture

For this dissertation we developed a web application to visualize and interact with the results. This web application was designed to be more accessible and understandable. The web page was designed on HTML5, CSS3.0 using Groundwork framework[2] and Javascript. In this web application we resort, also, to Open Street Maps using Leaflet framework[3] to present the tweets

[2] https://groundworkcss.github.io/
[3] http://leafletjs.com/

on the map. The web server was developed in Python using Flask framework[4]. This server is responsible to connect the web page and the data stored on DB. It receives requests from the web page, which it relays to the DB and, also, it receives responses from the DB and sends it to the web page. Figure 3.4 shows the System Architecture of the system. When the user wants some information on the web page, for example, he selects a date to see, the web page sends a HTTP Request to web server where it following connects do DB and formulate a JSON query to obtain information. After this, web server sends a JSON response with all data retrieved from DB to the web page that, finally, shows information to the user.

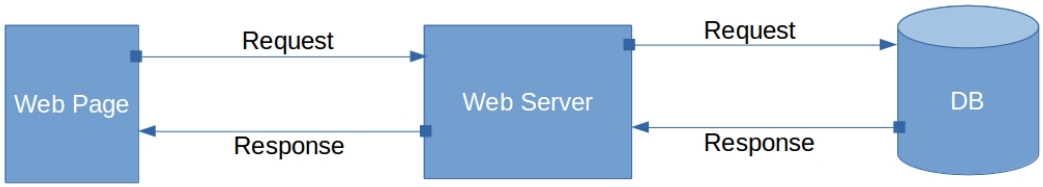
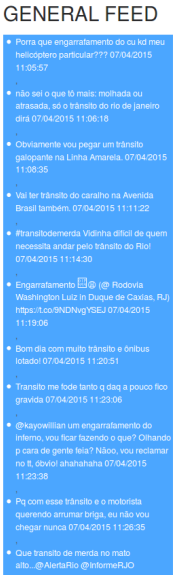


Figure 3.4: System Architecture of the system.

3.3.2 Appearance and Functionalities

The Web page is divided in three major areas: the general feed, the map and the official feed, from left to right respectively. On both the general and official feeds we present the text of each tweet of that category. On the map field we present a map centered on Rio de Janeiro, Brazil. This web page presents the user with the ability to select data by a specific day and hour or to present all tweets passing through all the days and hours and adding those tweets to map. In Figure 3.5a and 3.5b is perceptible that in the feeds area user can see the text of tweet and the temporal information of that tweet.

[4]http://flask.pocoo.org/



(a) General Tweets Feed.

(b) Official Tweets Feed.

Figure 3.5: TwitterJam's Feed Areas.

In Figure 3.6 is showed the area in TwitterJam where the user can select a day and hour to see the road traffic tweets.

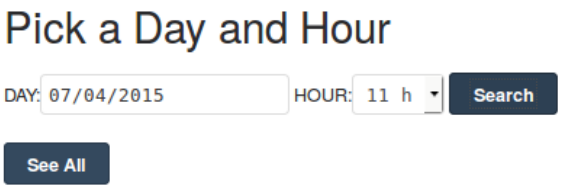


Figure 3.6: Selection of Day and Hour in TwitterJam.

At last, Figure 3.7 shows the full interface for our web application.

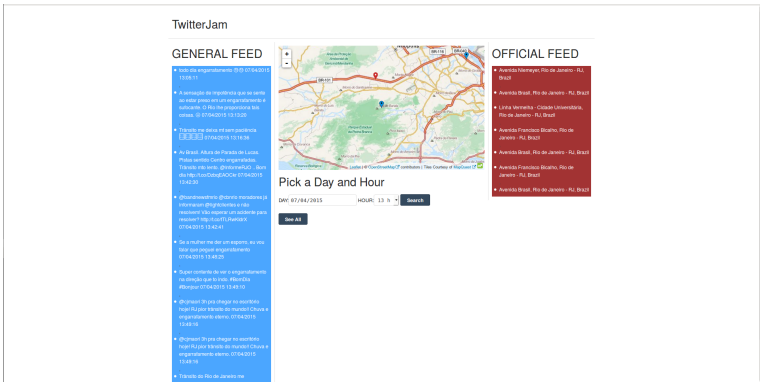


Figure 3.7: Screenshot of Web Application.

In the next Figures is exposed more features implemented in web application. Figure 3.8 and Figure 3.9 show what user can see when click in tweets points in map.

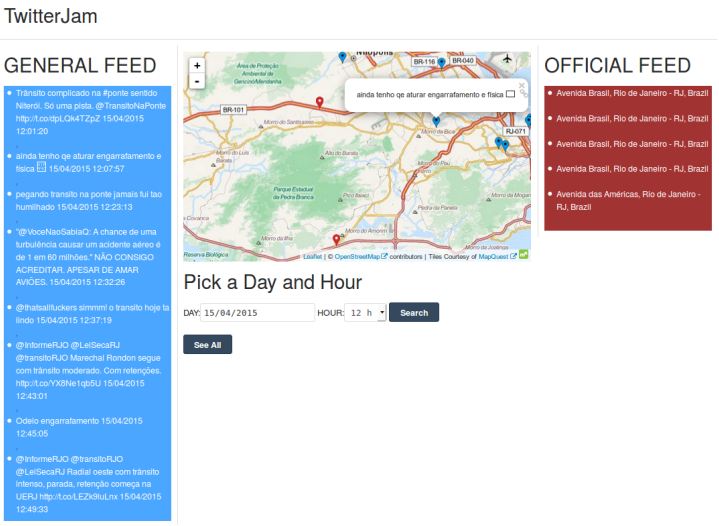


Figure 3.8: Selecting General tweets in Map.

