Masters' Degree in Informatics Engineering
Dissertation
Final Report

Sensing the Buzz

Fábio Miguel Ferreira Pedrosa
fmfp@student.dei.uc.pt

Advisors:
Francisco C. Pereira
João P. Vilela

August 31, 2012

FCTUC DEPARTAMENTO
DE ENGENHARIA INFORMÁTICA
FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

Acknowledgements

I am grateful to my advisors, Francisco C. Pereira and João P. Vilela, who have supported me with their knowledge and guidance throughout this dissertation while simultaneously allowing me the room to work in my own way.

I would like to express my regards to my colleagues at the AmiLab (Ambient Intelligence Lab) for the contributions and opinions.

I would like to show my gratitude to my friends, particularly Marco Simões, Nuno Lourenço, Vitor Silva, Ivo Gonçalves, Alexandre Martins and Nuno Martins for all the time that we have spent together, and the constant encouragement.

To my family and friends. Thank you for the support and encouragement.

Fábio Pedrosa
Coimbra, August 2012

Abstract

Events are a strong indicator of life in a city, and with the rise of real-time ubiquitous social-networks in recent years, never before the dynamics of a city were so publicly exposed. The amount of data being generated is actively growing both in online sources, like social-networks, and offline with sensors from traffic, public transport and taxis becoming more and more available.

The aim of this dissertation is to conduct analysis and apply machine learning techniques at both online and offline data looking for knowledge of how they relate and recognize what are their implications on mobility. We were able to show significant buzz in the Twitter network around some big events, find some correlation between traffic phenonoms, namely incidents, and discover significant flow effect following some big music concerts.

This dissertation presents an effort to relate large amounts of data from both online and offline sources, most from gathered unreliable sources such as the traffic data generated from road sensors, and of questionable quality, such as the twitter data with its poor language use. Work here is presented in the same way it was performed, in a exploratory sense without a good precise knowledge of which dataset or which technique would guarantee the best results.

Keywords: Social data mining, Collective Intelligence, Ubiquitous Computing, Information Retrieval

Contents

List of Figures

List of Tables

Chapter 1

Introduction

We live in a world where the amount of digital information has never been so large. The number of people with Internet access worldwide has more than doubled from 1,043 billion in 2006 to 2,267 billion in 2012 [23].

Accompanying this Internet population growth has been the mass adoption of social platforms such as Facebook, Twitter, Google+, LinkedIn, MySpace, FourSquare, GoWalla and Wikipedia. These services work around data about people such as information on their *personal life*, their *location*, their *professional network* and even their *taste* on music, movies, etc. Furthermore, there are services like Upcoming, Zvents and Eventful reporting *events* that will happen in the real physical world.

Following this mass usage there is also a large trend on adopting GPS-capable mobile devices as platform for accessing all these online services, and not only bringing this access to their every-day life but also making location matter. People can now create georeferenced data everywhere they go, and location-aware applications have closely followed this trend with inovative and creative uses of this information. There is now large amounts of real-time georeferenced data being created everyday, and most of it is publicly available.

Services like Facebook have become very popular by allowing users to share information about their life and get in touch with their friends. Facebook currently report about 955 million active users (Figure 1.1) with more than 50% entering the website in any given day. Nearly 45% of this users access the website through their mobile device[1]. Twitter is another social platform currently reporting a number of 400 million tweets per day being shared (Figure

[1] http://www.facebook.com/press/info.php?statistics

1.2) by asking its users the question "What are you doing?".

Unlike bloggers that create a new post every few days, twitters share several tweets a day.

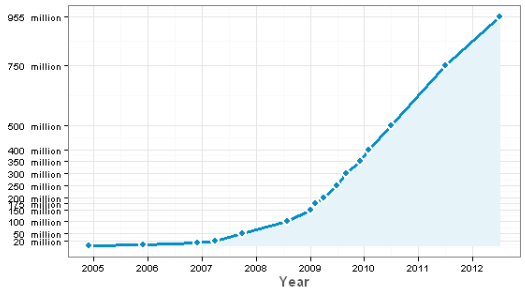

Figure 1.1: Active users on Facebook

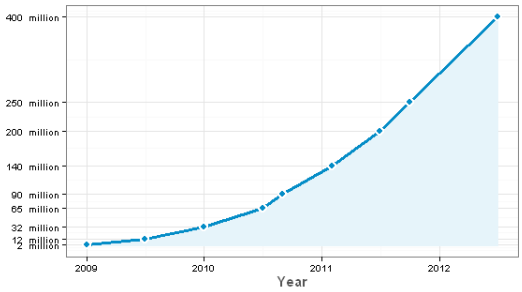

Figure 1.2: Tweets sent per day

On par with this visible growth in Internet usage there is now increasing deployments of sensors, specially in big cities, allowing new approaches to be taken for the study of the environment. Cities like Singapore and London are data sinks with large amounts of elements being tracked real-time: car-counting loop detectors, cellphone usage, public transport tracking, car parks usage, to name a few.

Ideally, by leveraging the fact that through this data we get a near real-time updated view of current transport demand, we could automatically inform traffic prediction systems. A pragramatic example would be our system detecting a trending event located somewhere in the city and informing the network of public transports so that they could avoid that specific zone. Ultimately this would work to provide intelligence and make general mobility smarter.

We found a clear **motivation** provoked by this large growth in data available and made it our **goal** to explore the connection between these two different kinds of data to first understand what can be extracted from this data, and secondly understanding how they relate to everyday life behavior.

1.1 Objectives

Our **objective** is to create a system able to retrieve the popularity of real-life events from unstructured real-time data with the purpose of bringing that information into understanding what the mobility implications might be.

A methodology was developed for this system to gather all the data available in several online sources, and extract relevant information to measure the popularity of events. Mixing real-time information generated in services like Twitter.com and private data provided by traffic entities, we aim to collect more knowledge about the mobility in a city.

Our objectives can be summarized by the following list:

- **Retrieve data** from relevant sources such as: mobility traces from traffic authorities, real-time georeferenced content from social networks and collection of events happening in the city;
- Analise and assess the quality of the collected data;
- A methodology to **semantically enrich** the collected data;
- Estimate **real world participation** numbers though online popularity;
- Study the relation between sensory data and online data;

This report starts by presenting a review of the literature on fields related to our work, followed by an overview in the approaches we use to attain our objectives. We follow by delving into several experiments exhibiting our approaches with the available data and presenting our results. We finish with our conclusions for this dissertation and present some thoughts on how we get further results continuing the work presented.

Chapter 2

State of the Art

This chapter presents the theoretical foundations and important work related to the subject of this dissertation. We start by introducing *Information Extraction*, focusing in the various stages required to transform unstructured web data into useful structured data. In the following sections we describe two other relevant research fields related to information extraction, being *Topic Modeling* the field interested into modeling hidden thematic information from documents, and *Semantics* the field concerned with extracting meaning from text.

2.1 Social Networks

The last few years have definitely been tremendous in the growth of both social networks and data published on the web, specially georeferenced content. We believe this growth can help us with knowledge extraction by using this data as real-time sensors in the physical world. At the same time, traffic authorities are leveraging emerging technologies like WiFi, RFID and Bluetooth to improve the quality of data from the city sensors. By merging the knowledge extracted from social content with a updated view of traffic and transport demand we believe to be possible the development of a system able to inform traffic prediction systems on mobility effects.

Currently few social platforms exist with this mass-adoption effect, namely Facebook, Twitter and Google Plus (Google+). These are social networks built around the concept of microblogging where users share with followers a short status of text or other popular format like pictures, video and webpages.

There are a lot more social networks but they are mainly concerned with smaller communities like Flickr (photos), LinkedIn (professional), Last.fm (music), MySpace (music), Youtube (movies) and QQ (only China).

Because the knowledge we want to gather requires almost real-time updating we can filter this list to the general networks: Facebook, Twitter and Google Plus. We briefly describe them bellow:

Facebook

Facebook is clearly the most popular and biggest social network in the world with 800 million active users. According to Facebook there are more than 2 billion posts liked and commented and 250 million photos uploaded every day [15].

There is an API to make data access easier, but because users in this network actually share a lot of information about themselves, Facebook implements several restrictions on data access by machines.

By default a user can only access the stream of posts (called *Wall*) from another user if both actually form a friendship (both parties need to agree), making most user's information unavailable for machine crawling. Facebook also does not allow crawling for information by detecting patterns in the way the requests are made, making it difficult to crawl users using a tree transverse pattern like depth-first or breadth-first search.

There is some interesting related research being developed inside Facebook like measuring a USA political candidate buzz using topic modeling and sentiment analysis, creating a national happiness index by segmenting posts based on location and sentiment analysis, and characterizing the baseball fan base [14].

According to Gilbert et al. [18], Facebook users create relationships in a loosely form [20] by adding people they do not interact with. Research has shown that using the frequency of interaction between two people was a better way to quantify relationships than the list of friends they would add. This can be seen across other networks, but specially on twitter where a user can follow a user status without him following back.

Google Plus

Less than a year old, Google+ reports currently more than 90 million users [21]. With more than 65% of users being from USA and a male to female ratio of 66%/34%, this network still has not seen mass adoption in other countries and both sexes. The network should require more adoption until is possible to extract meaningful and less biased information from their data.

Google+ offers an API to access the data but limits queries to 1000 per day.

Twitter

A tweet consists of a small text (maximum of 140 characters) supposedly answering the question "What's happening?". Most of the time users rely on this service to share messages about their own personal life, share public information and keep conversations with other users. Unlike Facebook, users on Twitter can follow another user without his consent, making the friendship not mutual.

According to Naaman Mor et al. [34] almost 80% of users spend most time talking about themselves and only 20% spend more time sending tweets that are informational in nature. The latter have a significant larger number of friends (people subscribing to their messages). Informative tweets are usually about timely topics [29], meaning the subject is often the same on other sources at the same time, like local or world news.

Just during the *Fifa Women's World Cup* Twitter received about 8900 tweets per seconds and this social platform even keeps a higher number of people posting information using their mobile, 46%.

Three mechanisms that were introduced by twitter must be mentioned:

- Users can mention each-other by using an at sign "@" before the username. For example: "@fabiopedrosa is mentioned in the tweet";
- Tweets can be re-tweeted, allowing for users to share an existing tweet to their friends. It usually follows that retweeting makes known the user's interest in the subject of the tweet;
- Because of their size some ways of shortening were created. HashTags are the most popular by using a Number sign before a topic. For example: "Académica playing tomorrow #football". Emoticons and external links are often used too.

Twitter has been the subject of plenty of research, some have analyzed the structure of this microblogging platform [29, 16, 47], some focused on analyzing the diffusion of content through the network [51, 55] and several have leveraged the fact that we can get data in nearly real-time to research applications like messages of public interest [24], earthquake prediction [43] or epidemics detection [3].

Because tweets are by design very short, they provide too little content for most text analysis. For instance, comparison between tweets becomes difficult. On

Original	Post-Processing
@user3419 nay lol y u say dat? &wat u doing 2day?	No, why did you say that? What you doing today?
1001 colors: Contemporary art from Iran <URL> #Iran #culture #Art	1001 colors: contemporary art from Iran <URL>.
it's soo quiet, it's like I'm goin die	It is so quiet, it is like i am going to die.
#worstfeeling buyin a fresh laptop..then ur screen blowz out :((	worst feeling is buying a fresh laptop.. then your screen blowz out.
This is superb Grape+apple splash with manggo juice, super!	This is superb grape + Apple splash with mango juice, Super!
@user31903 u n ur fam can n if u interested ill b n touch w u bout it	You and your family can and if you interested Ill be and touch with you about it.
RT: @user4191 BEAUTIFUL CREATURES has a new #website designed by @user4192!	Beautiful creatures has a new website designed by @user4192!

Table 2.1: Text normalization example

top of that, Twitter is very noisy as a source of data, containing lots of OOV words (out-of-vocabulary), acronyms, words shortened (e.g. *nite* instead of *night*), emoticons, frequent misspellings, unfamiliar entities, slang, links and other forms of text lacking any understanding.

Handling this noise poses a challenge and is currently subject of wide consideration. A frequent technique for this kind of problem is transforming in a way this text to make it more consistent. This process is called Text normalization and is usually applied before other Natural Language Processing (NLP) processing is done. Because this process is similar to that of a translator algorithm, is often evaluated using a metric called Bilingual Evaluation Understudy (BLEU). Quality under BLEU is considered to be better if the number of words in the machine translation matches the words given by human translation (each word in the candidate translation must be present in the human reference translation). Because this metric will not be used in this work, for a more extensive description we refer to [37].

Influenced by the research done on normalization of short-message phone texts (SMS) [5] and the lack of good results in applying Named Entity recognition to tweets [32], Kaufmann and Kalita have applied *Moses*, a statistical machine learning package, to improve BLEU precision scores from 68% to 79% [27]. Table 2.1 presents the result of this normalization to several tweets:

Another trend of research gives less value to the text in the tweets and focuses on other related information. For instance Sriram et al. [48] combined the user profile information with tweets to classify the text according to a predefined set of classes.

Others have focused their research on the graph created by the social network. Weng et al. [54] created an extension to the popular PageRank algorithm formulated by Google, called TwitterRank. In Google PageRank every node has a rank determined by the sum of rank the nodes linking to it have, formulating a recursive algorithm. TwitterRank measures the influence in the same way by taking into account both the topical similarity between users and the link structure.

