Federal University of Pernambuco
& University of Pisa

Doctoral thesis proposal

Mining human mobility data and social media for smart services

Author:
Vincius Cezar Monteiro de Lira

Supervisors:
Prof. Rossano Venturini University of Pisa
Dr. Chiara Renso ISTI - CNR
Prof. Valeria Cesário Times Federal University of Pernambuco

October 2016
Contents

1 State of the Art 5
  1.1 Mobility Data ................................................................. 5
    1.1.1 Trajectories of Moving Objects ..................................... 5
    1.1.2 Human Activity Analysis in Mobility Data ...................... 7
  1.2 Carpooling ................................................................. 7
  1.3 Social Media ............................................................... 9
    1.3.1 Mobility and transport analysis using Social Media .......... 9

2 Work Plan 11

3 Preliminary Results 13
  3.1 Exploring Human Behavior for Activity-based Carpooling .......... 13
    3.1.1 The COMEWITHME system ........................................... 14
    3.1.2 Experimental Results ............................................... 18
  3.2 Classifying Attendance to Social Events ............................. 18
    3.2.1 The Event Attendance Classifier ................................. 20
    3.2.2 Experimental Setup ................................................. 22

Bibliography 25
Chapter 1

State of the Art

We present in this section the start-of-the-art of the main topics covered by this thesis proposal. Initially, we present in the Section 1.1 a brief overview on mobility data analysis and mining to discover mobility behavior of groups of moving objects (e.g. people). After in the Section 1.2, we focus on specific applications of this movement mining for carpooling services, where the analysis of trajectory data is a prerequisite to find the car ride possibilities. Lastly, in the Section 3.2, we discuss some works where social media have been exploited in synergy with mobility data and specifically in transportation services.

1.1 Mobility Data

The constant acquisition of a moving object’s positions has become technically possible thanks to the positioning sensors that are now of common use to many people like the common GPS (Global Positioning System) enabled devices. The increase in the use of these devices enables the tracking of moving objects and tend to generate a large amount of mobility data. Moving objects are physical object in the real world equipped with a device that allows the tracking of its geographical position at a given time and include people, animals, vehicles, vessels, etc. The moving object movement, as detected by a device, is generally called trajectory which can be defined as the evolution of an object’s position in space over a given interval time [56, 48, 51].

Data analysis of trajectories has been shown to be a highly multidisciplinary field, ranging from Physics to Sociology, Transportation Research and Computer Science [31, 30, 67, 51]. In this sense, much efforts in the community have been done to develop new techniques to support better understanding of human mobility.

This explosion of GPS-enabled devices comes together with the growing of social media with positioning services the so called Location Based Services Networks (LBSN) like FourSquare check-ins, Twitter, Instagram, Flickr, Facebook. These data combine the positioning location with a user action (check-in to a venue, small text as tweets or photos tags and comments on Instagram, Flickr or Panoramio) associated to social features of the tracked person like the network of friends, links, preferences, pictures.

1.1.1 Trajectories of Moving Objects

Several works in the literature address the analysis of trajectory data. Even the definition of what a trajectory is can have several variants. In [56, 48, 51], trajectory can be defined as “the user defined record of the evolution of the position (perceived as a point) of an object that is moving in space during a given time interval in order to achieve a given goal”.

The potentialities offered to several application domains by the analysis of huge amount of positioning data has opened new opportunities for developing analysis methods of this new form of data. Mobility data analysis has become a hot research topic since several methods on data
mining and statistical techniques, tailored to trajectory data, have been proposed in the literature, like [31, 30, 67, 51]. The task of analyzing large trajectory datasets can be carried out towards different directions. Basic statistics may be applied to trajectory data mainly to discover the distributions of people presence and origin-destination matrices [10]; other studies focus on trajectory data mining aiming at finding correlations in large datasets of positioning data [31, 65]. Techniques to extract movement patterns include: (1) clustering discovery - finding groups of objects moving together – the authors in [46] propose a time-focused clustering of trajectories based on OPTICS algorithm [7]; (2) sequential pattern discovery - finding the most frequent sequences of places visited – the authors in [47] propose an algorithm to discovery T-Pattern from a trajectory dataset; (3) flock detection - extracting the convergence of people moving together for a certain amount of time [31, 62]. In [60], a software called M-Atlas encompassing a series of trajectory data mining algorithms is presented. Mining methods based on unsupervised machine learning for mobility data have also seen the interest of the research in the last few years [6, 65]. A typical task is the prediction of the next location based on the user location histories, or the classification of the type of user (e.g. tourist or resident). This is usually done as a classification task. The idea is to train a model in a supervised way where objects are described with a number of features and a class label. The classifier, trained on such set, can be applied to a new unlabeled dataset to label new objects with a certain accuracy. The prediction accuracy is a quantitative measure for the classification performance that refers to the proportion of the correctly predicted examples.

Trajectories are well fitted for applications that aim only at locating some moving object or computing statistics on the spatiotemporal characteristics of trajectories [48]. However, most application analysis require complementing raw data with additional information from the application context. For example, interpreting trajectories of persons within a city requires some knowledge about the features of the city (e.g. map, points of interest). Thanks to city information, spatiotemporal coordinates can be replaced with street and crossing names, or with names of places of interest, such as shops, restaurants, and museums [48].

Adding knowledge to raw trajectories is known as a semantic enrichment process [48, 15, 38]. A semantic trajectory is the representation of the trajectory and a set of interpretations on the moves and on the stops made by the moving object during its journey. The same trajectory can have different interpretations depending on the context of the application and the modeled domain. Figure 1.1 shows an example of a trajectory with implicit semantic information. In this figure, the trajectory A shows how the data is usually collected in its raw format. With this representation, the trajectory has no semantic aggregate and it can only be seen as a set of points that form a route without associated information about the path taken. Regarding the two other trajectories (B and C) of Figure 1.1, the data from the same trajectory are overlayed with spatial information from different domains to semantically enrich the data. The notion of semantic trajectory can be extended to include several orthogonal aspects related to movement, from the activity performed, the goal of the movement, the transportation means used, the weather, pollution and all the contextual information that we can be related to the movement and interesting for a given application. Analyzing semantic trajectories facilitates the understanding of the mobility data and patterns, therefore we have seen an increase of methods proposed in the literature to enrich trajectory data and to investigate analysis methods for this complex form of data.

When tracking human beings, the privacy of the individuals comes to our attention. Indeed the location of a person may disclose private information that a user may not wish to make public. Let’s think about the home and work addresses, the places visited may include health related location or religious habits and preferences. Privacy as been recognized immediately as a main concern when dealing with large amount of mobility data [31, 51]. For this reason a number of data mining methods have been studied to find a trade off between the need to get useful mobility patterns and the need to preserve the privacy of tracked individuals. These methods are referred to as “privacy-aware trajectory data mining” [31, 51]. Although this thesis proposal does not specifically deal with privacy preserving methods for mining mobility data, we need to be aware of the privacy problem when dealing with the analysis human traces and take it into account in developing our methods.
1.2. Human Activity Analysis in Mobility Data

Most existing human mobility studies generally focus on the raw location of the individual, and not on the human activities performed during the movement. One reason for this is the lack of explicit large scale activity information data [63]. Usually activity data is collected in forms of interview to a sample of population when collecting Census data, with the obvious drawbacks of not being geolocalized, not covering a large population, not covering a long period of time (usually they refer to one day) and not updated. The mobility data, on the other hand, while providing the exterior characteristics of movements like geographical heterogeneity and distance decay, population density, geographical and social distances, urban morphology and the spatial distribution of places, they neglect activities, the driving force that underlies human movements.