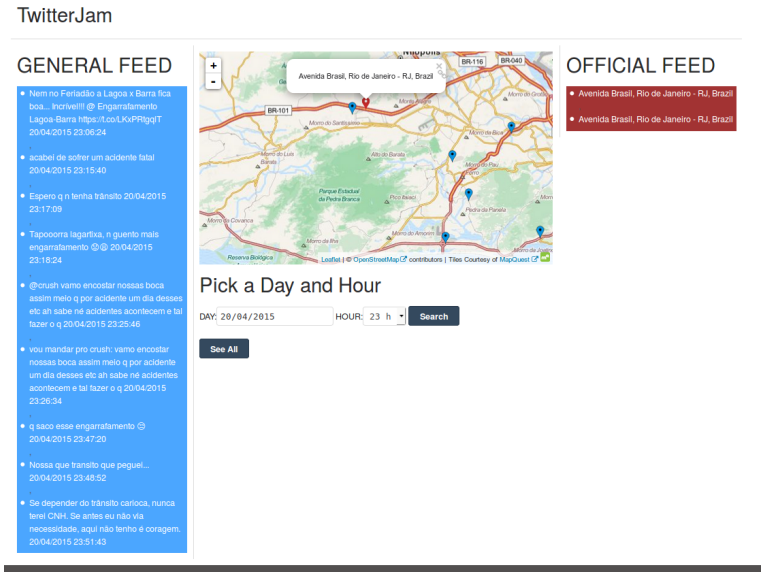


Figure 3.9: Selecting Official tweets in Map.

Figure 3.10 is how user can select a date to explore.

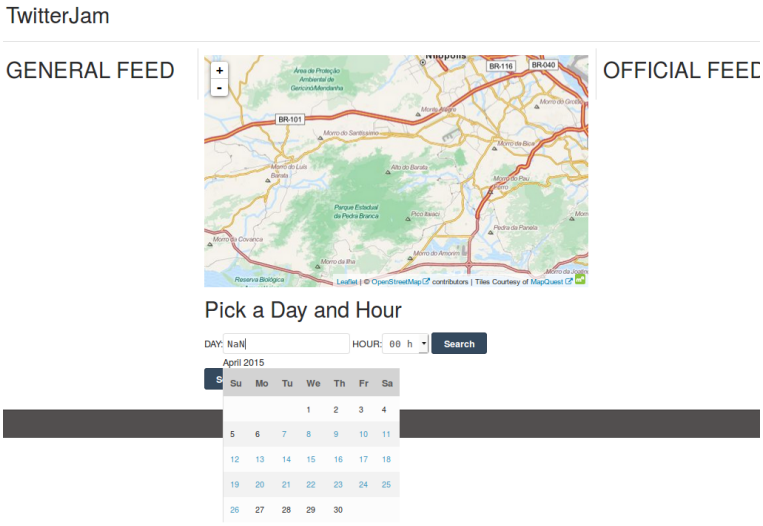


Figure 3.10: Selecting Date to pick tweets.

Chapter 4

Data Analysis

In this Chapter, we present a primary data analysis. These analysis will consist on analysing number of general and official tweets by hour and day.

At the end of collection we had 5.091.055 general tweets located in Rio de Janeiro, and 6.812 official tweets. After the data processing stage, which was explained on Section 3.1.2, we were left with 1.580 general tweets and 1.552 official tweets with relevant traffic information.

4.1 General Tweets Analysis

In this primary data analysis we try to find some patterns by week day and by hours of day.

Analysing the Figure 4.1 we can find two peaks of road traffic tweets, one in the morning and the other in late afternoon. Those peaks confirms what was expected because they match rush hours of the city .

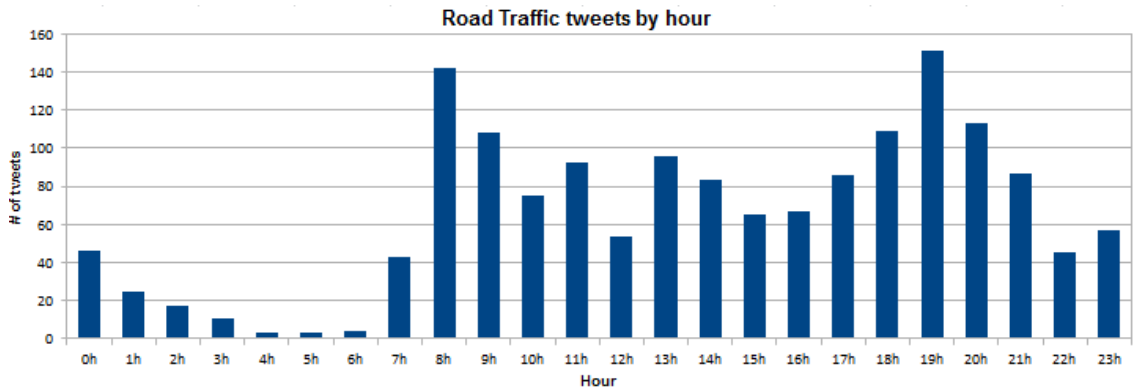


Figure 4.1: Number of Road Traffic general tweets by day hours.

