Università degli Studi di Udine
Dipartimento di Scienze Matematiche, Informatiche e Fisiche
Dottorato di Ricerca in Informatica

Ph.D. Thesis

Effectiveness of Data Enrichment on Categorization: Two Case Studies on Short Texts and User Movements

Candidate
Marco Pavan

Supervisor
Prof. Stefano Mizzaro

Co-Supervisor
Dott. Ivan Scagnetto

Cycle XXVIII — December 2016

Institute Contacts
Dipartimento di Scienze Matematiche, Informatiche e Fisiche
Università degli Studi di Udine
Via delle Scienze, 206
33100 Udine — Italia
+39 0432 558400
http://www.dimi.uniud.it/

Author's Contacts
Dipartimento di Scienze Matematiche, Informatiche e Fisiche
Universita degli Studi di Udine
Via delle Science, 206
33100 Udine — Italia
+39 0432 558457
http://marcopavan.net
marco.pavan@uniud.it

Ai miei genitori

Abstract

The widespread diffusion of mobile devices, e.g., smartphones and tablets, has made possible a huge increment in data generation by users. Nowadays, about a billion users daily interact on online social media, where they share information and discuss about a wide variety of topics, sometimes including the places they visit. Furthermore, the use of mobile devices makes available a large amount of data tracked by integrated sensors, which monitor several users' activities, again including their position. The content produced by users are composed of few elements, such as only some words in a social post, or a simple GPS position, therefore a poor source of information to analyze. On this basis, a data enrichment process may provide additional knowledge by exploiting other related sources to extract additional data.

The aim of this dissertation is to analyze the effectiveness of data enrichment for categorization, in particular on two domains, short texts and user movements. We describe the concept behind our experimental design where users' content are represented as abstract objects in a geometric space, with distances representing relatedness and similarity values, and contexts representing regions close to the each object where it is possibile to find other related objects, and therefore suitable as data enrichment source. Regarding short texts our research involves a novel approach on short text enrichment and categorization, and an extensive study on the properties of data used as enrichment. We analyze the temporal context and a set of properties which characterize data from an external source in order to properly select and extract additional knowledge related to textual content that users produce. We use Twitter as short texts source to build datasets for all experiments. Regarding user movements we address the problem of places categorization recognizing important locations that users visit frequently and intensively. We propose a novel approach on places categorization based on a feature space which models the users' movement habits. We analyze both temporal and spatial context to find additional information to use as data enrichment and improve the importance recognition process. We use an in-house built dataset of GPS logs and the GeoLife public dataset for our experiments. Experimental evaluations on both our studies highlight how the enrichment phase has a considerable impact on each process, and the results demonstrate its effectiveness. In particular, the short texts analysis shows how news articles are documents particularly suitable to be used as enrichment source, and their freshness is an important property to consider. User Movements analysis demonstrates how the context with additional data helps, even with user trajectories difficult to analyze. Finally, we provide an early stage study on user modeling. We exploit the data extracted with enrichment on the short texts to build a richer user profile. The enrichment phase, combined with a network-based approach, improves the profiling process providing higher scores in similarity computation where expected.

Contents

List of Figures

List of Tables

1

Introduction

In the following we will introduce both the scientific background and the aims of this thesis. In more detail, in Section 1.1 we will analyze the interplay of mobile devices and social networks as the generator of a new kind of data, with rather peculiar features w.r.t. other "more classical" computing environments like, e.g., the standard Web with desktop clients. Indeed, we will see that such data are often fragmented, ill-formed, short and, in general, difficult to understand without some preprocessing and treatment. In Section 1.2 we will consider *enrichment* techniques as possible solutions to the problem, taking the *context* as a suitable source of enriched objects. Our research aims and contributions will be detailed in Section 1.3 and 1.4, respectively. Finally, in Section 1.5 we will outline the contents of the other chapters of this thesis.

1.1 Mobile Devices and Social Platforms as New Data Sources

Nowadays the wide spread of mobile devices is quickly changing both the nature and the processing of digital information. First, social media have acquired a key role, enlarging their communities and providing a huge amount of new data, also thanks to mobile devices which allow users to post content from everywhere and at every time. Second, users are pushed to produce shorter and much more fragmented content compared to the past. For instance, the well known Twitter platform allows users to write and share short texts with a limited length (140 characters)[1]. This restriction, combined with a very frequent quick writing activity carried out by moving users, often with pervasive abbreviations and new coined acronyms, opens new challenges in the field of text processing. Indeed, social network contents are analyzed for several purposes: identifying trends, categorizing and filtering news, measuring their importance, spread etc. Other researchers try to categorize short texts posted on social networks (e.g.,

[1] In a lot of cases the posted texts have a very low number of characters: several surveys show that the mode of characters is 28 [6].

tweets), using content taken from the World Wide Web, to understand user interests, to build user models etc.

The increasing pervasiveness of mobile devices have also made location-acquisition systems available to everyone. Such systems can be easily embedded in popular apps and services, being very often active during many users' daily activities. This evolution allows to collect large datasets with spatio-temporal information, and in particular it has increased the interest of researchers on studies about user movements, behaviors and habits. Several mobile applications have been developed with the aim to exploit information extracted from raw location data. Some of those track users movement during sport activities in order to monitor their performance and to give suggestions about the next training. Other applications use GPS data to track users current position for navigation systems. Some companies use location data as a feature for social network based applications, in order to give new services to users based on their check-ins. Well known examples are Foursquare [3] that bases its entire service on users location information to give suggestions about points of interests, and Facebook [2] and Twitter [4] that allow users to add their location while posting a new message on their account, in order to add more information for other users. The spread and popularity of this kind of mobile apps give people the possibility to track their location data in a lot of different ways, also associated to useful services, and to share with their friends this increasingly important source of information, often in combination with the short texts they post.

With these premises it is clear that there is a new important source of potentially interesting information to exploit. Whence, it is of utmost importance to design and implement an effective extraction process to get the right information from the collected raw location data. Moreover, it can be useful to envisage some post-process analysis, in order to infer additional knowledge which could improve the results.

1.2 Enrichment and Context

The content generated by users with mobile devices are a new and very important source of information, but often the data created are poor or even incomplete due to the mobile approach which led users to make actions in hurry or in locations where embedded sensors cannot generate proper data; or on the other hand could be the presence of a too large amount of generated data, e.g., a continuous GPS log. For instance users could write texts very quickly, with the consequence to have messages not easy to understand, or mobile devices could track positions during daily routines generating too much data which make difficult the important places analysis. To overcome these kind of issues, additional data could be extracted, often from external sources, but also from related objects in the current environment, in order to improve the data processing phase.

In the literature there is a common consensus about the usefulness of exploiting an additional source of information to enrich the data. This phase, called *Enrichment*, can be applied on several kind of data, and it seems particularly useful with user generated data to overcome the brevity and sparsity issues of short texts, and with users' locations and trajectory analysis. To enrich data there is the need to find a proper source of new information which is related to the currently analyzed data, in order to exploit the properties they share. This relationship is very important to take in care because, for instance, in cases where the problem is the lack of information, this approach could

complete the missing part, or in cases where the problem is a classification task, it could improve the results. On this basis, we assume that the analyzed data could be surrounded by other data which are close in terms of relatedness. Abstracting the concept of distance among objects we can think about user content as objects into a geometric space where the dimensions could be the aspects we want to consider when we analyze such data. In this way we are able to define how close the objects are, and to determine a region from which to extract the additional data for the enrichment phase. That region is commonly defined as *Context*, and it is a concept very frequent in scientific literature when analyzing data and users.

Context is an ever-present factor; it is the information surrounding objects, and even users while they make actions, communications, in every situation. Contextual data can be produced by several sources, starting from information as time and place, but also user movements, news, locations, and even text posted on social networks. This kind of data can be single-source, such as properties from single contextual object, or multiple-source, like group-generated properties or behaviors. Context is everything strongly related to the currently analyzed object, it could help providing information which allow us to get new knowledge about it, and even in particular cases influence it.

1.3 Aims, Motivations, and Challenges

The motivations behind our studies are related to the importance of user generated content and the value of the new knowledge that it is possibile to extract from it, in order to design and build systems more and more able to customize services. Our analysis on user generated content led us to focus on data strictly related to the user daily activities, therefore we chose what can mostly characterize users habits. Opinions and argumentations expressed in textual form (often in social networks), and the geographic locations visited by users are the two types of data that we consider for that purpose. The nature of this data make them often not ready-to-use, therefore the study of how to "mine" them is a necessary task. In these two particular domains data are often not well-formed, not generated with specific intervals, not consistently with high quality, mainly due to the "human component". For these reasons, our studies rely on these two kinds of data, compared to others that, although they can be suitable to analyze the effectiveness of data enrichment, they may have less impact on the characterization of users. Moreover, the possible connections and relationships between these two layers of data allow us to plan further studies.

In this work we aim to study and evaluate the effectiveness of the enrichment process applied on systems which mine user generated data in two different domains: (i) short texts posted on Twitter platform, and (ii) user movements generated by GPS sensors embedded in mobile devices. In particular we conduct a first study on short texts addressing the problem of categorization, in order to identify the topics discussed in the analyzed tweets, and to compare different settings to run an extensive analysis of what role plays the data enrichment. Our second study is devoted to addressing the problem of users' locations categorization, in order to identify what places are important (or not) in their daily routines, and compare different algorithms which make different use of contextual data as enrichment.

In the first study we focus on the text enrichment issues and, in particular, on the

external knowledge which is used in that process, in order to overcome the brevity and sparsity issues of short texts. Indeed, in the literature there is a considerable number of research works which confirm the usefulness of exploiting an additional source of information. Since short texts posted by users are often related to recent events (sharing their opinions and thoughts with friends), the novelty of our approach is to use news collections instead of generic web content in the categorization process.

We start investigating the effectiveness of the enrichment process, by evaluating different strategies that can be adopted in the choice of the final set of words to use in the categorization phase. More precisely, we compare, by running a set of experiments, three different enrichment strategies: *No-enrichment* (indeed, this is mainly used as a baseline strategy), *Append*, and *Merge* which differ in how to determine the final set of words combining original and enriched data. Then, we carry out an experiment to determine when our system can stop adding new words in the enrichment process, i.e., to determine the best *Cut-Off* strategy. Coming to the categorization step, we study how the accuracy of the whole process varies depending on the chosen set of categories from Wikipedia. Another important experiment is about the impact of different *Types* of documents (e.g., generic web content, news articles, blogs, etc.) on the effectiveness of the enrichment process. We run a set of experiments and analysis which aim at that goal: more precisely, we determine which kind of collection (among news articles, blog articles, general documents from all web, and a mixed sample with the combination of news and blog articles) mostly improves the effectiveness of the enrichment process.

On the basis of the above mentioned experiments, we proceed to study how the choice of the news collection affects the results: in particular, how different news collections with different properties impact the categorization effectiveness. More specifically, we analyze, by means of several experiments, other properties of news collections: *Volume*, to see how different numbers of news provide different sets of terms for the enrichment phase and, consequently, affect the categorizations; *Variety*, to see how news of different nature impact the enrichment process; *Structure*, to see if there is a structural component of the documents, e.g. title or content, that could contain more relevant terms for the enrichment purpose, and *Freshness*, to highlight the different effectiveness by using news from different time windows (i.e., same temporal context, 1 year old, 2 years old etc.).

Our second study is focused on user movements, and in particular on a novel proposal for users' locations categorization: we pay attention on how users move during their daily activities, in order to recognize the importance of places they visit according to different points of view, such as the frequency or intensity of visits. Indeed, we observed that some meaningful locations are related to users' main activities, thus they spent a lot of time in specific delimited geographic areas, such as their office or home. Other locations, instead, have been visited several times during the analyzed days, but with not the same intensity as home or office. An example of this kind of places may be the newsstand or the supermarket. In order to categorize user locations in Personal Points of Interest (PPOIs) or not, we must first be able to detect the so-called *Stay Points* (SPs), i.e., locations where the users "may stay for a while" (see [52]). Not all stay points can be considered important places, but they are good candidates and effective off-the-shelf tools are available to extract them from raw data (whatever the source, like, e.g., a GPS-device). The candidate stay points need then to be filtered to provide the final set of PPOIs. We remark here that our proposal is technology-independent, being based

only on raw data: neither we carry out any enrichments of positional data nor we use any external knowledge sources (like, e.g., georeferenced posts or resources published on Twitter, Facebook or other social networks). Urban computing [106] and trajectory data mining [111] are two research fields which can benefit from this kind of work.

In the literature, earlier approaches focus on the density of detected positions inside a delimited area, and on time thresholds to check when changing area, in order to recognize the locations which might have particular meaning for users. However, this is not enough to ensure a good selection, which should also take care to discard all "false important places" (e.g., crossing at intersections or stops at traffic lights) and, at the same time, should not miss relevant locations. Indeed, grid systems which exploit density, but are based on cells of fixed dimensions, cannot always guarantee a correct recognition due to the location distribution on the geographic space: the cell bounds might overlap an important place and, as a consequence, the latter will be divided and wrongly processed as two or more distinct places.

Further complexity comes into play since users movements are affected by other factors, such as speed/acceleration, heading, relations between locations, and also by the changes of the accuracy of GPS devices during subsequent detections. Many approaches considering the speed parameter tend to identify stay points when the measurement of speed is (nearly) zero. However, this assumption is again not enough accurate (it is sufficient to think, e.g., of a walk in a park). Therefore, to properly understand users behavior and habits it seems more appropriate to analyze their movements by considering a set of combined elements to infer the right information about the way they move.

On this basis the novelty of our approach aims at overcoming the above mentioned issues and at refining the whole identification process. First of all, our method is modular; we exploit some state-of-the-art algorithms to do an initial filtering of the raw positional data. Then, we carry out a deeper analysis, taking into account some user-related measures as further steps to refine the recognition task. Namely, we consider the area covered by a stay point, the time spent in a given location and the frequency of visits. This second phase improves the final outcome in terms of precision (paying a little cost in terms of recall). In particular, our approach allows us to infer a description of places in terms of a set of *features* more related to users routine activities. Mapping the physical locations into an abstract space based on those features helps us to carry on a deeper analysis which allows us to observe if a place is repeatedly visited. Moreover, we can identify locations (e.g., rendez-vous points, newsstands, bus stops to name a few) which are visited several times during a longer period, but not with a sufficient "intensity" to be found by previous techniques.

Finally, we want to remark how user generated content are important to study and model users, and also to analyze their behavior and habits during their daily routine. We chose to focus on short texts, posted on social networks, and movements data due to their strong relationship with users preferences and opinions, and it is clear how they are raw data often "not ready to use". In this thesis we want to highlight how an enrichment process exploiting context could help in mining and categorizing elements in those domains, and, with our studies, we aim at providing information about what parameters, settings, or kind of data are suggested to get the most benefit from this process.

1.4 Contributions

The main contributions of this dissertation are: (i) the study of the impact and effectiveness of data enrichment phase included in systems which aim to analyze user generated data, and (ii) the analysis of the properties which characterize the additional data provided by enrichment process. Moreover, our research contributes to design and develop novel approaches to build systems which include a data enrichment phase. We investigate on data enrichment in two different domains: texts categorization, and in particular focusing on short texts posted by users on Twitter; and user movements analysis, more specifically GPS trajectories generated by users which share the locations traces they visit during daily routines. On this basis the studies involved in our research are detailed as follows.

Studies on Short Texts

- We study the categorization of short texts, posted by users on social networks and microblogging platforms, in particular focusing on Twitter, and we propose a novel approach which exploits the data enrichment. Since short texts do not provide sufficient word occurrences, and they often contain abbreviations and acronyms, traditional classification methods such as Bag-of-Words have limitations. We propose a method to enrich the original short text with a new set of words extracted from news articles of the same temporal context. Then we use those words to query Wikipedia, as an external knowledge base, with the final goal to categorize the original text using a predefined set of Wikipedia categories.

- We study the effectiveness of the enrichment process in mining short texts by comparing different enrichment strategies: No-enrichment, Append, and Merge, which differ in how to use the additional data.

- We study different properties of news datasets to observe how they impact the short text categorization results. In particular we focus on five properties characterizing the datasets: Volume, Variety, Type, Structure and Freshness.

- We study a method for computing user similarity based on only content posted by users. We propose an approach based on a network representing the semantic relationships between the words occurring in the same short text, and the related topics, posted by users. Our approach is social network platforms independent, therefore not reliant on following/being followed social relationships nor on the peculiar structure of short texts (e.g., links, hashtags etc.).

Studies on User Movements

- We study a model to represent user habits and behaviors during their daily routines. We propose an approach based on a feature space which allows to model aspects/measures that are more semantically related to users and better suited to reason about their similarities and differences than simpler physical measures (e.g., latitude, longitude, and timestamp).

- We study a methodology to identify the stay points in a user movement trajectory which exploit temporal and spatial context to improve the analysis of regions where users have been stationary.
- We study the problem of places categorization focusing on the recognition of important locations, i.e., places where people spend a fair amount of time during their daily activities. We pay attention on how users move during their daily activities, in order to recognize the importance of places they visit according to different points of view, such as the frequency or intensity of visits.

1.5 Outline

The dissertation is divided in three parts:

- Part 1: *Background and Related Work*, in which we present the conceptual framework and the research fields related to our work. It is divided in three chapters:
  1. Chapter 2: Concept, in which we describe an abstract representation of user generated content in a geometric space to highlight contextual regions and relationships among content and users. We highlight how data produced by users can have properties which make them similar from point of views more related to user behaviors and habits. This concept has stimulated the idea of finding new useful information to exploit and enrich data.
  2. Chapter 3: Short Text Mining, in which we present research works in the field of text mining, and with particular attention on short texts. We highlight the role of data enrichment phase and the interests showed by research community, in particular in recent publications. We describe the importance of enrichment listing works where it has been used to solve tasks in particular settings or critical situations, or to improve the systems performance. We delineate the value of the new data extracted with this process and the contextual information.
  3. Chapter 4: User Movements Mining, in which we present research works in the field of movement data mining, in particular related to GPS data generated by users movements using mobile devices. In particular, we highlight the researchers' interest in this kind of data and specifically in elaborating them to get valuable knowledge. We describe considerable results related to the analysis of mobile spatio-temporal data, focusing on the study of social habits and behaviors. We provide a general perspective for studies on human mobility by depicting and comparing methods and algorithms, highlighting some critical issues with information extraction from spatio-temporal data.
- Part 2: *Case Study 1: Short Text Analysis*, in which we present the study we have conducted on short texts with related proposed approaches and experimental evaluations, and the user model. It is divided in two chapters:
  1. Chapter 5: Short Text Analysis, in which we first present the aim of our study on short text enrichment and categorization, highlighting how different enrichment strategies could be applied and different data sources could impact

the results. We then propose a novel approach to enrich the original short text with a new set of words extracted from news articles of the same temporal context. To understand the effectiveness of the enrichment process and to evaluate different strategies that can be adopted, we run a set of experiments to compare different enrichment settings and to analyze news datasets with different properties and observe how they impact the categorization results.

2. Chapter 6: User Modeling Exploiting Short Texts, in which we present a novel user model based on content posted on online social networks and exploiting a network structure which emphasizes the relationships among words used in users' texts. By using several network centralities we compute scores for each node, in order to have a vector representing the user and to enable a comparison exploiting "classical" similarity functions.

- Part 3: *Case Study 2: User Movements Analysis*, in which we present the study we have conducted on user movements with related proposed approaches and experimental evaluations. It is composed of one chapter:

  1. Chapter 7: User Movements Analysis, in which we present a problem statement regarding open challenges in important places recognition tasks. We then propose a method to identify a set of users' candidate stay points, and then to map them onto a feature space having as dimensions the area underlying the stay point, its intensity (e.g., the time spent in a location) and its frequency (e.g., the number of total visits).

Finally, in Chapter 8 we summarizes the work done with final remarks and discussion on results. We sketch out future works originated by the researches described in this thesis.

I

Background and Related Work

2

Concept

In this chapter we describe how user generated content (i.e., short texts and positional data) could be represented as abstract objects in a geometric space, and how they are related in terms of distances. We first show a general representation to highlight distances and contexts, then we present two adaptations based on the data types we study to illustrate the type of experiments we focus on.

2.1 General Idea

During daily activities a large amount of data is generated by users who use mobile devices to track locations and post texts on social networks. These data could share some properties, such as the timestamp or the place when they are created, or other things more related to the specific domain from which the data came from, for instance speed or acceleration, if we are considering positional data, or topics and sentiment polarity if we are analyzing texts. User generated content could be intended as abstract objects which we can be placed in a geometric space, defined by a set of dimensions representing the properties we are interested in, to analyze them.

In Figure 2.1 there is a representation of objects in a two-dimensional space generated by two different users. In particular, the objects have different distances based on their properties, defined by the two dimensions. Objects created by a user could be strongly related and close if they have very similar properties, and the same happens among objects from different users. This fact leads us to consider the region surrounding an object as the area suitable for extracting additional information in order to enrich the knowledge we have about the currently analyzed object (we call such extraction process *enrichment phase*). That region is what is commonly called *context*. The object's context contains other objects with different relationships depending on the distance and it could be analyzed with different wideness to fit the needs of the enrichment task to perform. Moreover, defining the space with different dimensions, which characterize the objects under other aspects, the context will change and allows us to get information from another set of objects.

We can summarize the key elements involved in the presented concept as follows:

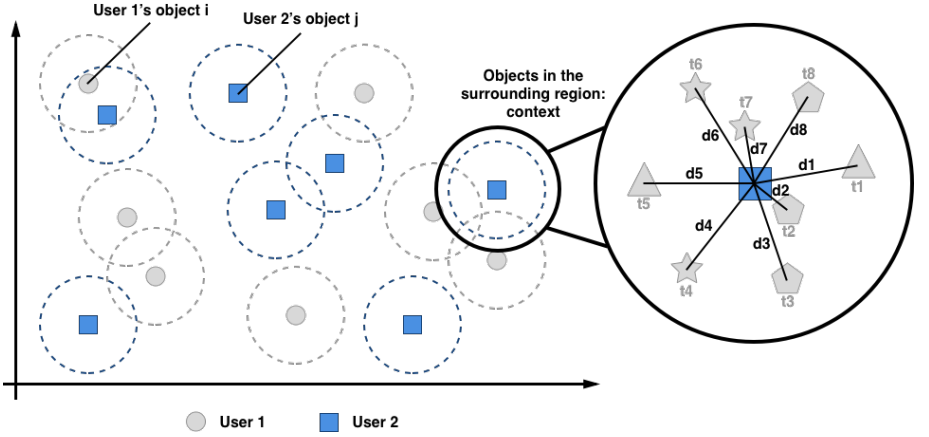

Figure 2.1: Geometric space with users objects and contextual enrichment.

- *User Object*: abstract representation of user activity which generated data useful to study his behavior. It consists in a set of raw data with properties and values, often "not ready to use".
- *Geometric Space*: abstract space where to place users objects in order to define an environment where they get a "location" which allow us to measure distances and regions.
- *Context*: it is a region surrounding a user object, suitable for extracting additional information related to it.
- *Enrichment*: it is an information extraction process which consists in analyzing data in an object context in order to add information to the currently considered object.

With the concept we described, composed of the elements we listed, it is possible to model the studies of enrichment effectiveness we focus on, analyzing the related objects in the context under several point of views:

- *Quantities*: we can use different quantities of additional data to get new information to exploit, therefore an interesting aspect to study is the presence (or not) of an ideal amount good enough to provide a considerable improvement, or if each enrichment task is a particular case with different needs.
- *Distances*: we can exploit the context with several sizes, in which data have different distances from the analyzed object. This fact begs the question whether there is an ideal size and a distance beyond which the enrichment phase could lose its effectiveness.
- *Properties*: we can use additional data with only some properties or specific types, in order to make the enrichment phase more specific for the domain analyzed. It is possible to study how the enrichment phase could be affected in different ways with data of different nature.

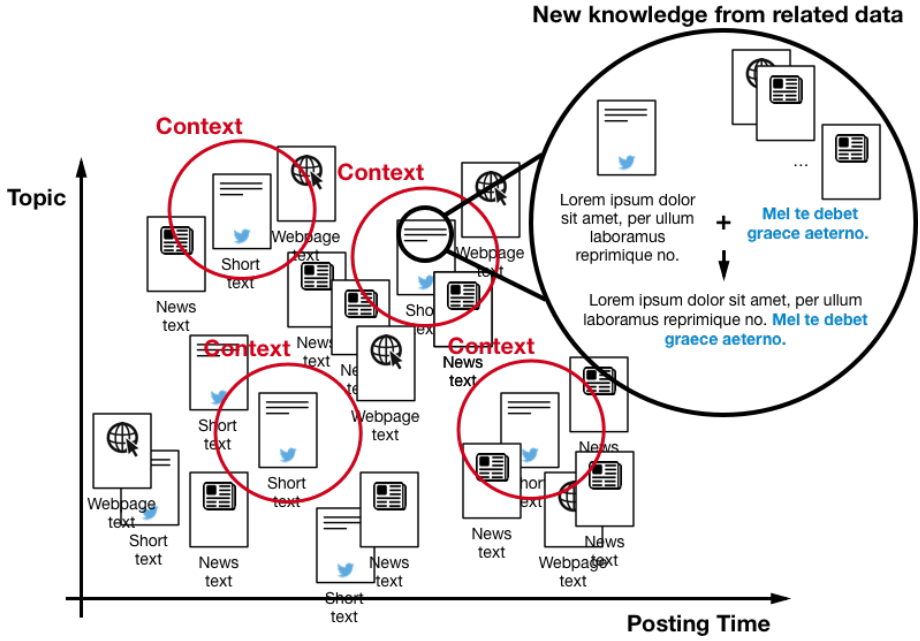

Figure 2.2: Short Texts with related documents as enrichment source.