Daniel Romero et al. [42] also created an algorithm in similar fashion to determine the influence and passivity of users based on their information forwarding activity. This algorithm is similar to HITS, a predecessor of PageRank, which separated the web pages into two classes, hubs and authorities. Hubs work as directories for other pages, while authorities are pages that are hyper-linked by many hubs. When applied to Twitter, results show that popular users can get the attention of millions of people while unable to spread their message very far. Showing that popularity and influence correlate weekly.

Interest

By looking at these networks we can easily imagine the large amount of data currently flowing in them and keep in mind that all these platforms (Twitter, Facebook and Google+) were created only after 2004, making them at most 8 years old. Research has shown increasingly over the last few years that this data can help predict real-life phenomenons and people's behavior.

Twitter in itself holds the better privacy model for research purposes, as it allows access to all public profiles (they are public by default) and its usage is mainly mobile thus creating more georeferenced content. Facebook and Google+ have a private by default privacy model severely restricting the amount of data we could possibly retrieve.

2.2 Information Extraction

Information Extraction is a particular type of Information Retrieval (IR) whose main goal is to transform unstructured machine-readable data into structured information [10]. The most common techniques rely on NLP which will be discuss in section 2.2.2 after presenting useful methods to retrieve our information from online sources in section 2.2.1.

2.2.1 Data Gathering

Data availability is an important requirement for our work progress. Working with private and public data using various formats, methods for information retrieval are critical. This section will briefly describe our methods for gathering data that is available publicly on-line.

Two main types of access exists for data online: data available using a public API (Application Programming Interface) and data only available through scrapping of web pages.

Web scraping relies on a custom retriever application that fetches web-pages using the commonly used HTTP protocol, usually requiring the client to interact with a server simulating the behavior of a Web Browser in order to get documents in HTML format [44]. This interaction can often get complex to handle when the website requires authentication, a particular order of requests or even manipulation of cookies. To extract useful information from this HTML documents we can use one of the following scrap methods:

- **Manually**: Although tedious and time-consuming, it is very easy and less prone to undetected errors in data.
- **String Manipulation**: Achievable in any programming language, often offers little flexibility to changes in the source.
- **Regular Expressions**: Requires some experience to write a simple and fast expression, but is often the most flexible method.
- **DOM transversing**: DOM is short for Document Object Model, a HTML standard to represent XML/HTML documents in memory allowing interaction using a commonly known API. This method is not considered most of the times because it requires a large amount of memory and an almost perfect document structure (which does not happen often with web pages).

2.2.2 Natural Language Processing

NLP is a cross disciplinary field combining Computer Science with Linguistics, with the general goal of allowing machines to understand human generated language. Using techniques developed in this field we aim to collect well-defined data that can be used to allow logical reasoning and inference based on the relevant content of the document.

Research in this field has resulted in several techniques, some of which are not suitable for our work, like for example statement parsing (requires a good grammatical structure) or automatic summarization (content is small already). As shown before on section 2.1, the data collected on social networks requires some special consideration because it is composed of a large quantity of documents but with a very small length individually.

NLP is usually segmented in a collection of subtasks in pipeline fashion. In our work the following procedures should be considered:

- **Content Noise Removal** is required in any kind of human generated text specially in social networks where noise is high. Usually follows that URLs, symbols, hashtags, and other easily recognized useless data is removed from documents. In Kaufmann et al. [27] tweets are normalized in this stage, before any other processing. Cui et al. [12] have found that emotion tokens (e.g. emoticons) provide the most significant feature to predict tweets sentiment, something that should be taken into consideration on noise removal.

- **Tokenization** is the process of splitting a document into sentences and words. This is a critical phase with implications on the following stages because useful information could be lost in the process. On documents like tweets this becomes harder because we need to break the text into smaller meaningful parts but almost no character is safe to do this besides the space (e.g. “tl;dr”, “2day”, “@john”). Laboreiro et al. in 2010 presented results using a SVM classifier trained with manually tokenized tweets achieving an F-score of 0.96 [30]. This work tackles difficult cases like “...bring the phone.and never...”, “Thank you last.fm!” and “Be a bone marrow donor-Show up in...”. It is critical to find a good way to tokenize this kind of erroneous text.

- **Stop-Words** are common words in all documents that need to be removed before further processing. Typically this corresponds to words like *the*, *is*, *at*, *which* and *on*. The list creation cannot usually be automated and is done by looking at the frequency of occurrence of words in

a big dataset of documents within the same language domain. Extra care should be made to leave collocations (words which commonly co-occur) unaffected. Examples of this are *new* in "New York" or *big* in "Big Mac". Named entities are usually unaffected as they are generated by looking at the raw tokens.

- **TF-IDF** stands for term frequency-inverse document frequency and is used as a weight in statistical measure for each word. We want each word to increase in importance proportionally to the number of times it occurs in the document, but decrease by the frequency it occurs in the full dataset/corpus. This is calculated with formula 2.1:

$$w_{i,j} = tf_{i,j} \times log\left(\frac{N}{df_i}\right) \tag{2.1}$$

  where $tf_{i,j}$ equals the number of occurrences of terms i in j, df_i equals the number of documents that contain the term i and N is the total number of documents. Logarithmic scale is used to match word frequency distribution that follows a Zipf law (power law).

- **Part-of-Speech (POS)** tagging is useful for detecting syntactic word classes such as noun, adjective, verb, etc. Two main approaches exist: rule-based and stochastic. An important work in the area is that of E. Brill [9] which combined these two approaches together and achieved 96.5% overall precision.

- **Name Entity Recognition (NER)** identifies proper names within documents. A proper name can either be people, places, companies, brands, and others. Some NER work without any prior POS information and work directly with raw tokens, thus able to recognize a larger set of distinctive features (capitalization, adjacent words, etc). Laferty et al. presents the state-of-the-art method using a statistical modeling method called Conditional Random Fields (CRFs) [7].

Summarizing this whole process, NLP is usually done in these stages: removing noise from the data and tokenizing every document available. After this NLP labelers such as POS and NER are usually applied to obtain information in structured form from these documents.

A number of NLP frameworks already exist with the capacity to work with well defined corpus, such as:

- **Kusco** [2] is a Framework developed by Ana Alves, a Phd student from our workgroup AmILab at Universidade de Coimbra. This framework

is able to process data in the same pipeline fashion presented before, including POS tagging, Noun phrase chunking and NER. And also offers the ability to match named entities using WordNet and Wikipedia as common sense resources. The final result is a ranked list of concepts called semantic index. Figure 2.1 displays the system architecture.

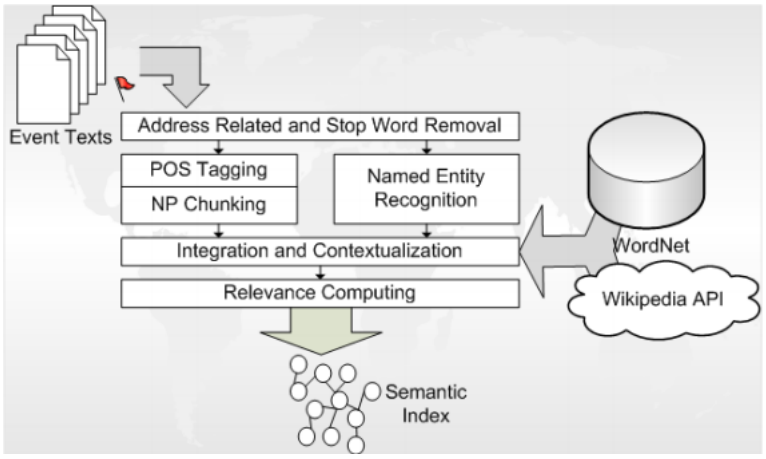

Figure 2.1: Kusco Architecture [2]

- **Stanford CoreNLP** [17] is another framework developed by the Stanford NLP Group that provides NLP tools for several tasks, including the ones mentioned before. It includes a coreference resolution system able to detect word dependencies such as "Obama is running. He is tired." where the subject is the same in both sentences.
- **NLTK**[1] is a module created for Python containing most NLP tools mentioned before, including useful tools like NER, wordnet connector, corpus readers and weka interface.
- **OpenCalais**[2], **Extractiv**[3], **AlchemyAPI**[4] and **DBpedia Spotlight**[5] are examples of public available web-services where it is possible to get entities, facts and events from unstructured text. Since most are closed source we cannot get a deep knowledge of how they work and actually achieve their results.

[1]http://www.nltk.org
[2]http://www.opencalais.com
[3]http://www.extractiv.com
[4]http://www.alchemyapi.com
[5]http://spotlight.dbpedia.org

2.2.3 Representation

Vector-Space Model is the standard representation for documents in NLP processing [46]. It works by creating an algebraic model of representing all terms in vectors such as:

$$D = (t_1, t_2, t_3, ..., t_i) \tag{2.2}$$

where each element t represents a dimension corresponding to whether that term is present or not in the document. Because some terms are more important than others we also include a weight for each term:

$$D = (t_1, w_1; t_2, w_2; t_3, w_3; ....; t_i, w_i) \tag{2.3}$$

These weights can be calculated using TF-IDF or some other weighting algorithm.

Comparing between two documents is also easier using this representation. By calculating the cosine of angle between two vectors we can obtain the similarity between them, called the *cosine vector similarity*. The expression for this is:

$$similarity(Q, D) = \sum_{i=1}^{l} w_{Qi} \times w_{Di} \tag{2.4}$$

where Q and D represent two documents [46].

2.2.4 Sentiment Analysis

Evaluating the positiveness of the content can be a useful metric for some cases. This area, focused on identification of emotion or sentiment by a piece of text, is referred to as Sentiment Analysis. It is often called opinion mining because of the popularity it has gained in real world applications.

In 2002 D. Turney [50] achieved an average accuracy of 70% when evaluating 410 online reviews according to their semantic polarity. The classification is predicted by the average semantic orientation of the phrases in the review that contain adjectives or adverbs. A phrase has a positive semantic orientation when it has good associations (e.g., "loved it") and a negative semantic orientation when it has bad associations (e.g., "sick feeling"). The semantic orientation of a phrase is calculated as the mutual information (computed using statistics gathered by a search engine) between the given phrase and the word "excellent" minus the mutual information between the given phrase and the word "poor". A review is classified as recommended if the average semantic

orientation of its phrases is positive. Pang et al. still in 2002 [36] improved this result by applying a SVM classifier in the same context of movie reviews, achieving an accuracy of 82.9%.

When applied in Twitter, emoticons have been shown [19] to provide the best feature for polarity identification achieving 81.3% with Naive Bayes and 82.2% with SVM. This research goes to show that the most distinct feature in this case is reduced to 2 to 4 characters in a document, fully ignoring the context of the message. Wang et al. [52] also researched sentiment analysis applied to Twitter content but through graph networks, by considering two users sharing the same tweet hashtag (e.g. "#politics") as edges between them (thus sharing the same topic), and messages as the source of polarity. Figure 2.2 show an example of this.

Figure 2.2: Mapping sentiment analysis to hashtags across twitter network. Red hashtags present negative sentiment, green present positive sentiment.

2.3 Topic Modeling

Social networks like Twitter and Facebook allow users to exchange messages that are mostly topical, following a common subject or even sometimes multiple [47]. Topic modeling is a statistical model frequently used as NLP tool to discover hidden topics in a collection of documents.

First presented by D. Blei et al. [8], the Latent Dirichlet allocation (LDA) is a generative model that has been used to gather thematic information on many different cases like detecting disaster-related tweets [28], finding the most influential people to each topic on Twitter [54] or even identifying trends on research by looking at papers submitted to conferences [22]. We briefly present this model bellow.

2.3.1 LDA

Each document is represented as a probability distribution over some topics, while each topic is represented as a probability distribution over a number of words. The topic distribution is assumed to have a Dirichlet prior. Common words in language will tend to have similar probabilities across all topics (making them irrelevant) while some topic-sensitive words will have different probabilities for specific topics. It assumes a generative process for generating each document as follows:

1. Choose $\theta \sim Dir(\alpha)$
2. For each of the N_d words w_n in the document:
   (a) Choose a topic $z_n \sim Multinomial(\theta)$.
   (b) Choose a word w_n from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

Figure 2.3 presents the model using the plane notation, as they have been presented in [8]. In this notation, an arrow corresponds to a conditional dependency between two variables and boxes indicate repeated sampling with the number of repetitions given by the variable in the bottom of the corresponding box.

Formally, each of a collection of D documents associated with a multinomial distribution over T topics, which is denoted as θ. Each topic is associated with a multinomial distribution over words, denoted as ϕ. θ and ϕ have Dirichlet prior with hyper-parameters α and β respectively. For each word in one document d, a topic z is sampled from the multinomial distribution θ associated with the document, and a word w from the multinomial distribution ϕ associated with topic z is sampled consequently. This generative process is repeated N_d times. The model has two parameters to be inferred from the data, i.e. document-topic distributions θ, and the T topic- word distributions ϕ [54].