Interpreting Figure 4.2, we can see that the biggest number of tweets during the working days. Weekend have less tweets.

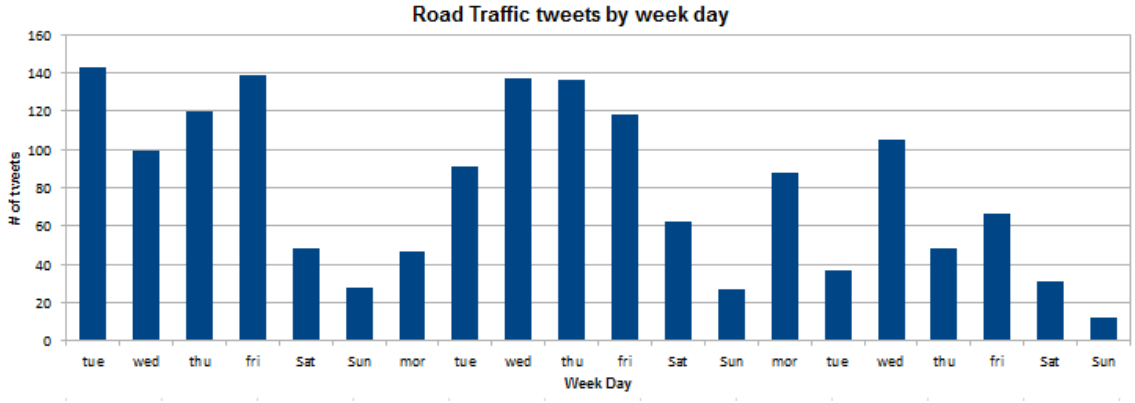


Figure 4.2: Number of road traffic general tweets by week days.

4.2 Official Tweets Analysis

Analysing the Figure 4.3 we can find two peaks of road traffic tweets, one in the morning and the other in late afternoon. The peak in the afternoon it is not so pronounced as in the morning. The peak of the afeternoon can reveal that can be less persons working for "Centro de Operações da Prefeitura do Rio de Janeiro" These peaks match the rush hours.

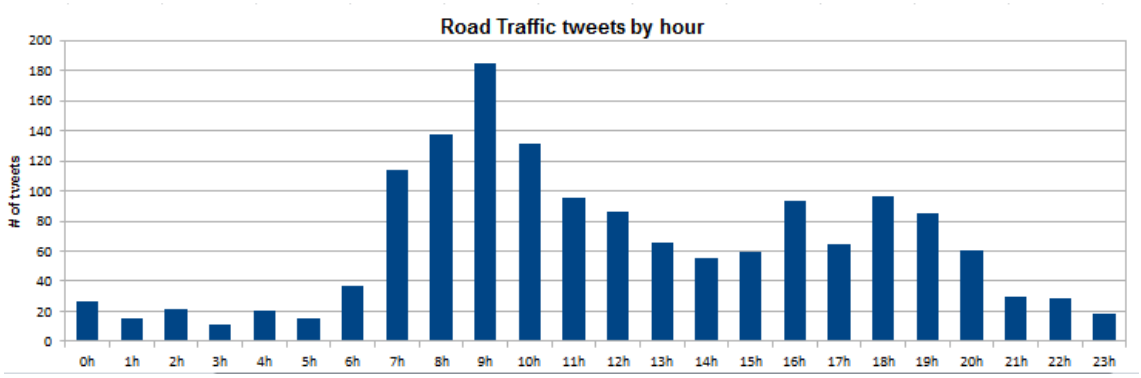


Figure 4.3: Number of Road Traffic official tweets by day hours.

Analysing the Figure 4.4, it can be seen that the working days are the most tweeted days. Weekend days have less tweets.

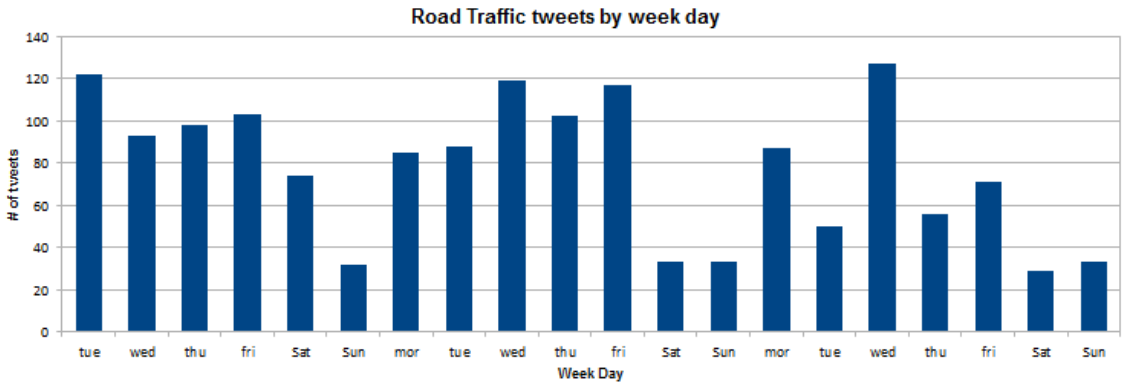


Figure 4.4: Number of Road Traffic official tweets by week days.