Based on the concept we described, we design our studies so that we can conceptually place the objects from the two fields we analyze in a space which allows us to exploit the context as enrichment source. We use that information to improve the performance of the categorization tasks and measure how the enrichment phase impacts the results.

2.2 Applying the Concept to Real Word Scenarios

Regarding the short texts field, we want to analyze different documents with different properties and distances. In Figure 2.2 is illustrated an example of tweets which take place among other related documents in a space where they are distant according to some dimensions. Around a tweet we can find the contextual region which we use to extract related documents to exploit. The concept we described, applied to this field, allows us to enrich the original text with more words extracted from the more related and relevant documents, close to the short text. Defining a distance for context region selection, it is possible to retrieve relevant documents in order to extract additional text to exploit and enrich the tweet with more information. Figure 2.2 is just an example of what kind of data we can use as dimensions, also not necessary with continuos values. Posting Time could be a continuous timestamp used to place documents in a time flow, but Topic could be a dimension segmented in regions based on a previously selected set of topics to analyze, which do not have a unique order rule. Using different kind of data it is possible to have all or only some continuous values, also, increasing the number of dimensions it is possibile to analyze context under several points of view. Figure 2.2 is

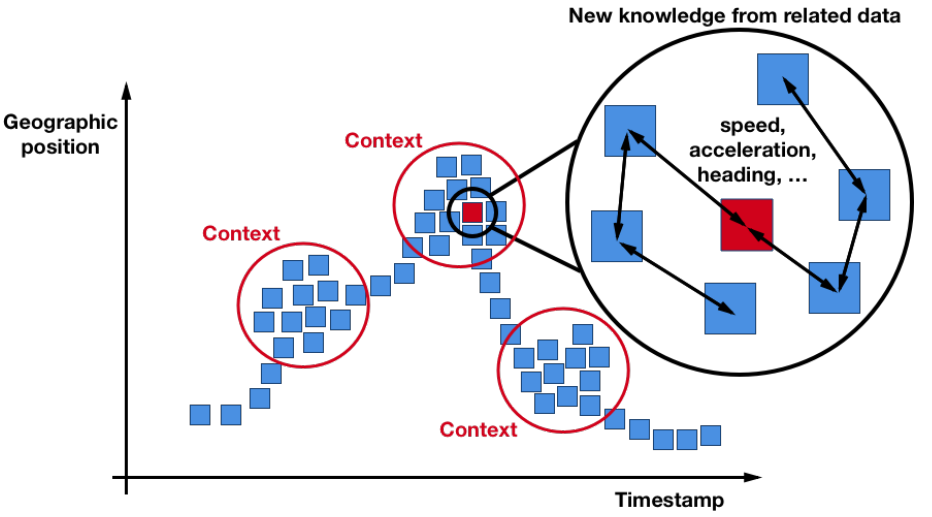

Figure 2.3: User Movements with visited locations and related places as enrichment source.

supposed to be just an abstract representation, not a real implementation.

Regarding the user movements field, we want to study how data of different nature, and from different sensors, could differently impact the place enrichment phase. Figure 2.3 shows an example of user movements as raw data placed in a space where the distances are defined by some dimensions, in this example timestamp and geographic position. Around each positional object it is possible to analyze the contextual region to understand how the user moved in that location. Exploiting the relationships among objects and additional data values we can enrich the currently analyzed location with data from different sensors. The contextual locations with their sensors values differences allow us to extract the new knowledge to use as enrichment. The image depicted in Figure 2.3 is just an abstract representation of objects geographic positions in one dimension, which are instead technically composed of two coordinates. The image simply provides an example of what it is possible to do with positional data. A model with more dimensions will be needed in order to described this kind of objects in a real implementation.

The concept we presented aimed to explain how real-world user actions can be translated in objects in a geometric space which emphasizes relationships through distances and measures. Moreover, it allows us to model two real world scenarios, in the two fields we want to analyze, to setup and run the experiments which constitute our studies. In particular in Part II of this thesis we conduct our studies on short texts, and in Part III we describe our studies on user movements. Both studies rely on the notions of context and enrichment explained with this concept.

3

Short Text Mining

In this chapter we provide a general state of the art of research work related to short texts mining and categorization. In the subsequent sections we will provide a survey of the main aspects which are considered in the literature, and which are somehow related to our work. We provide a general perspective on studies on short text with particular emphasis on short text categorization, enrichment, short text and external knowledge, short text and news, and short text for user modeling.

3.1 Short Text Categorization

Short texts generated by social networks are indeed an interesting source of information; for instance, many researchers focus on the aim of inferring users' interests from their posts on social platforms. To achieve such goal they often study and develop automatized software solutions being able to classify short texts, identifying the topics related to their content (e.g., politics, travels, literature, etc.).

Machine learning can play a fundamental role in classifying short texts: for instance, in [28] supervised SVM (Support Vector Machine) techniques are used to classify tweets into 12 predefined groups tailored for the online community of Sri Lanka. In [112] a completely automatized unsupervised bayesian model is used. In particular only tweets related to events are selected, exploiting a lexicon built from news articles published in the same period.

The authors of [102] observe that usually there are ordered subsets in short texts: such order relationship gives rise to the notion of *information path*. Hence, rather than classifying the whole dataset, it is easier to exploit the above mentioned information path and classify each subset separately, after the classification results about previous subsets are known. The approach proposed does not rely on any external knowledge data source, being robust on different datasets.

In [57] the classification of tweets is carried out by means of another social network, namely, YouTube. Indeed, classes assigned to videos published on the latter social medium are used as labels of tweets linking to such videos, providing a very huge set of labelled instances that can be used as training data. The experiments carried out

by the authors have shown the effectiveness of this classifier w.r.t. an analogous one trained with manually labelled data.

The proposed approach of [24] is to introduce the notion of *meta-hashtag* to overcome the bias introduced by considering freely user-defined hashtags as class labels of tweets. In particular, the authors use a *bag-of-words* model to represent tweets and the SVM method to cluster similar ones and determine the corresponding meta-hashtags.

Since the cited bag-of-words model is not entirely appropriate for short texts (due to their shortness and sparsity), in [38] the authors adopt dimensionality reduction techniques, to reduce accuracy and performance problems. Instead, the authors of [91] propose a novel *bag-of-concepts* model for the task of classifying short texts. Indeed, they leverage on a large taxonomy knowledge base to associate a concept model to each category and a set of relevant concepts to each short text. Then, the latter are classified with the most similar category. This approach benefits also from the fact that it facilitates ranking after classification, which is a good thing for many applications, e.g., recommendation systems.

Another solution to overcome the limits of the bag-of-words model in the case of tweets classification is proposed in [85]. The authors leverage on some domain-specific features inferred from the user's profile and texts to train their learning model. Such features include the author name and some of them are very specific to Twitter (e.g., "@username" at the beginning of the tweet or within the tweet). So doing, the authors show that they can effectively classify the text to a predefined set of classes, namely, News, Events, Opinions, Deals, and Private Messages.

In [23] a semi-supervised learning system for "micro-text" classification is introduced, with the interesting feature of combining machine learning with human interaction. The latter reveals to be very effective when decisions based on sparse data have to be made. Indeed, if the number of texts to classify is at a medium level, both machine learning and pure human analysis are clearly unfit. On the other end, according to the authors, their combination is quite effective in this scenario. A completely supervised approach is instead adopted in [80], where the authors exploit the hash-tags as indicators of topics, obtaining good results.

In [31], to avoid both the high costs and the difficulty of finding and retrieving a suitable external knowledge for enrichment-based classifiers (see Sections 3.2 and 3.3), the authors propose a novel classification technique. The latter exploits a structured sparse representation of short texts (as feature vectors in $\mathbb{R}^m$, using raw term frequency) and classify them applying a *convex hull vertices selection* process, i.e., solving an optimization problem in $\mathbb{R}^m$.

Several works use the clustering approach with the aim of measuring similarity between texts and grouping those that deal with the same topic. In [105] users' dynamic topic distributions are captured with the introduction of a dynamic user clustering topic model. The latter allows one to adaptively track changes of each user's time-varying topic distribution exploiting a previously estimated distribution and the short texts posted by the user during a given time period. In particular, word pairs from each user are sampled by a Gibbs sampling algorithm and the clustering results have the pleasant property of being explainable and human-understandable.

Other works exploit document expansion in order to overcome problems with lack of information in short texts. In [29], starting from the hypothesis that short docu-

ments tend to be about a single topic, authors submit documents as pseudo-queries and analyze the results to learn about the retrieved documents themselves. In [13] the authors propose an expansion technique, called TIDE, that can be applied on several Machine Learning, and Information Retrieval tasks on short texts (such as short text classification, clustering, entity disambiguation) without using task specific heuristics and domain-specific knowledge for expansion, extracting topic and using word embedding. In [82] authors expand short texts using word embedding for classification, and in particular they use a language model, which learns a new word embedding space, by using word correlations, such that semantically related words also have close feature vectors in the new space.

3.2 Enrichment-based Approaches

Text enrichment is a common technique used to semantically understand and classify short texts. The underlying idea is to infer additional information (by querying a web search engine or some other external information source) to overcome the possible sparsity and brevity of the original texts.

Tang et al. [86] propose a framework which performs multi-language knowledge integration for clustering. Sahami et al. [81] address the problem of measuring the similarity of short text snippets by leveraging on web search results, to infer a wider context for each short text (so doing, they can more easily solve ambiguity issues). In a recent paper Meng et al. [61] propose a method to expand short texts with the help of public search engines, by crawling related pages and getting contents as background knowledge of the original short text.

In [9] news articles are linked to tweets to enrich and to help inferring the semantic meaning of the latter. The final aim is to build user profiles representing individual Twitter activities.

In [59] the authors investigate a task of reputation dimension classification: one of the most interesting observations is about the contemporary nature of the web corpus used in the enrichment process of tweets. This temporal relationship between source documents and Twitter posts allows them to achieve very high scores in the experimental evaluation of the system.

A cluster-based representation enrichment method (CREST) is introduced in [25]: such system enriches short texts by incorporating a vector of topical relevances (besides the commonly adopted tf-idf representation). Finally, topics are extracted using a hierarchical clustering algorithm with purity control.

Enrichment techniques can also be quite sophisticated like, e.g., in [92] where a short texts are classified exploiting link analysis on topic-keyword graphs. In particular, after the initial topic modeling phase, each topic is associated to a set of related keywords. Afterwards, link analysis on a subsequent topic-keyword bipartite graph is carried out, to select the keywords most related to the analyzed short text.

In [101] the authors use semantic enrichment to understand tweets, but only after a *conceptualization* phase, where they assign related concepts to each term recognized by the Probase software. So doing, they are able to contextualize each term, eliminating typical issues due to polisemy.

3.3 Short Text and External Knowledge

Dealing with short texts unavoidably requires the use of some external sources of knowledge to overcome the lack of information. For instance, Wikipedia is often exploited either to query its huge archive of articles for enriched words or as a semantic platforms like, e.g., in the INEX Tweet Contextualization Track [1]. This approach turns out to be quite effective since articles are interlinked and tagged, according to a category graph. The latter can thus play the role of a semantic network which can be very useful to classify short texts. Indeed, in [55] the authors profile Twitter users and re-rank tweets in their timelines, measuring the relevance between each new tweet and user's interests. The above mentioned profile is built as a set of concepts taken from Wikipedia.

In [97] the authors classify generic online text documents, by adding a semantic context and structure, using Wikipedia as a knowledge source.

Banerjee et al. [14] propose a system for clustering similar items in the feed reader, to make the information more manageable for users, by enriching their representation with additional features from Wikipedia. Also Hu et al. [37] rely on Wikipedia as an external knowledge-base for document clustering, by mapping texts to Wikipedia concepts and categories.

In other proposals, Wikipedia is exploited to compute semantic relatedness between words or texts, like in [100], and more recently to identify the word sense with a disambiguation process, as described in [48]. Another recent use of Wikipedia knowledge is to enrich the semantic expression of a target commercial advertisement, as presented by Xu et al. in their work on contextual advertising [96].

DBPedia is instead used in [69] as a source for an enrichment procedure allowing the authors to rank in real time tweets which are conceptually related to a given subject, in order to improve the accuracy of information extraction.

WordNet and SentiWordNet are used in [65] to propose a novel unsupervised method for polarity classification in Twitter. More in details, they perform random walks over WordNet obtaining some PageRank scores which are subsequently used to weight synsets values.

The authors of [49] start from the observation that human beings, when interpreting short texts, resort not only to the content words, but also to their background knowledge like, e.g., semantically related words. Hence, they introduce a system based on the Dirichlet Multinomial Mixture model to improve topic modeling for short texts exploiting the background knowledge about word semantic relatedness acquired from a huge external corpus of documents.

3.4 Short Text and News

The key observation that many short texts posted in social networks are triggered by real world events has led some researchers to investigate the relationships and the interplay between short texts and news (since the latter are usually written to tell events).

For instance, in [33] a framework for linking tweets to news is provided together with a dataset of tweet-news pairs. The interesting consequence of this work is that the authors succeed in finding text-to-text correlations, exploiting hashtags (i.e., a tweet

specific feature), named entities (i.e., a news specific feature) and temporal constraints. Whence, they build a rather complete semantics of tweets.

In [9] the authors introduce several enrichment strategies (i.e., entity-based, topic-based, tweet-based and news-based) to relate tweets and news articles belonging to the same temporal context, in order to assign a semantic meaning to short messages.

Moreover, the ephemeral nature of Twitter posts begins to suggest to take into consideration the temporal dimension. For instance, Cataldi et al. [20] propose a technique to detect the most emergent topics expressed by the community on Twitter. They consider as emerging a term that frequently occurs in a specified time interval but it is rare in the past, and also take into account the source, by analyzing the author and his social relationships.

3.5 Short Text for User Modeling

In [87] the authors introduce TUMS (Twitter-based User Modeling Service), namely, a web application being able to build semantic profiles (in RDF format) starting from the messages a user posts on Twitter. TUMS seems very similar to our system since it also features topic detection and text-enrichment, allowing one to link tweets to news articles describing their context. The inferred profiles can be based on entities, topics or hashtags. TUMS uses the Friend-Of-A-Friend (FOAF) [17] vocabulary and the Weighted Interest vocabulary [16] for inferring user interests (while we use the category hierarchy of Wikipedia).

User similarity is also used in [103], exploiting both textual data (tweets, including URLs and hashtags) and social structure (following and retweeting relationships), in order to discover communities of users in Twitter.

When dealing with social networks, user profiling and modeling is often a research activity very tailored to the specific platform addressed (e.g., Twitter or Facebook) and, usually, the resulting profiles are not interchangeable nor interoperable. Instead, in [71] the authors propose a novel framework for automatically creating and aggregating several distinct user profiles by means of semantic technologies. Thus, they provide a tool to build larger and more general profiles starting from (possibly) unrelated and specific ones.

Due to the widespread of social networks and the huge amount of data they are generating, there is a growing interest to search for a suitable notion of *similarity* between users. Indeed, knowing if two given users are similar or not may help to improve ranking and recommendation systems in the task of filtering data.

For instance collaborative filtering methods are among the most used in suggesting recommendations based on similar users (where the notions of similarity are based upon "classical" measures such as cosine, Pearson correlation coefficient, and mean squared difference). In [53] a new user similarity model is presented, combining the local context information of user ratings with the global preference of user behavior, in order to help the system when only a few ratings are available (a sort of *cold start problem*).

In [51] the authors go further, developing a complex framework with a *non-linear multiple kernel learning algorithm*, in order to combine several notions of user similarities coming from different social network theories. Their experiments on a movie review data

set show that their system provides more accurate recommendations than trust-based and collaborative filtering approaches.

In [60] the author revisits the well known Page Rank algorithm and the related notion of random walks on a network, in order to improve ranking and recommendation systems based on the analysis of users' interactions carried out in the World Wide Web (e.g, records of friendship relations in social networks, e-commerce transactions, messages exchanged in online communities, etc.).

User similarity is also exploited in [46] for sharing training data, in order to build personalized classification models. So doing, people are freed from the daunting task of collecting and annotating data from their devices. Of course, only users with strong similarities are allowed to share training data, otherwise the classification model cannot be tailored to a specific user.

Another field which may benefit from user similarity models is related to social media applications, where users provide evaluations of one another. In such contexts it is well known that the comparative levels of status between two users influence the evaluations that one user gives to another. Interestingly, in [10] the authors show that, according to their experiments, evaluations are less *status-driven* when users are more similar to each other. In particular, they become very low when users are almost equal. Hence, similarity between users can be used to predict evaluation outcomes. A comprehensive survey of user modeling techniques in social media websites is available in [8].

Beside information filtering and recommendation systems, the "quest" for similarity models stimulated also the research branch studying networks built from unstructured data. In particular models and techniques borrowed from graph analytics (e.g., centrality analysis, path analysis, community detection and sub-graph isomorphism) have proven to be very effective tools in understanding and mining social networks [19].

An interesting social network analysis on a subset of tweets generated by a microblog group of 1082 users is carried out in [98], where the act of *following* has been characterized as the out-degree of nodes, while the act of *being followed* has been characterized as the in-degree of nodes. The resulting model and the simulation results carried out by the authors seem to cope well with real-world situations.

Recently Pavan et al. in [73, 74] addressed the problem of expert finding and the identification of similar professionals. They present a first attempt to create an expert search system to support users (such as researchers, students, authors) in finding experts to get in contact or to start a cooperation with in the field of textbook research. Hereby they semantically enrich user profiles building a Community Knowledge Graph (CKG) which defines relationships among users and related items.

4

User Movements Mining

In this chapter we provide a general state of the art of research work related to moving objects analysis, in particular users, by focusing on trajectories modeling, and then highlighting recent interests in more user-oriented methodologies, to analyze social habits and behaviors. We describe social approaches used to extract new semantic knowledge, combining those that are often separate literatures about the study of trajectories and the study of social habits and behaviors as further step on trajectory modeling. Moreover, we provide a general perspective on studies on human mobility by depicting and comparing methods and algorithms focusing on two significant aspects as place categorization and important place recognition.

4.1 Trajectory Modeling

From raw data, depending on needs and aims of the specific application, several reconstruction trajectories algorithms can be defined. They integrate different processing steps that allow to obtain trajectories on the different modeling levels. It is possible to apply data mining to get only conceptual trajectory, or to analyze data deeper to reach the semantic level.

4.1.1 Cleaning and Filtering Raw Data

Anyway, data sets collected by mobile sensors are often imprecise and incorrect due to noise. Raw data are exposed to two different kind of errors [41]: systematic errors derived to limitations of system positioning (a low number of satellites while detecting position, a low accuracy due to signal problems, etc.) that affect the final quality of data; random errors due to external reasons as clock and receiver issues, atmospheric and ionospheric effects, etc. Usually, different methods based on several parameters as time, speed, etc., or geometrically regression models are used to solve these problems. Yan et al. [99] describe a data preprocessing layer for cleaning data that applies velocity threshold to remove points that do not give us a reasonable correlation with expected velocity to solve systematic errors and a gaussian regression model is used to deal with random

ones. Marketos et al. [58] define a trajectory-reconstruction algorithm that, starting from raw data, uses maximum speed and tolerance distance between two timestamped positions to eliminate noise and redundant data.

Working in a network (e.g., road and rail networks), different map-matching algorithms can be used to replace or clean GPS positions of an object by a point on the network. These algorithms can be divided into geometric, topological, probabilistic and advanced [78, 89]. While geometric and topological algorithms, that use geometric and topological information, are simple, fast and easy to implement in real-time, probabilistic and advanced ones, that use probabilistic information and more refined concepts as mathematical theory of evidence, fuzzy logic models, etc., offer an higher accuracy but, they are generally slow and difficult.

Moving data grow progressively and intensively as the tracking time goes by and data compression is an essential task, that can be applied directly to raw data. Working with raw data, the compression consists of a reduction of the points used to describe a trajectory. Different algorithms, trying to balance the trade off between accuracy (and information loss) and storage size, consider different spatial and temporal parameters. Muckell et al [68], proposing a new approach to trajectory compression, called SQUISH, perform a comprehensive evaluation and a comparison of several of them: uniform sampling, Douglas-Peucker, opening window and dead reckoning.

4.1.2 Structuring Trajectories

Except for simply maps of movement, raw trajectories are insufficient and usually not useful for meaningful trajectory applications. For this reason, basic analysis can be performed to structure trajectories in episodes, sequences of GPS points with common properties.

This step, usually called segmentation, can be defined using different features associated to GPS points. A frequent example distinguishes between two states in a trajectory: stops and movements. It can be automatically obtained, as in [99], determining a speed threshold and analyzing the velocity associated to each position in a trajectory. In [50], Li et al. identify two categories of stay points: points where a user remains stationary for a time period exceeding a threshold and points where a user moves around within a certain spatial region for a period. Other authors consider different properties of movements. Buchin et al. [18] propose an algorithmic framework that segments any trajectory into a minimum number of segments under one or more criteria. They distinguish between basic attributes as location, heading, speed, and velocity, and other ones as curvature, sinuosity, and curviness.

The segmentation can also include the identification of different trajectories, subsets of GPS points, in continuous movement of an object. Different policies can be applied to divide consecutive trajectories as large temporal (and spatial) intervals or temporal periods (e.g., daily and weekly trajectories). This operation is applied from Marketos et al. in [58] where, in addition to attributes used to clean raw data, temporal and spatial gaps and a maximum noise duration are used to divide the movement of an object in different trajectories.

4.1.3 Semantic Enrichment of Trajectories

Semantic trajectories allow, through annotations, to enrich data with additional information depending on the specific aim of the application, and on the desired granularity level of information. To annotate each point is not usual because it can cause a big amount and redundant data. As described in [72], annotations are usually associated to episodes or to whole trajectories. Starting from contextual data repositories (e.g., OpenStreetMap and GoogleMaps), map-matching algorithms based on topological relationships allows to associate episodes of a trajectory with points (e.g., restaurants and shops), lines (e.g., walking streets and train rails) or regions (e.g., building and administrative areas) of interest [99]. Moreover, depending on similar associations and additional observations, activities or transportation modes allow to motivate and describe episodes. More general annotations can characterize the whole trajectories (e.g., work and touristic trajectories).

In many applications moving objects are restricted to move within a given network (e.g., vehicles on the road network). Particular kinds of annotations can be defined in a network. For instance, Richter et al. [79] define a semantic trajectory as a sequence of points localized in a transportation network annotated with specific events as origin, destination, intersections or stops. Several proposals combine map-marching algorithms on networks to data compression. In effect, focusing on semantic trajectories, acceptable information loss can be obtained achieving only interesting and significant points in a transportation network [44, 79].

In a more abstract description, aiming for the user level, the spatial details about movements from one place to another one and the specific geographic positions of those locations can be lost mapping trajectories in a graph structure keeping just the relations between nodes and attributes for edges. Abandoning the bond with a geographical map led us to focus on the elements that define user behaviors and her habits, in order to build a more generic model that allows to analyze users to find similarities, even if they live in different countries, but with same life style, i.e., same places and movement types. For instance, Zheng et al. [94, 110] build graphs among users' locations connecting nodes (i.e., clusters of positions with semantic annotations) with directed edges to study sequences of locations.

4.2 Towards Social Habits and Behaviors

The mobile devices and user social activities opens new challenges in trajectory mining and led researchers to work on new systems based on social approaches, with focus on users behaviors, in order to define that new layer of information. It is clear how this new source of information, resulting from this new user oriented approach, could be important to exploit, to improve current methodologies used in research to understand people and their behaviors. To design and implement an extraction process in an effective way, it is very important to get the right information from the collected raw location data. Also, a further refinement has considerable value to make deeper analysis, in order to infer additional knowledge about users.

Several researchers focus on recognizing patterns in mobile environments to analyze user communities. Karamshuk et al. [43] present a survey on existing approaches to

mobility modeling. Hui et al. [39] propose a system for the analysis of human mobility considering the community structure as a network, in order to emphasize the relationships and improve the understanding of behaviors. Laxmi et al. [47] presented a study that analyzes the behavior of user patterns related to existing works from the past few years. In this direction other authors present their work on analysis of user communities, in order to build human mobility models. Noulas et al. [70] analyze a large dataset from Foursquare to find spatio-temporal patterns and to observe how users make us of check-in feature provided by the social platform. Their results are useful for urban computing to study user mobility and urban spaces. Mohbey et al. [64] propose a system based on mobile access pattern generation which has the capability to generate strong patterns between four different parameters, namely, mobile user, location, time and mobile service. They focus on mobile services exploited by users and their approach shows to be very useful in the mobile service environment for predictions and recommendations. Zheng et al. [107, 109, 110] developed a brand new social network system, called GeoLife. It is based on user locations and trajectories, which aims to mine correlations between them.