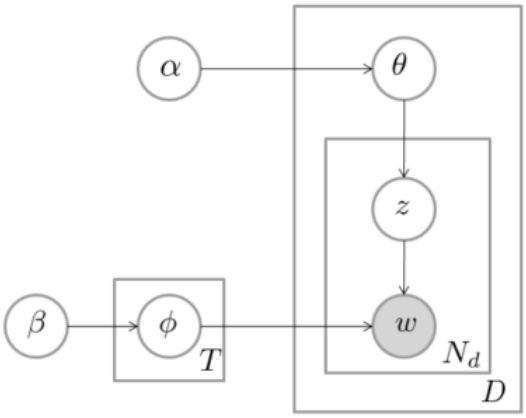

Figure 2.3: Graphical representation of LDA Model.

Because this is a generative model, it is useful for generating documents according to predefined topics distributions. But we are interested in reverting this process, inferring the set of topics that was used to generate specific documents, like public messages on social networks. Figure 2.4 illustrates how this process is inverted.

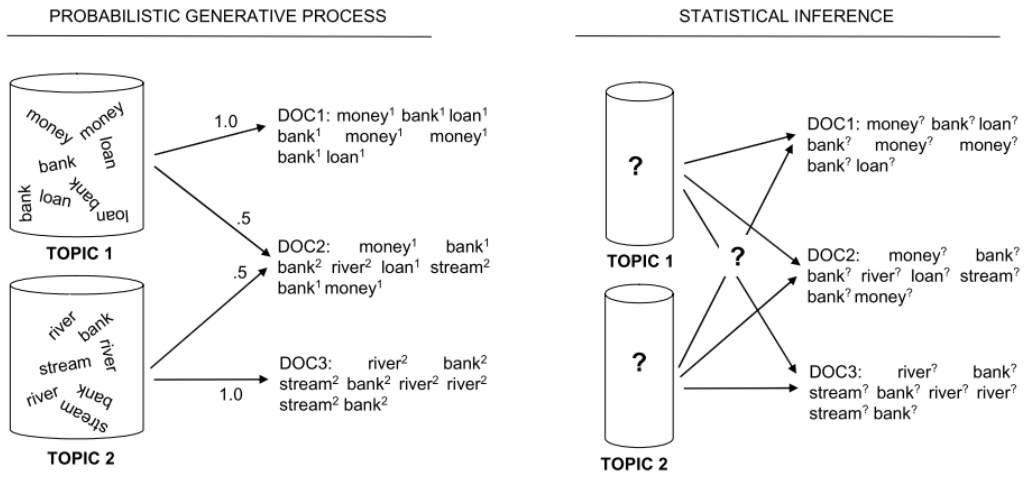

Figure 2.4: Inversed role in using a generative model to infer underlying topics [49]

This inference has first been covered in the research using a variational Bayes approximation [8], but more recent research has shown Gibbs sampling to be faster and better [25, 22, 49]. This inference can be done following this process:

- Initialize the word distribution on all topics by randomly assigning a topic to each document.
- For each word w in document d compute for each topic t the following:
  - $P(t|d)$, the proportion of words in document d that are currently assigned to topic t.
  - $P(w|t)$, the proportion of assignments to topic t over all documents that come from w.
- Assign w a new topic $t2$ chosen with probability $p(t2|d) \times p(w|t2)$. We reassign the word to the topic most likely to have generated this word.

This learning algorithm is repeated a large number of times until is found convergence in the quality of topic assignment. Convergence can be validated using perplexity, a measurement of information entropy, measuring the amount of uncertainty. Perplexity can be defined as:

$$\exp\left(-\sum_{d}^{M} \log w_d / \sum_{d}^{M} N_d\right) \tag{2.5}$$

for a set of M documents of length N_d. The lower perplexity we achieve for the model, the better fit the model is to the data [8].

Pozdnoukhov and Kaiser presented a methodology similar to this dissertation. They leveraged the use of LDA as topic modeling tool to presented several case studies where topics were extracted from a georeferenced streaming of tweets to develop event tracking such as a local music festival and the release of a Harry Potter movie, a global event causing spatially heterogeneous response. They conclude that topics found in streams are related to various events and inhere typical spatial patterns [40].

2.4 Semantics and Concepts

Figuring out the relations between words, phrases and symbols to understand the meaning of them is called Semantics. Because of the amount of concepts that exist, and the large quantity of ways humans can express their message, this is something considerably hard for machines to do well.

All the NLP tools mentioned previously in our work do not inherently require a *deep* semantic understanding of the data to work. They are able to leverage basic syntax and tokens to extract key features like the most significant terms

and even abstract topics across multiple documents, but without a clear relation between words/phrases there will always be an inability to understand what is being said and how they relate. More specifically, without deep semantics we are unable to disambiguate between similar entities and thus incapable of knowing if two entities with the same syntax refer to the same concept.

Wordnet[6] is often used as a lexical database to contextualize the use of words. Through the graph of relations between nouns, adjectives and adverbs, research works have shown the ability to disambiguate most terms [35, 11, 24, 39, 53].

A bigger effort is being led by the W3C called the Semantic Web, which promotes the inclusion of semantic content in data online. Tim Berners-Lee defined it as "a web of data that can be processed directly and indirectly by machines." [6]. This movement has seen some progress but it is still not fulfilled, and some say it will never be [33]. The main limitations of this work have been the standards used for data exchange and the need to have a complete database of concepts, ontologies.

Projects like Wikipedia[7], OpenCyc[8] and ConceptNet[9] have been used as *commonsense knowledge databases* because they contain most general knowledge people possess represented in ways that the computer can use to make inferences about the concepts expressed by language [41, 11, 32].

Although these databases have been shown to produce good results in the enrichment of events information [35], it is not known to which capacity we can use these in micro-blogging messages. Because of the size tweets have, context is very low and thus make the task of inferring the semantic meaning prone to greater error.

Motivated by research that have shown that over 85% of tweets posted everyday are related to news [29], Abel et a. [1] have devised a way to gather more context for tweets: by linking them with news articles. This linkage happens using several strategies, namely by the following:

1. Tweet contains at least one URL from selected mainstream news publishers.
2. Tweet is reply or re-tweet from another tweet caught by strategy 1.
3. Tweet is related to the article in which the TF-IDF is maximized with the article title.

[6] http://wordnet.princeton.edu
[7] http://www.wikipedia.com
[8] http://www.cyc.com/opencyc
[9] http://conceptnet5.media.mit.edu

4. Tweet is related to the article by maximizing the TF-IDF of its hashtags and the article title.

5. Tweet shares common entities with news article.

Results show that the first strategy proved to be the most relevant, and an approximate 15% of tweets were linked to news articles[1]. This type of user modeling could be useful in our work, by finding links between tweets and other dataset richer in semantic context like the events description or Wikipedia.

2.4.1 Text Classification

Documents are often labeled with a pre-defined set of categories in which the documents belongs to zero, one or more, and the task of making that assignment is called Text Classification. This categorization is usually used to enhance the information retrieval algorithms [10].

The task of text classification can be similarly to the task of determining the topic of the text (topic modeling) but in which the topic is selected from a pre-defined set of categories through learning by example, in a supervised fashion.

Although this categorization can often happen manually, we are focused on discussing the automatic supervised learning approach. Because this problem is mostly found when we have new unforeseen documents needing categorization, we are specially interested in high accuracy induction on new data.

A large amount of work has already been done applying statistical tenchinques and machine learning approaches such as the Naive Bayes, Linear Regression, Expectation-Maximization, K Nearest Neighbour (kNN) and Support Vector Machines (SVM) [26, 56, 13, 45].

The most popular evaluation criteria for this task is the average between precision and recall, where precision is the proportion of items in the category that are really in the category and recall is the proportion of items in the category that are actually placed in the category. Recall and Precision are defined as:

$$Recall = \frac{\#ofcorrectpredictions}{\#ofpositiveexamples} \tag{2.6}$$

$$Precision = \frac{\#ofcorrectpredictions}{\#ofpositivepredictions} \tag{2.7}$$

The Reuters-21578 dataset was made available by Reuters, Ltd. and is commonly used as the main benchmark for text classification evaluation. This is a collection of 21,578 newswire articles which were assigned classes from a set of 118 topic categories.

Using this dataset, state of the art work averages 92% breakeven point (when recall equals precision) for the 10 most frequent categories and 87 for all 118 categories [13].

Chapter 3

Approaches

3.1 Sources

Given the current position hold by my advisor, Francisco C. Pereira, as a researcher for the Singapore MIT Alliance for Research and Tecnhonolgy (SMART) Center for Future Mobility (FM) we have decided to focus this dissertation on the data available to this country, as it offers an opportunity to get access to otherwise restricted data.

For this dissertation we have collected data from these different sources:

- **Events** happening in Singapore during 2011. These events can have many different categories (e.g. Music, Art, Sales, Sports, etc) happening in a particular venue, on a specific date(s);
- **Tweets** from users living or currently in Singapore. Because not all users make the location available, we have only collected users tweeting in the country, or with this information on the profile;
- **Traffic Data** This dataset contains various types of information, mainly retrieved from road counters;
- **Weather Data** from different weather stations across Singapore, including temperature, wind speed and rain rate.
- **Incidents Data** detailing every road incident occurrence in year 2011;
- **Disruptions** on a speed prediction algorithm found in related work [4].

Next, we describe with more depth each of these datasets and how they were collected.

3.1.1 Events

Extracting events' information required the use of web scraping techniques already discussed in section 2.2.1 because most Singapore websites for this information do not provide a public API. Events were collected from the following websites:

- **Eventful.com** (eventful.com), **Eventbrite** (eventbrite.com) and **Zvents** (zvents.com) are web communities for events happening all over the world usually populated by advertisers.
- **Asia-City** (is.asia-city.com), **inSing** (insing.com), **Timeout** (timeoutsingapore.com) and **YourSingapore** (yoursingapore.com) are cultural magazines popular in Singapore making its databases of cultural events available in the country.

To collect this information several applications were developed and deployed to a server where they are periodically executed looking for new information on each of these websites. These applications were developed in *Python*, and because they require different scrapping methods in each website, it required different applications for each source of data.

A total of 23920 events were collected from these sources for the year 2011. Figure 3.1 presents the actual count for each source. We also reduced the duplicates by applying the rule of eliminating events *happening on the same day and same venue* thus reducing the count to 9202 unique events.

Information for events is of different quantity and quality across these sources, but we can reduce all to the common fields: name of the event, name and location of its venue, description, category and date(s).

3.1.2 Tweets

We already presented Twitter as a good source of real-time georeferenced content in Section 2.1 and guided especially by the ease of access to users data we will focus mostly on Twitter. With twitter we can have access to a stream of

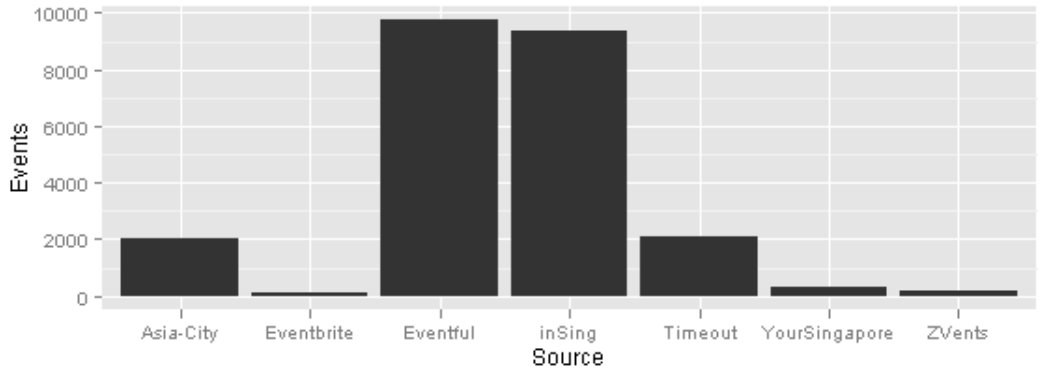

Figure 3.1: Events retrieved for each Source

updates with high throughput with both real-time characteristics and georeferenced content.

Because these systems are built to allow users to subscribe to the updates of their friends, and sometimes even strangers, the amount of information they see is often too large. That is why these networks force users to condense their message in the form of short message. Tweets are allowed a maximum of 140 characters, and these limits are not usually reached because users are already used to this style of short messaging.

About 155 million tweets were collected across the year 2011. Appendix A contains detailed demographics on this data.

This data works as sensors into the real-world, from which we extracted information about the popularity of events, performers and venues. Because we intend to apply sentiment analysis to this content, we can consider polarization affecting the popularity as perceived from the public (because not all press is positive press).

There are some inherent bias in the sampling of this data which is not necessarily reflective of the true population, mainly because we are limited to polling data from individuals who participate in this social network. Given these flaws, we still believe they should be able to reflect the actual popularity outside the online world.

3.1.3 Traffic Data

This part of the document is under a confidential agreement and can not be reproduced here.

3.1.4 Disruptions Data

One of the problems found in the performance of traffic prediction algorithms is the ability to recognize new unforeseen events thus making their prediction fail. This characteristic is commonly attributed to its robustness.

Asif et al. [4] working with the same traffic data and a smaller set of Singapore links have made available 53015 points of 5-minute intervals during March and April of 2011 where their prediction algorithm has been seen to fail with a large amount of error (above 3 times the value of its standard deviation). Figure 3.2 shows this disruptions projected over the Singapore map.

By making this data part of our analysis we hope to discover a connection between this algorithm performance and possibly new information gathered from our online source, and by making so, incorporating that online information as features to help this classifier.