4.3 Geolocation Analysis

Figure 4.5 reaffirms the rush hours already seen in Figure 4.3. This figure is a bar chart, where the number of tweets by hours for the nine avenues selected for official tweets analysis can be seen. In terms of road traffic, Avenida Brasil (colored dark blue in the chart) is the location with the biggest amount of road traffic.

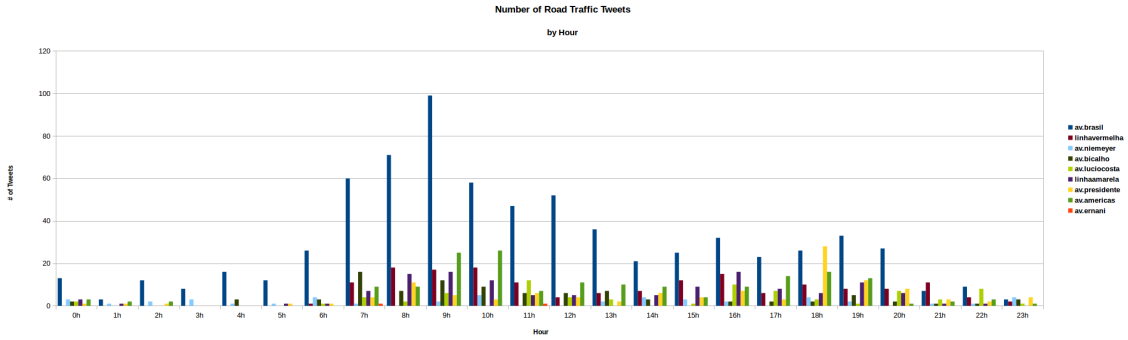


Figure 4.5: Number of road traffic tweets by hour in different avenues.

Figure 4.6 reaffirms the analysis done in Figure 4.4. This figure is a bar chart, where the number of tweets by weekdays for the nine avenues selected for official tweets analysis can be seen. In terms of road traffic, Avenida Brasil (colored dark blue in the chart) is the location with the biggest amount of road traffic.

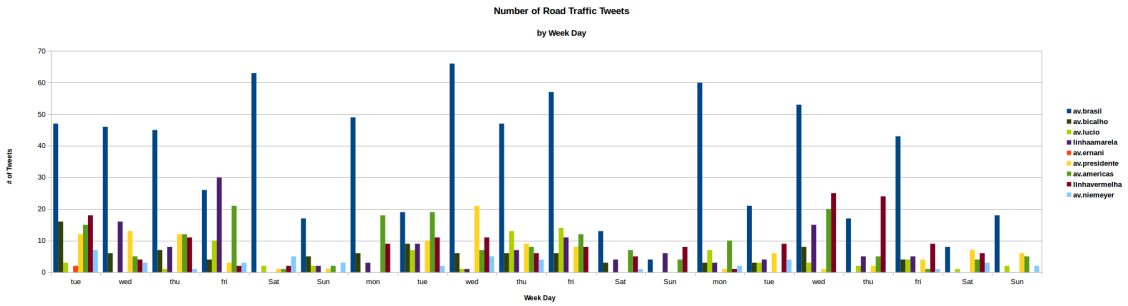


Figure 4.6: Number of road traffic tweets by week days in different avenues.

Most of the tweets are located in the north-east part of Rio de Janeiro. As we can see in Figure 4.7, blue points are general tweets and red ones are official tweets.

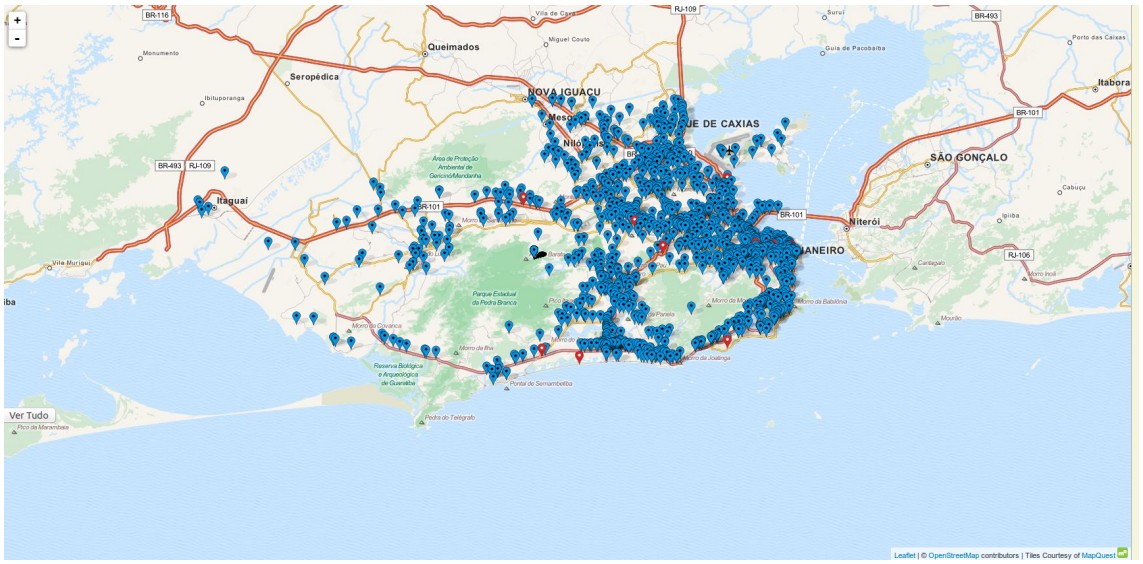


Figure 4.7: Location of Tweets.