The interest in these issues is strong, therefore some researchers also work on fundamental problems related to information extraction. A good starting point is to recognize important locations for the users, such places can tell a lot about their routine, namely, daily behavior and habits, thus, a sort of personal POI. This process aims to identify places which have particular meaning for users, such as home, work, or any place where they spend a considerable amount of time during the day or which they visit with regularity.

4.3 Human Mobility

Some authors focus on analyzing patterns in mobile environments. A study, presented by Laxmi et al. [47], analyzes the behavior of user patterns related to existing works from the past few years. Noulas et al. [70] analyze a large dataset from Foursquare in order to observe user check-in dynamics and find spatio-temporal patterns. Their results are useful to study user mobility and urban spaces. In this direction other authors present their work on analysis of user communities in order to build human mobility models. Karamshuk et al. [43] survey existing approaches to mobility modeling. Hui et al. [39] propose a system to improve the understanding of the structure of human mobility by analyzing the community structure as a network. Mohbey et al. [64] propose a system based on mobile access pattern generation which has the capability to generate strong patterns between four different parameters, namely, mobile user, location, time and mobile service. They focus on mobile services exploited by users and their approach shows to be very useful in the mobile service environment for predictions and recommendations. Zheng et al. [107, 109, 110] developed a brand new social network system based on user locations and trajectories, called GeoLife, which aims to mine correlations between them.

Other researchers focus on locations analysis for destination and/or prediction of places of interest (POIs); Avasthi et al. [12] propose a system for user behavior prediction based on clustering. They analyze the differentiated mobile behaviors among users and temporal periods simultaneously in order to make use of clusters and find

similarities. Zheng et al. [108] perform two types of travel recommendations by mining multiple users' GPS traces: top interesting locations and locations which match user's travel preferences. In [56] the authors combine hierarchical clustering techniques, to extract physical places from GPS trajectories, with Bayesian networks (working on temporal patterns) and custom POIs databases to infer the semantic meaning of places. Thus, they are able to discover in an effective way users PPOIs. Scellato et al. [83] developed a framework called NextPlace, a novel approach to location prediction based on time of the arrival and time that users spend in relevant places. Liu et al. [54] propose a novel POI recommendation model, exploiting the transition patterns of users' preference over location categories, in order to improve the accuracy of location recommendation. Another work in the direction of providing personalized (i.e., more accurate) POI recommendations is [21] where personalized Markov chains and region localization are used to take into account the temporal dimension and to improve the performance of the system. In [30] Gao et al. leverage on content information available in location-based social networks, relating it to user behaviour (in particular to check-in actions), to improve the performance of POI recommendation systems. In [34] authors study how personal context has a strong influence on mobility, in addition to personal preference and spatiotemporal factors such as time and distance. An individual's familiarity with an area is an interesting context because it can bias the influence of certain factors. For example, the mobility patterns of two persons who have similar preferences are different when their familiarity with the area is different, even in the same area. In [90] authors present a novel framework for estimating social point of interest (POI) boundaries utilizing spatio-textual information based on geo-tagged tweets. They fist define a social POI boundary as one small-scale cluster containing its POI center, geographically formed with a convex polygon. Then they find the radius of a circle such that a newly defined objective function is maximized, based on geo-tags in the POIs dataset. Others in [104] aim at associating tweets that are semantically related to real-world locations or points of interest (POIs). Tweets contain dynamic and real-time information while POIs contain relatively static information. The tweets associated with POIs provide complementary information for many applications like opinion mining and POI recommendation; the associated POIs can also be used as POI tags in Twitter.

4.4 Place Categorization and Importance Recognition

One of the most important issues underlying the systems that analyze user behaviors and habits is the recognition of users' important places. Several studies focus on this topic to propose new approaches on this recognition process, and thus provide novel algorithms to use on more complex systems. Passing from raw information about coordinates to semantically enhanced data, e.g., shop, work, bar, is an important aspect in the task of discovering important places.

Kang et al. [42] introduce a time-based clustering algorithm for extracting significant places from a trace of coordinates. They then evaluate it using real data from Place Lab [84]. Montoliu et al. [66, 67] propose a system based on two levels of clustering to obtain POIs: first, a time-based clustering technique which discovers stay points, then

a grid-based clustering on the stay points to obtain stay regions. Isaacman et al. [40] propose new techniques based on clustering and regression for analyzing anonymized cellular network data usage to identify generally important locations.

Hightower et al. [36] exploit WiFi and GSM radio fingerprints, collected by mobile devices, to automatically discover the places visited by people, associating semantics to coordinates, and detecting when people return to such locations. Their BeaconPrint algorithm, according to the authors, is also effective in discovering places visited infrequently or for short time. De Sabbata et al. [26, 27] provide an adaptation of the well-known PageRank algorithm, in order to estimate the importance of locations on the basis of their geographic features, focusing on aspects as contiguity, and the movements of users. In particular, in the calculus of the importance score for each location, the speed can be used to highlight either places where the user has stopped or places where there is a high traffic density. Thus, the notion of importance of a location can be customized by considering the current needs or situation.

Many of these approaches base their algorithms on the number of user detected positions within a geographic area, and in some works with attention to the elapsed time between a detected position and the next one. For instance, in [88] Umair et al. introduce an algorithm for discovering PPOIs, exploiting a notion of "stable and dense logical neighborhood" of a GPS point. The latter is automatically determined using a threshold based approach working on space, time and density of detections. To improve the recognition process, other factors and parameters are taken into consideration to enhance the algorithms. Li et al. [50] mine single user movements in order to identify stay points where users spend time; then, by analyzing space and time thresholds, they compute a similarity function between users based on important places that represent them. Xiao et al. [94] add semantics to users' locations exploiting an external knowledge based on a database POIs, in order to understand user's interests and compute a similarity function between two of them without overlaps in geographic spaces. Recently Bhattacharya et al. [15] extract significant places exploiting speed and the bearing change during user movement. More recently, Pavan et al. [76, 77] propose a novel approach based on a feature space for mapping stay points. They first identify locations where users remain stationary, with state-of-the-art algorithms, then they define a new space composed of features more related to users, by considering parameters which describe users' behaviors and habits. The feature space has, as dimensions, the area underlying the stay point, its intensity (the time spent in a location) and its frequency (the number of total visits). This approach allows to model aspects that are more semantically related to users and better suited to reason about their similarities and differences than, e.g., latitude, longitude, and timestamp.

Hang et al. [35] adopt a different perspective presenting Platys, an adaptive and semisupervised solution for place recognition based on user labeling. It makes minimal assumptions about common parameters, such as types and frequencies of sensor readings, which are usually tuned up manually in other systems. Platys lets users to label the place at any time, assuming that important locations are those visited sufficiently often.

The results of these recent works have built the foundation for the next step in that direction: understanding users' behaviors, in order to predict their future interests in terms of destinations they would like to visit in the next future.

II

Case Study 1: Short Text Analysis

5

Short Text Analysis

In this chapter we describe our studies on short texts analysis, related to our research works presented in [63, 75]. We introduce our novel approach to enrich short texts with a new set of words extracted from documents of the same temporal context. To understand the effectiveness of the enrichment process and to evaluate different strategies that can be adopted, we run a set of experiments. We compare different enrichment alternatives: *No-enrichment*, *Append*, and *Merge* which differ in how to use the additional data. We analyze news datasets with different properties to observe how they impact the categorization results. In particular we focus on five characteristics of the datasets: *Volume*, *Variety*, *Type*, *Structure* and *Freshness*. We obtain information about the ideal setup which could maximize the contribution of the text enrichment phase. We show the performance of the three enrichment strategies, highlighting the best one to use, and we demonstrate how all the properties have a significant impact on categorization accuracy.

5.1 Aims

5.1.1 Motivations and Research Questions

In the following we will use the expression "text enrichment" to denote the process of modifying a text, adding terms extracted from a dataset by submitting a query to a search engine. Obviously, this is a rather broad definition which leaves unspecified some aspects related to, e.g., the sources of terms, the methodology used to extract new terms, and the techniques used to determine the final set of terms. In particular, we want to study different enrichment strategies (i.e., how and how many new terms to extract and combine in the original text), in order to adopt the most effective one. Then, we plan to research the interplay between text enrichment and the peculiar choice of the set of categories. Finally, we aim at investigating the impact of the nature of the dataset on the whole categorization process, by analyzing a set of crucial properties.

More specifically, we focus on the following Research Questions:

Q1: What is the best methodology to properly enrich a short text?

- **Q1.1:** How the new extracted terms can be integrated with the original short text? Is it always useful to use the new terms to enrich the short text?
- **Q1.2:** How many terms it is reasonable to add to enrich the short text?
- **Q1.3:** Does the text enrichment always provide improvements also with different sets of categories?
- **Q1.4:** What kind of documents are the most relevant to be used as external dataset of new terms? What role do the News articles play?

Q2: How the nature of the enrichment dataset affect the results, and how the properties of news articles impact the performance?

- **Q2.1:** What is the ideal amount of documents to use to ensure a proper text enrichment?
- **Q2.2:** How important is the variety of the documents in the external dataset used for the text enrichment?
- **Q2.3:** What impact have the (terms present in) the different parts of documents (i.e., title and content)?
- **Q2.4:** How important is document freshness and how does it impact the enrichment performance?

5.1.2 Enrichment Strategies and News Collections Properties

To answer these questions, we defined a set of experiments to study different enrichment strategies and to analyze the collections properties. The first group of tests we designed has the objective of answering the first set of research questions from Q1.1 to Q1.4 and consequently the main question Q1. We use the Google search engine and the related archive of indexed web pages as dataset, which offers a wide variety and number of documents useful to have the heterogeneous environment for the analysis we plan to run. The enrichment strategies and processes we want to analyze are defined as follows:

- *Enrichment*: we want to see the impact of different methodologies to use to combine the new terms with the original analyzed short text. In particular if they are always useful and if it is reasonable to also discard some terms from the short text.
- *Cut-Off*: we want to study how different numbers of new terms affect the results, analyzing several cut-off approaches to choose the right point where to cut the list of new terms found querying the external knowledge.
- *Category Tree*: we want to study the interplay between text enrichment and the categorization process; in particular, we will experiment some scenarios with different sets of final categories, in order to see if the benefits of text enrichment are remarkable anyhow, or if there are some strategies where they are not tangible anymore.

- *Type*: external datasets often provide different kinds of documents, e.g., blogs, forums, social platforms' messages, news, etc. Therefore, we want to measure the different impact that each of them has in the whole process, when it is taken as the only source of documents.

The second group of tests aims to study the impact of a set of properties which characterize the documents used as enrichment source, therefore to answer the second set of research questions from Q2.1 to Q2.4 and consequently the main question Q2. We use three different document collections, which differ in number and kind of documents included, have different sizes, span from 2011 to 2013, and also have some temporal overlaps to make possible several comparisons. They allow us to analyze the following four key properties:

- *Volume*: we want to see the impact of news datasets with different number of elements. We sample a dataset extracting either a given number of news or a fraction of the dataset. With this test we aim to measure how the amount of news correlates to the final enrichment effectiveness, and if there are particular sizes where the improvement gets a considerable increase.

- *Variety*: differently from the type test, here we are focusing only on news as the possible source for datasets. However, even news alone may come from different kinds of documents, e.g., blogs, forums, online newspapers etc. Indeed, here we study how the different variety of news, i.e., the number of sources we consider, affects the text enrichment process.

- *Structure*: news structure is always characterized by different components, such as Title and Content. We study how the text inside different parts has different impact when used as enrichment.

- *Freshness*: short texts are often related to recent events; therefore, it is interesting to study how important it is to have the publishing time of the news close to the publishing time of the short text being enriched, and how the enrichment effectiveness changes using increasingly older news.

5.2 Short Text Enrichment and Categorization System

Our proposal consists in a two-phase approach which combines a text enrichment algorithm which exploits additional data extracted from news articles, and a text categorization algorithm, based on Wikipedia category tree, as external knowledge, to find the topic discussed in the analyzed short text. In the following sections the two phases are described in detail.

5.2.1 Text Enrichment

The enrichment process consists in querying an external source with the short text in order to get relevant documents to exploit to extract additional text. We can choose

to use both online and offline datasets, therefore we can query a search engine like Google to get documents from the web, or particular offline datasets, often provided by public research institutions, composed of documents with specific properties, such as only news articles, or blog posts, etc. In both cases we pay attention on the temporal context to have documents more relevant for the current analyzed text, in terms of temporal closeness. Such documents are then used to infer other terms to add to the original short text. Then, the enriched sentence will be used in the next phase, described in Section 5.2.2, to categorize the short text with the reasonable hope to obtain a more precise topic as result.

Often, the texts posted by users on social networks, and in particular on Twitter, are ephemeral and strongly connected with events and news very close to the posting time; therefore, a key property of our system is to query the search engine of the external source with a temporal parameter in order to select only document within a certain temporal range. On this basis, we chose to set this parameter to 1 week to get documents with that maximum temporal distance.

For our query q we define $D = \{d_1, d_2, ..., d_n\}$, the set of n retrieved documents[1], and $K = \{k_1, k_2, ..., k_m\}$, the set of all terms extracted from each $d_i \in D$ (by removing stopwords). We compute the *tf* weighting factor, as usual, for each term for each document, but we are interested in how frequent is a word inside the entire collection to understand if the contents are homogeneous in terms of semantics. With this approach we can identify if the original text has meaning, or if it is a set of "random" words, not related with each other or with events or news. To achieve that, we compute the average *tf* vector as follows:

$$tf_i = \frac{1}{n} \sum_{j=1}^{n} d_{ji}^{TF} \tag{5.1}$$

where d_{ji}^{TF} is the *tf* weighting factor for the term k_i in the document d_j.

We define the relevance score by also considering the document frequency as an indicator of homogeneity, as follows:

$$r_i = tf_i \cdot log(df_i), \forall i \in [1, m]. \tag{5.2}$$

The use of document frequency, in place of inverse document frequency, emphasizes terms that appear in many documents, therefore once again in favor of the homogeneity, that guarantees a meaningful text.

Finally, to refine the ranking function, we tune up the terms weight by considering the word frequency into the corpus of natural language.[2] We define the wf vector where $\forall wf_i$, with $i \in [1, |K|]$, wf_i is the frequency of terms k_i into the English language corpus.[3] Therefore, the ranking function is $r'_i = r_i - \alpha \cdot wf_i$ where $\alpha \in [0, 1]$ is a constant to tune the frequency (we use $\alpha = 0.2$, set empirically). Thus, we get the following ranking function that emphasizes terms if the collection is homogeneous and

[1]We selected the first 10 documents retrieved by the search engine, in order to have an adequate number of terms to analyze.

[2]Zipf's law states that given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table.

[3]Data extracted from `https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists#English`.

penalizes very frequent terms:

$$r'_i = tf_i \cdot log(df_i) - \alpha \cdot wf_i. \tag{5.3}$$

At the end we normalize all terms scores in order to fit the values in $[0, 1]$.

5.2.1.1 Terms combination methodologies

After getting the final ranked list of new words, we define several methodologies to combine the terms present in the original text with the new ones extracted, in order to have different approaches to compare. The aim is to study the impact of variants which differ in the number of provided words to understand if more data are always more information, or if there is significant amount. The methodologies are detailed as follows:

- *No-enrich:* this is the baseline approach which does not make use of any additional terms; it keeps the original set of words from the short text. We want to keep it as term of comparison to analyze the difference with the other approaches and see how the enrichment phase impacts the results. The cardinality of the final term set T_e with this approach will be $|T_e| = |tweet|$, where $|tweet|$ is the number of words in the original short text.
- *Append:* this approach is based on the addition of new terms to the initial set of words used in the text. The original words are all kept, therefore there is an extension of the original set with n new terms. The definition of n, namely how many terms to keep, is explained in details in Section 5.2.1.2. It is important to emphasize that the words from the ranked list could contain terms already present in the original text, therefore, in that case, the selection process scans the list until it finds n new terms. The cardinality of the final term set T_e with this approach will be $|T_e| = |tweet| + n$.
- *Merge:* this approach is based only on the final ranked list of words. The original words used in the text are not considered and we scan only the ranked list to keep the first n terms. In this set of n terms might be the presence or not of the words that were contained in the original text, and when we cut the list at a specific number of terms, some of the original words could be discarded, due to their low position in the ranked list. Therefore, there is also a cleaning phase embedded in this approach, that allow us to have a small set of words with only the most important included. The cardinality of the final term set T_e with this approach will be $|T_e| = n$.

5.2.1.2 Terms list cut-off methodologies

To choose a proper number of new terms to use in the text enrichment process, we define several cut-off methodologies to study how different sets of words provide different contribution to the enrichment phase. The different proposed cut-off approaches are described as follows:

- *Threshold:* this approach consists in selecting terms with a score greater than a chosen threshold. Each term in the ranked list has a score $r' \in [0, 1]$, as described

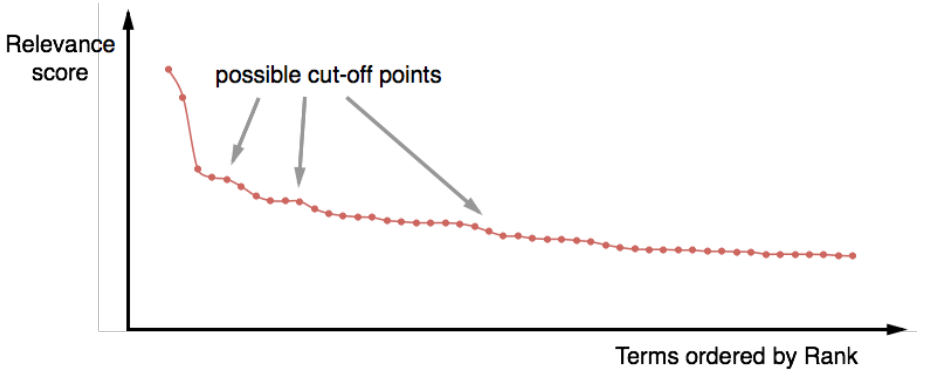

Figure 5.1: Terms relevance scores distribution.

in Section 5.2.1, and we set empirically a threshold equal to 0.75 to limit the terms selection.

- *Dynamic:* this approach takes in care the distribution of relevance scores that the words can get to find a reasonable point where to cut the list based on the decay of that score. Figure 5.1 shows an example of a typical terms relevance distribution with a considerable decay after few words, then a slow descent. It is not easy to identify a proper cut-off point due to a slight difference among scores, therefore we emphasize the score decay trend using the Relevance Gap distribution, as depicted in Figure 5.2. The blue line in chart represent the difference between the scores of each words couple, highlighting, with local maximum points, relevant positions to consider as cut-off points. We chose to cut at the first one in order to keep only the words with highest score and not to introduce too much noise in the enrichment set. Details about this process are described in Algorithm 1.

- *Fixed:* we define a threshold equal to 5 to extract a fixed amount of new words. To choose that threshold we run a preliminary test consisting in a sample of 1000 randomly selected short text from Twitter enriched with the previous described approaches. Table 5.1 shows the mean and the median of the number of words extracted. This approach has been developed to test the enrichment effectiveness with no sophisticated cut-off techniques, so that we can see and measure the differences with other cut-off approaches.

Table 5.1: Analysis on the number of new words selected for the enrichment.

Cut-off	Mean # words	Median # words
Threshold	5.39	5.00
Dynamic	4.87	4.00

All phases of the Text Enrichment process described in this section are summarized in a workflow depicted in Figure 5.3.

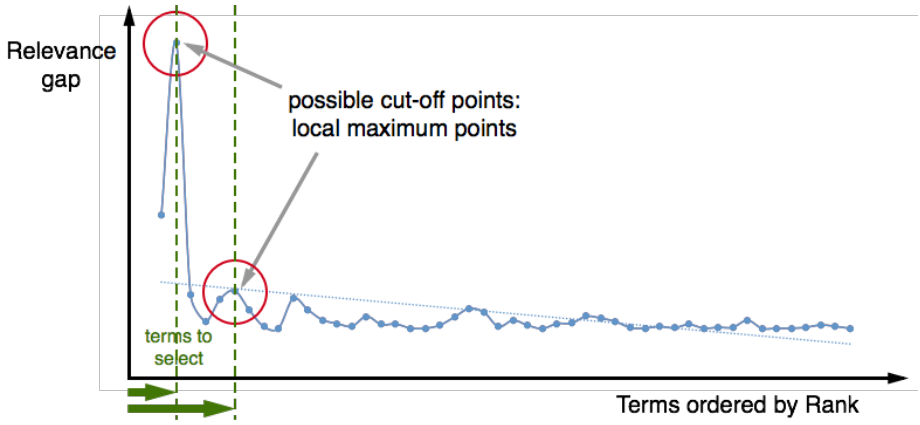

Figure 5.2: Dynamic cut-off based on the local maximum points on the Relevance Gap distribution.

5.2.2 Text Categorization

The Text Categorization step exploits the relationships within the Category Tree of Wikipedia for extracting the appropriate set of categories for each given short text. We preferred to work with Wikipedia since it is continuously updated with articles about news and popular events (being those the main topics of new tweets). However, the software architecture of our system is source-independent and we could easily switch to other databases (in the same way that we could change the search engine). Moreover, another reason of our choice is to show how it is possible to exploit Wikipedia in a different way, w.r.t. related works, by using techniques based on the Category Tree.

5.2.2.1 Articles selection

First, we queries Wikipedia APIs with each pair of words (bi-gram) from the short text. We use bi-grams and not single terms to get a set of articles more homogeneous, and to avoid too much generic articles, due to the words polysemy. With single-term queries it is difficult to focus on one or few topics, and we lose the semantic relations defined by the user who posted the text. On the other hand, with more than two words, i.e., n-grams with $n > 2$, it is likely to merge too many words that maybe are not all closely related to an article.

We define Q as the set of queries to perform, with $|Q| = \binom{|A|}{2}$, where A is the set of words extracted from the short text. $\forall a \in A, w(a)$ is the weight of the word in the original text. In this case we always set that weight to 1, for this first version where we do not compute the relevance score of each word. $\forall q \in Q$ we have a query weight defined as follows:

$$w(q) = \sum_{a \in q} w(a). \tag{5.4}$$

Hence, in this particular case $w = |q|$. By performing this set of queries to Wikipedia,

Algorithm 1 Dynamic cut-off.

Input: A set of words from external source $T_e = \{t_1, t_2, ..., t_n\}$
Output: An integer number of words to select *wordCounter*

```
RG = {}                                   ▶ list of Relevance Gaps between each couple of words
for t_i in T_e do
    g = |t_i.score − t_{i+1}.score|
    RG.insert(g)
end for
avgGap = avg(RG, 50)                      ▶ average gap among the first 50 words
currJumps ← 0
wordCounter ← 1                           ▶ we get at least 1 word
jumpThreshold ← 1                         ▶ we cut at the first considerable score dacay
while currJumps < jumpThreshold & wordCounter ≤ 10 do   ▶ we keep at most 10 words
    if |diffGap(g_i, g_{i+1})| > avgGap & |diffGap(g_{i+1}, g_{i+2})| > avgGap &
    diffGap(g_i, g_{i+1}) > 0 > diffGap(g_{i+1}, g_{i+2}) then
        currJumps = currJumps + 1
    end if
    wordCounter = wordCounter + 1
end while
return wordCounter
```

we obtain a set of articles, ranked by the relevance computed by the Wikipedia search engine.[4] For all $q \in Q$, there exists a (possibly empty) set R_q of relevant articles for q. We define $i_q(x) \in [0, |R_q| - 1]$ as the index of each article $x \in R_q$, and then we define the article weight as follows:

$$w_q(x) = \begin{cases} \frac{|R_q| - i_q(x)}{|R_q|} & x \in R_q \\ 0 & \text{otherwise.} \end{cases} \tag{5.5}$$

We combine all resulting articles in order to obtain a final set X with distinct entries as follows:

$$w(x) = \sum_{q \in Q} w(q) \cdot w_q(x). \tag{5.6}$$

Therefore, for a query $q \in Q$ and an article $x \in X$, $w_q(x) = 0 \Leftrightarrow x \notin R_q$, hence the query q does not change the final score of x. Also, the higher the number of queries with $x \in R_q$, the higher the weight $w(x)$ will be.

5.2.2.2 Category selection

As second phase, in order to have a set of possible topics to associate with the short text, we extract a set of Wikipedia categories. The selection is based on the categories associated to each Wikipedia article selected in the previous step (the categories related to an article are listed at the bottom of every Wkipedia page, as described by

[4] `http://en.wikipedia.org/wiki/Help:Searching#Search_engine_properties`

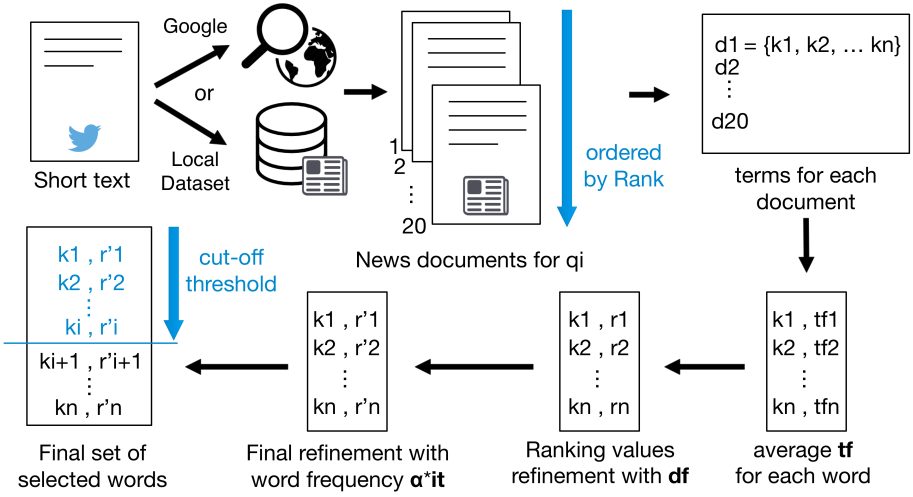

Figure 5.3: Workflow of the enrichment process.