Human mobility datasets have recently received increasing attention by the research community as a source to understand human activities. Authors of [19] propose an algorithm for automatically annotating raw trajectories with the activities performed by the users. They analyzed the stop points trying to infer the Point of Interest (POI) that the user has visited. Based on the category of the POI and a probability measure based on the gravity law, they inferred the activity performed.

In [34] and [63], authors used the activity category as a dimension in their analysis. In [34], they discovered a scaling law showing the relationship between the popularity of a place and the probability to select this place as a destination. Their results showed a strong influence of urban contexts on peoples’ activity participation and destination choices. Paper [63] proposed an activity-based model composed of two parts. For the first part, they find the transition probability between activities varying over time, and then they construct a temporal transition probability matrix to represent the transition probability of travel demands during a time interval. In the second part of the work, they suggest that the travel demands can be divided into two classes, locationally mandatory activity and locationally stochastic activity, according to whether the demand is associated with a fixed location or not. A recent paper [68] proposes an Bayesian network to infer the user activity and mobility on a LBSN based based on the the current location and time. They show that their approach outperforms the state-of-the-art approaches for the two problems and also captures individual check-in behavior at the activity level against the spatial-temporal sparsity issue of check-in records in LBSNs.

1.2 Carpooling

Mobility data has been extensively exploited for studying and proposing carpooling systems. The basic idea of car pooling is the sharing of a car ride (or part of) with other people, typically also sharing the expenses. Carpooling tends to reduce the circulating vehicles and, as a consequence, the traffic and the CO2 emissions, two important problems in today’s cities.

Carpooling is in fact one of the many travel alternatives promoted by transport policies to reduce the amount of vehicles on the road. It was promoted during World War II to deal with oil and rubber shortages and during the oil crisis of the 1970s [32]. Nowadays, carpooling is promoted by mobility management policies to put more emphasis on the issue of sustainable transport. The main targets here are a reduction of transport-related pollution, noise nuisance reduction and decrease of congestion levels and minimize the necessity for parking spaces [61, 25].

Furthermore, the terms carpooling, ridesharing and car-sharing may or may not be used interchangeably. As common sense, ride-sharing exists when two or more trips are executed simultaneously, in a single vehicle. We use the term carpooling since it is widely known in the literature [18, 49]. The term car-sharing is regularly understood as a service in which a car, provided by a company, can be booked by people who only occasionally need a ‘rental’ car for a short distance[61].

A particular set of carpooling determinants are the (dis)incentives present in mobility management schemes which aim to increase the popularity of carpooling. The rationale behind the promotion of carpooling is that every carpooling employee implies one car less on the road. In addition to the benefits to the environment, the quoted benefits of carpooling are self-evident: driving costs may be shared, commuters are not dependent on schedules and/or public transport networks [61]. Typically carpooling can also be motivated by an incentive such as a faster HOV (high-occupancy vehicle) lane or a toll reduction.

Casual carpooling (also called “slugging”) is a system of carpooling without trip-by-trip pre-arrangement [8]. In others words, casual carpooling refers to the sharing of a ride with a driver and one or more passengers, where the ridesharing between the individuals is not established in advance but coordinated on the spot [39]. It is the carpooling practice in ad hoc mode, informal carpools for purposes of commuting and eventual rides.

Some works in the literature on carpooling have focused on the real time constraints of their services [55, 54, 24, 25, 50]. In [24], they describe the design concepts, distribution and cloud computing strategies for future global carpool and ride-sharing solutions, making it scalable and ubiquitous enough to successfully reach and serve a global user base. [55] and [54] addressed the problem of optimizing dynamic requests processing for setting up an optimized dynamic carpool service. [54] proposes a distributed dijkstra for the implementation of a real time carpooling system based on the multi-agent concept. In a later work, [55] adopt a subdivision principle and the multi-agent concept that permit, jointly to a real time locating module, to perform a distributed process. The latter mainly concerns the optimized dynamic assignment of available cars’ offers to instantaneously issued users’ queries while ensuring traceability, communication and security services.

Minett and Pearce [41] aims to find out if casual carpooling reduces energy consumption, and if so, how much. Their results estimate that casual carpooling in San Francisco is saving about 1.7 to 3.5 million liters of gasoline per year, or 200-400 liters for each participant, much of which comes from the impact on the rest of the traffic. In [37] a technology is introduced that will allow casual carpooling to function in areas without high-occupancy vehicle lane (HOV), by providing an administrative system that records actual carpooling behavior so that the access to an HOV lane can be made available. The author also addresses some of the current shortcomings associated with casual carpooling such as personal safety, the “free rider” problem, and the disincentive to maximize the number of passengers sharing a ride.

There are some carpooling system currently available on the web [3, 4, 2, 1]. These systems have common functionalities like: search for occasional lifts and regular lifts; match of the lift for groups of users (school, events or companies). They are typicalle for pre-arranged c ride sharing. In these systems the users can have two different roles: driver or passenger. They fill their profile providing some personal information like name, gender, age and information about their trips, such as address of departure and arrival. The ride matching is based on information inputted by the users and notifications services keep the contact between the driver and passengers.

There are several works that address the carpooling issue, but few works are related to the computer science field by focusing on the casual carpooling and other restrict group consider semantic information in their methods. In addition, most of the work related to carpooling, do not consider the activity as an attribute in their models. One recent work done by Cho [17] follow
the activity-based approach since the authors use an ontology in an activity-based microsimulation by providing a carpooling case. They introduce related studies and basic knowledge about using methodologies, and provide an example of using ontology in an agent-based carpooling simulation. While no explicit evidence is presented in this paper, they focus in recognizing that the ontology is a useful and appropriate method for the activity-based microsimulation research. However, only a conceptual design and framework are suggested, and this study is a clearly preliminary step.

A recent study is about casual carpooling (slugging) was done by Ma et al. [39]. In this work, the authors formally define the slugging problem and its generalization. The authors provide proofs of their computational time complexity. For the variants of the slugging problem that are constrained by the vehicle capacity and travel time delay, they prove NP-completeness and also propose some effective heuristics. They performed the experiments in a GPS trajectory data set containing 60 thousand trips, and their results showed that their heuristics can achieve close-to-optimal performances, which means as much as 59% saving in vehicle travel distance.