4.4 Comparison of General and Official tweets

Analysing the the Figures 4.1 and 4.3 we can see that the two peaks of tweets remain in the same hours but official tweets are mostly later than general tweets. If there is a peak of road traffic, the people on the street can tweet immediately, but for official tweets the information must be confirmed in loco by traffic police agents or by cameras but only after being communicated some event to authorities. Due to this fact official tweets are emitted later that general tweets. These can be confirmed on Figure 4.8.

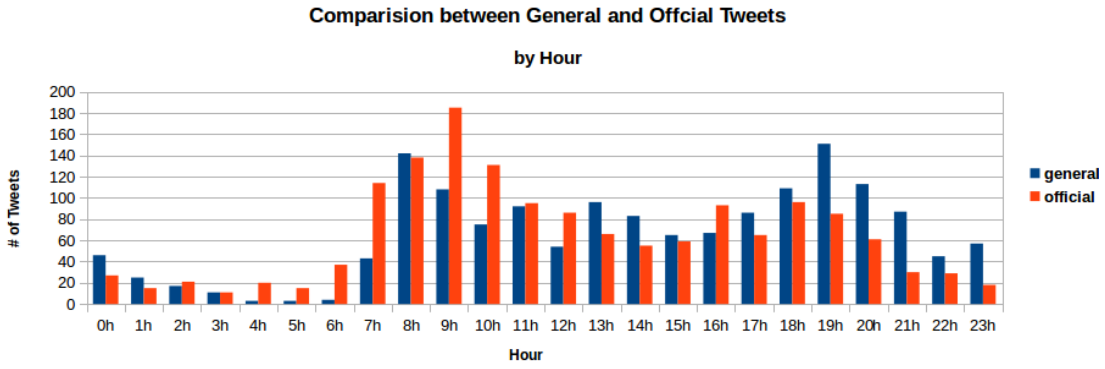


Figure 4.8: Comparison of General and Official tweets by hour.

Reviewing Figures 4.2 and 4.4 we can see that weekends are the days with less tweets. in Figure 4.9 is a bar chart of the number of tweets by during each day of the week for both official and general tweets. One more time, week days have less tweets of both types.

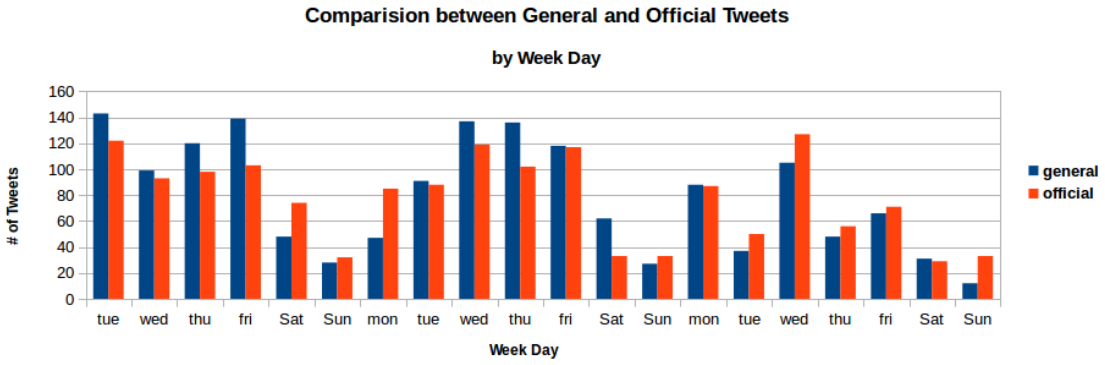


Figure 4.9: Comparison of General and Official tweets by week day.

Chapter 5

Model Validation

In this Chapter we will present how we validated the collected data and the results obtained with all tested correlations. This validation will use the rule of thumb, for interpreting the size of a correlation coefficient, proposed in [9]. The Table 5.1 show the interpretation of each interval.

Size of correlation	Interpretation
0.90 to 1.00 (-0.900 to -1)	Very high positive (negative) correlation
.70 to .90 ($-.70$ to $-.90$)	High positive (negative) correlation
.50 to .70 ($-.50$ to $-.70$)	Moderate positive (negative) correlation
.30 to .50 ($-.30$ to $-.50$)	Low positive (negative) correlation
.00 to .30 (.00 to $-.30$)	Negligible correlation

Table 5.1: Rule of Thumb for interpreting the size of a correlation coefficient [9]

Table 5.2 is a summary of the results of correlation coefficients and interpretation for each other. This results were obtained using data set referred in Chapter 4.

Correlation coefficient	Value	Interpretation
ρ_1	0,5188	Moderate positive
ρ_2	0,6649	Moderate positive
ρ_3	0,0140	Negligible
ρ_4	0,5873	Moderate positive

Table 5.2: Results of all correlation coefficients

5.1 Normal Correlation

In this Section, we will discuss the effects of the approach presented on Section 3.2.3.1. Equation 3.1 was used to calculate the PCC between the general and official tweets. . For calculate these correlation between the number of general and official tweets we use PCC because was a simple way of correlate two data series. Applying Equation 3.1, yields $\rho_1 = 0,5188$ which, according the rule of thumb we established previously, corresponds to a moderate positive correlation. This

result was expected, since the number of general and official tweets after all filters applied were close. It is not a perfect correlation because in some hours we had general tweets but we had not official tweets or vice versa.