Wikipedia guidelines), but we also exploit the Wikipedia category graph. The Wikipedia categories graph is organized so that each category is connected with each of its subcategories; therefore, the distances between nodes also represent the semantic relation values. More precisely, we selected 3 subsets of macro-categories in English language to properly classify the tweets extracted for the experimental evaluation with different levels of granularity. We first selected a set of 20 categories based on the popularity and the extensive use made in software platforms that offer applications, web pages, or multimedia contents to users, in order to have a set that can represent what users are interested in, and therefore most likely discuss on the web and especially on social networks. Then, we aggregated them, based on their semantic affinities, to obtain a smaller set composed of 12 categories, and again to get set with 5 categories. This choice is driven by the fact that we want to study the enrichment effectiveness on different set of categories. The 3 category sets are listed in Table 5.2.

More formally, the categorization process is as follows. Let $G = (C, E)$ the categories graph, where $C = \{c_1, c_2, ..., c_n\}$ is the set of all categories, and E the set of directed edges. We say that there exists $e_{c_i,c_j} \in E \Leftrightarrow c_j$ *isSubcategoryOf* c_i. Let $L \subset C$ be one of the set of macro-categories selected for text categorization, listed previously in Table 5.2. Let $x \in X$ be an article extracted during the previous phase, we define $C_x \subset C$, the set of categories directly related to the article, as our starting set. Then, for each $c_i \in C_x$, we define $C_{c_i} \subset C$, the set of categories reachable with a path from c_i. We are interested in just few of those, specifically if they are in our selected set (namely, L), therefore $L_{c_i} = C_{c_i} \cap L$.[5] At this point we have restricted L to L_{c_i}, and we denote by l_i the labels extracted form L_{c_i} as follows:

$$l_i = l \in L_{c_i} : sp(l, c_i) = \min_{l \in L_{c_i}} sp(l, c_i), \tag{5.7}$$

[5]This set can also be empty. In that case the category c_i does not affect the labels detection.

	Category Set 1		Category Set 2		Category Set 3
1.	Science	1.	Science	1.	Technical/scientific disciplines
2.	Computer Science				
3.	Economics	2.	Economics & finance		
4.	Finance				
5.	Medicine	3.	Medicine		
6.	Meteorology	4.	Meteorology		
7.	Politics (Politics, Law)	5.	Politics	2.	Humanities disciplines
8.	Literature	6.	Literature & philosophy		
9.	History				
10.	Philosophy				
11.	Entertainment (Hobby, Entertainment)	7.	Entertainment	3.	Fashion art & entertainment
12.	Sports	8.	Sports & Engines		
13.	Engines (Automobiles, Auto racing, Motorcycle sport)				
14.	Fashion	9.	Fashion		
15.	Photo and Video (Photography, Film)	10.	Multimedia	4.	Multimedia
16.	Music				
17.	Videogames				
18.	Places & free time (Tourism, Geography, Travel)	11.	Places & free time	5.	Health & free time
19.	Food and Drink	12.	Health		
20.	Health and Fitness (Health, Physical fitness)				

Table 5.2: Wikipedia categories used in our systems, in English language. The notation X=(Y,Z,...) denotes the category names we made to group categories about related topics.

where $sp(l, c_i)$ is the shortest path from l to c_i. The shortest path may not be unique, so there may be more than one l that satisfies the condition. In that case we keep all the retrieved categories. Let L_x the set of $l_i \in L$ selected with this approach, we define the category relevance value as follows:

$$r(l) = n(l) \cdot \frac{1}{\overline{sp(l)}}, \tag{5.8}$$

where $\overline{sp(l)}$ is the mean length of all shortest paths from l to the associated categories, and $n(l)$ the number of these categories. By selecting the category with the max $r(l)$ we get the most relevant one for that article,[6] as follows:

$$l_x = l \in L_x : r(l) = \max_{l \in L_x} r(l). \tag{5.9}$$

By repeating this process for each extracted article we obtain the set L_X of all the categories which potentially represent the topic discussed in the short text. We define a new ranking function for categories to select the most relevant as follows:

$$\forall l \in L_X, r'(l) = \sum_{x \in X_l} w(x), \tag{5.10}$$

where $X_l \subset X$ is the set of articles with the specific category l, and $w(x)$ the weight of article $x \in X$. With this final ranked list, by selecting the first category, with the highest relevance score, we obtain the topic that is the best match for the analyzed short

[6]The category with max value may not be unique. In that case we keep all the categories with max value.

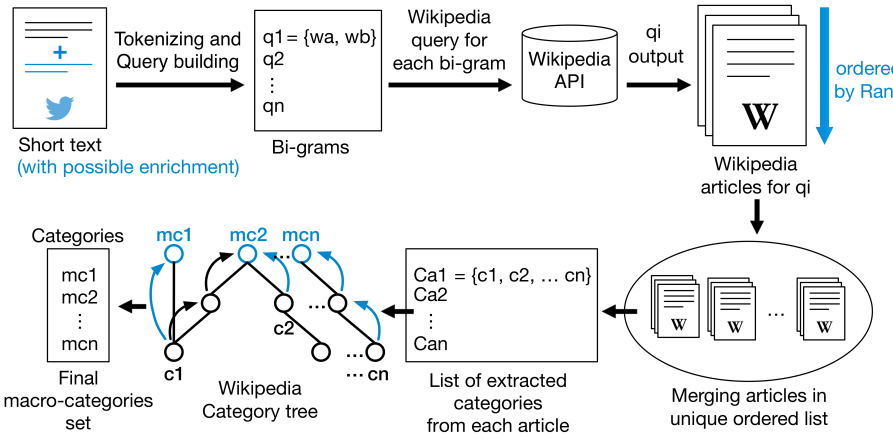

Figure 5.4: Workflow of the categorization process.

text. However, we prefer to keep a set of 3 categories (at most), with related relevance scores, in order to analyze eventual subtopics discussed in the text. Figure 5.4 shows an overall representation of the categorization process.

5.3 Extensive Study on Short Text Enrichment

5.3.1 Collections, Datasets, and Methods

To evaluate the impact of different enrichment strategies and the news properties on the categorization process we selected a set of three document collections, as enrichment source, and we defined two tweets sources to compose proper short text datasets for each experiment. The three collections have been selected from well known institution and research groups in order to have document sources structured, well known, and certified. The two tweets sources have been designed to provide two natures of short texts which differ in comprehensibility of text, ease of reading, presence of acronyms and/or neologisms and number of words, in order to have a first data source with tweets easier to categorize, and a second one with more problematic short texts. We used a Python wrapper [5] around the official Twitter API [7] to retrieve tweets, in order to compose datasets of 1000 randomly selected short texts to enrich and categorize for each test. In the following paragraphs we describe in details the three collections, then we define two tweets sources used to compose the datasets needed for each experiment, and finally we explain the metodologies we adopt to benchmark the experiments.

Collections: We use three different document collections in order to have several offline sets of articles extracted from news, blogs, forums, etc. to use as enrichment source. They differ in number and kind of documents included, and also in what properties characterize them. The first collection is NTCIR Temporal Information Access 2012 (Temporalia), which uses a web corpus, called "LivingKnowledge news and blogs annotated subcollection", constructed by the LivingKnowledge project and distributed

by Internet Memory. The collection is ~20GB uncompressed and over 5GB zipped. It spans from January 2011 to December 2013 and contains around 2M documents collected from about 1500 different blogs and news sources. The data is split into 970 files, named after the date of that day and some information about its sources (there might be more than one file per day). The second used collection is Knowledge Base Acceleration 2012 (KBA), which is part of TREC tasks and challenges. We used the third stream corpora, provided by the KBA institution, composed of ~930GB of text from news, blogs, forums, and social networks; around 20M documents. It spans from October 2011 to May 2013. The third collection is SignalMedia dataset for NewsIR'16 workshop (SignalMedia), which were originally collected by Moreover Technologies (one of Signal's content providers) from a variety of news sources for a period of 1 month (1-30 September 2015). It contains 1 million articles that are mainly English, but they also include non-English and multi-lingual articles. Sources of these articles include major ones, such as Reuters, in addition to local news sources and blogs. Table 5.3 shows all details about the three collections, illustrating name, size in terms of number of documents and bytes in memory, kind of documents, and the timespan the articles cover.

Table 5.3: The three news collections used in the experiments

Acronym	Name	# of docs/ size	kind of docs	Timespan
Temporalia	NTCIR Temporal[a] Information Access 2012	~2M / ~20GB	blogs, news	Jan2011 – Dec2013
KBA	Knowledge Base[b] Acceleration 2012	~20M / ~930GB[c]	blogs, news, forums, social	Oct2011 – May2013
SignalMedia	SignalMedia dataset for[d] NewsIR'16 workshop	~1M / ~1GB	blogs, news	Sep2015

[a] http://ntcirtemporalia.github.io/NTCIR-12/collection.html

[b] http://trec-kba.org/

[c] Data extracted from the 3rd stream corpora http://s3.amazonaws.com/aws-publicdatasets/trec/kba/index.html

[d] http://research.signalmedia.co/newsir16/signal-dataset.html

Figure 5.5 shows a representation of the three collections distributed over time and tweets as short texts to analyze. The different experiments related to research question Q2 are depicted with colored arrows in order to show how tweets and documents from the same temporal context are used. The Volume test, highlighted in orange, aims to compare the categorization results with samples of news from the same collection but with different sizes; the Variety test, in green, compares results among news samples with same cardinality but with different kinds of news; the Structure test, in red, aims to analyze samples composed of different news components; and the Freshness test, in purple, exploits news from the same collection but in different years. The figure shows only some examples; the details of all the experiments are described in the next section. On the other hand, for the group of experiments related to research question Q1 we use Google Search Engine to retrieve web pages as documento source for the defined enrichment tasks.

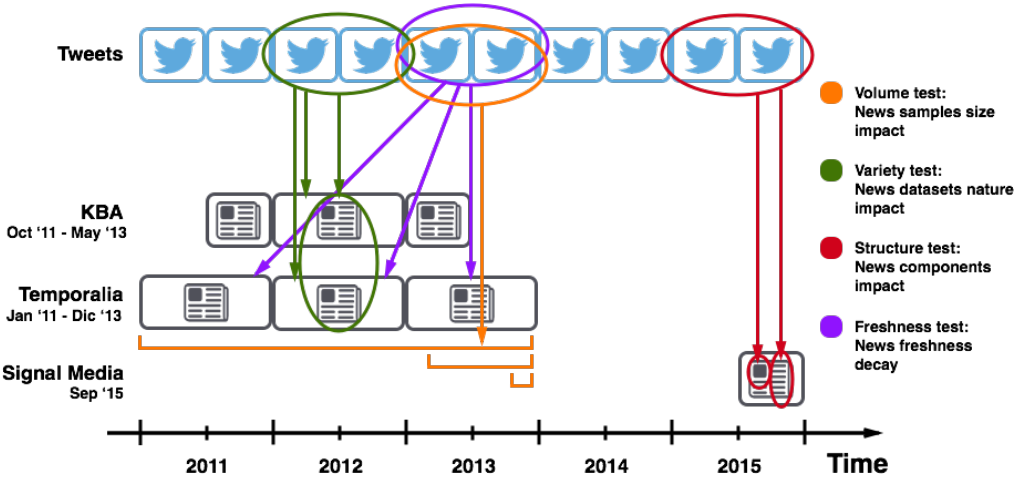

Figure 5.5: News collections distribution with properties based tests

Data Source 1: We selected a popular Twitter account for each of the 20 categories described in Section 5.2 in order to have short texts which span on different topics and cover the most common and discussed arguments. We are aware that not all the tweets of an account are related to the same topic, indeed for instance a politician could talk about only politics, instead a newspaper could argue on different topics. However, we used this approach to be able to provide a dataset of short texts as heterogeneous as possible. Table 5.4 shows all the account names with related categories. The link between the two has been defined by analyzing what makes famous the chosen account, for instance Spotify is a well known music platform therefore the best match is with the Music category.

Using the Data Source 1 we extracted a set of 50 random tweets from each account in order to compose a dataset with 1000 short texts. This dataset is suitable for benchmarking the different enrichment strategies in a heterogeneous environment with not only particular or critical texts, but with also "well formed" expressions.

Data Source 2: We selected a set of 7 popular Twitter accounts of very active users, famous in different fields, which often post tweets which contain problematic texts, not well formed, therefore not easy to categorize. We chose this kind of critical short texts because they are very representative of what is the common style of writing in social networks used by ordinary people who also discuss personal events or publish opinions on recent facts, which therefore differs from an editorial style, used by newspapers for instance. On this basis we selected David Cameron for Politics, Harry Kane for Sport, Bill Gates for Science, Neil Patrick Harris for Entertainment, Rihanna for Music, Jamie Oliver for Food & Drinks, and Donald Trump to have a very active user with provoking and/or aggressive tweets (due to his impact in recent events). Table 5.5 lists the selected accounts with related categories.

We extracted several sets of tweets from each account in specific time windows in

Table 5.4: Data Source 1 - Twitter accounts and related categories.

Twitter account	Category: [Wikipedia categories]
wired	Science: [Science]
linuxfoundation	Computer science: [Computer_science]
TheEconomist	Economics: [Economics]
business	Finance: [Finance]
RedCross	Medicine: [Medicine]
weatherchannel	Meteorology: [Meteorology]
BBCPolitics	Politics: [Politics & Law]
OliverBooks	Literature: [Literature]
britishmuseum	History: [History]
OUPPhilosophy	Philosophy: [Philosophy]
IMDb	Entertainment: [Hobbies & Entertainment]
sportingnews	Sports: [Sports]
AutoMotoWorld	Engines: [Automobiles & Auto_racing & Motorcycle_sport]
Fashionista_com	Fashion: [Fashion]
20thcenturyfox	Photo & Video: [Photography & Film]
Spotify	Music: [Music]
E3	Video games: [Video_games]
NatGeoTravel	Places: [Tourism & Geography & Travel]
jamieoliver	Food & drink: [Food_and_drink]
FitnessMagazine	Health & fitness: [Health & Physical_exercise]

order to build datasets composed of 1000 random short texts, according to the tests we planned to run and described in the next sections. Using Data Source 2 we are able to build datasets more difficult to categorize therefore more suitable to benchmark how different collections properties affect the enrichment phase and consequently impact the categorization results.

Table 5.5: Data Source 2 - Twitter accounts and related categories.

Twitter account	Category: [Wikipedia categories]
David_Cameron	Politics: [Politics & Law]
HKane	Sports: [Sports]
BillGates	Science: [Science]
ActuallyNPH	Entertainment: [Hobbies & Entertainment]
rihanna	Music: [Music]
jamieoliver	Food & Drink: [Food & Drink]
realDonaldTrump	Politics: [Politics]

Methods: In order to measure the effectiveness of the enrichment strategies and the impact of the collections properties, we carried out several expert evaluations over sub-samples extracted from the previously defined short texts datasets.

We used our proposed system described in Section 5.2 for the categorization of short

texts which provides, as final outcome, a list of categories extracted from Wikipedia category tree. The system, which analyzes texts, enriching and categorizing them, is suitable for benchmarking the enrichment phase relying on the evaluation of the text categorization. The choice of using only our system is driven by the fact that other systems in the literature are less flexible and suitable to the experiments at the basis of our studies. Techniques based on machine learning approaches need a training set, time and resources for the training process, and also they result to be dependent on the domain of the dataset in which they made training. Moreover, they are not comparable with our system. Our approach does not rely on pre-processed structures, short texts sources, or elements like hyperlinks, hashtags, etc., therefore it is more flexible, it does not suffer from the cold start of the problem, and it is more oriented to be used on real world applications.

The texts have been submitted to our proposed system with different settings and news collections according to the tests described in the following Section 5.3.2 and Section 5.3.3. For each test, in order to assess the impact of the enrichment process, the set of categories yielded by the system has been evaluated by expert users. The latter assigned a relevance judgment indicating how the categories properly represent the topics discussed in the tweet as follows:

- *0 - Not relevant*: The category has no relation with the topics discussed in the short text.
- *1 - Relevant*: The category either is related to the main topic discussed in the short text, or it represents a secondary topic.
- *2 - Highly relevant*: The category properly represents the main topic discussed in the short text.

The expert users during the evaluation assigned a relevance score $r \in \{0, 1, 2\}$ to each category provided by the system for each short text. This approach is similar to the experimental evaluation carried out in [80]. In general we repeat the evaluation over 3 different samples composed of 100 tweets each. In particular for the tests which need multiple sampling, we run the evaluation several times, with news samples randomly rebuilt each time, where we used only a portion of the entire collection. We kept the average ratings obtained with different sub-collections, avoiding bias due to the random set of news. Specifically we run the evaluation 1, 3, or 5 times depending on the sample size, approximating the average ratings to the nearest integer value. The nature of short texts posted online led us to often have short texts with more than one category as representation of the topics discussed. On this basis, the expert users specify what category is the best representation of the main topic discussed in each short text, a possible second category as secondary topic, and a possible third one, in order to provide a "Ground Truth" for the evaluation phase. The category set used for the following experiments is the one composed of 20 categories, defined by us and illustrated in Table 5.2. To evaluate our results, we used the relevance judgments to compare the provided categories with the ones indicated as Ground Truth. We then calculated some standard IR metrics, such as Precision (P) and Normalized Discounted Cumulative Gain (NDCG). These two specific metrics allow us to compare the different settings in the experiments under the two points of views we are interested in. We want to know if

the first category provided is the proper main topic discussed in the short text, thus we use Precision to measure that aspect. Secondly, we want a global score which considers also the possible second and third categories provided as secondary topics. The nature of Twitter texts makes them most likely dense of argumentations, often with more than one topic, therefore NDCG is fit for the purpose. We then defined the benchmarks as follows in the next sections.

5.3.2 Experiments on Enrichment Strategies

In this section we describe all tests which aim to analyze the different enrichment strategies. In particular in the following sections we define the experiments setups and we illustrate the results related to the first research question Q1 introduced in Section 5.1.1.

5.3.2.1 Enrichment Experiment

Setup We design this test to compare the effectiveness of the three different system strategies for enrichment terms combination, *No-enrich*, *Append*, and *Merge*, as described in Section 5.2.1.1. We define "Experiment 1a" and "Experiment 1b" to run two different analyses on different enrichment sources. With Experiment 1a we focus on news articles provided by the offline datasets introduced in Section 5.1.2, with Experiment 1b we run the analysis using online news extracted with Google search engine as documents source. The 2 experiments are defined as follows:

Exp 1a: Tweets posted in whole range 2012-2013, offline categorized with Temporalia, KBA, and SignalMedia complete datasets, used together as a single source to exploit the strength of all documents included, with *No-enrich*, *Append*, and *Merge* approaches.

Exp 1b: Tweets posted in whole range 2013-2016, online categorized with Google search engine, with *No-enrich*, *Append*, and *Merge* approaches.

Results Table 5.6 shows the average NDCG and P@1 scores obtained by the three compared system strategies with offline and online news. It is clear how there is a considerable improvement when the system exploits the enrichment source; the help of external knowledge results useful for both tests 1a and 1b. In particular we can observe how the *Merge* approach seems better than *Append*, probably due to the embedded "cleaning" phase which reduces the number of terms non strictly related to the original analyzed short text keeping only a low number of word high semantic value for that purpose.

Table 5.6: Enrichment impact: Average NDCG and P@1.

	1a - Offline		1b - Online	
Approach	NDCG	P@1	NDCG	P@1
No-enrich	0.473	0.263	0.580	0.393
Append	0.533	0.340	0.711	0.527
Merge	**0.553**	**0.380**	**0.720**	**0.560**

Moreover, during the analysis of the results of this experiment, we observed that in some cases the *Append* approach performed better than *Merge*. The reason behind this issue could be due to the nature of the documents used in the enrichment phase, because with some texts probably to use different kind of documents, or documents with different properties, could lead us to get different results. On this basis we investigate deeper on this aspect in the set of experiments described in Section 5.3.3. We kept the *Merge* approach as general setting for terms combination, for the next experiments, due to its better overall performance.

5.3.2.2 Cut-Off Experiment

Setup An important phase to analyze in an enrichment process is the cut-off methodology used to get a limited set of new terms to add to the original text. We define "Experiment 2" to compare the simpler approach based on a fixed number of terms, with a threshold-based one, and with a more sophisticated dynamic technique, as described in Section 5.2.1.2 as *Fixed*, *Threshold*, and*Dynamic*. The Experiment 2 is defined as follows:

Exp 2: Tweets posted in whole range 2013-2016, online categorized with Google search engine, with *Fixed*, *Threshold*, and*Dynamic* approaches.

Results Table 5.7 shows the results of Experiment 2, with average NDCG and P@1 values for each compared Cut-Off approach. The highest values for *Dynamic* Cut-Off highlight how every text enrichment needs a different analysis in order to get the right set of new terms, and there is no fixed number suitable for each case. Neither the threshold-based approach, which exploits the relevance score got by the terms in our system, allows to improve the performance considerably, indeed we got just a little increment on both the average NDCG and P@1 values.

Table 5.7: Cut-Off impact: Average NDCG and P@1.

Approach	NDCG	P@1
Fixed	0.692	0.557
Threshold	0.695	0.570
Dynamic	**0.754**	**0.607**

The results discovered with this test led us to add the *Dynamic* Cut-Off to the list of settings to take in care in order to get higher performance, and suggest that it is reasonable to further investigate in that direction studying other different dynamic methodologies.

5.3.2.3 Category Tree Experiment

Setup We define "Experiment 3" to study the performance of text enrichment with different set of categories as final outcome of our system (sets defined in Table 5.2). Since we use our proposed categorization process as a tool for evaluating the effectiveness of the enrichment phase, we want to analyze if there is an improvement regardless of which

are the details of the categorization phase. We are aware of the fact that increasing the number of categories causes a worsening in performance of the system making the task more difficult. On this basis, we want to assess whether the enrichment helps at each level of categorization and also if it contributes to lessen the performance degradation by comparing the precision obtained with the other two baseline systems. The first baseline system (Rnd) is based on a random approach to guess the right category, and the second one (Crd) is based on the category with the max number of tweets related to that topic, in order to exploit the maximum cardinality to reduce errors. We want to measure if our proposed system (Prp) reduces the score degradation thanks to the embedded enrichment process. The Experiment 3 is defined as follows:

Exp 3: Tweets posted in whole range 2013-2016, online categorized with Google search engine, with 20 categories, 12 categories, and 5 categories.

Results Table 5.8 illustrates the average NDCG and P@1 scores obtained by our Prp system categorizing tweets with the three different set of categories. As expected the highest values have been obtained with the lowest number of categories, and both the scores decrease when the number of categories increases. In particular, our Prp system obtained a precision level similar to state-of-art solutions without the help of supervised approaches like, e.g., [80]. Therefore, we investigate deeper by looking at the details of the precision scores.

Table 5.8: Category Tree impact: Average NDCG and P@1.

Approach	NDCG	P@1
5 categories	**0.894**	**0.707**
12 categories	0.845	0.653
20 categories	0.808	0.603

In Table 5.9 we can see the precision scores obtained by the three compared systems in three different levels of precision: P@1, to benchmark how much the systems provided the right category as first proposed category, namely a category evaluated with a "2" by users (as defined in Section 5.3.1); P@2, to see the systems' performance when the right category is in the "top 2" proposed categories; and P@3, to check the "top 3" categories.

Table 5.9: Category Tree impact: Precision comparison.

#	Rnd			Crd			Prp		
Categ.	P@1	P@2	P@3	P@1	P@2	P@3	P@1	P@2	P@3
5	0.200	0.400	0.600	0.290	0.540	0.780	0.707	0.937	0.977
12	0.083	0.167	0.250	0.210	0.380	0.510	0.653	0.850	0.930
20	0.050	0.100	0.150	0.170	0.280	0.360	0.603	0.790	0.863

As expected the values of all three metrics for all systems follow the trend we previously discussed, namely the precision got lower score when we increase the number of categories. Our proposed system got the highest values compared with the other ones

at every level of precision and with every number of categories. This fact highlights how the enrichment process helps to increase the performance at each level providing considerable contribution which allows to outperform the baseline systems.

Moreover, we analyze the precision score differences between the precision with 5 and 12 categories, and then between 12 and 20, in order to see how the score decays differ among the systems. In Table 5.10 these details are illustrated, and it is clear how our system with the enrichment phase keeps the decay in general lower compared with the other systems. In particular, we can observe how the performance worsening is reduced when we increase the number of categories, and even when we increase the number of retrieved categories (as highlighted by the P@2 and P@3 scores).