3.1.5 Incidents Data

This dataset contains 117653 incidents records that have happened during 2011 in Singapore Expressways with link identification, textual description of the occurrence, and type (can be an accident, roadwork, vehicle breakdown, etc). There is an average of 165 incidents per day (stddev=41.6). Figure 3.3 shows this incidents projected over the Singapore map, and Figure 3.4 shows how they are distributed along time of day.

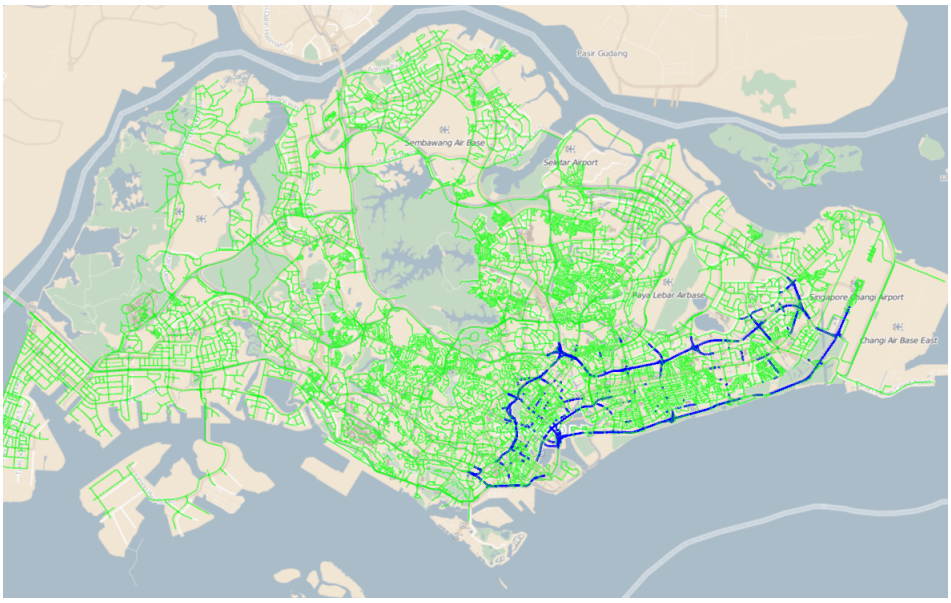

Figure 3.2: Disruptions projected over Singapore

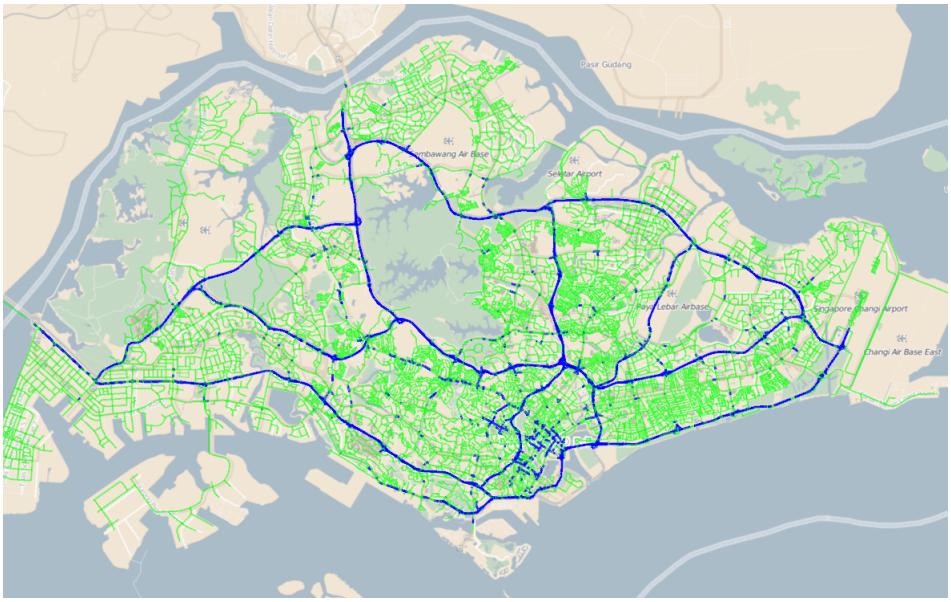

Figure 3.3: Incidents projected over Singapore

3.1.6 Weather Data

Through Nanyang Technological University (NTU) in Singapore we were able to obtain weather data for the months of traffic data we have. This weather

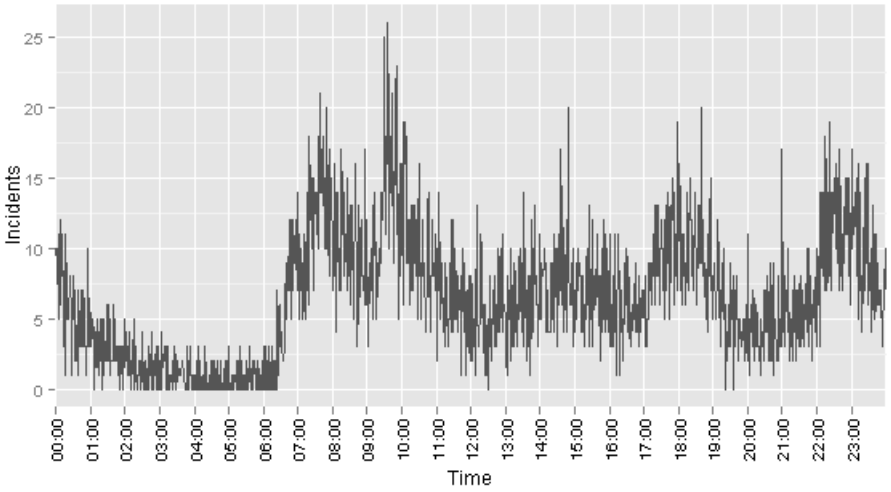

Figure 3.4: Incidents starting time

data contains values for rain rate (mm/hr), temperature, wind speed and humidity for 6 different stations located across Singapore sampled on 1 minute basis. Table 3.1 lists this stations, and the coverage of this data, while Figure 3.5 shows this stations on the Singapore map with a 5km arbitrary radius.

Name	Samples present	Moments of rain
Intellisys	55.77%	127
RI	45.76%	85
RVHS	23.00%	5
JSS	10.03%	4
NSS	4.93%	6
NYGH	5.30%	12

Table 3.1: Weather stations data coverage

We can see there is only two stations, Intellisys and RI, providing a reasonable amount of sensor samples. Figure 3.6 shows the number of samples with positive rain rate over these months, indeed confirming that only these two stations have consistent information over the date interval.

By considering a raining moment more than 10 consecutive samples with positive rain rate, we find an average of 25 raining moments per month during March and April, lasting an average of 242 minutes each.

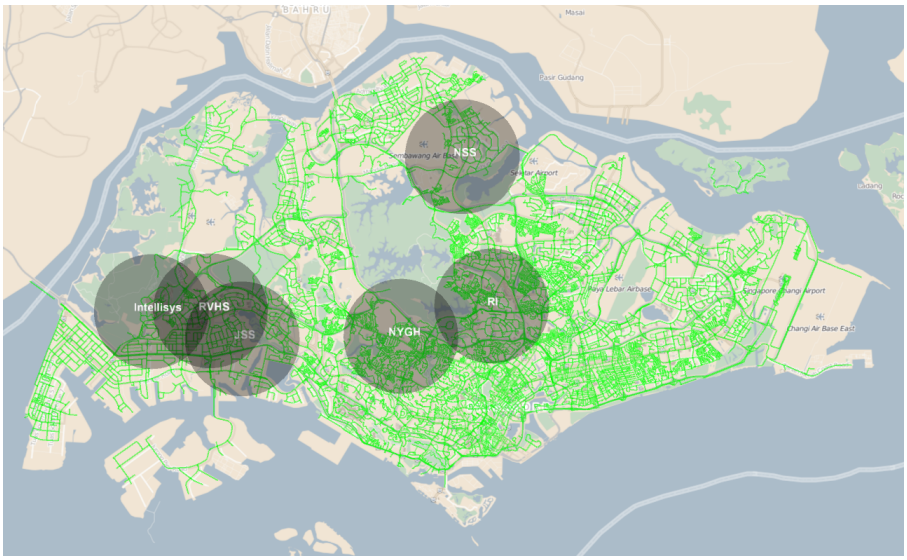

Figure 3.5: Singapore Weather stations

It is our objective to analise this weather data as to find possible association with Incidents (does weather impacts accidents?) and Disruption data (is weather a good feature for traffic prediction?).

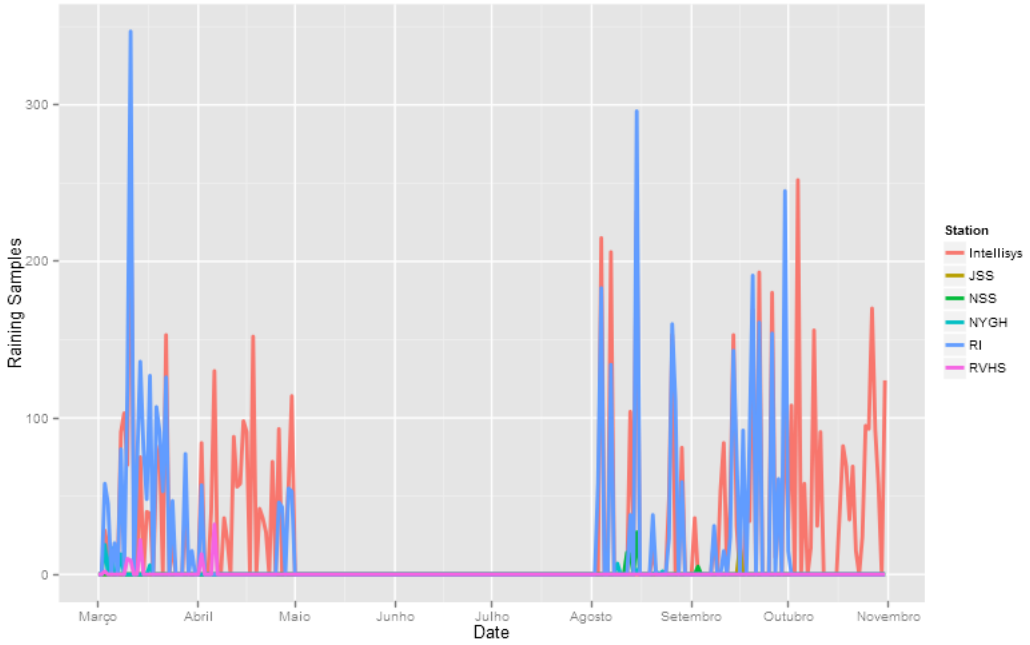

Figure 3.6: Samples with positive rain rate for each weather station

Figure 3.7 shows how small the area of overlap is if we consider a 5km radius around the weather station.

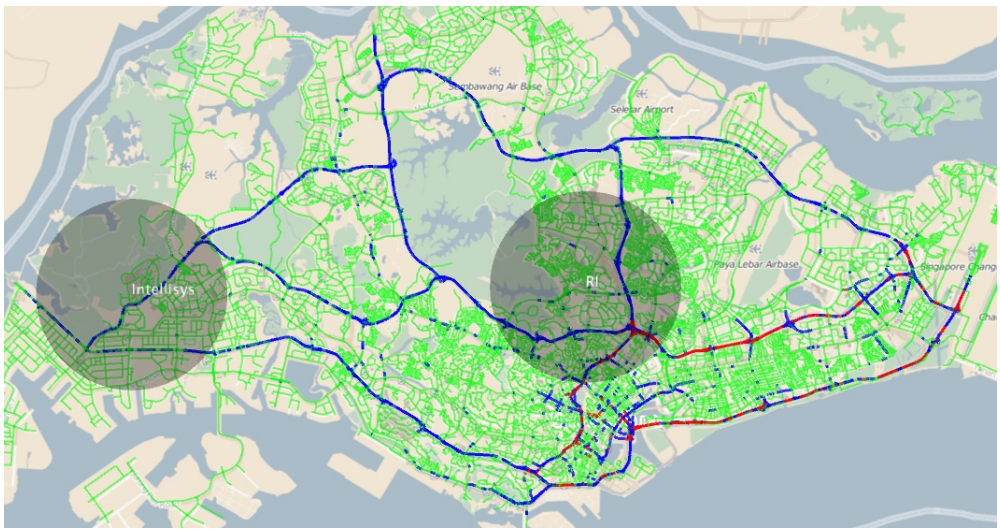

Figure 3.7: Incidents (blue), Disruptions (red) and Weather stations

Figure 3.8 shows the histogram of incidents varying in hours since it last rained

and distance to the closest station.