5.2 Spatial Correlation

In this Section, we will discuss the effects of the approach presented on Section 3.2.3.2. Comparatively to Section 5.1 the Equation 3.5 for spatial correlation presents a difference. The substitution from Equation 3.1 to Equation 3.5 is the haversine calculation. The results for this correlation coefficient were expected because we only collected tweets from Rio de Janeiro. Figure 4.7 shows that most tweets are located on the same part of Rio de Janeiro. The result of Equation 3.5 is $\rho_2 = 0,6649$.

5.3 Space extended correlation coefficient

This sections showcases three attempts to create a two-dimensional correlation coefficient relevant for the problem at hand. The next two sections (sections 5.3.1 and 5.3.2) detail the two possible approaches that were devised.

5.3.1 First Approach

This approach is expressed in Equation 3.6, in section 3.2.3.3. It relates Equations 3.1. and the haversine distance (c.f. Section 3.2.3.2) between general and official tweets. This hypothesis is achieved by creating a link between Equation 3.1 and the haversine formula. This correlation coefficient has a new element $haversine(G_{ij}, O_{ij})$. This element inserts a new measure for correlation that is the distance between every midpoint of general and official tweets by day and hour. It yields a space extended correlation coefficient, with a value of $\rho_3 = 0,0140$.

Given the negligible correlation, this approach was tested in four different tests. Table 5.3 describes the different datasets created to evaluate this approach.

	# of General tweets	**# of Official tweets**	**Average distance between Gen. and Off. tweets**
Dataset 1	111	111	2,2662 kms
Dataset 2	111	58	1,3983 kms
Dataset 3	111	111	30,0611 kms
Dataset 4	111	58	30,7853 kms

Table 5.3: Datasets for the first approach test of space extended correlation coefficient

Figure 5.1 and Figure 5.2 show the location of general and official tweets in the datasets created. First two datasets have the locations of Figure 5.1 and the last two datasets have the locations of Figure 5.2. Blue marks are general tweets and the red ones are for official tweets.

Figure 5.1: Dataset 1 and Dataset 2 tweets locations.

Figure 5.2: Dataset 3 and Dataset 4 tweets locations.

The results were contradictory. It was expected that the test with dataset 1 would yield the largest correlation coefficient but, surprisingly, it actually yielded the lowest one. The remaining datasets had the expected results. The correlation coefficient with dataset 2 was larger than dataset 3 because the distance between the general and official tweets is lower (1,3983 Km) compared to the dataset 3 (30,0611 Km) and the difference between the number of tweets is not too big. Also, dataset 3 yielded a slightly higher correlation than dataset 4 (as expected) because the number of official and general tweets in the later dataset is different. In Table 5.4 can be seen the results of each test.

	Results
Dataset 1	0,1968
Dataset 2	0,3172
Dataset 3	0,2117
Dataset 4	0,2035

Table 5.4: First approach of Space extended correlation coefficient results

5.3.2 Second Approach

This correlation coefficient as presented on Section 3.2.3.3 and Equation 3.7 is result of the geometrical average between Equation 3.1 and Equation 3.5. This, as said in Section 3.2.3.3, joins

two dimensions space and quantity. The result is $\rho_3 = 0,5873$ and express a moderate positive correlation.

Chapter 6

Conclusions

In this chapter we present a summary of the project and achieved goals, a discussion of some decisions and results and some limitations of both the correlation methods and the visualization platform.

6.1 Summary

This dissertation goals were to develop a correlation method that confirms if general tweets can reliably inform about road traffic. The purpose of creating a system with all this information is to support the users and to support transportation planning and management. In order to accomplish these goals, we developed a system divided in two parts: creation of a correlation method with two dimensions (quantity and spatial) and a web application that allows the visualization of the evolution of road traffic information and to see this information segmented by day and hour. For each stage we had shared tasks: data collection and its posterior processing.

Collection of data: We were able to collect tweets by using SocialBus and a new implemented Geolocation filter. It enables to use a static data set for data processing and is a new contribution for this project.

Data processing: Before using the extracted data each tweet needs to be filtered and pre-processed so we can obtain the desired tweets with road traffic information. For General tweets, we use NLP for filtering only tweets that contain keywords that represent road traffic. For official tweets we use regex expressions with the selected avenues.

Correlation: With the filtered tweets, we proceeded to correlate them through different methods. In this point of the project we propose two approaches to correlate general and official tweets. The first two tested coefficients are an direct application of PCC. Both of them treats a single dimension. The remaining correlation coefficients are attempts to combine two dimensional data.

Web application: The web application, visualization tool, includes three main areas. Map, to visualize tweets and its context. General Feed that exhibits general tweets context, date and time and, lastly, official Feed that does the same as General Feed but for official tweets. All these

widgets were implemented using various JavaScript libraries and a Python web server connecting web application and DB's.

This dissertation is part of the specification design and implementation of an integrated transport analysis platform, MAS-Ter Lab that will make use of the knowledge extracted from unstructured sources such as Twitter [34] [35]. The web application will be integrated in a multi-modal transportation dashboard to monitoring of transport service levels and network status [36] and was developed designed making use of the concept "Artificial Transportation Systems" for the integrated analysis of transport and mobility systems [37] [38].

All the goals set for this dissertation where met. However, in the following section we will be discussing some of the decisions made and results obtained.