In a recent paper [11] proposes GOTOGETHER, a recommender system for car pooling services that leverages on learning-to-rank techniques to automatically derive the personalized ranking model of each user from the history of her choices (i.e., the type of accepted or rejected shared rides). Then, GOTOGETHER builds the list of recommended rides in order to maximize the success rate of the offered matches using social media check-in information from Foursquare.

The social aspect is also taken in consideration in paper [33] where authors introduce a measure of enjoyability based on people’s interests, social links, and tendency to connect to people with similar or dissimilar interests. They compute enjoyability from crowd-sourced data, and they show how this can be used on real world datasets (Rome and San Francisco) to optimize for both mobility and enjoyability.

1.3 Social Media

Unprecedented amount of user-generated data on human movement has been collected through the introduction of location-based services in social media applications of smart-phones provided by some virtual social networks (e.g., Facebook, Foursquare, Instagram, Twitter etc) [16]. It has enabled people to share their activity related choices performing “check-ins” during their visits to venues. This data contains detailed geo-location information, which reflects extensive knowledge about human movement behavior. On recent years the amount of “geo-social information” increased substantially contributing as an important source of knowledge about the human mobility behavior. In 2010, the geo-tagging feature was also added to Twitter by generating a large amount of geo-information daily. On the first year the average number of Tweets sent per day was 50 million while in December 2012 it has increased to 350 million [35]. Nowadays people are considered as sensors, producing signals on events they are directly involved in or they are present.

Social media has been used for the analysis of collective sentiments, for understanding the behavior of groups of people or the dynamics of public opinions [57, 14]. It has become an important source for unprecedented reach, speed and democratization of communication. Due this vast applicability, social media analysis is a fast growing research area aimed at extracting useful information from this amount of data.

Social media plays a crucial role for understanding of the modern life, including potentially for analysis with focus on transport [26, 28, 22] and human mobility analysis [40, 14, 35].

1.3.1 Mobility and transport analysis using Social Media

Social media posts are often tagged with geographical coordinates or other information (e.g., text, photos) that allows identifying users’ positions. This information may enable mobility pattern analysis using trajectory mining techniques. Therefore, social media users moving through a set of locations produce a huge amount of geo-referenced data that embed extensive knowledge about human dynamics and mobility behaviors [14]. In the latest years, there has been a growing interest in the extraction of trajectories from geotagged social data using trajectory mining techniques [66].
Using millions of check-ins gathered from foursquare, [21] introduced a model for the structure of local urban areas that groups nearby foursquare venues into clusters called Livehoods. Those clusters reflect the dynamic nature of activity patterns in the lives of city inhabitants. In [14], the author described a methodology and main results of an experimental study aimed at discovering behavior and mobility patterns of Instagram users visiting EXPO 2015 in Milano. They were able to discover how the number of visitors changed over time, identify the most frequent sets of visited pavilions, which countries the visitors came from, and the main flows of destination of foreign visitors to Italian regions and cities after their visit to EXPO. A similar work was previously presented [13] presenting analysis of geotagged tweets carried out to understand the behavior of people attending the 2014 FIFA World Cup in Brazil. The authors monitored the Twitter users attending the World Cup matches to discover the most frequent movements of fans during the competition. Original results were obtained in terms of number of matches attended by groups of fans, clusters of most attended matches, and most frequented stadiums.

Social media has valuable information for transport policy that can be harvested. Its has been recently employed as a source of information for event detection, with particular reference to road traffic congestion and car accidents, as well as it may represent a new and effective means for organizations to communicate with customers and citizens. Paper [28] explores two sides of engagement with social media – firstly the potential uses of social media by transport service suppliers and secondly the potential value to policy development of shared transport related information by the public. The findings give insights into the practices of organizations of different size, function and longevity of social media use. The early results of a study to harvest freely available transport information from the public and transport system users are also presented, demonstrating that transport policy relevant information can be harvested from online social media sources. The authors also proposed in [27] an approach for detection of presence of policy-relevant transport information in the texts collected from social media and a posterior classification of the text on three main categories: expressing a need to travel from origin to destination, updating on the current status of the transport network, and expressing an opinion about a transport service.

A group of works have been done for real-time monitoring system for traffic event detection from Twitter stream analysis. Paper [5] has as main contribution an automatic tweet interpretation tool, based on Machine Learning techniques, that achieves good performance for traffic-related tweets distributed by traffic authorities and news agencies. Given a traffic-related tweet, their tool uses named-entity recognition techniques to identify the location of the event the tweet describes and relation extraction methods to capture relations between the components of the event. In [22], they propose a system that fetches tweets to perform the classification aiming to assign the appropriate class label to each tweet, as related to a traffic event or not. They employed a SVM classification model achieving an accuracy value of 95.75% by solving a binary classification problem (traffic versus non-traffic tweets). Their model is also able to discriminate if traffic is caused by an external event or not, by solving a multiclass classification problem and obtaining an accuracy value of 88.89%. In [26], the authors are interested in the people’s thoughts about carsharing, bikesharing, electric vehicles. However, the use of social media was at an early stage of the research.
Chapter 2

Work Plan

Smart cities are nowadays becoming a trendy topic proposing this vision of ultra modern urban area addressing the needs of business, institutions and citizens. Today the cities are claimed to be the major contributors to the climate problem consuming 78% of the world’s energy, producing the 60% of the CO\textsubscript{2} but covering less then 2% of the planet surface [52]. Smart cities are also defined in [58] as “places where information technology is combined with infrastructure, architecture, everyday objects and our bodies to address social, economic, and environmental problems”.

Figure 2.1 shows a comprehensive smart city model with the role of the different and related “smart” component like mobility, environment, living, people, economy and governance.

![Figure 2.1: The smart city model proposed in [52]](image)

This figure shows how the different aspects are strictly interconnected under the smart city vision and advances in one sector unavoidably affects the others. As a consequence, we think of an integrated vision where smart mobility is really related to all the other aspects. In cities becoming more and more “smart” we will see the proliferation of ubiquitous computing and the popularization of smart-phones and wearable devices favoring the collection of enormous amount of data, the so-called \textit{Big Data}. Great part of these data are generated by individuals who track and share their experience and opinions daily about the “physical” and “on-line” world. This type of sensing has been denominated \textit{social sensing} since the citizen acts as human sensors reporting information about the their lives and the surrounding environment in a “near real time” fashion.

Our proposal is to take advantage of this social sensing, or wisdom-of-the-crowd represented by social media, to propose new services in the context of smart cities and in particular smart mobility. The challenges become how to effective harvest and mine crowd data such as the social
media, to shape mobility recommendations according to users’ tastes and context. We highlight some targeted researches as proposals for novel smart services based on the analysis of human social mobility data.

- **Exploring Human Activity Behavior for Smart Carpooling.** The idea of this line of research is to exploit the mined knowledge about human mobility to propose a new smart way to share car rides. Here the idea is to exploit the human tendency to be flexible in choosing their destination when to perform a given activity. By choosing alternative destinations we boost the ride possibilities thus increasing at high rate the carpooling potential. This line of research has already produced the first results published in two papers [43, 44], a mention in the prestigious IEEE Spectrum magazine [64] and it is detailed in the next chapter of this proposal.