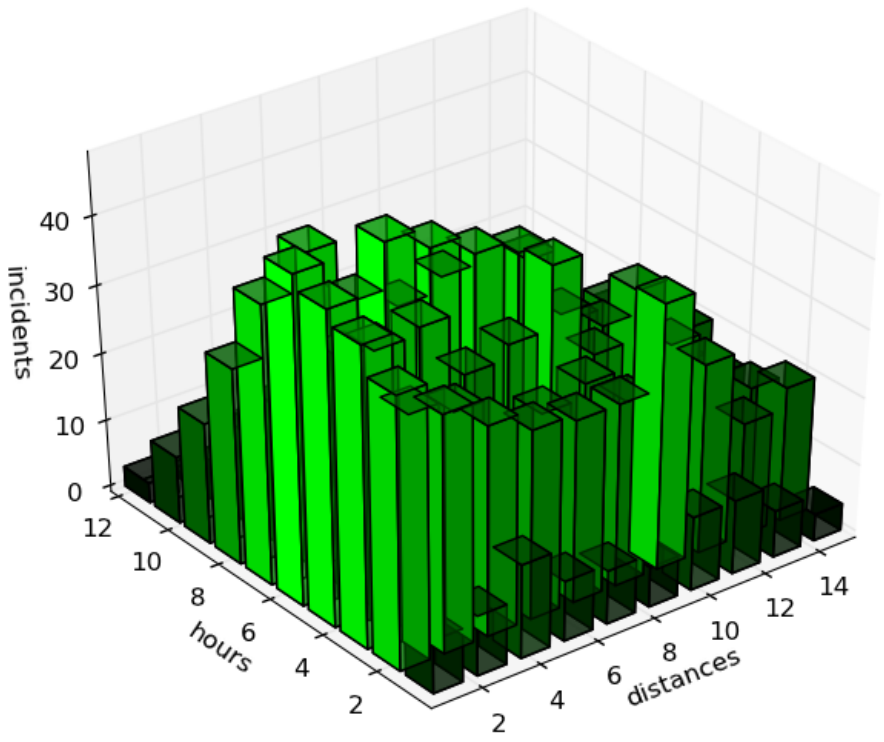

Figure 3.8: Time since last rained before an incident, and distance to the closest station

We discarded the data for stations with less than 30% samples, leaving Intellisys and RI. For the purposes of this dissertation it was our hope that correlating this weather data with the incidents data would bring some insight on their dependence, and a quick analysis showed that within the 9273 incidents in March and April 42% (3903) were within 10km of the closest station and had a raining moment in the previous 6 hours before it occurred. But due to time constraints we did not followed further this analysis.

3.2 Information Extraction

As we have discussed before, we are working with noisy sources of data, especially with Twitter data. Information Extraction plays an important role in the task of retrieving relevant information from these. Because we consider tweets looking to connect them with the collection of events we have gathered, the first step is be the enrichment of these events.

Using Kusko as NLP tool, we extract a ranked list of concepts (semantic index) from the general event information, resulting in a collection of entities like the performers, the events venue and other important features. After this, these entities are enriched using Wikipedia, providing us with structured connections between articles/concepts. Figure 3.9 presents this process.

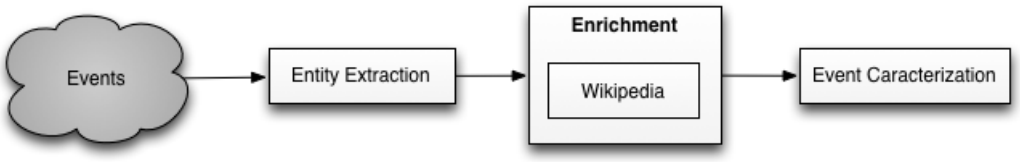

Figure 3.9: Information Extraction stages

With this structured event information, we can look to extract more information from Twitter. Because this is a very different source of data, requiring some extra precautions because of the noise and special entities like hash-tags and URLs, we combine tools for text normalization in micro-blogging context [27] and text tokenizer's able to handle this kind of text [30].

What constitutes a relevant tweet in terms of popularity might be summarized in the following list:

1. Tweet shares key entities with the event information (e.g. event name, description and performers). If the tweet contains URLs, we can follow this links and expand the tweet with this content, making the context easier to extract, even if we add redundancy. TF-IDF can be devised to work in measuring entities relevance, in contrast to the full corpus, by maximizing not only the number of shared entities but also the uniqueness of these entities. From this tweets, commonly shared hash-tags can also be extracted.

2. Re-tweets of a tweet considered relevant by step 1. Because of their similarity in content, they are very easy to detect and works as an indication

of the user's shared interest.

3. Tweets sharing hash-tags considered relevant in step 1. These hash-tags play an important role because they are easier to extract from tweets as a result of their format, and should be less prone to NLP errors.

These steps should help to gather a collection of tweets relevant enough to connect with a scheduled event.

Using sentiment analysis we can measure the polarity in these tweets towards an event to find if they help to gather more or less affluence. We believe negative statements, if linked to an existent scheduled event, can work to bring less affluence.

The extraction of relevant hash-tags could be enhanced by following Twitter links contained in the Wikipedia page of its performers. It is often clear from a Wikipedia page what the performers twitter account name is, making it easier to extract relevant hash-tags by weighting the fact that performers already add their own hash-tags as extra annotation to their tweets.

3.3 Event Demand

With timestamped georeference content linked to an existent scheduled event, we can start to recognize the pattern over time to recognize the trend in popularity. These trends, shown as incremental references to the event, define its online popularity.

By looking at the patterns shown in traffic flow data during these events, we can see how this online popularity relates to real-work event participation. From this relation, we can start characterizing both place and time to recognize what could be used as a good predictor of this affluence.

Figure 3.10 shows a illustration of our system architecture.

The result of this system should be not only the accurate estimation of participation, but also a characterization of participants. We can see how the venue affects its popularity, and how having big performers can boost attendance, and how the radius of the spatial dimension in online popularity have affected this affluence. Depending on the quality of traffic data we get from the LDA, some transport characterization could also be made, such as the type of transport used.

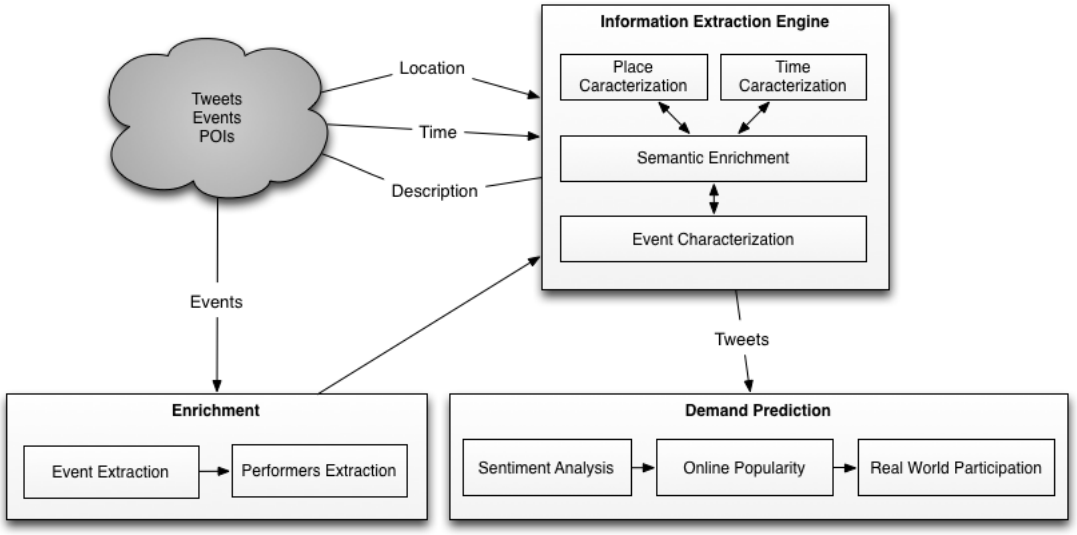

Figure 3.10: System architecture

3.4 Emergent Events

Although there are large quantities of information about events happening in the future through sources like *Upcoming* and *Zvents*, we still cannot account for a multitude of events, namely private events (e.g. political summit) and spontaneous events (e.g. car crash, store sale) happening in the city.

Research has already recognized that people are topical when using social networks [47]. Given a spatial segmentation within a city creating distinct zones of greater granularity, we could relax the definition of an event to a distribution of topics characterized by their coexistence in that specific zone.

A practical example of this would be the occurrence of topics like "sale" and "h&m" found coexistent in a certain zone in the city. Given the distribution of topics, an event might be inferred. Because every event has its own spread radius, the issue of choosing the best zone granularity would certainly be important. We could devise an algorithm to explore different zone granularities, detecting events of different magnitudes.

LDA described in section 2.3 presents itself as a helpful tool to explore this approach. By examining the topic distribution of conversation in a specific zone in a limited time-frame, we can gain some insight into what is happening on that specific place and perhaps explain traffic effects as they occur.

Chapter 4

Experiments

This dissertations sets its goal to explore the relation between online information and its real-life mobility implication. That was done through several experiments dictated by the conditions of the data and the intermediary results. Most experiments are done with a smaller contained set of data aiming to maximize its results and prove worthy of more exploration, but that was not the case for many of them. In this chapter we present those experiments with the purpose of showing the knowledge it has brought.

4.1 Twitter Buzz

For this experiment we hand picked 10 events from Singapore during the second half of 2011. They were picked so they would be the 10 biggest events from this time-frame. This was possible by looking at some indicators like the event venue (number of seats), the price-range of the tickets (all of the 10 have tickets costing over 100 Singapore Dollars) and the fact they show on all event sources. Table 4.1 presents these events.

We can see the names are quite distinctive, which is great for matching with tweets over this period. We picked all tweets from three weeks before the event to one week after and processed it with the most basic NLP tools, namely Python NLTK to tokenize and Levenshtein distance to look for words within 3 edit-operations. Figure 4.1 shows the frequency of mentions during this period for these events. We can see an increase in mentions on the days before the event.

Date	Name
29-06-2011	Kylie Minogue
01-08-2011	Cranberries
03-08-2011	SUEDE
21-08-2011	PARAMORE
05-10-2011	Alice Cooper
21-10-2011	Yanni
29-10-2011	Faye Wong
03-10-2011	Westlife
22-10-2011	Taiwan Golden Melodies
11-11-2011	ABBA MANIA

Table 4.1: Events chosen for experimenting with data

Looking at this preliminary data we can see a spike in buzz concerning the event sometime prior to the date. Some more analysis should be made to assess if these spikes are significant for our purposes and further validation to make sure they are more than a random result provided by noise.

On average 28 tweets mentioned the respective event (with 80 thousand overall tweets average per day). Because these even were chosen to maximize the number of mentions, this number points to a considerable risk of not being significant.

An 18% increase from two days before to the day prior to the event was found in these events, meaning we saw a short spike 24h before the real event. We believe this increase rate can be enough indication of the upcoming spike, which will require a more throughly analysis and systematic approach to conclude if this holds for a bigger sample and if it is statistically significant. Because of the Twitter massive dataset size and large computation time, this analysis did not take place.

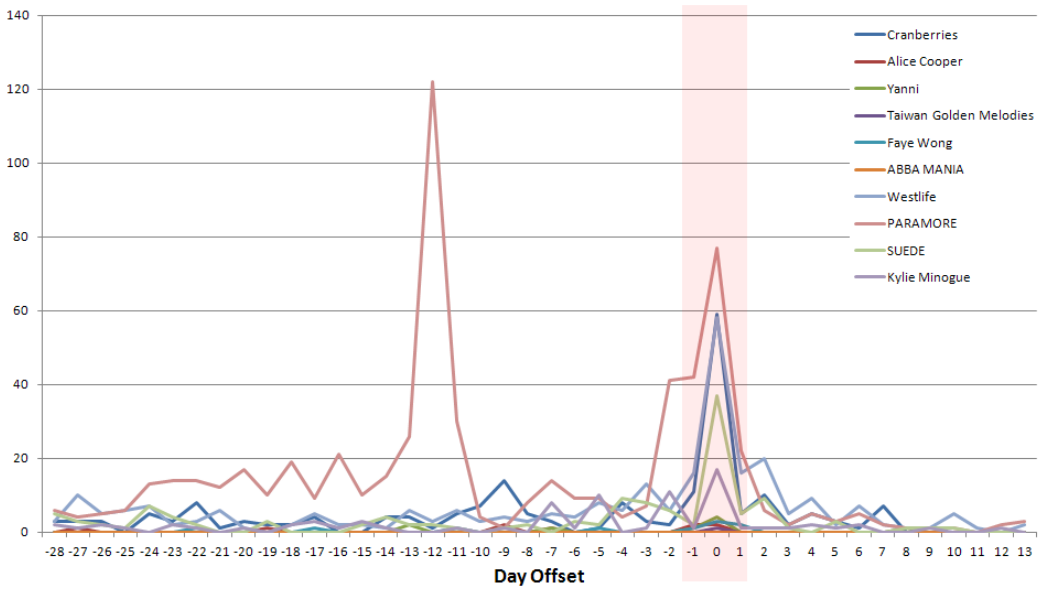

Figure 4.1: Number of mentions per day by music concert

4.2 Twitter Connection

Using the same list of events we did a brief analysis on the connection between the performers and Twitter. By getting the Wikipedia page for each performer we were able to find a path through links between the performer and its Twitter account for most events in this list. The following table shows this results:

This presents a 70% rate of getting a connection to Twitter within a distance of 1 link. If we added to this, a search using Google, we could increase this number to 90% by only following the search first hit, leaving only *Taiwan Golden Melodies* without a Twitter's account. Although this is a very small sample, goes to show that if we used Wikipedia to extract this trust-worthy connection into the performers Twitter account, we could extract important and relevant information from this account.

A practical example of this would be the extraction of the Twitter account so that we could lookup the performer's hash-tag usage. For instance, the rock band *The Cranberries* often share tweets that include their own hash-tag "#thecranberries". This authoritative information could be helpful in connecting tweets that include this hash-tag, knowing that they most likely refers to the band.