6.2 Discussion

In the developed system, there are some theoretical decisions that influenced the performance and results, such as:

- **Data:** As explained before, we choose Rio de Janeiro instead of other cities. We chose Rio de Janeiro because tweets with exact geolocation were available in greater number than on the other cities and also because road traffic in Rio de Janeiro is high, as referred in Section 3.2.1.

- **Correlation coefficients:** The spatial correlation coefficient is influenced by the Geolocation filter implemented in SocialBus. We had a large area to cover ($1.255 kms^2$) so we used geographic coordinates with extreme precision. In the Space extended correlation coefficients, the first approach failed because the new element inserted is working like a constant and influencing negatively when distances are high. The second hypothesis is the best approach for this problem. It gives us a geometrical average between both dimensions and it is a more scientifically strategy, as the results have shown.

- **Database:** In order to store the results of the used NLP process MongoDB was used. SocialBus uses MongoDB to store results, so we decided to maintain our results saved in the same platform. Even though this was an effective solution for our problem, it is not the perfect solution as the system response time is sometimes affected by the large number of requests. one possible solution is to use MySQL database.

- **Visualization widgets:** The widgets used in this web application reveal appropriate and useful but many others widgets can still be incorporated in system. Could be inserted a counter to count the number of tweets in the same avenue.

Chapter 7

Future Work

Although we feel the goals for this dissertation were met, there are some aspects that can still be improved. We believe that the project TwitterJam can always be improved and can always be better, whether changing the spatial extendend correlation coefficient to confirm data or by extending implemented features in the web application. Some of the most important aspects to be improved are:

- **Data:** Improve the size of collected data. With a larger dataset, a deeper analysis can be done. Also, we would like to select other regions to analyse since it could be an improvement to the project. The selection of a region where there are official information or road traffic sensors would be excellent.
- **Data processing:** Improve the identification and categorization of general tweets with road traffic information. This could, possibly, enable an increase of accuracy in the system. In general tweets besides getting a greater number of avenues, a future work could also get a better analysis of tweets context to create degrees of road traffic intensity and correlate this degree with the number of official tweets.
- **Correlation:** About the correlation coefficients, there is also some improvements to be implemented. The first hypothesis of space extended correlation coefficient can be improved, inserting new elements. to get more accurate results and the second hypothesis is influenced by the first to correlation coefficients, so to improve results data collection must be bigger.
- **Wep application:** For the web application a big improvement is to transform the platform from off-line to on-line mode and constantly updated with road traffic information. Also, it would be a good improvement to create a widget to measure the most mentioned words on general tweets and a widget that shows the intensity of road traffic in combination with degrees presented in Data Processing point. An example is to implement a choropleth map marking the more intense regions.

References

[1] Twitter. The Streaming APIs. Available at. https://dev.twitter.com/docs/streaming-apis. [Online; accessed 10-01-2015].

[2] Sri Krisna Endarnoto, Sonny Pradipta, Anto Satriyo Nugroho, and James Purnama. Traffic condition information extraction & visualization from social media twitter for android mobile application. In *Electrical Engineering and Informatics (ICEEI), 2011 International Conference on*, pages 1–4. IEEE, 2011.

[3] Spatial Data Mining and Visual Analytics Lab. Flow Mapping with Graph Partitioning and Regionalization, howpublished = "http://www.spatialdatamining.org/software/flowmap", note = "[online; accessed 10-01-2015]".

[4] Axis Maps LLC. Choropleth maps, howpublished = "http://indiemapper.com/app/learnmore.php?l=choropleth", note = "[online; accessed 10-01-2015]".

[5] Xiaotong Liu, Yifan Hu, Stephen North, and Han-Wei Shen. Compactmap: A mental map preserving visual interface for streaming text data. In *Big Data, 2013 IEEE International Conference on*, pages 48–55. IEEE, 2013.

[6] Alexander Boettcher and Dongman Lee. Eventradar: A real-time local event detection scheme using twitter stream. In *Green Computing and Communications (GreenCom), 2012 IEEE International Conference on*, pages 358–367. IEEE, 2012.

[7] Xiying Wang and Dan Cosley. Tweetdrops: a visualization to foster awareness and collective learning of sustainability. In *Proceedings of the companion publication of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 33–36. ACM, 2014.

[8] Alan M MacEachren, Anuj Jaiswal, Anthony C Robinson, Scott Pezanowski, Alexander Savelyev, Prasenjit Mitra, Xiao Zhang, and Justine Blanford. Senseplace2: Geotwitter analytics support for situational awareness. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, pages 181–190. IEEE, 2011.

[9] Dennis E Hinkle, William Wiersma, and Stephen G Jurs. Applied statistics for the behavioral sciences. 2003.

[10] Tiago Cunha, Carlos Soares, and Eduarda Mendes Rodrigues. Tweeprofiles: detection of spatio-temporal patterns on twitter. In *Advanced Data Mining and Applications*, pages 123–136. Springer, 2014.

[11] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65. ACM, 2007.

[12] Aaron Smith and Joanna Brenner. Twitter use 2012. *Pew Internet & American Life Project*, page 4, 2012.

[13] Alex Burns and Ben Eltham. Twitter free iran: An evaluation of twitter's role in public diplomacy and information operations in iran's 2009 election crisis. 2009.

[14] Lev Grossman. Iran protests: Twitter, the medium of the movement. *Time Magazine*, 17, 2009.