- **Investigating Social Media for Classifying Attendance for Social Events.** Here the idea is investigate social media data to identify and predict the attendance of social events such as musical performances, conferences and festivals. These kind of social events are attractors since thousands of individuals move to a specific location (the place where the event is held) at a specific time (when the event starts until its end). The motivation for our research is the vast amount of smart services that can be provided to attendants such as: (1) transportation planning like the allocation of buses from certain areas (2) the recommendation of carpooling or taxisharing services for attendants (3) ubiquitous services for recommendation of third-party services like uber or taxi services. Based on this idea our target is to develop a framework for classifying the user attendance to social events from social media. The current status of the framework and some preliminary results are presented in the next chapter and in a submitted paper [42].

- **Pushing the role of social aspect more in the carpooling practice.** It is clear that the acceptance of a ride depends mostly on a trust relationship of the passenger with the driver. One possible line of research is to investigate how group recommendation can play a role in creating groups of friends that can share a ride, thus combining the individual preferences in choosing the best ride with the possibility to share the ride with people the passenger may trust. We can take inspiration by the work [9] where a framework to recommend the best group of friends to enjoy an item is presented. The challenge can be to redefine this problem taking into account the spatial and temporal aspects relative to the ride sharing trying to balance the friendship relationship with the car rides preferences.

- **Advancing in the recommendation and personalization aspects.** The ride recommendation may be combined with the location recommendation, so that the best ride combines also the best Point of Interest where to go based on the user preferences, history, friends preferences, etc. How to explore these aspects is a topic to be investigated.

This links to one of the biggest trends in the past year that is the importance of the visual content: “a picture is worth a thousand words”. In 2014, visual-centric social media platforms like Pinterest and Instagram increased their user count and the engagement rate as a skyrocket. On the second semester of 2014, 70 million photos and videos are shared on Instagram daily and 50 percent of Pinterest’s 30 billion pins were added in the last six months 1. It has been reported that 63 percent of social media is made up of images2. This leads to a new kind of analysis of social media based on images and videos. How to fruitfully combine this topic with smart mobility services is a hot research topic. Linked to our current ongoing work on investigation of attendance in social media, we can use posted photos for the extraction of new features to increase our accuracy.

The research directions presented above are just a few example of interesting ideas we can investigate for mobility analysis from crowd data to propose smart services.

---

1http://expandedramblings.com/index.php/pinterest-stats/
2http://www.skyword.com/
Chapter 3

Preliminary Results

We present our preliminary results in two topics discussed in the Work Plan section: (1) Exploring Human Activity Behavior for Smart Carpooling (2) Investigating Social Media for Classifying Attendance to Social Events. In Section 3.1 we present the system ComeWithMe, published in papers [43, 44] for activity-based carpooling. In Section 3.2, we present some initial results on the mining of social media to classify event attendance. These results are detailed in a submitted paper [42].

3.1 Exploring Human Behavior for Activity-based Carpooling

In big cities, the traffic congestion becomes a factor of decision in the people's life. When planning their daily activities people begin to consider where to go, which time to go and how to go, how much delay is supposed to be wasted given the traffic congestion. In some cases, people are willing to accept other more easy reachable options. For example, a person may wish to eat at a famous restaurant located at the opposite side of the city, but the possibility to get stucked in traffic jams, the consequent waste of time to arrive and the stress of driving in heavy traffic conditions may convince him to change destination to closer or better reachable restaurant. The development of the cities allows people to have more and more venues where to perform theirs activities. The accessibility of these places in terms of ease of transportation is crucial when deciding which place to visit. Usually people can find several options of restaurants, gyms, cinemas, supermarkets to name a few common activities, not being restricted to move always to the same place.

The analysis of the human mobile behavior in relation to the activities has been described in [23] where we discover how people tend to be regular or flexible when choosing a place associated to a given activity. The regularity analysis was conducted over two different perspectives: spatial and temporal. We discovered that each person has his/her own preferences about which place (spatial perspective) and which time (temporal perspective) to carry on their activities. The regularity in visiting a specific place depends on many factors, such as a working relationship, proximity to home or work, convenience, pleasantness, the effort needed to reach the place to name a few. During this research we therefore noticed that a person does not need to go always to a specific place to perform a given activity since some activities are not strictly associated with a unique place. A typical example is a person willing to have dinner at a restaurant: an urban area typically provides a considerable amount of places where eating out and therefore restaurant from which to choose. Some activities can be also be performed during different intervals of the day, and this is the gym example: the place can be fixed but users can change the time to go according to the daily routines and commitments.

We explored this idea of flexibility in choosing the destination in a carpooling context in the ComeWithMe carpooling system [43, 45, 64]. The idea is to take advantage of users flexibility
to offer to a passenger rides to alternative places where to perform the desired activity as an alternative to the desired destination (and/or at the desired time) when a ride is not available.

The method works using the semantic information about the venues to check the possible activities that may be performed and their corresponding opening hours.

As a carpooling system, the idea is to have drivers offering rides and passengers requesting rides. Then the systems expands the requested destination into a sent of alternative destination where the intended activity can be still carried out.

A routine trip, corresponds to a regular trip made by a driver and it is represented as a sequence of spatio-temporal points. A common example of a routine trip is the daily path from home to work. Routine trips are considered as fixed routes during the day and they represents the possible rides for offering carpooling. The user who performs a routine trip assumes the role of a potential driver.

A carpooling request represents the need for a ride from a passenger \( u \) and can be represented by a tuple \( cr_u = \langle \text{DeparturePoint}, \text{DepartureTime}, \text{DestinationPOI} \rangle \).

**Problem definition: Activity-based Carpooling.**

Given a set of routine trips \( RT \), a user \( u \) with her preferences \( \text{prefs}_u \), a carpooling request \( cr_u = \langle \text{DeparturePoint}, \text{DepartureTime}, \text{DestinationPOI} \rangle \), we want to find and rank all the routine trips \( r = r_1, \ldots, r_m \in RT \) such that \( r \) allow the passenger \( u \) to reach either the desired destination \( \text{DestinationPOI} \) or an alternative destination satisfying her preferences \( \text{prefs}_u \). The ranking \( \text{Rank}(u, r_i) \) is defined such that \( \text{rank}(u, r_i) > \text{rank}(u, r_j) \), if \( r_i \) is more relevant than \( r_j \) respect to \( \text{prefs}_u \).

### 3.1.1 The ComeWithMe system

ComeWithMe comes as a possible solution to the activity-based carpooling problem. It receives as input a dataset of routine trajectories, dataset of Point Of Interests, a set of intervals, a spatial grid, a carpooling request and produces as output a list of possible matches including the alternative destination for an activity-based carpooling.

This process is illustrated in Figure 3.1 and detailed as follows.