Name	Distance
Kylie Minogue	1 (Official Page)
Cranberries	0
SUEDE	1 (Official Page)
PARAMORE	1 (Official Page)
Alice Cooper	1 (Official Page)
Yanni	1 (Official Page)
Faye Wong	Not found (Official Page)
Westlife	0
Taiwan Golden Melodies	Not Found
ABBA MANIA	Not Found (Official Page)

Table 4.2: Link distance between the performers Wikipedia page and its Twitter account

4.3 Traffic Data and Events

4.4 Incidents and Disruptions

Considering road links where both incidents and disruption have happened we can for each get a vector with a positive value at a given 5 minute interval if there is any incident and disruption event at that time respectably. So for each link we consider these 2 vectors and by computing the correlation between these two vectors we can find if there is some dependence between them.

To compute this correlation we use the Pearson product-moment correlation coefficient [31] formula as:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{4.1}$$

where X and Y are our two vectors. This gives us a value between -1 and 1 and tells us the direction and magnitude of their association, where 0 means there is no association between the two.

Table 4.3 shows the coefficient values for 501 links where both incidents and disruptions data exists.

]-0.1,0.0[	]0.0,0.1[	]0.1,0.2[	]0.2,0.3[	]0.3,0.4[	]0.4,0.5[	]0.5,0.6[
453	16	10	11	6	2	3

Table 4.3: Correlation histogram between Incidents and Disruptions

Although there is considerable correlation in some links, it amounts to a small percentage of the links we have, and should not significantly help our efforts to explain disruptions using incidents data.

4.5 Event Categorization

One of the problems with data we gathered online about social events happening is the severe lack of categorization. From the 7 different sources we collected events, to a total of 23920, only 2 of these sources have categories for the events, giving us only 11520 categorized events (48.1%). But categorized events include 9401 from a single source (inSing) that does not contain other important information like the geo-position of event venue.

Using text categorization we intend to use this 9401 events and their categories to learn and categorize the remaining non-categorized events.

We intend to use the LDA model covered before in the topic modeling section 2.3 to perform as features of this classification. We start by applying state-of-the-art classifiers to a standard benchmark dataset to establish a reliable comparison with our own classification technique.

4.5.1 Reuters Dataset

The Reuters-21578 collection, commonly used as the main benchmark for text classification evaluation, contains 21,578 newswire articles which were assigned classes from a set of 118 topic categories. Each article can be present in none, one or several categories, with a existent average of 1.24 classes per document [10].

We decided to apply different state of the art learning methods to this dataset so we could compare to the use of topic modeling in the task of text categorization. We used three commonly used classifiers: Naive Bayes, Linear Support Vector Machine (SVM) and K Nearest Neighbors (kNN).

Following similar work we have decided to perform our experiments on the top 10 occurring classes, thus reducing our set to 8598 instances. Table 4.4 presents the occurrences for each class.

Class	Occurrences	Class	Occurrences	Class	Occurrences
acq	2193	corn	10	crude	500
earn	3763	grain	513	interest	273
money-fx	622	ship	219	trade	483
wheat	22				

Table 4.4: Reuters top 10 occurring classes

Table 4.5 shows both the precision, recall and micro-averaged performance results for 80/20 testing (1720 instances shuffled, 80% training data, 20% data) and 5-Fold Cross Validation accuracy. This results follow that covered by the state of the art in Section 2.4.1.

Classifier	Precision	Recall	F1	5-Fold Accuracy
NB	0.91	0.90	0.90	0.899 ($\sigma = 0.005$)
NB + TFIDF	0.84	0.83	0.80	0.798 ($\sigma = 0.002$)
10-NN	0.82	0.82	0.81	0.821 ($\sigma = 0.004$)
10-NN + TFIDF	0.90	0.90	0.90	0.902 ($\sigma = 0.003$)
SVM	0.93	0.93	0.93	0.934 ($\sigma = 0.003$)
SVM + TFID	0.95	0.95	0.95	0.945 ($\sigma = 0.004$)

Table 4.5: Classification results with 20% test data and 5-Fold CV

Using the LDA model with Gibbs sampling we were able to reduce the dimensionality of this data to 10 topics distribution and perform the same classification this as features. For better results, documents were stripped of any stop words and named entities were identified and considered as distinct single term. Table 4.6 shows the results with the same testing procedure.

From this we observe that the performance with the smaller dimensionality we get from the LDA topic distribution has not reached that of common state of the art classifier.

Classifier	Precision	Recall	F1	5-Fold Accuracy
NB	0.64	0.70	0.66	0.718 ($\sigma = 0.004$)
10-NN	0.77	0.76	0.76	0.772 ($\sigma = 0.003$)
SVM	0.73	0.75	0.72	0.751 ($\sigma = 0.003$)

Table 4.6: Classification results with 20% test data and 5-Fold CV with LDA features

4.5.2 Events Dataset

The events dataset contains 6 distinct categories with a total of 9401 instances. Table 4.7 shows this categories occurrences.

Class	Occurrences
Art & Performing Arts	2643
Fairs & Festivals	1464
Hot Events (sales)	735
Kids & Family	940
Leisure & Sports	1260
Music & Nightlife	2340

Table 4.7: Events dataset classes

We wanted to see how the same classifiers performed against this dataset. Table 4.8 shows the results for the same testing procedure.

Classifier	Precision	Recall	F1	5-Fold Accuracy
NB	0.71	0.72	0.71	0.707 ($\sigma = 0.004$)
NB + TFIDF	0.72	0.61	0.55	0.596 ($\sigma = 0.002$)
6-NN	0.48	0.49	0.48	0.484 ($\sigma = 0.004$)
6-NN + TFIDF	0.68	0.68	0.67	0.676 ($\sigma = 0.002$)
SVM	0.68	0.67	0.68	0.681 ($\sigma = 0.003$)
SVM + TFID	0.73	0.73	0.73	0.724 ($\sigma = 0.004$)
NB + 6-LDA	0.34	0.49	0.37	0.484 ($\sigma = 0.000$)
6-NN + 6-LDA	0.42	0.44	0.58	0.456 ($\sigma = 0.007$)
SVM + 6-LDA	0.36	0.49	0.38	0.487 ($\sigma = 0.001$)

Table 4.8: Classification results with 20% test data and 5-Fold CV

Although the overall results decreased substantially for this dataset, we can see the LDA acting as features still does not perform that good.

Figure 4.2 shows how the classes are distributed across the 6 clusters in the 6-NN. Observing this figure we can see the overlap on the topic distribution between distinct classes.

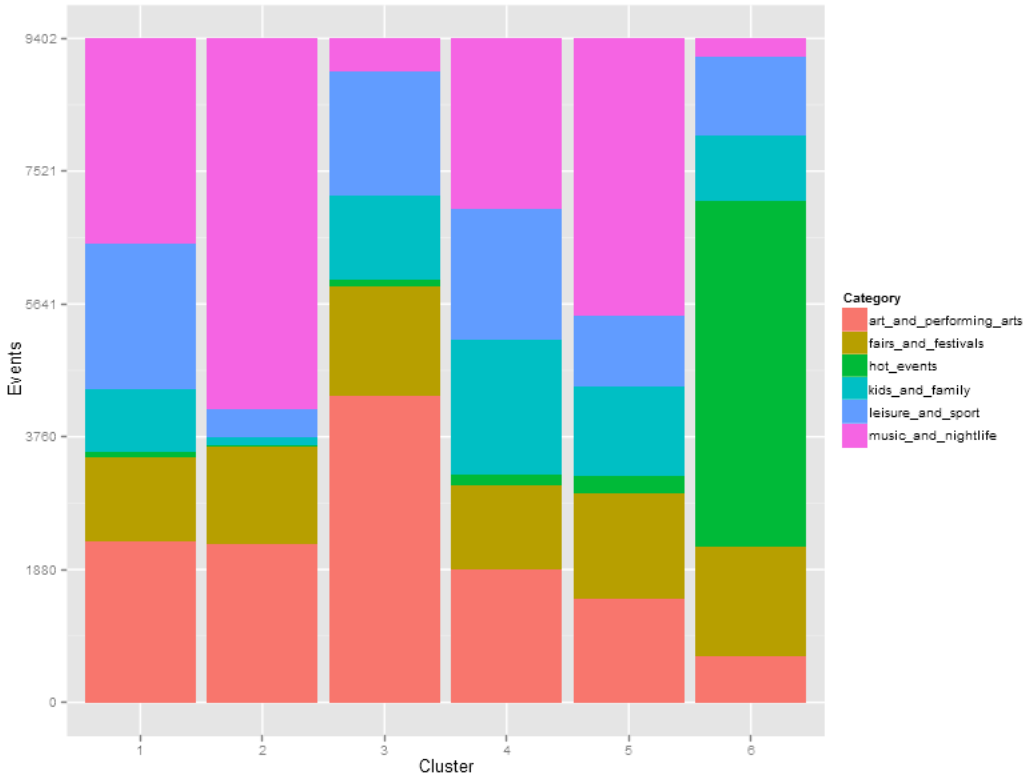

Figure 4.2: Events classes distribution across 6 clusters

Observing this we can not find a clear link between the clusters and the 6 already existent categories.

The following list presents the 10 most frequent words in each of the 6 topics inferred by the LDA model. On average these 10 terms summed frequency contribute 3% to its topics.

1. music, concert, world, band, musical, songs, night, rock, pop, performance;
2. bay, theaters, esplanade, dishes, cooking, food, home, chef, bring, culinary;
3. join, please, register, call, gov, local, fun, registration, email, school;

4. night, party, friends, time, special, good, fun, bring, dance, celebrate;
5. art, life, world, learn, exhibition, works, children, artists, film, story;
6. last, enjoy, stocks, terms, miss, sale, brands, brands, accessories, chance.

We can observe some overlapping terms being used in multiple topics, and some others that could easily be categorized as useless for discrimination.

Chapter 5

Conclusion

The main goal of this dissertation was to explore how online data from social networks and other sources of information could be helpful in estimating popularity and how they can be related to physical sensory data, believing we can extract knowledge about its relation.

By stating online their intention of going to a music concert, or any event for that matter, people are collectively sharing not only their personal taste, but also a portrait of how the masses react. But because social networks are both noisy and lacking context with their small messages, NLP tools have an harder job achieving good results, reflecting in poor semantic representations of the content being shared.

In our experiment in Section 4.1 we observed this by evaluating the number of mentions to some handpicked famous bands in our twitter dataset and found that although there is a noticeable effect prior to the event time, this effect can not be easily reproduced to a bigger set of events, specially smaller ones. We were not able to find significant quantity and relevance in Twitter messages to attest to the hypothesis that people do use this social networks to share this kind of information.

Following that experiment, in Section 4.2, we also verified in a small set of events that our approach to tackle the small context of twitter messages using Wikipedia was possible, but this was not done due to time restrictions.

In Section 4.3 we performed an analysis of limited scope in the traffic data hoping to find a big effect near a handpicked venue near the time of big music concerts. We were not able to reproduce a big effect, mainly due to several attenuating factors:

1. Although this venue isolation presents less influences nearby, it holds a metro station directly underneath, in a city where car ownership rate is 12 cars per 100 people;
2. A lot of links near the venue are inexistent in the data provided;
3. Public transport might be more suitable for several reasons (e.g. drinking, confusion);

In Section 4.4 we used the incidents dataset together with the disruptions dataset to shed some light on the usefulness of the incidents dataset. By finding good quality correlation between this two datasets, we have shown that having access to information on road incidents can indeed improve state of the art traffic prediction algorithms.

And finally in Section 4.5 we delved into the task of enriching our events dataset through categorization, solving the issue of unlabeled events and bringing useful features to this data. Although there was not enough time to further use this data, having this labels can provide help in the quest of determining events popularity.

We believed this dissertation has presented a dent in the goal of exploring the relation between the big amount of data that is being produced online namely on social networks and events websites, and offline with sensory data.

5.1 Future Work

There is considerable amount of improvement in this work, specially on exploring better what its possible with the traffic data and the social network data. Some of the ideas queued for lack of time were:

- Integrate the events topic distribution as additional features in existing classifiers using this data [38];
- Work with Origin-Destination matrix data inferred from the traffic data;
- Improve the text categorization by using Correlated LDA;
- Improve Twitter parsing, semantic extraction by leveraging Wikipedia as additional context, and content identification through graph models using HashTags as edges connectivity;
- Create a more systematic approach to detect the traffic effect of events.

Bibliography

[1] Fabian Abel, Qi Gao, Geert-jan Houben, and Ke Tao. Semantic Enrichment of Twitter Posts for User Profile Construction on the Social Web. *New York*, pages 1–15, 2010.

[2] A Alves, F Pereira, and A Biderman. Place enrichment by mining the web. *Ambient Intelligence*, 2009.

[3] Eiji Aramaki. Twitter Catches The Flu : Detecting Influenza Epidemics using Twitter The University of Tokyo The University of Tokyo National Institute of. *Computational Linguistics*, pages 1568–1576, 2011.

[4] M T Asif, J Dauwels, C Y Goh, and A Oran. Unsupervised Learning Based Performance Analysis of ν -Support Vector Regression for Speed Prediction of A Large Road Network. *Intelligent Transportation Systems Conference 2012*, (Section VI), 2012.