Table 5.10: Category Tree impact: Performance decay comparison.

Categ.	Rnd			Crd			Prp		
step	P@1	P@2	P@3	P@1	P@2	P@3	P@1	P@2	P@3
5-12	0.117	0.233	0.350	0.080	0.160	0.270	0.054	0.087	0.047
12-20	0.033	0.067	0.100	0.040	0.100	0.150	0.050	0.060	0.067

5.3.2.4 Type Experiment

Setup "Experiment 4a" and "Experiment 4b" aim to understand how different kinds of documents affect the enrichment effectiveness, and in particular to study the impact of news in this scenario where they represent a kind of document strongly related to what people post on social networks. With the Experiment 4a we analyze news articles, blog articles, and a mixed sample with the combinations of news and blog articles, categorized with an offline dataset. Experiment 4b is related to the same analysis but using an online search engine to retrieve, news articles, blog articles, general documents from all web, and a sample composed of news and blogs in combination with reports, scholarly articles, social media posts, forums, and tech articles. The 2 tests are defined as follows:

Exp 4a: Tweets posted in September 2015, offline categorized with SignalMedia News, SignalMedia Blogs, and SignalMedia (complete dataset).

Exp 4b: Tweets posted in whole range 2013-2016, online categorized with Google News, Google Blogs, Google AllWeb, Google Articles[7].

Results Table 5.11 shows the results obtained with the offline source of documents. The average NDCG and P@1 values highlight how the News are more suitable for the purpose of enriching short text, compared to the Blog articles. Moreover, the combination of the two decreases a little the performance. This issue is further proof of how much of the credit goes to News, and the addition of contents from Blogs could have worsened the global quality of texts.

[7]The "Article" set of documents is defined by schema.org institution `http://schema.org/docs/full.html`

Table 5.11: Type impact - Offline: Average NDCG and P@1.

	4a - Offline	
Approach	NDCG	P@1
News	**0.479**	**0.350**
Blogs	0.347	0.223
News+Blogs	0.467	0.313

Table 5.11 shows the results obtained with the online source. Also in this case News are more effective than Blog articles. The use of all web documents improves the results compared to News, but the increase is not so much higher than the value obtained with only news, therefore it is a further confirmation of how important is the role played by news articles. A remarkable value is obtained by "Articles", which is the sample composed of the combination of News, Blogs, with also few other forms of social and technical articles.

Table 5.12: Type impact - Online: Average NDCG and P@1.

	4b - Online	
Approach	NDCG	P@1
News	0.640	0.537
Blogs	0.385	0.283
Articles	**0.727**	**0.630**
AllWeb	0.708	0.590

These results confirm the relation between the short text published on social network and the News articles. They suggest to use only this kind of documents in order to optimize the enrichment phase with a lower number of documents, and related computation cost, but keeping a considerable effectiveness. On the other hand, we discovered that there are also other particular kind of documents that would be worth analyzing in order to study the better performance obtained by "Articles" over News.

5.3.3 Experiments on News Properties

In this section we describe all tests which aim to analyze the different news collection properties. In particular in the following sections we define the experiments setups and results related to the second research question Q2 introduced in Section 5.1.1.

5.3.3.1 Volume Experiment

Setup To measure the impact of collections volume we defined 4 tests: "Experiment 5a" and "Experiment 5b" focus on different fractions of documents in a dataset to use. On the other hand, "Experiment 5c" and "Experiment 5d" study samples with fixed number of documents with different order of magnitude. We analyzed samples using news subsets with different cardinality extracted form Temporalia and KBA datasets. With these experiments we aim to see how changing the amount of news affects the results, with a comprehensive view from both "percentage" and "order of magnitude"

point of view; and also if the results will generalize across different collections. The 4 tests are defined as follows:

Exp 5a: Tweets posted in whole 2013, categorized with Temporalia 1%, Temporalia 10% and Temporalia 100%.

Exp 5b: Tweets posted in whole 2013, categorized with KBA 1%, KBA 10% and KBA 100%.

Exp 5c: Tweets posted in whole 2013, categorized with Temporalia 1K news sample, Temporalia 10K news sample, Temporalia 100K news sample, and Temporalia 1M news sample.

Exp 5d: Tweets posted in whole 2013, categorized with KBA 1K news sample, KBA 10K news sample, KBA 100K news sample, KBA 1M news sample, and KBA 10M news sample.

Results Table 5.13 shows the results related to Experiment 5a and 5b. The average NDCG and P@1 values highlight how for both collections the amount of documents used is an important property to consider to measure the volume impact. We can observe a noticeable improvement with Temporalia 100% compared to smaller samples. We notice a slighter difference between Temporalia 1% and 10%, where the news increase in number from an order of magnitude 10K to 100K, compared to the other couple of samples. The same with KBA, where we can observe the most noticeable difference between KBA 10% and KBA 100%. This fact emphasizes how increasing the sample sizes has considerable effects on the results only when a certain amount of news is reached, but we need to deeper analyze different sample sizes in order to understand if it could be a unique and fixed number, or rather, it depends on the dataset.

Table 5.13: Volume impact - Percentage of documents: Average NDCG and P@1.

Sample	NDCG	P@1
Temp 1%	0.254	0.117
Temp 10%	0.294	0.142
Temp 100%	**0.403**	**0.223**
KBA 1%	0.361	0.200
KBA 10%	0.395	0.239
KBA 100%	**0.436**	**0.263**

The results related to Experiment 5c and 5d are illustrated in Table 5.14, where it is possible to see how the order of magnitude of the document samples affects the results at every step for both datasets Temporalia and KBA. With Temporalia, the most significant improvement has been detected from 100K to 1M, instead with KBA the best improvement is from 1M to 10M. These results are aligned with what discovered with Experiment 5a and Experiment 5b, where we learned that in any case we need to go over the 10% of documents, and suggest us that reaching the 50% of a dataset allows us to get the first significant increase in performance.

In general, increasing the number of documents seems to improve the results, but this particular fact raises interesting issues concerning the "nature" of the dataset, and led us to investigate on other properties which make a dataset more o less suited to be used as enrichment source.

The diverse impact of Temporalia and KBA is probably also due to other factors than the only difference in size. Of course the same percentage, applied to collections with very different sizes, yields sets of extracted documents whose cardinality is very different; whence we can also expect a different variety of such sets. Moreover, for instance, KBA does not fully cover year 2013, whence the effectiveness could be affected by the publishing date of the analyzed short texts. Such aspects are taken into consideration in the remaining experiments.

Table 5.14: Volume impact - Order of magnitude of documents: Average NDCG and P@1.

Sample	Average NDCG	P@1
Temp 1K	0.210	0.083
Temp 10K	0.235	0.095
Temp 100K	0.283	0.128
Temp 1M	**0.335**	**0.197**
KBA 1K	0.276	0.104
KBA 10K	0.294	0.147
KBA 100K	0.358	0.194
KBA 1M	0.384	0.206
KBA 10M	**0.421**	**0.250**

5.3.3.2 Variety Experiment

Setup "Experiment 6a" and "Experiment 6b" aim to measure how the variety of news inside a collection could impact the enrichment phase and consequently the categorization process. We use two approaches to obtain a measure of variety, which is of course a delicate concept. First, the two used collections Temporalia and KBA are composed respectively of news from 2 and 4 different sources, as explained in Table 5.3, therefore it is reasonable to assume that KBA has more variety than Temporalia. Second, we seek anyway for a more quantitative comparison and we compute the term frequency for each term, after stopword removal, in: (i) a sample composed of 400K random news from Temporalia, (ii) a sample of 400K random news from KBA, and (iii) a mixed sample composed of 200K news from Temporalia and 200K news from KBA. We represent the three samples with their frequency distributions, and we compare each pair of distributions using the Jensen-Shannon Divergence to have an indicator of "distance" JSD.

Table 5.15 shows the comparison values obtained by each couple of term frequency distribution with the Jensen-Shannon Divergence. Analyzing the JSD values it is clear that the variety of the samples follow this rule: $Temp < KBA < Temp + KBA$. The highest JSD value, got by the samples Temp and KBA+Temp indicates that their distribution are the most distant, and it reflects what we expected because Temporalia

is the dataset with less variety and KBA+Temp has the highest variety due to the combination of the two datasets.

With Experiment 6a we analyze news samples with the same cardinality from different collections and from a time window equals to 1 month, in order to see the effects of changing news varieties in a short period of time. With the Experiment 6b we focus on a time window equals to 6 months, to study if on a wider time window we have the same effects we get on only 1 month. The 2 tests are defined as follows:

Exp 6a: Tweets posted in January 2013, categorized with Temporalia Jan 2013 (60K news sample), KBA Jan 2013 (60K news sample) and Temporalia+KBA Jan 2013 (30K+30K news sample).

Exp 6b: Tweets posted in the second half of 2012, categorized with Temporalia Jul-Dec 2012 (400K news sample), KBA Jul-Dec 2012 (400K news sample) and Temporalia+KBA Jul-Dec 2012 (200K+ 200K news sample).

Table 5.15: Variety analysis with Jensen-Shannon divergence on the term frequency distributions.

	Temp Jul-Dec '12	KBA Jul-Dec '12
KBA Jul-Dec '12	0.0045825	-
Temp+KBA Jul-Dec '12	0.0053596	0.0031857

Results Table 5.16 shows how the variety of news inside the analyzed samples affects the enrichment effectiveness. For both experiments there is a noticeable difference among the samples which highlights how increasing the variety of news allows to improve the final categorization also on different time windows. The differences between Temporalia and KBA are higher in Experiment 6a, where the time window is smaller. This fact highlights how important is to increase the variety of news in order to improve the set of words to use as text enrichment, especially when the dataset is small.

Table 5.16: Variety impact: Average NDCG and P@1.

	6a - 1 month		6b - 6 months	
Sample	Average NDCG	P@1	Average NDCG	P@1
Temp	0.316	0.159	0.460	0.227
KBA	0.427	0.236	0.475	0.327
Temp+KBA	**0.465**	**0.279**	**0.566**	**0.357**

5.3.3.3 Structure Experiment

Setup Another aspect we consider is if there is a structural component of the news, e.g. title or content, that could be more effective for the enrichment purpose. "Experiment 7a" and "Experiment 7b" aim to compare the effectiveness of the enrichment

phase exploiting terms from (i) only the news title, (ii) only the news content, (iii) from both components weighted 50% each.[8]

To run further analysis on this aspect we built an "artificial" news title composed of words form the news content, with different cardinalities: the same one of the original title (N), twice (2N), four times (4N), and an artificial title with all content terms with duplicated removed (MAX). Based on this approach we define "Experiment 7c"[9] to study if the different writing style between the two news components affects the effectiveness, and how the number of terms from contents impact the results. It is important to point out that using the entire news content involves duplicated terms that are emphasized during the retrieval process with the search engine, thus giving a higher contribution to the identification of important terms. The 3 tests are defined as follows:

Exp 7a: Tweets posted in whole 2013, categorized with Temporalia 2013 Titles, Temporalia 2013 Contents, and Temporalia 2013 weighted Titles+Contents.

Exp 7b: Tweets posted in whole 2012, categorized with KBA 2012 Titles, KBA 2012 Contents, and KBA 2012 weighted Titles+Contents.

Exp 7c: Tweets posted in whole 2013, categorized with Temporalia 2013 Artificial Titles N, 2N, 4N, MAX.

Results Table 5.17 shows the results with different news components, on both Temporalia and KBA datasets. In both cases using only the title we obtained the worst performance, and even in combination with the content the title has a bad effect.

Table 5.17: Structure impact - Components comparison: Average NDCG and P@1.

	7a - Temp		7b - KBA	
Sample	NDCG	P@1	NDCG	P@1
Titles	0.258	0.127	0.371	0.207
Contents	**0.313**	**0.153**	**0.399**	**0.287**
Titles+Contents	0.260	0.150	0.395	0.253

On this basis the news content seems to be the best container of good terms to use as enrichment source, but by analyzing the result of Experiment 7c, depicted in Table 5.18, we can observe how the performance decreases when the length of the artificial title is 2N, then a minimal improvement with 4N, and finally a slightly larger score when all news content is used. The results suggest that the frequency of terms in the news content plays an important role in the selection of the most suitable terms for the enrichment, perhaps by discarding a lot of noise due to a large amount of terms with lower weight.

[8] We used Solr search platform based on Lucene IR system.

[9] Experiment 7c run over only Temporalia dataset as the only one with the structured data for that purpose.

Table 5.18: Structure impact - Artificial titles comparison: Average NDCG and P@1.

Sample	NDCG	P@1
N	**0.122**	**0.053**
2N	0.114	0.047
4N	0.117	0.050
MAX	0.126	0.040

5.3.3.4 Freshness Experiment

Setup To benchmark how the news freshness is important we performed 2 experiments, "Experiment 8a", "Experiment 8b", based on different news "aging" run on different collections. These tests aim to analyze the difference between enriching the tweets with news extracted from different time windows increasingly large and distant from the tweet posting date. We call *contextualized* sample a set of news with dates with this kind of temporal relation with the analyzed short text. The 2 tests are defined as follows:

Exp 8a: Tweets posted in whole 2012, categorized with Temporalia 2011-2012 - *contextualized* in a time window equal to 1 week, 1 month, 3 months, 6 months and 12 months.

Exp 8b: Tweets posted in whole 2012, categorized with KBA 2011-2012 - *contextualized* in a time window equal to 1 week, 1 month, 3 months, 6 months and 12 months.

Results Table 5.19 shows the results related to Experiment 8a and 8b. By looking at the average NDCG and P@1 scores related to the test on Temporalia, it is possible to notice how the news freshness affected the results effectiveness decreases when the news get older, especially going from 1 week to 1 month. Temporally contextualized news allow us to get the best effectiveness.

By analyzing the average NDCG and P@1 scores related to Experiment 8b on KBA, we can observe a similar trend in performance which decreases due to the news "aging" effect. A noticeable fact to observe is how, even if there is a worsening in performance with larger time windows, there is an increase with 12 month-old news. This is probably due to the well known cyclical nature of the news.

As revealed from the previous tests in this section, tests on different datasets could be subject to the effect of other properties, therefore in this case the KBA Volume and/or Variety could be some of the factors behind this difference. In general these tests highlighted how the news freshness considerably affects the results if the documents are very close to the short text posting time, therefore this is a confirmation of the relation between the short text published on social network and the recent events.

5.3.4 Final Remarks and Discussions on Short Text Analysis

In this section we want to remark the main results and the most important discoveries obtained with our experiments on short texts, in order to have an overview of our studies which allow us to stimulate discussions and new observations. In Table 5.20

Table 5.19: Freshness impact: Average NDCG and P@1.

	8a - Temp		8b - KBA	
Sample	NDCG	P@1	NDCG	P@1
1 Week	**0.739**	**0.427**	**0.623**	**0.440**
1 Month	0.286	0.167	0.302	0.200
3 Months	0.282	0.167	0.305	0.193
6 Months	0.254	0.150	0.294	0.180
12 Months	0.107	0.053	0.333	0.200

we summarize all experiments on short texts listing the subject involved, the aims, the datasets used in the experiments, and finally the results and observations for each of them.

Experiment 1 - Enrichment: The enrichment process results to be a useful phase which contributes in raising the system effectiveness. In particular the *Merge* approach got the best scores due to its terms cleaning phase, although in some cases it proved to be too restrictive. In particular cases, when the enrichment source provided not so suitable documents, or the text was particularly difficult to categorize, it would be better to use the *Append* approach avoiding to discard too many words.

Experiment 2 - Cut-Off: The compared cut-off methodologies showed how each text generates its new terms list with different relevance score distributions, highlighting how there is no specific fixed number of terms to keep, or relevance score to use as threshold. For that reason the *Dynamic* approach results to be the best, and it seems reasonable to further investigate in that direction.

Experiment 3 - Category Tree: The proposed approach outperformed the compared baseline approaches and it placed not so distant from other state-of-the-art systems. The enrichment phase affected the categorization results with all sets of categories, helping to slow down the performance degradation when the number of categories increases.

Experiment 4 - Type: News articles used in the enrichment phase got the highest scores providing the best contribute as enrichment source. Blogs are unable to help to raise the score when used in conjunction with news, probably due to their limited or bad contribution. The comparison with *AllWeb* and *Articles* highlighted how the news performance are not so distant from the results got by using *AllWeb*, proving how the news play an important role in the document kinds to use as enrichment source. An interesting finding is the one related to the result obtained by the *Articles* kind of documents which got the best scores, and it suggests us to deeper investigate on what other useful documents could be integrated to further improve the impact of the enrichment phase.

Experiment 5 - Volume: The number of documents results to be a crucial aspect to analyze. A low number of documents do not provide sufficient data to improve the original short text; indeed, the results related to the dataset percentage highlight how with the 10% is not enough to reach the critical mass. With the second test, related to order of magnitude of documents samples, we can confirm that small amounts of data give a low contribute. The scores obtained by increasing order of magnitude demonstrated how there is a major increment when the 50% of the dataset is reached for both ones we used in the study. Moreover, we observed some differences between the two analyzed datasets,

and in particular KBA got in general higher scores compared to Temporalia. This fact confirms how the news sets have other properties which characterize the samples and make the performance different, therefore worth to study.

Experiment 6 - Variety: The variety seems to be a relevant feature to take in care, indeed on both analyzed time windows it has a considerable impact on the results. Moreover, the Jensen-Shannon Divergence (JSD) confirmed the higher variety for KBA compared to Temporalia, and even higher when the datasets have been combined. In particular the results emphasize how the variety difference between Temporalia and KBA is more noticeable on the small time window, highlighting how with smaller amounts of data variety has a stronger impact.

Experiment 7 - Structure: The News contents play the most important role as source of additional texts to exploit, compared to News titles, and even with the weighted combination of titles with contents. The use of an artificial title composed of the most frequent words in the content highlights how it is not enough, and how the word scores obtained using the word frequency is the crucial point.

Experiment 8 - Freshness: The News freshness results to be an important property to take in consideration; indeed, it has a strong impact when the news publishing time is close to the short text posting time. Moreover, We discovered a sort of "cyclic effect" of news on KBA when we use 1-year old news, probably due to the presence of similar news contents in a similar period of time.

In general the enrichment process on short texts resulted to be a useful phase to improve the categorization process. The detailed analysis on the various properties taken into consideration emphasized how there is not a unique best setting to use, due to some of those not combinable. For instance, if we limit the document types or freshness, it is difficult to reach the proper volume of document needed to have a considerable improvement. On the other hand there could be sets of short texts which do not need big volumes of documents, but the benefits come with more freshness. On this basis, this study highlighted how it is necessary to define different settings to better fit the needs of each task. This study provides a sort of guide on the impact which the different document properties have, and how to combine them to find a proper setting to use in a real application. Moreover, differently that other solutions proposed in literature, our categorization approach can be easily integrated as mobile module to analyze user daily activity, because it does not rely on machine learning techniques which depend on training processes. It is also not dependant on any structure like hyperlinks or hashtags, therefore more flexible than other specific system proposed in other research works, and able to handle short texts from different sources.

Table 5.20: Overview of experiments results on Short Texts.

Experiment	Subject	Aims	News Dataset	Results	Observations
1a	Enrichment	NO-enrich. VS Append VS Merge	Temp+KBA	Enrichment useful. Merge: Best approach.	Merge: sometimes too strict terms cleaning.
1b	Enrichment	NO-enrich. VS Append VS Merge	Google	Enrichment useful. Merge: Best approach.	Merge: sometimes too strict terms cleaning.
2	Cut-Off	Threshold VS Dynamic VS Fixed	Google	Dynamic: Best approach	No fixed number of terms to keep
3	Category Tree	P@1,P@2,P@3 on 5,12,20 Categories	Google	Enrichment phase useful with all category sets	Enrichment helps to reduce the performance decay
4a	Type	News VS Blogs VS News+Blogs	SignalMedia	Best doc type: News articles	Blogs unable to improve the performance.
4b	Type	News VS Blogs VS Web VS Articles	Google	News plays an important role, not so distant from AllWeb	Best doc type: Articles. Other doc type could be useful
5a	Volume	Percentage of documents	Temporalia	Major increment: toward 100%	Small amount of data not sufficient.
5b	Volume	Percentage of documents	KBA	Major increment: toward 100%	Small amount of data not sufficient.
5c	Volume	Order of magnitude of documents	Temporalia	Major increment: toward 50%	
5d	Volume	Order of magnitude of documents	KBA	Major increment: toward 50%	Higher scores suggest other properties to study
6a	Variety	Different sources on 1 month	Temp+KBA 1m	JSD confirms Temp+KBA > KBA KBA > Temp	On smaller time window Variety has stronger impact
6b	Variety	Different sources on 6 months	Temp+KBA 6m	JSD confirms Temp+KBA > KBA KBA > Temp	
7a	Structure	Title VS Content VS Title+Content	Temporalia	Content provides best text	Title+Content worsened the performance
7b	Structure	Title VS Content VS Title+Content	KBA	Content provides best text	Title+Content worsened the performance
7c	Structure	Artificial title N,2N,4N,MAX	Temporalia	Increasing the length has no considerable effects	Word frequency is crucial to build word scores
8a	Freshness	News “aging” 1w,1m,3m,6m,12m	Temporalia	Strong impact from 1 week to 1 month	
8b	Freshness	News “aging” 1w,1m,3m,6m,12m	KBA	Strong impact from 1 week to 1 month	“Cyclic effect” with 1-year old news

6

User Modeling Exploiting Short Texts

In this chapter we describe our research work related to user modeling exploiting short texts and network structures, presented in [62]. We introduce our proposed method for computing user similarity based on a network representing the semantic relationships between the words occurring in the same tweet and the related topics. We use such specially crafted network to define several user profiles to be compared with cosine similarity. This approach relies on the methodologies, studies and proposed system described in Chapter 5, and aims to further study the effectiveness of enriched data, having a broader vision analyzing all tweets posted by a user as a whole representation of his/her interests. We also present a preliminary experimental evaluation on a limited dataset.

6.1 Network-based User Modeling Exploiting Short Texts

There is a growing interest in analyzing social networks content to produce user models and to measure user similarity. The latter has been traditionally exploited in filtering and recommendation systems, and more recently in web search. A common approach is to define user similarity by exploiting the graph of the social relationships between users, such as, friendship, sharing, liking, and commenting on Facebook, following, retweeting, and favoriting on Twitter. However, this approach has some drawbacks: the resulting system could be too strictly tailored against a peculiar social network and it may not easily adapt to other cases; it may fail in representing and comparing "lone" users, i.e., people not liking to follow other people or being followed; there might be a cold start problem. An approach based on social relationships may not be successful where these are weak or absent, for example messaging systems not relying on a social network.

The approach we present in this chapter is content based, indeed we try to predict user similarity by relying on contents only. While we focus on Twitter, our approach

is independent from the specific social network: we do not rely neither on the *following/being followed* social relationships nor on the peculiar structure of tweets (e.g., links, hashtags etc.). In our model each user is represented by a network linking the words most often posted, and other words from text enrichment procedures, with the tweets they occur in, and the latter with the topics the user is interested in. Topics are taken from the category hierarchy of Wikipedia, as explained in the proposed system described in Chapter 5, but we are free to switch to other equivalent knowledge sources. The network allows us to evaluate several distinct approaches to users profiling and similarity.

6.2 Proposed Approach

Our approach has the following overall steps:

1. The words in user's tweets are collected.
2. Text enrichment is used to add more words to each tweet and to associate *topics* to the tweets, obtained from the Wikipedia categories. We distinguish the most specific from the most generic ones (called macro-topics in the following) on the basis of Wikipedia category hierarchy.
3. Words, tweets, and topics are used to build the network.
4. Vector-based user profiles are defined, whose components are words weighted by network centralities.
5. User similarity is computed by using cosine similarity function.

In the following sections all details about the proposed approach are provided.

6.2.1 The Network-Based User Model

We first extract a set of tweets which are the contents posted by selected Twitter accounts, in order to have short texts to enrich and categorize using the methodologies described in Chapter 5. We then build a network with the set of words related to each user, with three layers of nodes, similar to those in [45, 93]. We define a first layer of nodes which represent the original words posted on Twitter and the additional words got from the enrichment process. Each node into this layer contains the string representing the word and an *ID* to identify if it was part of the original ones or those added. The second layer of this network is composed of nodes that represent *tweets*, with an *ID*, and a *timestamp*, to allow future temporal network analysis. Each word is connected to the corresponding tweet where it was published, and added to the network uniquely; therefore if a word is already present during the network building, a new edge will be added to the new tweet, in order to avoid duplicates and consequently emphasize the weight of that specific word. The third layer is composed of those words extracted during the labeling process, which represent the topics discussed in a specific tweet. We keep the relations among Wikipedia categories therefore we have a list of linked words connected to the related tweet. This structure creates paths from the words, through the tweets, and the topics, up to the macro-topics, which represent the most generic user interests.

Then we complete the network structure by assigning a weight to each edge of the graph. We denote the edge between the word i and the tweet j as e_{w_i,tw_j}, and define its weight as follows:

$$e_{w_i,tw_j} = 1/(|tw_j| - 1), \tag{6.1}$$

where $|tw_j|$ is the number of words that compose the tweet tw_j. With this approach we emphasize words contained in shorter texts, so as to give higher scores if a word strongly represents the semantics of the user's tweet. We consider in the computation also the new words added by the enrichment process.