![Figure 3.1: Schema of the ComeWithMe Process](image)

The dataset of POIs provides geographic (latitude and longitude) and semantic information (POI category and opening time) about location of interest in the area. The semantic information is used for identifying, through the categories, which activities can be performed in that POI and when during the day the POI is available. The POIs datasets can be downloaded using any service.
like Foursquare\(^1\), Google Places\(^2\), Yellow Pages\(^3\) or any other. What is crucial is to have the category of the POI since the category identifies the activities that can be performed and the opening hours. The set of intervals corresponds to slices of the day, while the spatial grid overlaps and slices the geographic region of interest in squared cells. The diagonal of the cell represents the maximum straight distance that a person should walk in the worst case to start the ride (in case of the pickup cell) and/or to reach the desired place from the drop-off point (in case of the arrival cell). The routine trips are obtained by mining historical GPS data by applying the clustering method proposed in [59].

Figure 3.2 illustrates the mapping of a routine trip (left) and a carpooling request (right) to a grid 3x3 according to the set of intervals.

A routine trip is mapped to the cells crossed during its path, while the carpooling request is mapped by its departure point, the destination POI and relative temporal intervals.

Each POI contains spatial (i.e. the geographic position) and semantic (the category and the business hours) properties. Using the spatial properties, we map the POIs to the corresponding cell in the grid. Using the semantic properties and the set of intervals, we also map the available cell according to the category and the business time of the place. Mapping the dataset of POIs to the spatial grid using these two semantic properties gives us a semantic grid describing the availability of the POIs in a cell for a given category.

ComeWithMe is designed to answer efficiently and effectively activity-oriented carpooling queries. Efficiency and scalability are needed since the possibility of reallocating space and time boosts the rides opportunities at a high rate. For example, the activity "eating Italian food" may result in hundreds of different destinations in a large city, which raises to thousands when a less specific activity is chosen, e.g., "eat at a restaurant". In general, the number of possible rides tends to exponentially increase with the generality of the activity performed. The additional flexibility in the management of the pickup-up points and the temporal constraints further increment the complexity of query answering. It is therefore crucial to handle the scalability as the rides possibilities increase and, at the same time, manage a relevance notion to present to the user the best rides ranked according to his/her interests. Indeed, ComeWithMe exploits information retrieval techniques to storing semantic grids, answering the carpooling requests with alternative destination and ranking these candidate rides according to the user’s context and preferences. A number of features affect ranking such as: user current location and schedule, her willingness to walk to reach the pick-up or destination points, her adaptability to anticipate or delay the preferred departure and arrival time. The personalization of ranking criteria is important to improve the quality of experience perceived by the user and increase the probability of the ride acceptance by the passenger.

The ComeWithMe passengers’ queries are answered by providing a list of ride opportunities ranked according to the user context and preferences. This task is accomplished by mean of three

---

1http://www.foursquare.com
2places.google.com
3http://en.wikipedia.org/wiki/Yellow_pages
modules: Inverted Index, Query Expansion and Ranking Model.

Inverted Index. Given the pick-up cell, the destination cells, and a time-window, ComeWithMe exploits an inverted index to store and retrieve the trips that cross these cells within the specified time window.

The index aims at speeding up query processing by improving the capability of ComeWithMe to handle a number of ride opportunities, thus making the system scalable. We have a vocabulary entry for each cell of the spatial grid. In turn, each cell is associated with a list of postings, where each posting stores the tripid and the estimation of the time at which the car will cross the given cell. To speed up query processing, postings are organized into buckets where each bucket contains the postings referred to a specific coarse time window. Within each bucket, the postings are ordered by tripid. The organization in buckets avoids the scanning of the complete postings list when searching for trips, while the postings ordering within buckets allows to perform efficiently postings intersection operations.

To retrieve the candidate rides for a user query, ComeWithMe uses the pick-up cell, the time window for the ride (hint for accessing the right buckets), and the possible destination cells. The queries have the following semantics:

\[(\text{pickCell AND} \ (\text{destCell}_1 \ OR \ \text{destCell}_2 \ OR \ \cdots \ OR \ \text{destCell}_n))\]

Trips matching the above Boolean query are retrieved from the inverted index and the ones not respecting the correct order between pickup and destination cells, or the user-specified time constraints, are filtered out from the results list.

Query Expansion. This module boosts the possibilities of car rides by exploiting a query expansion technique. The use of query expansion generally increases recall and it is widely adopted in many application fields [12]. The queries in our context are carpooling requests expressing the passenger’s intention to move to a venue to perform an activity. Given the destination Point of Interest (POI) specified by the passenger, the query is automatically expanded with places related to the same activity by using a hierarchical thesaurus (an example is shown in Figure 3.3). The specific POIs are the narrowest terms, while the intermediate layers represent different activities abstraction levels and thus possible query generalizations. For example, looking at Figure 3.3(left), when a passenger requests as destination “Da Gino”, we see that it is an Italian Restaurant and expanding the query over Italian restaurants we have “Ristorante Giannino” as an alternative destination. Abstracting again up to “Eating” we have all the venues where they serve food corresponding to “Pizzeria”, “Japanese restaurants”, etc. The more we expand the query to broader terms, the more rides possibilities the passenger can select from the driver offers.

![Figure 3.3: Structure of the thesaurus (left) and query expansion example (right)](image)

Each venue in the thesaurus is associated with a cell of the spatial grid indicating its location. Analogously, user queries are coded with the cells representing the pick-up area, the destination place, and a set of other cells representing alternative destinations. An example of the expansion process is illustrated in Figure 3.3 (right): the destination POI “Da Gino” is expanded with other possible venues (and cells) where the passenger can perform the activity “Eating”.

Ranking Model. We have already discussed how the query expansion consistently increases the possibilities of rides and, consequently, how crucial it is to propose to the user the rides that she is willing to accept.

The ranking score of candidate rides is thus computed as a linear combination of a set of features, mainly derived from the flexibility preferences the passenger can set: 1) a temporal tolerance indicating the delay of the departure time of the trip respect to the preferred time
indicated in the query; 2) a temporal tolerance on the possible anticipation of the trip respect to the indicated preferred time; 3) a spatial tolerance indicating how much the passenger is willing to walk to reach the pick-up point and/or the destination location. Other information considered in the computation of the ride score includes the trip duration and the semantic similarity between the actual destination of the ride and the one specified in the query. Intuitively, the duration of the trip should not be too long respect to the duration of the fastest of all the possible rides. On the other hand, the destination venue should be, in order of preference: close to the PoI chosen in the query; another PoI in the same thesaurus category (e.g., a different Italian restaurant when the requested venue was an Italian restaurant); a PoI in the more abstract category of the thesaurus.

The set of candidate rides retrieved from the index are thus ranked by considering:

1. the **distance** that the passenger is supposed to walk to reach the pick-up point and/or the destination location from the drop-off point. This distance is computed as the sum of the distance between the user location and the pick-up point, plus the distance between the drop off point and the destination venue. The obtained value is normalized with respect to the maximum distance the specific passenger declared to be willing to walk;

2. the **anticipation** of the trip respect to the desired pick-up time. This time is normalized with respect to the maximum time tolerance set in the user preferences;

3. the **delay** of the trip respect to the desired pick-up time (normalized as the previous feature);

4. the normalized **duration** of the trip. This feature has the value 1 when the duration is the minimum among all the retrieved rides, while it has the value 0 when the duration is close to the maximum duration of all the candidate rides;

5. the **semantic similarity** between the preferred destination PoI in the user query and the alternative PoI reached by the current ride. This value is set to 1 when the ride reaches the preferred destination, while it smoothly decreases to 0 as we consider PoIs in broader categories of the thesaurus.