[15] Amanda Lee Hughes and Leysia Palen. Twitter adoption and use in mass convergence and emergency events. *International Journal of Emergency Management*, 6(3):248–260, 2009.

[16] Rosaldo J. F. Rossetti Sara Carvalho, Luís Sarmento. *Real-Time Sensing of Traffic Information in Twitter Messages*. Proceedings of the IEEE ITSC 2010 Workshop on Artificial Transportation Systems and Simulation (ATSS'2010), Madeira Island, Portugal, September 19, 2010.

[17] Zafeiris Kokkinogenis, João Filguieras, Sara Carvalho, Luís Sarmento, and Rosaldo J. F. Rossetti. Mobility network evaluation in the user perspective: Real-time sensing of traffic information in twitter messages. *Advances in Artificial Transportation Systems and Simulation, Boston: Academic Press*, 12:219–234, 2015.

[18] Lúcio Sanchez Passos, Rosaldo JF Rossetti, and Zafeiris Kokkinogenis. Towards the next-generation traffic simulation tools: a first appraisal. In *Information Systems and Technologies (CISTI), 2011 6th Iberian Conference on*, pages 1–6. IEEE, 2011.

[19] Twitter. Twitter, REST API. https://dev.twitter.com/overview/documentation, 2014. [Online; accessed 10-01-2015].

[20] Twitter. Twitter, Streaming API. https://dev.twitter.com/streaming/overview, 2014. [Online; accessed 10-01-2015].

[21] Matko Boanjak, Eduardo Oliveira, José Martins, Eduarda Mendes Rodrigues, and Luís Sarmento. Twitterecho: a distributed focused crawler to support open research with twitter data. In *Proceedings of the 21st international conference companion on World Wide Web*, pages 1233–1240. ACM, 2012.

[22] Prakash M Nadkarni, Lucila Ohno-Machado, and Wendy W Chapman. Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, 18(5):544–551, 2011.

[23] Peter Jackson and Isabelle Moulinier. *Natural language processing for online applications: Text retrieval, extraction and categorization*, volume 5. John Benjamins Publishing, 2007.

[24] Gobinda G Chowdhury. Natural language processing. *Annual review of information science and technology*, 37(1):51–89, 2003.

[25] Napong Wanichayapong, Wasawat Pruthipunyaskul, Wasan Pattara-Atikom, and Pimwadee Chaovalit. Social-based traffic information extraction and classification. In *ITS Telecommunications (ITST), 2011 11th International Conference on*, pages 107–112. IEEE, 2011.

[26] Sílvio S Ribeiro Jr, Diogo Rennó, Tatiana S Gonçalves, Clodoveu A Davis Jr, Wagner Meira Jr, and Gisele L Pappa. Observatório do trânsito: sistema para detecção e localização de eventos de trânsito no twitter. *Simpósio Brasileiro de Bancos de Dados*, 2012.

[27] Mark Gahegan. 11 visual exploration and explanation in geography analysis with light. *Geographic data mining and knowledge discovery*, page 291, 2009.

[28] C Pettit, IVO WIDJAJA, P Russo, RICHARD SINNOTT, R Stimson, and MARTIN TOMKO. Visualisation support for exploring urban space and place. International Society for Photogrammetry and Remote Sensing, 2012.

[29] Gustavo Cardoso, Sandro Mendonça, Tiago Lima, Miguel Paisana, and Marta Neves. A sociedade em rede em portugal 2014 – a internet em portugal. Technical report, OberCom - Observatório da Comunicação, 2014.

[30] Kalev Leetaru, Shaowen Wang, Guofeng Cao, Anand Padmanabhan, and Eric Shook. Mapping the global twitter heartbeat: The geography of twitter. *First Monday*, 18(5), 2013.

[31] Nick Cohn. Tomtom traffic index: Toward a global measure.

[32] Comparision between Vicenty Formula, Haversine Distancte formula and Law of Cosines. http://jsperf.com/vincenty-vs-haversine-distance-calculations-test/3. [Online; accessed 15-04-2015].

[33] Samuel de Oliveira Carvalho. Sistema de monitoramento de veículos de transporte público. 2013.

[34] Rosaldo JF Rossetti, Eugénio C Oliveira, and Ana LC Bazzan. Towards a specification of a framework for sustainable transportation analysis. In *13th Portuguese Conference on Artificial Intelligence, Guimarães, Portugal*. Citeseer, 2007.

[35] Rosaldo JF Rossetti, Paulo AF Ferreira, Rodrigo AM Braga, and Eugénio C Oliveira. Towards an artificial traffic control system. In *Intelligent Transportation Systems, 2008. ITSC 2008. 11th International IEEE Conference on*, pages 14–19. IEEE, 2008.

[36] Ana Zaiat, Rosaldo JF Rossetti, and Ricardo JS Coelho. Towards an integrated multimodal transportation dashboard. In *Intelligent Transportation Systems (ITSC), 2014 IEEE 17th International Conference on*, pages 145–150. IEEE, 2014.

[37] Rosaldo JF Rossetti, Ronghui Liu, and Shuming Tang. Guest editorial special issue on artificial transportation systems and simulation. *IEEE Transactions on Intelligent Transportation Systems*, 2(12):309–312, 2011.

[38] Rosaldo JF Rosetti and Ronghui Liu. *Advances in Artificial Transportation Systems and Simulation*. Academic Press, 2014.