[5] A.T. Aw, Min Zhang, Juan Xiao, and Jian Su. A phrase-based statistical model for SMS text normalization. In *Proceedings of the COLING/ACL on Main conference poster sessions*, number July, pages 33–40. Association for Computational Linguistics, 2006.

[6] Tim Berners-Lee, James Hendler, and Ora Lassila. The Semantic Web. *Scientific American*, 284(5):34–43, 2001.

[7] David M Blei and John D Lafferty. Topic models. Technical report, Princeton University, 2009.

[8] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[9] Eric Brill. Some advances in transformation-based part of speech tagging. *Arxiv preprint cmp-lg/9406010*, 1994.

[10] H. Schütze C. D. Manning, P. Raghavan. *Introduction to Information Retrieval*. Cambridge Univerásity Pressn, 2008.

[11] Ana Cristina and Oliveira Alves. Semantic Enrichment of Places. *Word Journal Of The International Linguistic Association*, 2010.

[12] Anqi Cui, Min Zhang, Yiqun Liu, and Shaoping Ma. Emotion Tokens: Bridging the Gap Among Multilingual Twitter Sentiment Analysis. *next.comp.nus.edu.sg*, (20090002120005), 2011.

[13] Susan Dumais, John Platt, David Heckerman, and Mehran Sahami. Inductive learning algorithms and representations for text categorization. *Proceedings of the seventh international conference on Information and knowledge management - CIKM '98*, pages 148–155, 1998.

[14] Facebook. Data, 2012. [Online; accessed 24-January-2012].

[15] Facebook. Statistics, 2012. [Online; accessed 24-January-2012].

[16] Tim Finin and Belle Tseng. Why We Twitter : Understanding Microblogging. *Network*, 2007.

[17] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Artificial Intelligence*, (1995), 1997.

[18] Eric Gilbert and Karrie Karahalios. Predicting Tie Strength With Social Media. *Group*, 2009.

[19] Alec Go and Richa Bhayani. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 2009.

[20] S.A. Golder, D.M. Wilkinson, and B.A. Huberman. Rhythms of social interaction: Messaging within a massive online network. *Communities and Technologies 2007*, pages 41–66, 2007.

[21] Google. Google+ 2011 q4 earnings, 2011. [Online; accessed 24-January-2012].

[22] Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101 Suppl:5228–35, April 2004.

[23] Miniwatts Marketing Group. internetworldstats.com, 2011. [Online; accessed 05-January-2012].

[24] Divij Gupta and Chanh Nguyen. Detecting Real-Time Messages of Public Interest in Tweets. Technical report, Stanford University, 2010.

[25] Gregor Heinrich. Parameter estimation for text analysis. *Bernoulli*, 2008.

[26] Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. *Machine Learning: ECML-98*, 1998.

[27] Max Kaufmann and J. Kalita. Syntactic normalization of Twitter messages. In *International Conference on Natural Language Processing*, pages 1–7, 2010.

[28] Kirill Kireyev and L Palen. Applications of topics models to analysis of disaster-related twitter data. *on Applications for Topic Models*, 2009.

[29] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is Twitter , a Social Network or a News Media ? Categories and Subject Descriptors. *Most*, pages 591–600, 2010.

[30] Gustavo Laboreiro and Jorge Teixeira. Tokenizing Micro-Blogging Messages using a Text Classification Approach. *October*, pages 81–87, 2010.

[31] Ching-yung Lin. Improving User Interest Inference from Social Neighbors. *Human Factors*, pages 1001–1006, 2011.

[32] Brian Locke, James Martin, and D Ph. Named Entity Recognition : Adapting to Microblogging Named Entity Recognition , Adapting to Microblogging. *Processing*, 2009.

[33] Catherine C Marshall and Frank M Shipman. Which semantic web? *Proceedings of the fourteenth ACM conference on Hypertext and hypermedia HYPERTEXT 03*, pages 57–66, 2003.

[34] Mor Naaman, Jeffrey Boase, and Chih-hui Lai. Is it Really About Me ? Message Content in Social Awareness Streams. *Information Systems*, 2010.

[35] João Oliveirinha and F Pereira. Acquiring semantic context for events from online resources. *3rd International Workshop on Location and the Web*, 2010.

[36] Bo Pang and Lillian Lee. Thumbs up?: sentiment classification using machine learning techniques. *-02 conference on Empirical methods*, 2002.

[37] K Papineni, S Roukos, and T Ward. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, number July, pages 311–318. Association for Computational Linguistics, 2002.

[38] Francisco C. Pereira. Internet as a sensor : a case study with special events. 2500(August), 2011.

[39] Francisco C. Pereira, Ana Alves, João Oliveirinha, and Assaf Biderman. Perspectives on Semantics of the Place from Online Resources. *2009 IEEE International Conference on Semantic Computing*, pages 215–220, September 2009.

[40] Alexei Pozdnoukhov and Christian Kaiser. Space-time dynamics of topics in streaming text. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, page 8. ACM, 2011.

[41] Alan Ritter, Sam Clark, and Oren Etzioni. Named Entity Recognition in Tweets : An Experimental Study. *Library*, pages 1524–1534, 2011.

[42] D Romero, W Galuba, and S Asur. Influence and passivity in social media. *Machine Learning and*, 2011.

[43] Takeshi Sakaki. Earthquake Shakes Twitter Users : Real-time Event Detection by Social Sensors. *Earthquake*, 2009.

[44] M. Schrenk. *Webbots, spiders, and screen scrapers: a guide to developing Internet agents with PHP/CURL*. No Starch Press Series. No Starch Press, 2007.

[45] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, March 2002.

[46] Amit Singhal. Modern information retrieval: A brief overview. *IEEE Data Engineering Bulletin*, pages 1–9, 2001.

[47] Daniel Sousa and L Sarmento. Characterization of the Twitter @ replies Network : Are User Ties Social or Topical ? *on Search and mining user*, 2010.

[48] Bharath Sriram, David Fuhry, Engin Demir, and Hakan Ferhatosmanoglu. Short Text Classification in Twitter to Improve Information Filtering. *Performance Evaluation*, pages 4–5, 2010.

[49] Mark Steyvers. Probabilistic topic models. *Handbook of latent semantic analysis*, 2007.

[50] P.D. Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 417–424. Association for Computational Linguistics, 2002.

[51] Dong Wang, Zhenyu Li, and Kave Salamatian. The pattern of information diffusion in microblog. *of The ACM CoNEXT Student*, pages 2–3, 2011.

[52] Xiaolong Wang, Furu Wei, Xiaohua Liu, Ming Zhou, and Ming Zhang. Topic Sentiment Analysis in Twitter : A Graph-based Hashtag Sentiment Classification Approach. *Network*, pages 1031–1040, 2011.

[53] Gerhard Weikum and Martin Theobald. From Information to Knowledge : Harvesting Entities and Relationships from Web Sources. *Machine Learning*, 2010.

[54] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Q He. Twitterrank: finding topic-sensitive influential twitterers. *Proceedings of the third ACM international conference on Web search and data mining*, pages 261–270, 2010.

[55] Jaewon Yang and Jure Leskovec. Modeling Information Diffusion in Implicit Networks. *2010 IEEE International Conference on Data Mining*, pages 599–608, December 2010.

[56] Shibin Zhou, Kan Li, and Yushu Liu. Text Categorization Based on Topic Model. *International Journal of Computational Intelligence Systems*, 2(4):398, 2009.

Appendix A

Data Demographics

We have a total of 4 different types of data, mentioned before in section 3.1, which are: Tweets, Events, POIs and Traffic Data, all from the country of Singapore. In this appendix we present the statistical characteristics of this data. We also briefly present some of the challenges found when working with massive datasets of data, like the Twitter dataset.

A.1 Twitter

There are currently three ways of getting data from Twitter: REST API, Search API and Streaming API. Because we used all of them for our work, we will briefly describe them:

A.1.1 REST API

This was the first API available from Twitter that made available public methods to access data from Twitter users, like their past tweets and connections to other users. This is done using the Representational state transfer (REST) architecture style which defines methods to request data using the commonly used HTTP protocol. It is also possible to access a user's private information by authenticating with the website credentials, something we do not need for our work.

The use of this API has some limitations, namely a rate limit of 150 hits per hour (350 when authenticated), a maximum of 100 users information can be

fetched in a single request, and a maximum of 3200 past tweets can be fetched from each user (requiring a minimum of 16 requests to do it).

Although these limits constrain the rate at which we can information, we were able to transverse the Twitter graph of users to gather a list of users with either their location or timezone set to Singapore. This was done by doing a breath-first search of the Twitter graph starting with one random user from Singapore (happened to be a famous writer named Charles Yeo with 263039 users following his tweets). Using this method 343911 unique users were gathered from Twitter.

The only way we could do this in a useful time-frame was creating several accounts on Twitter to be able to make several request simultaneously. This was possible because the limits on this API are based on the authentication used for requesting data.

From these 343911, only 179670 made publicly available their tweets. Because having access to their tweets is fundamental for this work, private users were removed from our dataset and are not considered for any further analysis. Figure A.1 presents this in a graphic fashion.

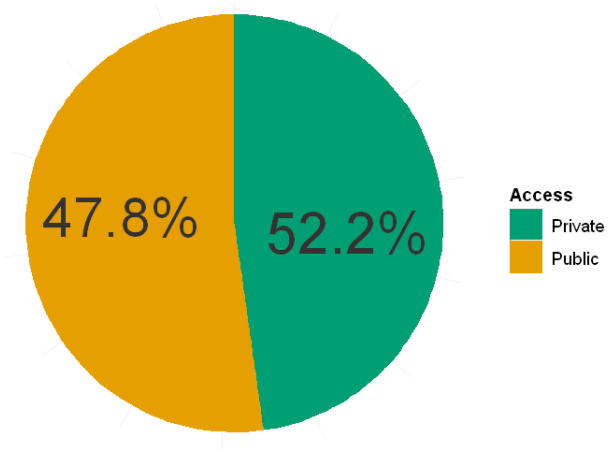

Figure A.1: Singaporean Twitter users access type

From these users, only about 45.5% (81805) stated in their profile information that they are located in Singapore. This low number can be explained by several factors, namely by the fact that this information is presented in a very privileged location on their profile webpage and people use this to present some other non-related or fictional information about themselves (e.g “neverland”, “home sweet home”).

Another data field relevant to the location might be the timezone used. Singapore has its own timezone (Singapore Standard Time (SST) or UTC+08:00) and 150794 users use this as their timezone. This timezone is shared with Hong Kong, Macau, parts of China, Malaysia, Brunei, Indonesia and Western Australia. But we get a specific timezone for Singaporean users because Twitter provides options where this different versions of the same timezone can be selected. Figure A.2 shows the users location with the SST timezone, and we can see that Philippines takes 8.6% of users because they share the same timezone but Twitter does not give a custom option for this country. Figure A.3 shows the users timezone, where about 5.1% of the users with Singapore in the location choose Alaska as their timezone because they share the same offset and this choice is on top of the list when registering a new account on Twitter, making it more likely that users will not look for Singapore official timezone, altough that option is present.

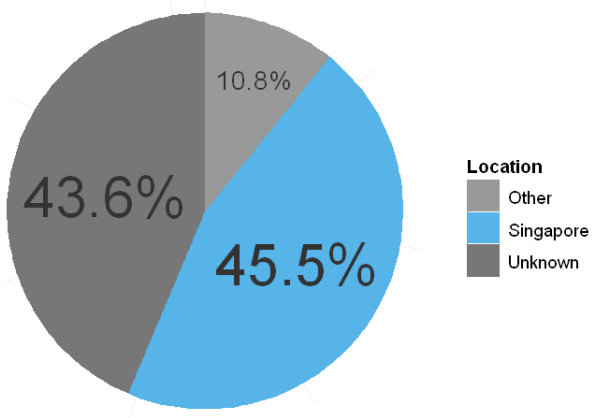

Figure A.2: Twitter users location

From these users, 99.1% are tweeting using the English language, only lowering slightly the final count of users, making them 178058 potential Singaporean users to retrieve tweets from.

On average, users have 192 friends but 218 followers, that means most people actually have more people following them than them following back, something we see often in celebrities and entities using Twitter. Figure A.4 shows how these two variables relate to each other. We can easily see the disproportion from followers to friends.

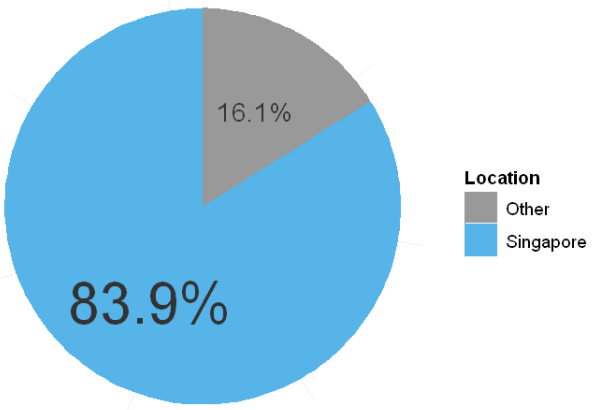

Figure A.3: Twitter users timezone

We also looked at the number of tweets an user has ever tweeted. An average of 2938 tweets were sent by these users since they started using Twitter. Figure A.5 shows this.

After about one month of retrieving tweets under the limits of the REST API, we have collected 160 million tweets for theses users. That is about 180 Gb of raw text data needing processing.