For the first three features lower is better, while for the last two the opposite holds. Candidate rides are ranked using a linear combination of these five features where each feature is associated with a weight $\omega_i \in [-1, 1]$. Since no golden standard is available to optimize weights, the current prototype uses a uniform weighting schema $[-\frac{1}{5}, -\frac{2}{5}, -\frac{1}{5}, \frac{1}{5}, \frac{2}{5}]$. The effect in the result list of manually changing these weights will be shown during the demo.

The **Mobile Application** **ComeWithMe** has been implemented as a mobile app and has two different profiles of users: the driver and the passenger. The passenger interface allows the user to request rides for a given destination, as shown in Figure 3.4. Once the required information is filled in, the user can submit the query and see the ranked list of rides offer.

In Figure 3.4, we see on the left our query example representing a user asking for a ride in the city of Pisa to go to the “Bella Napoli” pizzeria located in “via del Borghetto”. The query specifies also the temporal tolerance (delay 30 minutes or anticipate 30 minutes) from the desired departure time at 19.00 and the spatial tolerance indicating the maximum distance the passenger is willing to walk (up to 600 meters).

**ComeWithMe** returns, for each query, a ranked list of rides where the most relevant options are shown on the top. We used a test dataset with GPS trajectories in Pisa [44] from which we extracted the routine trips. During the specified temporal window (from 18:30 to 19:30), we have a total of 276 routine trips, 23 of which spatially matches the query from the pickup point to at least one “Pizzeria” among the 121 in the dataset. Since each routine trip can pass through many cells where “pizzeria” places are located, the system retrieves and ranks a total of 156 rides to “pizzeria” alternative destinations (see Figure 3.4 (left)). Observing the results of the query, we notice that the first two rides are to the intended destination pizzeria “Bella Napoli”, while other destinations are “Panuozzo” and “La Greppia”.

The passenger can select a ride from the ranked list and visualize some information about the driver and other details about the ride (e.g. the pickup address and time, the destination place,
the estimated arrival, etc). Once the user selected and confirmed a ride, ComeWithMe notifies the driver about the upcoming request. Symmetrically, as shown in Figure 3.4 (right), a driver can see the list of passenger requests and she can select one to visualize the details. From the details interface the driver can accept or decline the request, she can call the passenger, start a chat and visualize the trajectory on the map.

A demo video of the system is available at the web site http://comewithme.isti.cnr.it

![Figure 3.4: Passenger Interface (left) and Driver Interface (right).](image)

### 3.1.2 Experimental Results

As reported in [43] we performed an evaluation of the system using historical data. The basic idea is to identify, from the dataset of GPS trajectories, the *Routine trips* and *Occasional trips*. Occasional trips are all the trips which are not routine. Then we make a direct correspondence of the occasional trips to *Carpooling Requests*. The intuition behind is that occasional trips can be considered as requests since they are not often performed by drivers and they can be avoided if a ride can be given. Naturally, this is only a simulation to evaluate the carpooling potential of ComeWithMe. Due to lack of space we do not report here the details of the evaluation which can be found in [43]. However, we think it could be interesting to comment here an interesting results which is the table fo the potentially saved kilometers with ComeWithMe. Table 3.1 shows the amount in kilometers of occasional trips potentially avoided from the supplied carpooling requests. As expected the values for our proposed algorithm reached values much higher than the baseline approach (BaselineM) which is the standard carpooling. These values represent a positive impact for the traffic and the environment. If we consider that using 1 liter of gasoline, a normal car performs about 10 km the lowest result (50.203 km) could save 5.020 liters of gasoline, while the highest (106.786 km), 10.678 liters. These gas savings represent less pollutant for the atmosphere and lower cost for the drivers, since they can share the cost of the trips.

As expected, ComeWithMe produced the best performance gains, indicating that, in almost all the cases, if the passengers have the flexibility to change the destination and the time, then ComeWithMe gives good results. In our experiments, the maximum gain produced when compared to the traditional carpooling was +83% (92% - 9%) with a spatial tolerance of 0.5 kms and a temporal tolerance of 0.5 hours.

### 3.2 Classifying Attendance to Social Events

The huge volume of user-generated data in social media platforms, such as Facebook, Twitter and Instagram, can be exploited to extract valuable information concerning human dynamics and
3.2. Classifying attendance to social events

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>$tw$</th>
<th>$\alpha = 0.25$ km</th>
<th>$\alpha = 0.5$ km</th>
<th>$\alpha = 1$ km</th>
</tr>
</thead>
<tbody>
<tr>
<td>BaselineM</td>
<td>0.5 hour</td>
<td>5.333 km</td>
<td>10.260 km</td>
<td>18.030 km</td>
</tr>
<tr>
<td></td>
<td>1 hour</td>
<td>5.333 km</td>
<td>10.260 km</td>
<td>18.030 km</td>
</tr>
<tr>
<td></td>
<td>2 hours</td>
<td>5.333 km</td>
<td>10.260 km</td>
<td>18.030 km</td>
</tr>
<tr>
<td></td>
<td>4 hours</td>
<td>5.333 km</td>
<td>10.260 km</td>
<td>18.030 km</td>
</tr>
<tr>
<td>ComeWithMe</td>
<td>0.5 hour</td>
<td>50.203 km</td>
<td>72.281 km</td>
<td>86.727 km</td>
</tr>
<tr>
<td></td>
<td>1 hour</td>
<td>61.101 km</td>
<td>82.716 km</td>
<td>94.353 km</td>
</tr>
<tr>
<td></td>
<td>2 hours</td>
<td>75.998 km</td>
<td>94.634 km</td>
<td>101.248 km</td>
</tr>
<tr>
<td></td>
<td>4 hours</td>
<td>91.137 km</td>
<td>105.061 km</td>
<td>106.786 km</td>
</tr>
</tbody>
</table>

Table 3.1: Kilometers of the OTs Potentially Avoided

Figure 3.5: Percentage of CRs Potentially Supplied ($Pot^{CR_{sup}}$)

behaviors [14]. Social media is now a platform for sharing virtually many kinds of information, and studies on Twitter data have demonstrated that aggregating huge amount of messages can provide valuable insights into a population [36]. Once a topic is raised by a social media participant, typically others react and a ‘conversation’ may develop. The exchange of transport-related information can be found in all forms of social media. Therefore, the contents of these new forms of social interaction represent a large variety of interests of the media users.

In general, as we have seen in section 1.3.1, most of the literature regarding the analysis of mobility using social media are: (1) geospatial analysis based on geotagged data [14, 13] (2) real-time application focused on traffic event detection [10, 22, 5] (3) sentiment analysis on quality of the transportation services [28, 27]. From a different perspective, our on going research investigates social media data to identify the users attendance in large events such as musical performances, conferences and festivals.