The edges that connect tweets with topics, are denoted as e_{tw_i,t_j}, and for the weight assignment we exploit the relevance scores obtained by the labeling process for the macro-topics. We propagate their values along the network emphasizing topics more distant from the macro-topic, in order to give higher weight to nodes that represent a more specific topic rather than generic ones. The edge weight as computed as follow:

$$e_{tw_i,t_j} = c_{t_{wi}} \cdot (steps(t, c_t) + 1), \tag{6.2}$$

where $steps(t, c_t)$ is the number of steps necessary from the topic t to get its unique reachable macro-topic c_t in the path. And $c_{t_{wi}}$ is the relevance score got by the Wikpedia macro-topic c_t related to that specific topic t and the tweet tw_i, during the labeling process.

With this approach the user model highlights the specific user interests, without losing information about the macro-topics, thanks to the network structure.

The edges between topics, that we denote as e_{t_i,t_j}, have a particular meaning from the user interest point of view. For each tweet we can extract the topics that deals, but each topic has a path to a macro-topic which identify the generic interest covered by the user, therefore we can again propagate the same relevance score to the path with the same rule previously described. In this way, each time a tweet propagates the score, we sum the values to raise up the weight of those edges. With this approach we can highlight the relationship between the specific topics dealed and the "super"-topics in the path, and define a network structure that makes possible the similarity computation between users at multiple levels. The final weight for these edges is computed as follows:

$$e_{t_i,t_j} = \sum_{tw_k \in T} c_{t_{wi}} \cdot (steps(t_j, c_t) + 1), \tag{6.3}$$

where $T = \{tw_1, tw_2, ..., tw_k\}$ is the set of all tweets connected with a path to the topic t_i and consequently t_j, related to current analyzed edge. We say that there exists $tw_k \in T \Leftrightarrow tw_i$ *isConnectedTo* t_i.

Figure 6.1 shows an example of network that model a user with words and topics as nodes and weighted edges.

6.2.2 User Profiling and User Similarity

Before comparing users, we need to define on which data (extracted from the whole user model) to carry out such comparison, i.e., we must define *user profiles*. In the

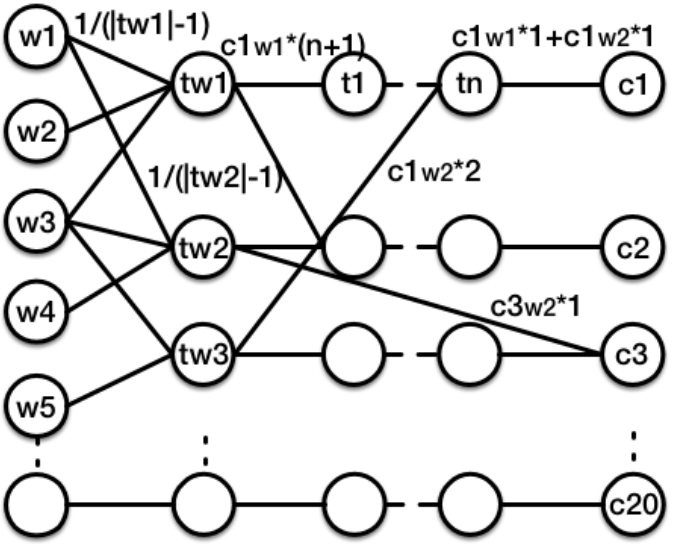

Figure 6.1: Example of Network-based user model

following we aim at evaluating several kinds of profiles, considering multiple elements, either separately or together, in order to experiment different points of view.

First, we set a baseline for our remaining approaches: we count only the set of words originally posted by the user, simply by considering how many times a word is connected to a tweet, and we repeat the process for the words added by the enrichment process, and finally for all the words in the network (both originally posted and added). The score list is normalised to have a final rank list. This step allows us to compare the sets of words, to evaluate if the enrichment process has led to improvements, i.e., if the added words with related scores better represent the analysed user.

Then, we build a profile based on the part of the model related to topics and macro-topics, to focus on the main interests of the user, leveraging on the scores obtained by the text enrichment and categorization (for the details see Chapter 5). If we want a coarse-grained profile, we restrict to macro-topics (identifying only the main interests of the user). Otherwise, we can resort to a fine-grained profile, considering the entire path of topics with related scores computed according to Formula (6.3). Fig. 6.1 (b) shows a profile built for a popular Italian Twitter account about politics called "tweetpolitica". We selected the first 5 words, with their scores, for each approach described above.

As second step we use network centrality measures, to assign scores to nodes taking advantage of our network structure and weights. We extract a subnetwork to make a first computation based only on the relationships between words and added words, and test the centrality-based profiling. We use both types of words since the enrichment process added a useful set of new terms to add information about user, while the original words preserve his/her original style of expression. In particular, we exploit: the strength centrality to see which are the nodes with higher degree, by analysing the edges weights (as defined in Section 6.2.1); the eigenvector centrality to emphasize the words often used

Table 6.1: User profiling based on centralities (user "tweetpolitica"), legend: s→strength, e→eigenvector, b→betweenness, w→original words, e→enriched words

	s (w+e)		e (w+e)		b (w+e)		s (topics)		e (topics)		b (topics)	
1	August	1	August	1	Italy	1	Pol. parties in Italy	1	Politics of Italy	1	Pol. parties in Italy	1
2	Italy	.660	Italy	.520	August	.957	Politics of Italy	.897	Pol. parties in Italy	.798	Politics of Italy	.917
3	politics	.451	sole	.483	politics	.449	Politics by country	.598	Politics by country	.654	Chronology	.500
4	sole	.380	politics	.466	euro	.369	Politics	.458	Politics	.196	Politics by country	.479
5	euro	.346	gov.	.448	sole	.279	Chronology	.306	Italian gov.	.098	Sport	.438

Table 6.2: User similarity comparisons, legend: s→strength, e→eigenvector, b→betweenness, w→original words, e→enriched words, macro-t.→macro-topics

	words	w+e	enrich.	topics	macro-t.	s (w+e)	e (w+e)	b (w+e)	s (topics)	e (topics)	b (topics)
matteorenzi - beppe_grillo	.028	.054	.131	.874	.904	.510	.602	.002	.868	.998	.757
matteorenzi - tweetpolitica	.068	.089	.148	.602	.790	.714	.766	.385	.447	.003	.654
matteorenzi - Pontifex_it	.029	.024	.082	.371	.514	0	.978	0	.308	.003	.610
matteorenzi - SerieA_TIM	.019	.023	.062	.175	.147	0	.786	0	.271	0	.685

in conjunction with the most used; and the betweenness centrality, to have information about words often present in tweets (i.e., to highlight the user style of expression). We adopt the same approach for the subnetwork composed of topics, to build a user profile based on the user major interests. Table 6.1 shows the final rank lists of terms with score computation based on network centralities for both words and topics.

Given two user profiles, we compute their similarity score from multiple points of view, like in the profile building process. For instance, if we consider only the macro-topics, we can say if two users have in common some general interests. Then, by considering all topics, we can get more detailed information. The similarity may be also computed by analysing just the words, to understand how users express their opinions and how they discuss their topics. Someone can use peculiar terms or grammar constructs to deal with the same topics. Users may satisfy different similarity notions.

On this basis, after focusing on a certain set of data, we build a list of terms with just those the two users have in common, with related scores. Hence, we build a geometric space defined by the features represented by the terms in common, and users are represented as vectors into this space. We compare them by using the *cosine* similarity function, to compute how "close" the users are.

6.3 Evaluation and Results

Being in a prototypical phase, we carried out an expert evaluation to assess pros and cons of our method. With a dataset of at least 30 tweets (carefully processed in their right one-month long temporal context) for each of 17 selected accounts, we built their user profiles, as described in Section 6.2.2. Then, we computed the cosine similarity over several couples of accounts to test if our approach properly assigns scores to similar users, and how the network-based user model provides information at multiple levels (e.g., to understand if two users have in common just the main topics, if they match deeper, or if they have a similar style of expression). We compared the account "matteorenzi", the Prime Minister of Italy, due to his well defined political focus, to four accounts with

different similarity w.r.t. him: "beppe_grillo", founder of the Italian political party Five Star Movement; "tweetpolitica", the account used in Section 6.2.2 focused on Italian political news; "Pontifex_it", the account of Pope Francis; and "SerieA_TIM", the top Italian football competition. Table 6.2 lists such pairs (ordered from the most to least similar, based on expert evaluation) with the computed similarity scores.

Scores based only on counting the words posted by users are very low, although the enrichment process has improved the computation. The combined solution (words + enrich.) seems to be the most reliable due to its mixed composition: enriched words make users more similar if they talk about the same topic, and the originally posted ones keep the users' style of expression. The labelling process provided a set of terms that well identify the trend of posts. The resulting scores are very high for the first pair, as we expected, still high for the second one, and lower for the remaining ones. We notice how "matteorenzi" and "Pontefix_it", apparently not related, have a considerable score. This fact is due to the nature of texts extracted during the test period; indeed, both users have talked about topics related to war and Iraq. The topics scores provide further information: users similar on macro-topics not necessarily are related also on more specific topics. For instance, the score is lower for the second and third pairs.

As to the network centralities, it is possible to see how the strength on words can give more semantics to what users post. The first pair, with high similarity got a good value also for the strength centrality: this fact indicates that the links in the network of words lead to high scores for both, representing a high correlation on expression. The users "matteorenzi" and "Pontefix_it" have a very low similarity if we consider strength on word and topics, but an high value on eigenvector for the same reason previously described. With our approach based on centralities we are able to grasp this kind of correlation, when users talk about related topics by using different modes of expression or with different purposes. The high scores for all pairs on betweenness with topics indicate a high presence of common sub topics. This fact is probably due to the extraction of "Locations" or "Geographic regions" as topics whenever texts contain names of states, regardless of their use. This is an issue to take in care for future improvements.

III

Case Study 2: User Movements Analysis

7

User Movements Analysis

In this chapter we describe our study on user movements analysis, related to our research works presented in [32, 76, 77]. We introduce a novel approach to address the problem of places categorization exploiting contextual data to enrich the user locations, in order to identify the important places visited by users during their daily routines. We present the two phases of our methodology: during the first phase, a set of candidate stay points is identified analyzing each user detected position and exploiting contextual data deriving from GPS-logs and sensors embedded in mobile devices used for tracking; in the second phase the candidate stay points are categorized by mapping them onto a feature space having as dimensions the area underlying the stay point, its intensity (e.g., the time spent in a location) and its frequency (e.g., the number of total visits). We conjecture that the feature space allows to model aspects/measures that are more semantically related to users and better suited to reason about their similarities and differences than simpler physical measures (e.g., latitude, longitude, and timestamp). An experimental evaluation on the GeoLife public dataset is presented to confirm the effectiveness of our approach.

7.1 Aims

7.1.1 Definitions

By observing a dataset of users' movements readings, it is possible to notice some coordinates where people remain stationary for long time periods, often inside buildings or delimited areas where they perform their daily activities. We call those locations *Stay Points* (SPs), as described and defined in [50]. Theoretically, a stationary user generates the same location data for all the stay time, i.e. the same point in the geographic space; we call those places *Natural Stay Points*, due to the nature of data that does not require any particular processing to understand the corresponding visited locations. However, in real situations there are several factors that affect the tracking of user movements. Due to technology limitations, there may be locations where the position detection is not possible, or the user moves in a way that the detection result cannot be so accurate.

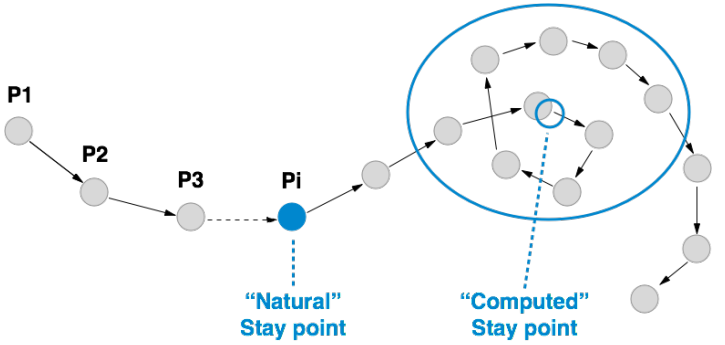

Figure 7.1: Stay point types from user positional data

For instance, if we use a GPS, there are places where there is no signal or where the accuracy is very low due to the transportation mode that varies from underground to surface. These issues led us to have data generated by several detections which do not properly match when the user is stationary. Instead, they yielded a group of points corresponding to a location with a high density of detections within a given (limited) range. This situation may also occur when users move inside a delimited area, such as their work place where they may move among offices, or during a walk inside a mall. As described in [50], for both these latter situations we can compute the mean point of that cluster of detections in order to determine the user's stay point. We call this kind of places *Computed Stay Points*, since they approximate the original real locations. Figure 7.1 shows an example of user's movement readings with the two types of stay points described above. The process to identify stay points from user movements readings helps to get the set of visited locations, but neither necessarily all of them are important for the user [50] nor they provide information. By analyzing just the density of detections, some locations may be recognized as stay points even if they are not strictly related to user's main visited places. Figure 7.2 shows how a road crossing, where users transit a lot of time during their activities, can generate a geographic region with high density of detections, and consequently a possible stay point. We call those places *False Stay Points*, because they identify locations that do not represent a user activity, and do not provide important information about user habits and behavior.

On this basis it is clear what kind of locations we consider *Important Stay Points*, namely *personal places of interest* (PPOIs): locations that can help to infer information about the user who has visited them, in particular the activities that may have been carried out at each location, the stay time, and how frequently it has been visited.

7.1.2 Challenges

To better understand what are the main problems and difficulties emerging with places categorization, and consequently importance recognition, we list a set of conceptual problems presented by the current state-of-the-art solutions. We have also run a preliminary experiment to analyze how much the conceptual problems do appear in practical scenarios, and which of them are addressed by the existing solutions; we discuss the

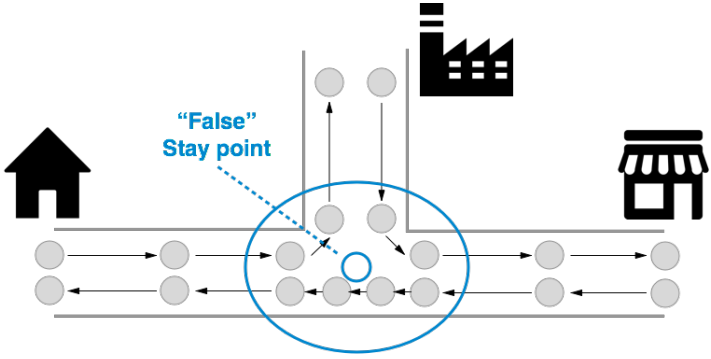

Figure 7.2: The density problem of important places discovering

results in detail in Section 7.2.

A first approach based on density may exploit a spatial subdivision of the territory where user moved to recognize the most visited locations, and consequently assign an importance value to them, but, as described in [50], this grid-based solution is affected by several issues. The cell definition during the spatial subdivision is not a technique that can be adapted to each case and to each user movement style. As represented in Figures 7.3 and 7.4 the cell might have a size not appropriate to analyze each user and each movement, causing the not-proper categorization of PPOIs due to what we call *boundary problem*, which might divide them (Figure 7.3), or include more than one of them (Figure 7.4). So two first conceptual problems are:

P1a: Boundary problem - undersized cells. With a grid based approach, cells can be too small, and thus wrongly split a stay point.

P1b: Boundary problem - oversized cells. With a grid based approach, cells can be too large, and thus wrongly identify false stay points that either merge two or more stay points, or even are created without any real stay point.

The technique used in [50, 110] for stay point computation avoids the static approach used in the grid-based solution, which mainly analyzes the user movements as an overview on a map, in favor of a dynamic approach that scans each detected position, in order to reproduce the user movements and get more information from user behavior. By using a dataset composed of users' GPS detected positions, it is possible to avoid problems related to grid cell size by focusing on defining thresholds, based on space and time, to recognize when users move and when they remain stationary.

As introduced in Chapter 2, the region surrounding the user position could be analyzed to extract additional information from related objects; in this case they are other user positions in the trajectory. For each point in the user trajectory we can exploit the context to study the user movement.

Let $P = \{p_1, p_2, ..., p_n\}$ the list of points corresponding to GPS readings ordered by time of detection, tT the time threshold and dT the distance threshold. By checking

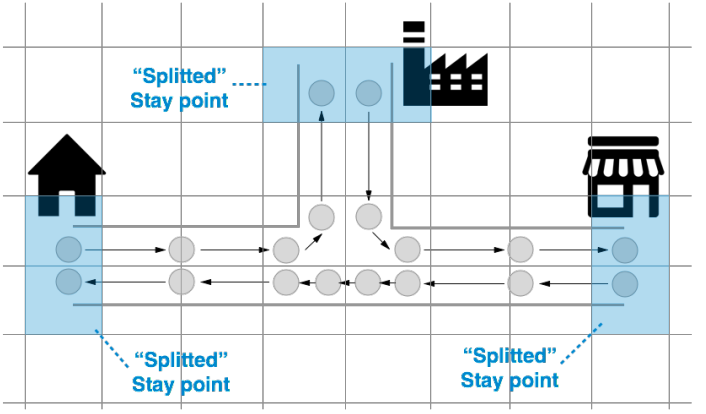

Figure 7.3: The boundary problem of important places recognition, with too small cells (P1a)

the time and distance thresholds between the point p_i and the point p_{i+1} it is possible to know if the user moved or not in that specific delimited geographic region. We call that space *segment*, since it approximates the original real user movement between the two analyzed points. If the user remains stationary (i.e., she does not exceed both the thresholds), this process can be repeated by keeping fixed p_i, scanning the next points $\{p_{i+2}, p_{i+3}, ..., p_n\}$ and stopping when the thresholds are exceeded, in order to detect when and where the user changes behavior. At the end of this process it is possible to compute a *Mean Stay Point* based on the current set of analyzed points from p_i to p_{i+k}, with $1 \leq k \leq (n - i)$, by calculating the average latitude and longitude of points.

This technique, based on space and time thresholds, is not affected by the issues related to the cell size, which is dynamically determined, but some problems are still present (and we will indeed observe its performance in our preliminary experiment in Section 7.2). On straight and long trajectories, where users move with no particular changes in speed, the dynamic approach performs a scanning which, after a certain number of points, computes a stay point based on the exceeded thresholds, i.e. the mean of points in the analyzed segment, and it repeats this process for all the trajectory length, thereby determining a set of consecutive false stay points. Figure 7.5 shows an example of this issue displaying a path between two important places segmented with false stay points. We call this problem:

P2: segmentation problem - constant speed. A trajectory between two distant important places is divided into several segments defined by the computed false stay points.

Other works in the literature [15, 107] introduce other parameters to use and improve the previous approach and minimize the segmentation problem. In particular they use new thresholds based on user speed, acceleration and even heading change, in order to better understand user behavior. More precisely, speed and acceleration thresholds are used in the same way as those about space and time, i.e., as soon as they are exceeded, the scanning process stops in order to compute the stay point. The heading threshold

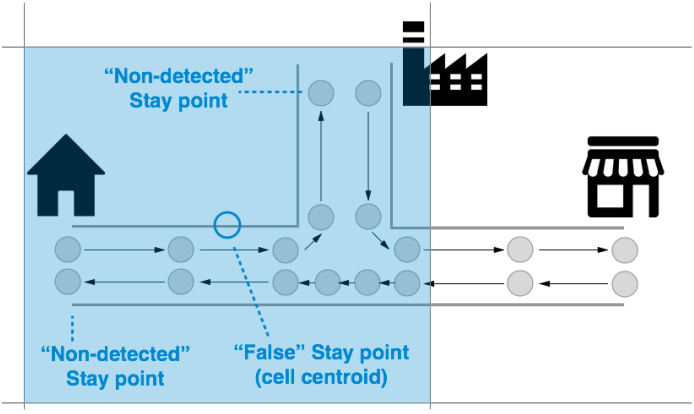

Figure 7.4: The boundary problem of important places recognition, with too large cells (P1b)

on the other hand is used in the opposite way: a constant heading indicates a movement from a SP to the next one. However, also with these approaches potentially there are some difficulties to avoid the computation of false stay points (and our preliminary experiment confirms that). For example, if a user moves with a high speed for a long time, i.e. while driving on an highway, she still exceeds the speed threshold, and after a certain amount of time the other ones, causing again the segmentation problem.

Whence, fixed thresholds may not be suitable for all user movements; indeed, some settings perfectly tuned for some users may be very wrong for others. As we will see in Section 7.4, by changing the thresholds values we observed how the recognition process varied the granularity (i.e., the number and the density of stay points) of the result, providing different set of stay points. This issue causes the computation of false stay points if the thresholds are not properly set considering the current user movements to analyze. User activities which involve several vehicles and in wider areas generate different datasets compared to users that move in small regions and mainly with one mode of transportation; whence the need of different analysis. Figures 7.6 and 7.7 show two examples where wrong thresholds raise the two last problems:

P3a: Fixed thresholds problem - slow speed. In Figure 7.6, it is possible to see how a region (delimited by the circle) where user moved with very slow speed, differently from the rest of the tracked movement, makes the threshold-based techniques unable to properly categorize the PPOIs, due to a too high threshold for the current tracked movement. Indeed, as soon as the speed exceeds the related threshold (changing from slow to high again), the whole slow speed region inside the circle will be processed in the same way of a walk inside a building, therefore generating a single false stay point. Moreover, the latter, whose position is the result of a mean of the coordinates of all the points inside the circle, can also be put in a totally wrong place, w.r.t. the progress of the path in the region.

P3b: Fixed thresholds problem - high speed. On the other hand, Figure 7.7 shows

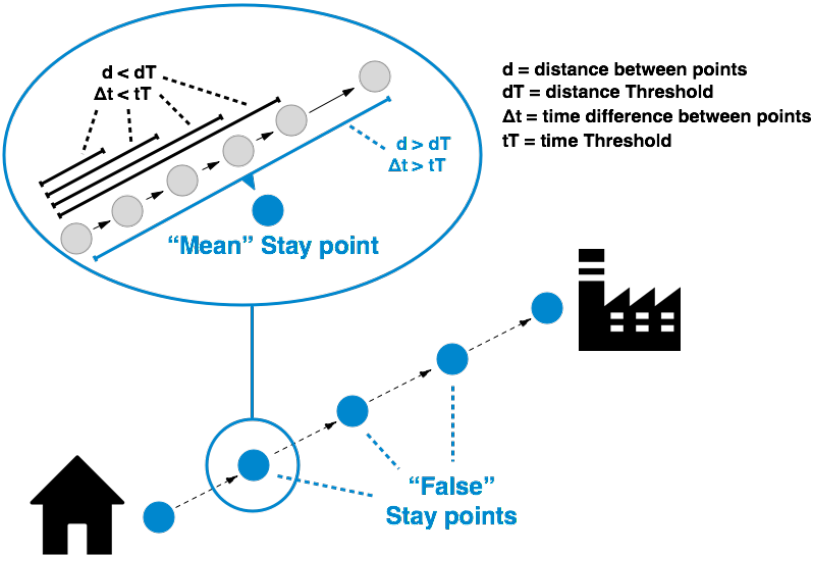

Figure 7.5: The segmentation problem of important places recognition (P2)

how regions with high speed (higher than the threshold set), in a trajectory between two locations, generate false stay points, again due to a not proper threshold value setup.

7.1.3 Motivations and Research Questions

The motivations behind this study are mainly related to difficulties emerging with the analysis of movements and places, which are based on data generated by users, manually or automatically with mobile devices. Raw data need several processing phases in order to clean noise, fix some detection, aggregating similar data, and finally get the set of positions which characterize an important place or a path between two of them. On this basis we want to study different approaches on movements analysis, in order to adopt the most effective one. We want to see the role of context in this domain, by analyzing how data of different nature, and from different sources, could be used as enrichment to impact the movements analysis, and in particular the important places categorization. Finally, we aim at investigating on how important is the new information obtained using data enrichment to better understand user habits.

More specifically, we focus on the following Research Questions:

Q1: What is the proper approach to enrich and analyze user movements data to understand habits?

- **Q1.1:** Is it sufficient to use a static approach with an overview over data, or is it more suitable a dynamic methodology to scan each position and analyze current context?
- **Q1.2:** How it is reasonable to cut-off the additional data from context? Dynamically computed thresholds are preferred over fixed ones?