After some data exploration we have determined that two problems existed with this data: duplicates and erroneous values. Duplicates were most likely the result of the time this operation took, making it possible for the system to change between requests. Erroneous values were detected mainly in the timestamps provided within each tweet, some even containing dates prior to the existence of the website. This is something the Twitter developers are aware but did not fix, blaming synchronization problems within complex cache systems. Because we are only working with tweets during 2011, this erroneous tweets are just ignored.

For the task of removing tweets in duplicate we knew each tweet had an identification ID, but required some way to store in memory every ID we had seen so that we could remove the duplicate. In the worst case we have to keep in memory 160 million different IDs (exactly 159232281). This proved to be difficult using a data structure like the hash table because of their large memory consumption. The solution was using a data structure called Bloom filter.

A Bloom filter is probabilistic data structure used to test if a element is member of a set. False positives can happen, while false negatives cannot. Works by

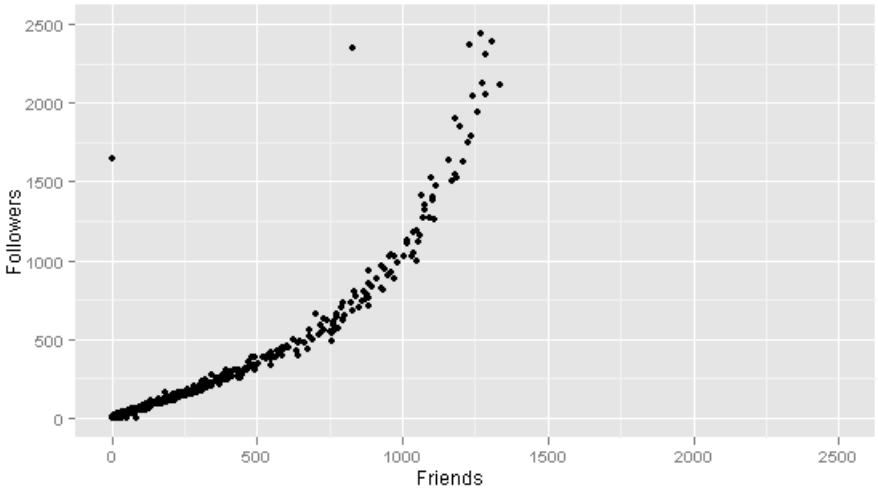

Figure A.4: Number of friends/Number of followers

using several different hash functions using a uniform random distribution to map the tweet ID to a different position in a bit array. To query if an ID already exists in this structure we simple hash the ID using all the hash functions and check if all positions given are set, if any of them is not, then the ID has not been set. By using a memory mapped filed to keep this structure in memory, we can have enough memory (about 8Gb) to hold the structure with an error of false positives bellow 0.1

Figure A.7 shows the histogram of all tweets gathered using this approach during the month of December and January 2012. A total of 145 million (145179048) tweets were created in 2011, an average of 397751 tweets each day. This is 91.2% of the total of tweets retrieved using the approach. Figure A.6 shows the histogram for tweets created in 2011.

Only 179671 tweets in this collection are georeferenced, that is only about 0.11% of this tweets. This is clearly the biggest problem with this approach, we can get a lot of content but is mostly without a precise location within the country. But there is still a lot we can do using data, leveraging the fact that we have so much time-referenced content.

Another big issue with this approach is the speed in which we can get this data. Gathering this data took more than one month of server time, only to gather 3200 of the latest tweets for each user. When this process ended, many

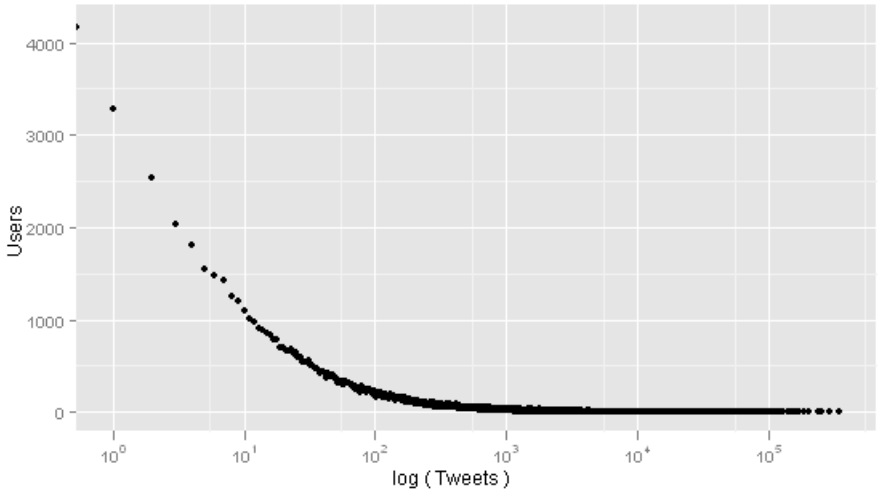

Figure A.5: Tweets sent since the user creation (logarithm scale)

tweets from very active users were already lost in this time-window.

A.1.2 Search API

Another approach created by Twitter to fetch tweets is the Search API. This is a dedicated API running searches against the real-time index of recent tweets limits the search of tweets to a window of 6 to 9 days. This API runs without authentication, but still limits the amount of requests based on the frequency they are made.

Unlike the REST API, this API limit is based to each IP, making it harder to get data faster.

The big advantage of this API is the chance to ask for georeferenced tweets located in a specific place. This is done by sending within the request coordinates for the bounding box containing the place we want to monitor.

Using this API we have fetched data by making requests every 5 minutes on the bounding box of Singapore. Because João Oliveira, a fellow student from the AmiLab, was already working with this API, we have collected data starting back on May 2011.

From May to December 2011, 8.5 million tweets (8525618) were collected using this method. Figure A.8 shows the histogram of this data. An average of 94747

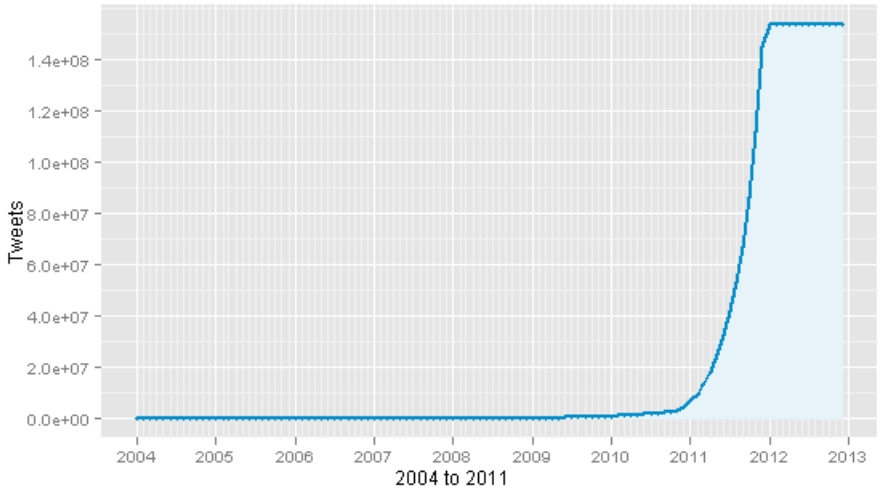

Figure A.6: Tweets created from 2004 to 2011 by all Singaporean users

tweets per day have been collected.

The biggest difference in this approach is clearly the georeference content, 23% of these tweets are georeferenced and are within 30km of Singapore (1995956 tweets exactly). Figure A.9 shows this best in visual way.

Just like the previously mentioned REST API, this approach also is prone to a lot of noise, most likely coming from the known caching problems in Twitter's system. Tweets from other parts of the world do come up in this data, something we would not consider because of the bounding box limiting our interest.

Something we found using this API has a larger diversity of spoken languages in tweets, namely 61.65% English, 12.44% Indian and 8.43% Indonesian. This is mostly because of the noise brought by the bounding box.

This approach is clearly a better source of georeferenced content, something we will require for bigger granularity in location. The only downside to this approach is clearly the big difference in tweets quantity, and the noise obtained by Twitter's cache systems.

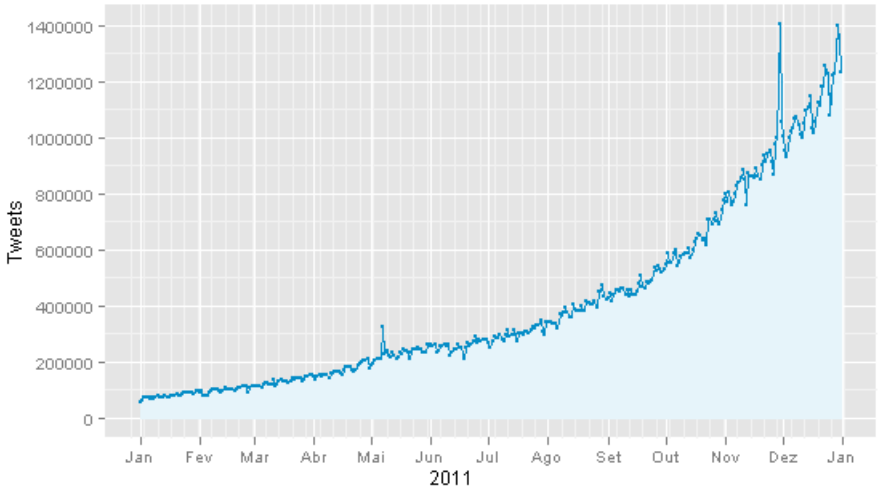

Figure A.7: Tweets created during 2011 by all Singaporean users

A.1.3 Streaming API

This is a new Twitter API introduced in 2010 allowing for high-throughput near-realtime access to Twitter data. Unlike previously mentioned approaches, this API does not require constant requests to their servers. This API relies on a indefinitely opened connection where Twitter's system will automatically send new data, as they arrive on their system.

This API was developed as an effort to eliminate the lack of uptime in Twitter's servers, creating a way for them to control the throughout of the data by eliminating the amount of connections clients often create.

Although this API offers the same kind of filtering we found in the Search API (by location), we have tested the use of this API to gather data in the same place and seen 10x less tweets after a full month of use. There were also several problems due to Twitter's lack of throughput, where we have restarted connections after a big period of time, to find a large speedup by doing so. We have attributed this fault to their bandwidth controlling algorithms. We have therefore decided to dismissed this approach.

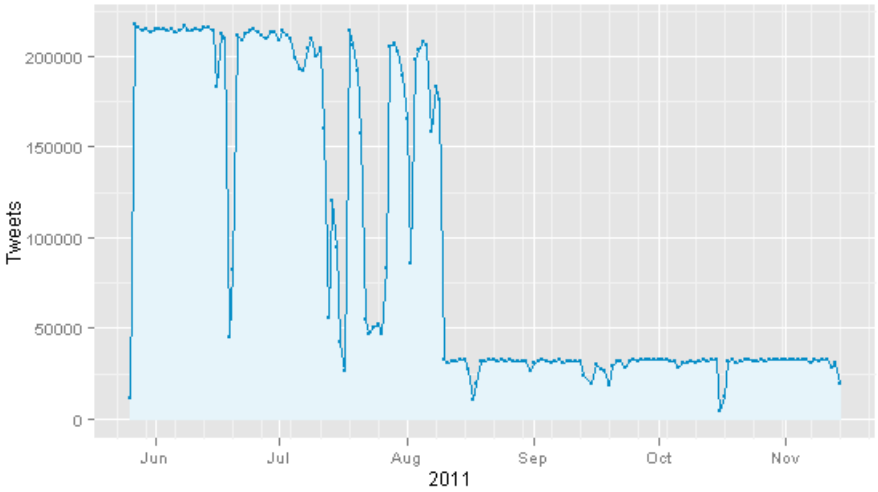

Figure A.8: Tweets over 2011

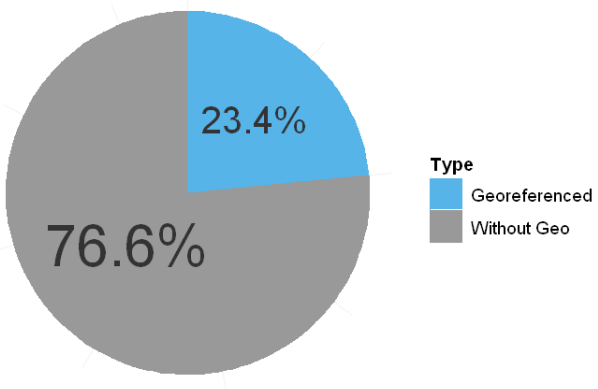

Figure A.9: Tweets georeferenced

A.2 Events

We have collected events from Singapore from several websites providing that information. This websites differ in the amount of information they present for each event. By common denominator we have the following information for each event: event name, description (from 10 to 200 words), date and venue name.

- *Eventful*: 9750 events throughout 2011.
- *inSing*: 9401 events throughout 2011.
- *Asia-City*: 2033 events throughout 2011
- *Timeout*: 2119 events throughout 2011
- *YourSingapore*: 325 events throughout 2011
- *Zvents*: 185 events throughout 2011
- *Eventbrite*: 107 events only after December 2011

Considering duplicate entries every two events happening on the same day and venue, we have found 9202 unique events in our dataset. Because some of these events are actually repeating occurrences, we have removed all repeated events and condensed their dates in a single entry, reducing these events to 7136 unique events.

A.3 Traffic Data