Musical festivals, religions celebrations and conferences can attract thousands of persons to a specific location (the place where the event is held) at a specific period of time (the time during which the events is held, can be one or more days) thus raising mobility issues.

There is a tendency for social media users to express, through the social network their feelings, experience or opinion about a social event. These thoughts can be shared before, during or after the event, and all together this generates a large amount of social media data related to the event. The social networks work as a channel between people and “the event”, a kind of virtual entity typically represented by the sponsor and organizers accounts. On Twitter, for example, people post text, links, photos or videos of the event by using hashtags (text beginning with the # symbol), or they explicit mention the event’s accounts (text beginning with the @ symbol) or they simply mention the name or acronym of the event.

The analysis of these event related data can generate valuable information when planning future events, but also to improve the services offered during the event itself. The whole data may contain implicit feedbacks or brief experience reports from the attendants, in addition to the general comments and feelings from the non participants. As we seen in the brief literature survey in Section 3.2, such data can be used to collect opinions of the attendants through sentiment analysis, while the spatial analysis, mainly based on the geo-tagged tweets, can be used to monitor traffic events.
3.2.1 The Event Attendance Classifier

Given this context, we propose a new investigation line where the main objective is to infer the event actual attendance of the media users through the analyses of their posts. We model this as a classification problem where the social media posts are analyzed by a machine learning algorithm to label the posts as “participants” or “non participants”. This machine learning problem has been split into three tasks depending on which posts are used in relation to the event: posted before, during and after the event. Figure 3.6 depicts these three temporally disjoint datasets. The content of the posts varies significantly during these intervals. Before the event users typically express expectation, during the event they may express opinion about how the event is going and after they can post general discussions and feelings.

![Figure 3.6: Time scale representation](image)

We therefore formalize some terms and definitions that will be used later.

**Social Event** \( (e_{social}) \). A social event is an entity with spatial and temporal properties. In the social media an event is associated to an identifier \( i \), such as an hashtags and/or a social network account. In the real world, a social event is usually held at a venue with a spatial location defined by geographical coordinates (lat, long) of the representative point, \( l \). An event has temporal duration which may vary from minutes to days or weeks and it is represented by a time-widow defined by the starting time \( (t_{e0}) \) and the ending time \( (t_{ef}) \). In a social network, the social event’s related posts can be identified by mentioning its identifier \( i \). We formalize a social event as tuple: 
\[
\text{e}_{social} = < i, l, (t_{e0}, t_{ef}) >
\]

**Social Media Post** \( (p) \). A social media post has a media content \( c \) such as text, links, emoticons, photos or videos, a temporal property \( t \) represented by the date in which it has been created/posted, a user \( u \) who created the post and it may have or not spatial information \( s \). We formalize a social media post as tuple: 
\[
p = < c, s, t, u >
\]

**Event Related-Post** \( (p_{ev}) \). An event-related post is any post \( p \) related to the event \( e_{social} \) through the mentions of any event identifiers \( e_{social}.i \) in the content of the post \( p.c \):
\[
p_{ev} = p \text{ where } e_{social}.i \text{ in } p.c
\]

Depending on on the the posting date, an event-related post can belong for one of these disjunct sets:

(a) **Before the Event** \( (p_{ev_{bef}}) \). When the post has been created before the Event starts. Formally:
\[
p_{ev_{bef}}, \text{ is any } p_{ev} \text{ where } p_{ev}.t < e_{social}.t_{e0}
\]

(b) **During the Event** \( (p_{ev_{dur}}) \). When the post has been created during the Event. Formally:
\[
p_{ev_{dur}}, \text{ is any } p_{ev} \text{ where } p_{ev}.t \text{ between } e_{social}.t_{e0} \text{ and } e_{social}.t_{ef}
\]

(c) **After the Event** \( (p_{ev_{aft}}) \). When the post has been created after the Event. Formally:
\[
p_{ev_{aft}}, \text{ is any } p_{ev} \text{ where } p_{ev}.t > e_{social}.t_{ef}
\]
3.2. CLASSIFYING ATTENDANCE TO SOCIAL EVENTS

We recall that our main objective is to perform classification of users into attendant or not by classifying their posts to social media and labeling them as positive (attend) and negative (not attend). Thus, for each temporal set (before, during and after) a specific classifier is modeled and applied. Therefore, we manage a total of three classifiers, one for each classification task. The classifiers are boolean since the labels of the posts are positive or negative.

Task 1. Classifying Attendance Before the Event
This task has the implicit meaning of predicting the attendance of a user to an event based on his or her post before the event $p_{\text{bef}}$. Through this classifier it is possible to detect the users likely to participate in the event, that, from a mobility point of view, people that are supposed to move from their residential address to the event location during a given temporal interval.

As we can see in the Figure 3.7a, the user implicitly expressed his not attendance in the event CreamFields, a music festival in UK. The user in Figure 3.7b instead clearly expresses his future participation in the event.

Task 2. Classifying Attendance During the Event
The implicit meaning of this task is to identify the user that are (current-time) or were (historical analysis) present during the event. People can express their current participation in the event through social network by different ways such posting photos, either making comments about the event. Therefore, the classifier for this task is in charge to detect the attendance (positive, as in the Figure 3.8b) or not attendance (negative, as in the Figure 3.8a) of a user in the event by his posts in a social network.

Task 3. Classifying Attendance After the Event
Even after the event people comment, express their opinion or publish photos. The interpretation of some posts after the event can give a clear idea of past attendance (positive, as in the Figure 3.9b) or not (negative, as in the Figure 3.9a).

Features Extraction. The classifiers we have built exploit several boolean features extracted from the posts, grouping them into four different categories: text features, temporal features, social features and media features.

(a) Text-features. They are extracted based on the text contained in the post. Each unigram and bigram in the text becomes a boolean feature.

(b) Temporal-feature. These features are extracted based on the date of the post. There features are specific for the task 1: is_six_months_before, is_three_months_before, is_two_months_before,

---

4http://www.creamfields.com/
CHAPTER 3. PRELIMINARY RESULTS

Figure 3.9: Classifying attendance after the event

(a) negative  (b) positive

is\_one\_months\_before, is\_one\_week\_before and is\_one\_day\_before.

These features are specific for task 3: is\_one\_month\_after, is\_two\_week\_after, is\_one\_week\_before.

For the task 2, there is no temporal features since we consider the social event as one single
time-window defined by the duration of the event. Thus all posts for this task has no different
temporal properties as their temporal interpretation is ‘during the event’.

(c) Social-feature. These features are extracted from the user’s friends social network con-
nections. The list of features: is\_follower, followers\_count\_gt\_750, followers\_count\_gt\_1500,
followers\_count\_gt\_3000.

(d) Media-feature. These features are extracted based on the media content of the post. They are:
has\_photo, has\_video, has\_instagram\_content, has\_foursquare\_content, has\_youtube\_content,
has\_facebook\_content.