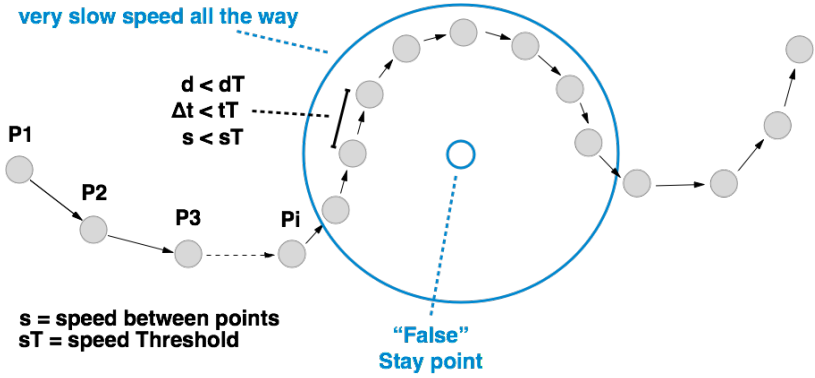

Figure 7.6: The thresholds problem of important places recognition, the slow-speed issue (P3a)

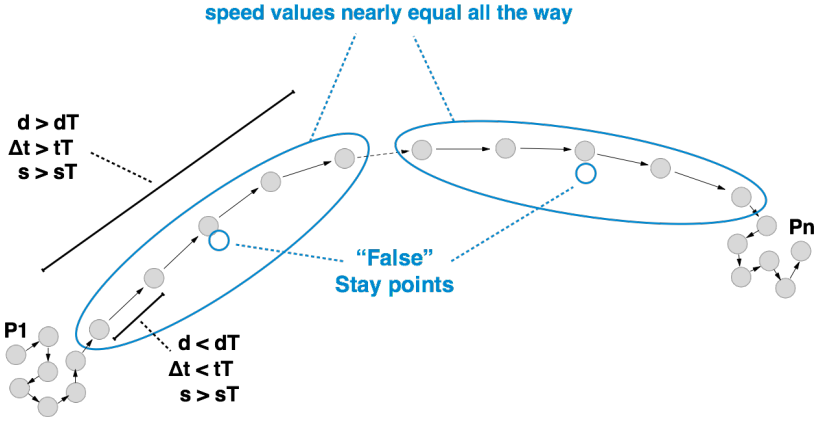

Figure 7.7: The thresholds problem of important places recognition, the costant-speed issue (P3b)

Q2: How the nature of the enrichment data affect the results?

- **Q2.1:** What kinds of contextual data are the most relevant to be used as additional data?
- **Q2.2:** Are all contextual data useful or there is some of those which could adversely impact the performance?

Q3: How important is the role of enrichment in terms of providing new information?

- **Q3.1:** Is it possible to exploit contextual data to define new features to better describe user movements and habits?
- **Q3.2:** Does the new information help to address the problems related to movements and places analysis, described in Section 7.1.2?

7.2 Preliminary Experiment

In order to understand the impact of problems described in Section 7.1.2 on real world user movements, and the effectiveness of the most used approaches in the literature, we have planned two evaluation tasks. The first one is based on an in-house dataset. Indeed, we built a mobile application (in two versions, for both iOS and Android smartphones) to gather real movement data from people. Basically, we needed a sequence of GPS points consisting in latitude, longitude, speed, timestamp and accuracy, to have a trajectory that represents how and where user moved. We have chosen a sample of 13 (Italian) users in order to collect a sufficient amount of GPS detections during 4 days of common daily activity. The second evaluation task involved the same group of 13 users, but on 4 days of movements related to 13 maps (one for each user) taken randomly from the GeoLife dataset [109]. The latter has been collected in (Microsoft Research Asia) GeoLife project by 182 users in a period of over three years (from April 2007 to August 2012: for the details see [11]).

Designing the two tasks, we paid attention to have different types of behavior, from frequent home-work travels to routines very stationary, also with different modes of transportation, e.g. motorized vehicles, bicycle, walk. We have estimated to collect (for the in-house task) and to choose (for the GeoLife task) data for 4 days for each user, in order to have enough detections to properly recognize behaviors and habits, since a lower number of days might not emphasize locations with high frequency and/or intensity.

Of course, a key difference between these two preliminary evaluation tasks is related to the users' knowledge about the datasets. In the in-house case each user evaluated the performance of the algorithms on its own data, being in the perfect condition to establish the ground truth. Instead, in the GeoLife case there was no ground truth about PPOIs available in the database, and our users were not acquainted with the Chinese regions of GeoLife. However, in each case users had the same skill and knowledge level in identifying the potential important places.

We implemented a set of algorithms to check what issues affect them. In particular we compare approaches with static and dynamics methodologies, with contextual information from the related positions and form different sensors, in order to study what contextual information could provide more benefits, and, on the other hand, what is not relevant. The first, named G, is a static approach based on the grid method described in [50], useful to see how the boundary problem affects the results on dataset with movements from different user behaviors and habits. The second one is based on a dynamic approach and only space threshold, exploiting context only to get objects distances, named S, as described in [52]. We have also implemented the T and V versions of threshold-based algorithms, since they have been often used in literature, even recently [50, 66, 88, 110], and they exploit more contextual information, such as, time and speed between couples of positions. Moreover, we have developed further versions of the latter algorithms with more parameters as thresholds and more data exploited as contextual enrichment, such as acceleration, and heading change, named A and H, respectively (like in [15, 107]), to see how the addition of parameters affects the PPOIs categorization.

We have run all algorithms to see the results on our datasets and make some considerations about the issues explained in Section 7.1.2. Observations on results showed

that:

- the static approach, namely the grid-based clustering, got variable performance due to different types of movement that need different cell sizes (P1a, P1b):
  - smaller cells allow us to recognize the right SPs, but adding a lot of false SPs;
  - larger cells generate the right number of SPs, but with wrong locations since the centroid is taken as the mean of all points in the cell;
- dynamic approaches with contextual enrichment fit well any type of movements readings;
- generally, to add new thresholds based on new parameters and new data from the context helped to discard false stay points;
- acceleration seems to be a too strict parameter, since too many points are discarded;
- heading change gives a low contribution to PPOIs categorization, anyway it helps to improve the precision of the importance recognition process;
- the segmentation problem (P2) is still present;
- to use fixed thresholds does not allows us to always have a perfect setup for all situations, due to the different types of movements (P3a, P3b);
- generally, the preliminary experiment encourages us to adopt dynamic approaches exploiting contextual enrichment with an automatic thresholds computation methodology (also helping to deal with "sensitive" parameters like, e.g., acceleration).

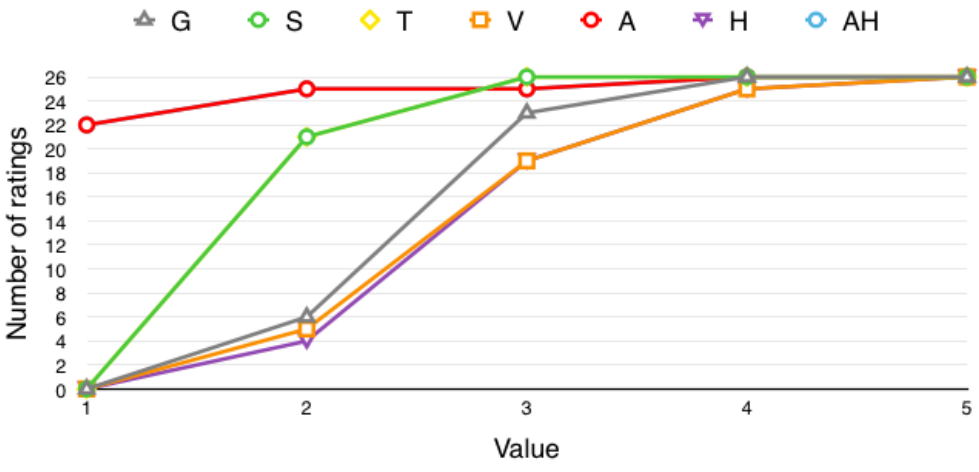

Figure 7.8: Preliminary experiment: cumulative rating distribution for all algorithms

Table 7.1: Preliminary experiment: algorithms comparison

Algorithm	Average rating	Precision	Recall	F-measure
G	2.88	**0.186**	0.806	**0.302**
S	2.19	0.026	**1**	0.051
T	2.19	0.041	**1**	0.078
V	**3.11**	0.138	0.846	0.238
A	1.23	0.125	0.063	0.084
H	**3.15**	0.160	0.835	0.269
AH	1.23	0.125	0.063	0.084

This preliminary experiment helped us to confirm how the context with additional data as enrichment plays an important role to improve the categorization process in all the above mentioned approaches, but some issues are still present and a further study could be performed to discover new improvements to adopt. Figure 7.8 illustrates the cumulative rating distribution for all the algorithms considered in the preliminary experiment (notice that in the figure the lines for S and T algorithms coincide, the same for A and AH). Table 7.1 shows the average ratings, precision, recall and F-measure reported by each algorithm. We can see that, despite the higher precision and F-measure of G w.r.t. V and H, users have preferred the latter two algorithms with better average ratings (3.11 for V and 3.15 for H vs. 2.88 for G). This can be explained considering that G does not discard any candidate SPs, but it simply clusterizes them. Hence, the user can be confused looking at the representation in the map, seeing many "spurious" points scattered around in a uniform way. Moreover, sometimes the grid-based approach does not identify the right coordinates of important places, due to the cluster centroid which is affected by the high number of points contained in the cell (which can be too large). Finally we ran the Wilcoxon test in order to verify if there are significant differences among the rating distributions got by the algorithms. The resulting p-values appear in Table 7.2. We can observe that there are statistical significances between several pairs of algorithms (where the p-value < 0.005). In particular, we can confirm again that increasing the number of parameters used as thresholds by the algorithms allow us to get a significant improvement, apart the cases of the threshold T which does not give any contribution and the threshold H which contributes slightly.

Table 7.2: Preliminar Wilcoxon test - p-values

	G	S	T	V	A	H
G	-	-	-	-	-	-
S	2.467e-05	-	-	-	-	-
T	2.467e-05	NA	-	-	-	-
V	0.01966	1.507e-05	1.507e-05	-	-	-
A	4.732e-06	3.69e-06	3.69e-06	4.1e-06	-	-
H	0.01073	9.044e-06	9.044e-06	1	3.586e-06	-
AH	4.732e-06	3.69e-06	3.69e-06	4.1e-06	NA	3.586e-06

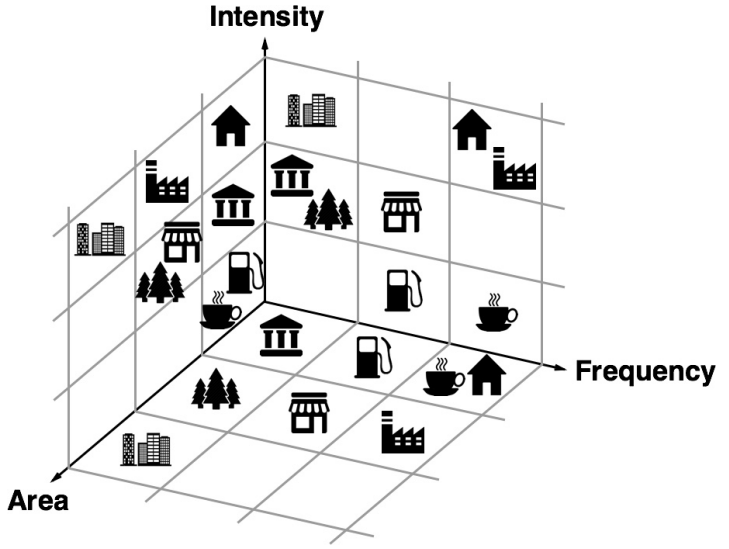

Figure 7.9: Some kind of important places positioned into the features space

7.3 Important Places Categorization System

The methodology we propose for the place categorization process is based on the discovering obtained during the preliminary test, so as to continue to exploit the context with the related data that provide additional information, but taking a further step towards the analysis of user context to improve the categorization process.

The key contribution and novelty of our approach rely on a mapping from the physical space (determined by raw positional data) to an *abstract* space, called *features space*. The latter allows us to consider as coordinates some *features* which are semantically related to users' habits and behaviors. To achieve this goal the context and data enrichment phase play an important role providing the new information which allow us to better understand new properties that characterize a place. Figure 7.9 shows an example of how places can be positioned in our features space, where the importance is defined by these new dimensions. The icons indicates some kind of common places, such as home, office, bar, mall and fuel station, but also a park, a travel to a city, or a visit to a museum. From a pragmatic point of view places represented in this space allow one to infer some similarities among them. In the example it is possibile to see how they may be related according to a couple of features. For instance, a bar and a fuel station have low values on *area dimension* and *intensity* but they differ in *frequency*. Or, home and bar may have the same *area dimension* and the same *frequency*, but people spend a different amount of time in those locations.

Thus, we can provide a deeper and more meaningful representation of PPOIs: for instance, in the following we will see how we can observe if a place is repeatedly visited or if it is visited several times during a longer period, but not with a sufficient intensity to be taken into consideration by previous techniques (this can be the case of, e.g., rendez-vous points, newsstands, bus stops etc.).

Our approach for PPOIs categorization consists in a method that analyzes a dataset of user movements readings, regardless of the type of technology used for tracking, even without any help from external knowledge sources. A point in the dataset just needs to be described by a set of coordinates to identify the location into a space, and a timestamp to understand the temporal order of detected points. On this basis it is possible to work on datasets with data gathered with various technologies, for instance WiFi triangulation inside a building, such as a mall or a museum, or by using a GPS for outdoor movements.

More precisely, we rely on a dynamic approach which analyzes user movements point by point, exploiting other points in the context, to identify stay points by checking thresholds, as described in Section 7.1. In this way we get a set of *Candidate Important Places*, due to the nature of user stay points which represent possible locations with particular meaning for the user. In more details, our approach is organized in two main steps, with a preliminary phase consisting in defining the values of the thresholds to use (based on the tracked user activity) as follows: the first one exploits a stay point computation algorithm to get a set of candidate important places; then the second step applies our feature-based technique to properly select the most important places for the current analyzed user. These steps are described in full details in the following sections.

We conclude this paragraph noticing that intensity, frequency or other similar features have already been taken into account in the literature. For instance, in [22] Chon et al. propose to combine external knowledge from crowdsourcing and social networks data, to automatically provide places with a meaningful name or a semantic meaning. In order to carry out this goal, they consider several factors such as *residence time* (indicating "stay behavior of users at a place tied with time-of-day") and *stay duration* (indicating "pattern of stay behavior without time-of-day"). More in general, the very concepts of features space and feature vector have been exploited in [95], in order to find similarities between users, starting from their location histories.

7.3.1 Preliminary Phase: Thresholds Definition

As preliminary phase, we address the threshold definition problem. As explained in Section 7.1, there are no fixed values for thresholds that fit perfectly for each user and for each dataset; therefore a brief reasoning may help to understand what kind of movements we are analyzing. During our preliminary experiment (see Section 7.2) we observed that changing the speed threshold highly affected the results for users with different use of the vehicles and transportation mode, and even the acceleration and heading change are strictly related to how users move routinely. Therefore, we define a method for extracting a good set of values for these three thresholds. We run a scan on the dataset in order to get information about the three parameters described above, paying attention on the median of non-zero values of speed, acceleration and heading change between each couple of points. This choice stems from the considerations discussed in Section 7.1; indeed we observe that the median of all speeds reached by the analyzed user, may be a good value to identify when user changes behavior. We adopt the same consideration for acceleration and heading change, in order to have a set of thresholds to use in the next step to build algorithm variants for comparison purposes.

About distance and time we keep the thresholds fixed. We set the distance threshold dT equal to 50 meters, and time threshold tT equal to 50 seconds. These are param-

eters set empirically, by observing a sample of user movements during the preliminary experiment (see Section 7.2), where we noticed that they do not strongly affect the stay points identification.

7.3.2 Step 1: Stay Points Computation

As second step we identify the user stay points, by using a dynamic approach which consists in a scan of all points in the dataset, in order to simulate and reproduce the movement, and exploits data in the context with thresholds based on some parameters to understand user behavior and recognize when and where users move or remain stationary in a location. To make possible a proper evaluation of our method, we implement several solutions of this dynamic and contextual approach; in particular, we want to compare earlier methods based just on space and/or time to others that also exploit speed, acceleration and/or heading change, as additional data to exploit from related points in the context. By observing the results of our preliminary experiment (see Section 7.2), we notice that space and time are not sufficient to properly determine the right set of stay points, and also other related works take into account other parameters [15, 107]. Moreover, acceleration and heading change were too strict as parameters of selections, in our heterogeneous dataset, and they have led the algorithm to discard too many stay points. Based on these observations, we chose to use space, time and speed parameters as thresholds for the stay point computation module. More formally, during the analysis of a point p_i and a point p_j, i.e. the next one in the user trajectory, we add the point p_i to the list of candidates for the stay point computation if one or more of the following constraints are satisfied:

$$\begin{aligned} distance(p_i, p_j) &\leq dT \\ timeDiff(p_i, p_j) &\leq tT \\ speed(p_i, p_j) &\leq sT \end{aligned} \tag{7.1}$$

During the computation we may also take into account the accuracy of coordinates detected during the movement tracking process. If the analyzed dataset provides the accuracy values for each point reading, it is possible to improve the parameters computation between two points. For instance, if we use a dataset with data gathered by using a GPS, we can discard coordinates with very low accuracy, in order to avoid weird values due to detection errors, or even we can exploit the instant speed detection, if the accuracy is good enough to make the value reliable. On this basis, our method checks the presence of the accuracy parameter into each entry of the dataset in order to exploit it for discarding data with low reliability, and to use the instant speed, if detected. If the dataset provides this additional information, we keep only data with $accuracy \leq 30$ meters[1], in order to avoid errors in distance computation and user speed analysis, due to problems with point data acquisition. For the speed computation we also take into account the instant speed as follows:

$$speed(p_i, p_j) = \begin{cases} \frac{segSpeed(p_i,p_j)+iSpeed(p_i,p_j)}{2} & p_j.acc \leq 10 \\ segSpeed(p_i, p_j) & \text{otherwise,} \end{cases} \tag{7.2}$$

[1] $accuracy \leq 30$ is a parameter set empirically, by observing the raw data.

where $p_j.acc$[2] is the GPS accuracy value for that specific detection, $iSpeed(p_i, p_j)$ is the average value of instantaneous speed detected by the GPS in points p_i and p_j, and $segSpeed(p_i, p_j)$ is the average speed from the point p_i to the point p_j in the user trajectory, i.e. the space segment $\overline{p_i p_j}$.

If $speed(p_i, p_j)$ is above the speed threshold, the user might be moving, thus we update the point scanning with $i = j$, in order to discard locations which could not be appropriate stay points. Otherwise, if user has low speed, we keep fixed p_i and perform a scan over the next points p_{i+k}, with $1 \leq k \leq (n-1)$, in order to detect locations to add to the list of candidates for the stay point recognition, focusing on distance and time thresholds, but also keeping checked the speed for the scan update. When the speed threshold is exceeded again, the list of candidates is processed in order to compute a *Mean Stay Point*, and the scan can continue with the next points. With this methodology we exploit the context related to the currently analyzed object in a dynamic way and under multiple point of views based on the chosen parameters to monitor. The context wideness is not fixed and it grows up until the thresholds will be not exceeded. In this way the context has been dynamically explored. Algorithm 2 illustrates the stay points computation process in detail.

Algorithm 2 SPs computation

Input: A set of user movement readings $P = \{p_0, p_1, ..., p_n\}$, a distance threshold dT, a time threshold tT, and a speed threshold sT
Output: A set of SPs SP

$i, j = 0$; $n = |P|$; $q = newPoint$
$CP = \{p_0\}$ ▶ list of candidate points
$SP = \{\}$ ▶ final list of SPs
while $i < n$ **do**
  $j = i + 1$
  while $j < n$ **do** ▶ $p_i, pj \in P$
    if $dist(p_i, p_j) > dT$ & $time(p_i, p_j) > tT$ & $speed(p_i, p_j) > sT$ **then**
      $q.coord = meanCoordInCP()$ ▶ $\forall p_k | i \leq k < j$
      $q.arrivalTime = p_i.time$
      $q.leaveTime = p_j.time$
      $SP.insert(q)$
      $i = j$
      $CP = \{p_j\}$
      break
    else
      if $speed(p_i, p_j) \leq sT$ **then**
        $CP.insert(p_j)$
      end if
      $j = j + 1$
    end if
  end while
end while
return SP

With the same methods described in the algorithm for the thresholds definition and in Algorithm 2 we implemented several versions of the stay point computation algorithm based on different thresholds, in order to compare the performance and understand what are the most useful set of thresholds for user behavior analysis. Table 7.3 shows all variants implemented for the comparison and evaluation process.

[2] $p_j.acc \leq 10$ is a parameter set empirically, by observing a set of GPS detections in several signal acquisition conditions.

Table 7.3: Stay points computation algorithms variants

Algo name	Thresholds
S	space
T	space, time
V	space, time, speed
A	space, time, speed, acceleration
H	space, time, speed, heading change
AH	space, time, speed, acceleration, heading change

7.3.3 Step 2: Places Categorization through Importance Recognition

The main idea inspiring this step of our method is to map physical locations to an *abstract* space defined by a set of features more semantically related to users' habits and behaviours. For instance, a candidate feature is the frequency of visits, since users tend to behave similarly in everyday's life. Thus, in order to define a procedure for the important places recognition, it is useful to observe users' movements across a period of time longer than a single day[3]. In other words we want to explore the possibility of superimposing the locations visited by user several times in order to extract additional semantic information, and possibly refining the results of the previous phase. Hence, to implement such strategy, we consider new parameters to describe locations, alongside latitude, longitude and timestamp, that may help to improve the recognition process.

First, we modeled the user movements readings into a three-dimensional space where a point is described by the three original raw data gathered by sensors (latitude, longitude and timestamp), in order to have a distribution of points into this space that reproduces the original user movements (see Figure 7.10 (left)). We observed how the data is divided into groups, nearly in layers, which approximately represent the days when user performed the activity. Therefore this aspect makes possible further analysis and helps to get more information form each place exploiting the context in which they are located. In this model the time component, combined with the other spatial components, plays an important role as contextual parameter, since it allows us to identify the relationships among locations in terms of repeated visits during the analyzed days. On this basis we define a set of three features to describe each important place (PPOI) as a vector $PPOI = \langle A, I, F \rangle$, where A, the $Area$ of the PPOI, is a value which indicates the diagonal extension of the rectangular region that spans over all points involved in the stay point computation. As explained in Section 7.1, when users visit locations tend to not stay perfectly stationary, but to move around a delimited area. We also keep the set of physical coordinates which describe it, in order to also represent it graphically for user-testing purposes, and for checking potential overlaps. The feature I, the PPOI $Intensity$, is a value which indicates how many times the user position has been detected inside the PPOI's area. Finally, the feature F, the PPOI $Frequency$, indicates how many times that location has been visited by the user, thus a parameter that increments its value each time the user came back for another visit in that place.

[3]Otherwise, activities of a single day may escape from the usual routine and could easily hinder the recognition process.

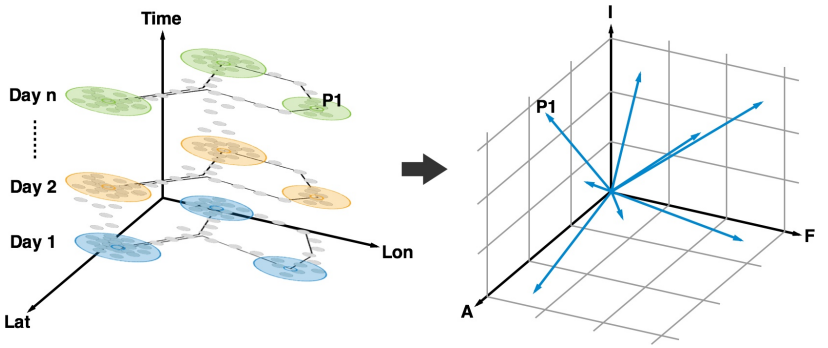

Figure 7.10: The feature-based approach that moves places into the feature space

Formally, we have the following map ($PhysSpace_{SP}$ is the physical space and $FeatSpace_{SP}$ is the features space):

$$feature \quad : \quad PhysSpace_{SP} \longrightarrow FeatSpace_{SP}$$

$$feature \quad : \quad \begin{array}{l} \langle sp.latitude, \\ sp.longitude, \\ sp.timestamp \rangle \end{array} \quad \longmapsto \quad \begin{array}{l} \langle sp.Area, \\ sp.Intensity, \\ sp.Frequency \rangle \end{array}$$

Figure 7.10 shows how we model movements data into the three-dimensional space, defined by *latitude*, *longitude* and *timestamp*, and how we map stay points from the physical space to the features space based on the new three dimensions A, I and F.

This feature-based approach, with the help of contextual enrichment, makes possible a refinement step to emphasize locations visited intensively and/or repeatedly, and also to filter out the false stay points that the previous phase was not able to discard. In this phase we first analyze the SPs computed during the previous step, in order to calculate the stay point area A based on the detections involved in the stay point computation. Then we extract the number of detections inside that area to compute the value of I, and get the PPOI intensity, and we also set the frequency of each SP to 1 as initial value. With all SPs described into our features space we can analyze both the A and I distribution over the entire dataset, in order to better understand user behavior and define new thresholds that may help to filter out some places not important for that user. We set an area threshold aT equal to 3 km to run a pre-filtering process which discards SPs with diagonal area ≥ 3 kilometers[4]. This operation helps to identify the kind of false stay points described in Section 7.1 as problems P3a and P3b, where a wrong threshold set may cause SPs generated by detections that span over a wide space. We also observed how a different use of vehicles and mode of transportation generate different density of detections, with the consequence of having higher I values for users that usually move slower. This issue led us to define an intensity threshold iT in order to discard SPs with I too low in proportion to the values obtained in the rest of the dataset.