These features have been used to train and develop the three Bayesian classifiers[53] to effect-
vively identify the attendance-related tweets each of our classification tasks.

3.2.2 Experimental Setup

We instantiated our attendance classifier the music festivals scenario. Particularly, we collected
Twitter posts about the CreamFields 5 festival, which was held in Daresbury, United Kingdom,
from August 25th to August 28th, 2016. The keyword ‘creamfield’ was used to collect tweets using
the Streaming API and the Full-Rest API provided by Twitter6. We crawled the data using the
Streaming API from August 10th to September 15th 2016 while we used the Full-Rest API to
collect also the past tweets from March 1th to September 15th 2016. The period for crawling thus
covers the period before, during and after the event.

Using both APIs, the collected data contains over 90 thousands of unique tweets related to this
event. The distribution in the temporal intervals are shown the Figure 3.10

Figure 3.10: Temporal distribution of the tweets $p_{ev}$ for the Creamfields dataset.

---

5 www.creamfields.com/
6 add site twitter API
3.2. CLASSIFYING ATTENDANCE TO SOCIAL EVENTS

<table>
<thead>
<tr>
<th>Task</th>
<th>Accuracy</th>
<th>N. Pos</th>
<th>N. Neg</th>
<th>Neg_{Fm}</th>
<th>Neg_{Prec}</th>
<th>Neg_{Rec}</th>
<th>Pos_{Fm}</th>
<th>Pos_{Prec}</th>
<th>Pos_{Rec}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.79247</td>
<td>163</td>
<td>92</td>
<td>0.7291</td>
<td>0.7093</td>
<td>0.7609</td>
<td>0.8311</td>
<td>0.8568</td>
<td>0.8107</td>
</tr>
<tr>
<td>2</td>
<td>0.73505</td>
<td>198</td>
<td>96</td>
<td>0.6851</td>
<td>0.5658</td>
<td>0.8750</td>
<td>0.7704</td>
<td>0.9170</td>
<td>0.6673</td>
</tr>
<tr>
<td>3</td>
<td>0.82909</td>
<td>207</td>
<td>91</td>
<td>0.7280</td>
<td>0.7152</td>
<td>0.7475</td>
<td>0.8749</td>
<td>0.8872</td>
<td>0.8649</td>
</tr>
</tbody>
</table>

Table 3.2: Classifier Evaluation

<table>
<thead>
<tr>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 1</td>
</tr>
<tr>
<td>{[(sell, number_int), neg), (followers_count gt 1500, neg), ((day, standard), neg), (standard, neg), (month_date, neg), (sell, neg), ((number_int, ticket), neg), (be, go), neg), (wish, neg), (sunday, neg), (fuck, pos), ((number_int, day), neg), (sunday, ticket), neg), (week, neg), (not, go), neg]}</td>
</tr>
<tr>
<td>Task 2</td>
</tr>
<tr>
<td>{[(followers_count gt 3000, neg), (wish, neg), (followers_count gt 1500, neg), ((wish, i), neg), ((i, be), neg), ((be, i), neg), (anyone, neg), (sell, neg), (man, neg), (why, neg), ((die, festival), neg), (some, neg), (today, neg), ((be, go), neg), (cop, neg)]}</td>
</tr>
<tr>
<td>Task 3</td>
</tr>
<tr>
<td>{[(look, neg), (followers_count gt 3000, neg), (can, neg), (followers_count gt 1500, neg), (wish, neg), (eric, neg), ((eric, prydz), neg), ((see, yo), neg), (ticket, neg), (when, neg), (wish, i), neg), (buy, neg), (when, yo), neg), (yo, know), neg), (nickyromero, neg)]}</td>
</tr>
</tbody>
</table>

Table 3.3: Most Informative Features

To generate our training set, we use a sampling technique [20]. For each classification task we selected a total of 300 tweets. To better represent the real scenario in our training set, the selected samples keeps the same ratio of tweets generated by accounts with more and less than 1500 followers. The observation behind this is that most of the accounts with more than 1500 Twitter followers correspond to entities like sponsors, organizers, sellers, bands, VIPs and they usually don’t represent real people potentially attending the event. Their posts correspond to advertisement or general information about the event, therefore not positive case of attendance for our analysis. In our dataset, 63% of the tweets related to the event Creamfields were posted by account with less than 1500 followers and the remain percentage, 37% by account with more than 1500 followers. Thus, for each of the three tasks, each tweet is individually classified as positive (participant of the event) or negative (not participant of event).

Our experiments have been driven by three research questions.

RQ1: How accurate is our event attendance prediction classifier?

We use a 5-fold cross validation process over our dataset to evaluate the performances of our classifiers. In particular, we use the following performance indicators: total accuracy, F-measure, precision and recall for negative and positive cases.

The Table 3.2 summarizes the results about the total number of positive and negative examples and the evaluation measures. We notice that we have higher accuracy for the tasks 1 (prediction) and 3 (past classification) than for task 2 (current attendance). One possible reason could be high number of posts contenting images or videos with few or none textual information. One strategy to improve our results is to use the information extracted from the published photo. Visual content is a trend in Social Media and could be more explored for our classification through the use of techniques of Deep Learning.

RQ2: What groups of features are necessaries to attain high accuracy prediction?

The most informative features used by the classifiers is showed in the Table 3.3. We see that the most top informative features identify mainly the negatives cases through features strongly related to advertisement tweets. A explanation is the lower diversity of the features for these kind of post, making using of common terms as “ticket”, “sell” and “you” by account with high number of followers. It suggests the identification of sponsor and organizers. Another notable set of features for the negatives cases, was the terms “wish”, (“be”, “go”) and “why” expressing regret or frustration for not attending in the event.

RQ3: How do these model generalize these events?
This a relevant question since we want our model not to over-fit this specific event but be as much general as possible. Our initial experiments were conducted in a dataset collected from the CreamFields music festival in UK but it is necessary to investigate the accuracy of our training set when applied to other music festivals or even to other kinds of events. Beside this, another issue to be better investigated is how the classifier generalizes among the different languages.

**A practical application of the results.**

In this section we show a practical application of our classifier. The motivation is the identification of attendance through the tweets collected in our dataset. Using our classification, we want to discover the users that were present in festival and through their profile, map their hometown in the map. In the Figure 3.11 is possible to see the origin of the attendants for the festival Creamfields in UK. Each circle, represents a cluster of users labeled with its size (amount of users). This analysis can be useful to help on future strategies for allocation of bus and other kind of transportation as discover possibilities for groups of carpooling.

![Time scale representation](image)

**Figure 3.11: Time scale representation**

More application and analysis for transportation services can be explored with the prediction of attendance. The broad idea is to propose a Framework for identification and classification of attendance in real-time using techniques of Learning Transfer. Transfer of learning is the dependency of human conduct, learning, or performance on prior experience. It explores how individuals would transfer learning in one context to another, similar context – or how ‘improvement in one mental function’ could influence a related one [29].
Bibliography