[4] $aT \geq 3$ is a parameter set empirically, by observing user movements during our preliminary experiment described in Section 7.2.

Moreover, with particular attention on the I values of adjacent SPs to recognize where the segmentation problem may have occurred (see Section 7.1). On this basis, we define the intensity threshold $iT = max3consec(intensities)$, where $intensities$ is the array with all intensity values of each SP, and the method $max3consec$ returns the maximum value of intensity that in the SPs sequence is present at least three times in a row (up to some tolerance threshold for dealing with measurement errors and small deviations[5] from the maximum value). This technique helps us to recognize where the user is moving, and also where he is generating the same intensity values. By selecting the maximum value, we can discard the false stay points induced by the scenario described in the conceptual problem P2. Moreover, automatically computing the intensity threshold as previously described, we avoid locations where users stopped just once and for an amount of time not so remarkable as the time spent in home, office, supermarket, etc. Such places may be intersections with traffic lights which block vehicles for a long time, traffic-clogged streets, or rail crossings. Based on these two thresholds we run a pre-filtering process, to have a more accurate subset of SPs and proceed to take into account the frequency of visits.

By analyzing SPs under the time point of view, it is possible to observe how the temporal context represents a region where the SPs present might be the same place visited in different day or moment, due to the user routine. Using the same visualization depicted in Figure 7.10 we can "flatten" the representation of SPs with related areas with raw points in order to analyze how close they are. It is reasonable to assume that if two SPs have a considerable amount of area in common they could be considered the same place, and therefore visited two times. The use of this additional information from the SP's surrounding area allows us to enrich the SP with new knowledge regarding the new feature we are interested in. On this basis, by analyzing SPs sequentially it is possible to check if their areas overlap with other ones very close, in order to get information about locations visited repeatedly. If the areas of two locations overlap with an intersection region $\geq 50\%$[6] of one of the current analyzed areas, they may be considered to represent the same place. In Figure 7.11 it is possible to see an example of two visits on the same geographic area where for the locations a and b there are very similar detections on both days, therefore they represent the same important place. In that case the intensity values will be summed, the area will be their union, and the frequency will get a value equal to 2 because of the number of visits. Otherwise, the locations c and d have detections with an area overlap $< 50\%$, therefore they will be considered as two separated places. After this filtering step, we repeat the process of merging areas several times until we get just separated regions, which identify our important places. As final phase, we run again the filtering process in order to clean out PPOIs that may have been generated with too large areas. All phases of our method named AIF are illustrated in Algorithm 3.

[5]For instance, the user slightly changes speed while driving along a highway.

[6]$overlap \geq 50\%$ is a parameter set empirically, by observing user movements during the preliminary experiment.

Algorithm 3 AIF computation

Input: A set of user stay points $SP = \{sp_1, sp_2, ..., sp_n\}$
Output: A set of important places $PPOI$

$aT, iT, fT = 0$; $q = newPoint$
$areas, intensities = \{\}$ ▶ arrays with all values
$PPOI = \{\}$ ▶ final list of important places
for sp_i *in* SP **do** ▶ pre-filtering
    $insertInAreas(sp_i.computeArea())$
    $insertInIntensities(sp_i.computeIntensity())$
    $sp_i.freq = 1$
end for
$aT = 3$ ▶ empirically set to remove SPs with area diagonal > 3 km
$iT = max3consec(intensities)$ ▶ the maximum value repeated at least three times in a row
$PPOI = preFiltering(SP, aT, iT, areas, intensities)$
$overlaps = true$ ▶ to check overlaps during the points scan
while $overlaps == true$ **do** ▶ points merging
    $overlaps = false$
    for p_i *in* $PPOI$ **do**
        for p_j *in* $PPOI \setminus \{p_i\}$ **do**
            if $overlap(p_i, p_j)$ **then**
                $overlaps = true$
                $q.area = mergePointsAreas(p_i.area, p_j.area)$ ▶ A
                $q.intensity = p_i.intensity + p_j.intensity$ ▶ I
                $q.frequency = p_i.frequency + p_j.frequency$ ▶ F
                $PPOI.add(q)$
                $PPOI.remove(p_i, p_j)$
                break
            end if
        end for
        if $overlaps == true$ **then**
            break
        end if
    end for
end while
for p_i *in* $PPOI$ **do** ▶ post-filtering
    $PPOI = postFiltering(PPOI, aT, iT, areas, intensities)$
end for
return $PPOI$

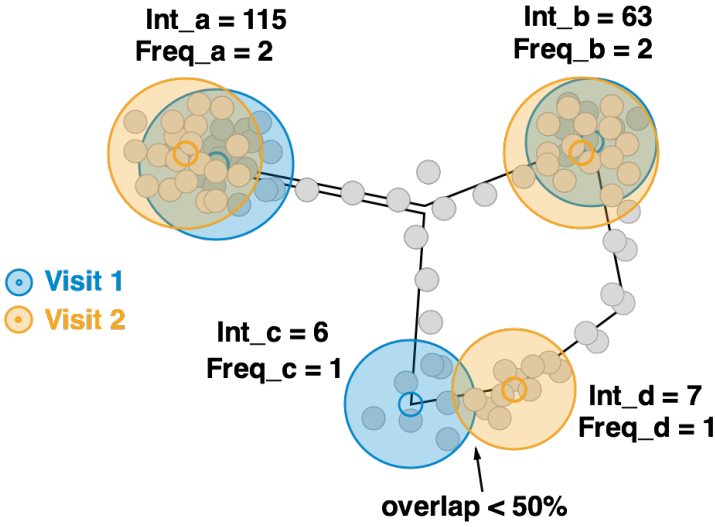

Figure 7.11: Example of overlapping activities

7.4 Experimental Evaluation

7.4.1 Experimental Design

To evaluate several approaches to important places categorization, and to benchmark our proposed solution with related contextual enrichment effectiveness, we run a set of algorithms over the GeoLife dataset. We implemented the set of threshold-based algorithms described in Section 7.2, to make possible a comparison among the approaches used in the literature. As final algorithm, we have implemented our solution *AIF*, as defined in Section 7.3.3, to compare it to the other approaches. Several other approaches in literature rely on users movements data obtained by using platforms or services which allow users to track specific positions, therefore not a set of raw data which represent the real complete movements. On this basis, this evaluation aims at comparing approaches which make only use of raw positional data, not affected by particular places tracked, or other information from other user activities.

We have selected a sample of 16 people to evaluate the results, distributed as follows:

- 62% men, 38% women (all Italians);
- 62% with age between 21 and 30, 38% more than 30;
- 88% with very good familiarity with smartphones;
- 56% with intensive use of map services, 25% with intermediate use and 19% occasional user;

As in the preliminary experiment, all of them were not familiar with the geographic regions in GeoLife, due to the different nationality: GeoLife data have been collected in China, while our participants were Italians. Since the GeoLife dataset does not contain

information about PPOIs, we asked all participants to indicate the most important places in order to build a Ground Truth to then evaluate the place categorization process. This fact yielded the positive effect that all the participants had the same skill and knowledge level in identifying the potential important places. We have defined a test protocol providing detailed instructions to participants so as to guide them during the PPOIs indication for the Ground Truth, and then for the evaluation, also providing information about the aspects to take into consideration. We have implemented a testing tool for them to show on a map some randomly selected sets of GPS detections form GeoLife dataset, with attention on choosing at least 4 consecutive days of movements readings. The participants had available a heatmap to better understand the original user movement and properly indicate PPOIs, and then evaluate the places categorized by the algorithms, showed as pins on the map. Each of the 16 users interacts with 4 maps, randomly extracted from a pool of 16 we selected for the experiment, in order to have a final evaluation composed of 4 different tasks for each map. During the evaluation of place categorization, the tool displays sequentially and randomly maps with pins computed by one of the algorithms previously described, in order to make not clear to participants how to associate the algorithms with the corresponding suggestions. This is a precaution to not affect them with clues during the test. During an evaluation a number between 1 and 5 indicates how they judge the overall PPOIs categorization. The meanings of the rate values are the following:

1 - SPs retrieved $\leq 20\%$;

2 - SPs retrieved $> 20\%$ and $\leq 50\%$ or a very high number of false SPs;

3 - SPs retrieved $> 50\%$ and $\leq 80\%$ or $> 80\%$, but with an high number of false SPs;

4 - SPs retrieved $> 80\%$ and $\leq 90\%$ and zero or a very low number of false SPs;

5 - SPs retrieved $> 90\%$ and zero or a very low number of false SPs.

Moreover, they were requested to indicate which pins properly represent PPOIs visited during the tracked activity, and also how many have been missed. To evaluate the results we used the participants judgments to compare the important places provided by the algorithms with the ones indicated as Ground Truth. We calculated some standard IR metrics, such as Precision, Recall and F-measure, (i.e., the harmonic mean of Precision and Recall) of each algorithm, in order to measure the aspects we are interested in. We want to analyze how many important places are properly categorized, how many are missed by the systems, and a overall evaluation of the performance; on this basis the three adopted metrics fit for the purpose.

7.4.2 Results

Results are reported in Figure 7.12 and Table 7.4. The figure shows the cumulative distribution of the ratings obtained by each approach; the table shows, besides the average rating for each algorithm, also its Precision, Recall, and F-measure. The rating distribution and the average ratings show how the S algorithm obtained many 1-value ratings, due to the low filtering that it applies with the single threshold approach, thus getting a mean rate equal to 1.16; the T solution has been evaluated slightly better

but most of rates still remains low; the adding of speed improved the performance as we expected; the A algorithm instead has worsened the identification process due to the acceleration parameter which has made too strict the PPOIs recognition process; H, based on the heading change parameter, obtained a good performance but also a minimal improvement over V; the algorithm AH has been penalized by the use of acceleration; finally, our proposed method AIF collected a lot of positive evaluations, obtaining a mean rate equal to 3.59, the highest score among all the compared algorithms.

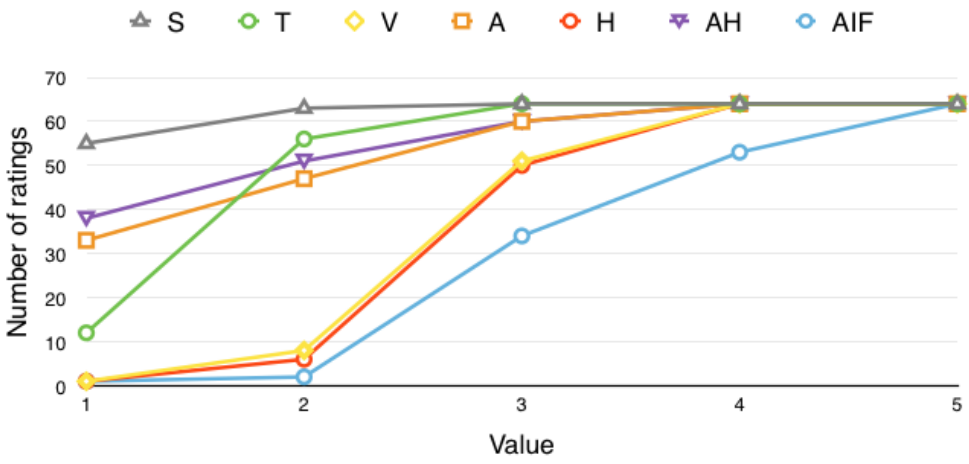

Figure 7.12: Cumulative rating distribution for all algorithms for important places identification

Table 7.4: Algorithms comparison

Algorithm	Average rating	Precision	Recall	F-measure
S	1.16	0.004	**0.992**	0.007
T	1.94	0.009	**0.992**	0.019
V	3.06	0.126	0.657	0.211
A	1.81	0.286	0.217	0.247
H	3.11	0.131	0.657	0.219
AH	1.67	0.172	0.161	0.166
AIF	**3.59**	**0.370**	0.606	**0.459**

The simpler methods, such as S and T, got the higher Recall values but with Precision very low, due to the filtering process that discards few false stay points, and provides a final set of PPOIs not so much different from the original set of movements readings. By adding more sensors' data from the context, and thus increasing the enrichment, the identification process improved, providing more accurate set of PPOIs. Moreover, the introduction of an automatic threshold algorithm computation has further improved the results. V and H solution increased the Precision, also keeping good Recall values. But

the use of acceleration has reduced a lot the Precision of algorithms A and AH, obtaining very low performance in every aspect. The AIF solution has proven to be the most accurate method, with the highest Precision and a good Recall, obtaining a good overall evaluation with the highest F-measure. The results confirm how AIF has improved the PPOIs categorization process by providing few false stay points, and guaranteeing not to lose too many important locations. Moreover, it provides a less confusing visualization on the map, and it is less affected by the type of users' movements. The context played a crucial role in this study, where space and time are the principal characterizations of user movements, and the enrichment phase results to be very important both in the two steps of our approach. During the first step the analysis of related locations led us to obtain a better SPs identification, then in the second step, the context and the data enrichment have been the key elements for achieving the goal we have set to overcome the problem stated.

7.4.3 Statistical Significance

Moreover, we have run a statistical test to determine whether there are any significant differences between the means of ratings got by the algorithms. Due to the nature of the data with non-normal distribution we run the Wilcoxon test in order to verify if datasets have significant differences.

Table 7.5: Wilcoxon test - p-values

	S	T	V	A	H	AH
S	-	-	-	-	-	-
T	1.6e-12	-	-	-	-	-
V	1.415e-13	6.343e-14	-	-	-	-
A	4.623e-07	0.1731	2.778e-11	-	-	-
H	1.109e-13	1.364e-13	0.1489	1.903e-11	-	-
AH	2.928e-06	0.004182	5.605e-12	0.003353	5.251e-12	-
AIF	1.317e-12	1.086e-12	5.844e-09	1.364e-13	2.752e-08	1.107e-13

Table 7.5 shows all the resulting p-values for each couple of algorithms to compare. We can observe that the most of them got very low p-value, lower than the standard Wilcoxon threshold 0.05. Therefore, this output indicates a statistically significant difference between means, and consequently a relevant improvement in performance for those algorithms. It is possible to notice how the low performances of T and A are not statistically different. Finally, the use of the heading change threshold did not bring a significant improvement when used in algorithm V, and it provided only a small noticeable improvement when used in algorithm A.

7.4.4 Final Remarks and Discussions on User Movements Analysis

In this section we want to remark the main results and the most important discoveries obtained with our experiments on user movements, in order to have an overview of our studies which allow us to stimulate discussions and new observations. In Table 7.6 we

summarize all experiments on user movements listing the subject involed, the aims, the datasets used, and finally the results and observations.

Preliminary Experiment: The first experiment let us to obtain an overview of the compared algorithms performance. The grid-based algorithm (G) suffered of the boundary problem, due to the different cell sizes which are not well suited for every case. It got low Precision, good Recall, but the main problem was related to the low ratings obtained by users who found it confusing in providing PPOIs. The Dynamic approaches S, T, and V resulted to be better as the number, and different kind, of additional data has been provided. V got the highest ratings among the three, Precision slightly higher, but lower Recall, due to filtering process which rewards the point cleaning phase at the expense of retrieving an high number of them. This fact confirms the enrichment effectiveness and how useful is to add more types of data. Acceleration (A) proved to be too strict, indeed it worsened the results when used in A and AH. The heading change resulted to be a useful additional parameter, but providing just a little improvement. Therefore it is reasonable to select a proper set of parameters to keep based on the task needs.

Extended Experiment: The second experiment aimed to compare the algorithms used in the preliminary one to our proposed approach AIF. S, T, and V confirmed again the effectiveness to have a dynamic approach with contextual information to use and data enrichment. Again, the use of multiple data types helped to raise up the final ratings obtained. The context analysis allowed us to get a better set of PPOIs which also made less confusing the evaluation for users. Acceleration remains too strict, and Heading change with a low contribution. Our proposed method AIF outperforms other methodologies in every metric: it got best ratings, higher P and R. The results confirmed how AIF provided a better set of PPOIs with few false points, and guaranteeing not to lose important places.

In general the enrichment process on user movements resulted to be a useful phase to improve the categorization process. The analysis of the context, with data of different nature, emphasized how new features could be extracted, in order to describe user movements under different points of view. The feature based approach allowed us to have a more social description of what places represent for users related to their daily routine. The study we presented highlighted how the use of data from different sensors not always contribute to improve the results; indeed the acceleration resulted to be a data type to avoid. Even in combinations with others does not provide useful information, lowering the performance. Moreover, the contextual data represented the elements that allowed us to make a step further toward a social analysis, and this seems the right direction to follow to improve user models with information more semantically related to users activities. Differently than other solutions proposed in literature, our categorization approach does not rely on positions tracked by users using particular services, therefore this fact makes the proposed approach flexible and suitable to be embedded in any real applications. Moreover, it is not limited to users but it could be also adopted for tracking vehicles and object, making the system ready to use for different purposes.

Table 7.6: Overview of experiments results on User Movements.

Experiment	Subject	Analysis	Results	Observations
Preliminary	Grid (G)	Static approach based on grid.	Got variable performance due to	Smaller cells get right SPs, lot of false SPs.
(Datasets:		Boundary problem	different types of movements	Larger cells get right SPs, wrong locations.
in-house	Space (S)	Dynamic approach with	Low average rating,	Dynamic better than Static.
and		only space data	high R but very low P.	too few data to have good P.
GeoLife)	Time (T)	Dynamic approach with	Low average rating,	Dynamic better than Static.
		space and time data	high R but very low P.	too few data to have good P.
	Speed (V)	Dynamic approach with	Higher average rating,	Enough data to get
		space, time, and speed data	higher P and R.	good performance.
	Acceleration (A)	Dynamic approach with	Low average rating,	Accel. is too strict,
		space, time, speed, and accel. data	lower P and very low R.	too many points discarded.
	Heading (H)	Dynamic approach with	Average rating a little higher than V,	Heading gives low contribution
		space, time, speed, and heading data	P and R a little higher than V	
	Accel. and	Dynamic approach with	Low average rating,	Accel. is too strict,
	Heading (AH)	all previous sensors data	lower P and very low R.	It cancels Heading benefits.
Extended	Space (S)	Static approach based on grid.	Variable performance,	Low filtering applied.
(Dataset:		only space data	many 1-value ratings.	
Geolife)	Time (T)	Dynamic approach with	Slightly better than S, with high R	Bad filtering.
		space and time data	but still low ratings and P.	
	Speed (V)	Dynamic approach with	All performance improved .	Speed helps in filtering.
		space, time, and speed data		
	Acceleration (A)	Dynamic approach with	Accel. worsened the general	Accel. is too strict.
		space, time, speed, and accel. data	performance.	
	Heading (H)	Dynamic approach with	Ratings, P, and R	Heading slightly helped.
		space, time, speed, and heading data	raised a little.	
	Accel. and	Dynamic approach with	Accel. penalized performance.	
	Heading (AH)	all sensors data		
	Area, Intensity, and	Feature based approach with	Highest scores for ratings,	Most accurate method thanks
	and Frequency (AIF)	Area, Intensity, and Frequency	P and R.	to few false SPs, higher P and R.

IV

Conclusions

8

Final Remarks and Discussion

In this dissertation we presented a study on the impact and effectiveness of data enrichment in systems which aim to analyze user generated data. In this final chapter we summarize our research, pointing out the main achievements with final remarks and discussions on both studies (short texts and user movements), and finally we present some future research directions.

8.1 Research Summary

In the first part of this dissertation, we introduced in Chapter 2 the concept which led us to design our studies on the effectiveness of data enrichment on the two analyzed fields: short texts and user movements. We described how we can place users contents in a geometric space to represent their relationships using context, distances, and comparing their properties. We highlighted how this concept models users' objects, and also the studies of enrichment effectiveness we focus on, allowing us to analyze and exploit the related objects in the context under several points of view, in terms of quantity, distance, and properties. Moreover, in the first part we provided a general state of the art of research work related to the two analyzed fields. We provided a general perspective on studies on short texts in Chapter 3, in particular regarding short texts categorization, enrichment, short texts and external knowledge, short texts and news, and short texts for user modeling. Then we presented research works related to user movements in Chapter 4, by focusing on trajectories modeling, highlighting recent interests in more user-oriented methodologies, to analyze social habits and behaviors. We described social approaches used to extract new semantic knowledge, showing how it is possible to combine literature about the study of trajectories and the one related to studies on social habits and behaviors, as further step on trajectory modeling. We illustrated studies on human mobility, place categorization with emphasis on importance recognition.

In the second part of the dissertation, we present the first case study: the short text analysis, and related reasoning about data enrichment use and effectiveness evaluation. We introduced our novel approach to enrich short texts with a new set of words extracted from documents of the same temporal context, then we presented in details all the

experiments we designed to extensively study the data enrichment effectiveness and properties. The first group of experiments, focused on different enrichment strategies, aimed to analyze No-enrichment, Append, and Merge alternatives, which differed in how they used the additional data. We highlighted how merging data resulted the best approach due to the terms cleaning phase which removes noise. Going deeper, the cut-off experiment confirmed the benefits obtained using dynamic approach, which keeps a limited set of terms with a cardinality dynamically computed. The second group of experiments was based on the properties of data used as enrichment. We analyzed a set of five characteristics of the datasets, namely, Volume, Variety, Type, Structure and Freshness, to study their impact on the results, and how we can setup the enrichment phase in order to maximize the performance. The extensive study conducted has revealed that as enrichment source it is recommended to have a dataset with high documents volume, high variety, but better with news documents, and possibly fresh. This issue does not allow us to have the perfect setup, because having only fresh news implies lower volume and vice versa, therefore the system has to be configured according to the needs. Moreover, in the second part we introduced in Chapter 6, a method for computing user similarity. Exploiting the text enrichment methodologies presented in Chapter 5, we built a model based on a network representing the semantic relationships between the words occurring in the same tweet and the related topics. This study led the data enrichment analysis a step further, evaluating its effectiveness on a set of short texts, not only one or once a time, which represent user interests. We concluded that chapter presenting a preliminary experimental evaluation on a limited dataset which confirmed again the effectiveness of the enriched data. The resulting model appeared richer, with more information about users. The use of network and related centralities emphasized similarities where we expected, better than using a simple count, and also providing additional information to better understand what users have in common, even under several point of views.

We dedicated the third part of this dissertation to the second case study: the user movements analysis, with investigations on the role and effectiveness of data enrichment and context. We introduced a novel approach in Chapter 7 to address the problem of places categorization. The use of contextual data to enrich the user locations, resulted to be a useful methodology which helped to identify the important places visited by users during their daily routines. The proposed approach has been presented in two steps to describe first the identification of stay points and then the recognition of locations importance to categorize places. A feature space has been presented as model to map places and to represent them with new dimensions, such as, the area underlying the stay point, its intensity and its frequency. We compared several algorithms which exploited different data from different sensors. The use of space, time, and speed combined values as enrichment helped in raising the places categorization performance. On the other hand, acceleration worsened the categorization, and heading change provided a minimal improvement; therefore, we preferred not to use those data. The proposed approach based on context and data enrichment, namely AIF, obtained the major number of positive evaluations and the highest mean rate among all the compared algorithms.

The results obtained in all our studies and experiments confirmed the effectiveness of data enrichment process and provided useful information on how to enrich the data, with details on quantities, distances and properties. Moreover, the context played an

important role in providing the right related data to use, and allowing us to analyze different portions of surrounding regions, with related effects on the systems. We conclude the research summary listing the main contributions of this thesis as follows:

- a new methodology and a system to categorize short text exploiting contextual enrichment and external knowledge;
- an extensive study on text data enrichment with reasoning about enrichment strategies and data properties more suitable for that purpose;
- a new network-based user model to compute similarities, which exploits data enrichment and network centralities;
- a new approach for places categorization through importance recognition, which exploits contextual data to enrich user locations;
- a study of user movements data enrichment with reasoning about data type suitable for enrichment and context wideness with distances.

8.2 Final Considerations

With our research work we provided an extensive study on data enrichment with details on how to properly apply that process with an efficient methodology, and also with a sort of guide on how several documents properties impact the results of categorization. This work allow us to have information about how to improve the categorization processes in the two analyzed domains which represent two important aspects of users daily routines, and led us to study their behavior and habits. In literature there is no extensive studies on enrichment process and in particular on these two domains; also other works which address this topic are not completely flexible. Some of those rely on pre-defined structures, others on some peculiarity of social platforms from which they extract data, or they exploit hyperlinks, hashtags or other textual elements.

In this work we also designed and developed new approches and prototype systems for short texts and places categorization, with the aim to be as flexible as possible and not dependent on any structure. This fact makes possibile to continue these studies in order to achieve real applications and systems to include in mobile devices. It will allow us to run further studies on larger datasets produced by larger user communities. Moreover, with a complete tracking systems which analyzes data on both domains it will be possibile to find relationships between what users posts and the places they visit, in order to expand the concept of context and to take a step further to the enrichment process cross-domain.

8.3 Future Research Directions

The contextual data enrichment resulted to be, on both studies, a powerful tool to improve the effectiveness of systems in which user generated data are involved. The context analysis and the enrichment phase design have a large scope and we have obviously not

addressed all the relevant issues. In this section we present some future directions and we discuss some issues that deserve further investigation.

To take our studies a step forward we can first expand the experiments on text enrichment in order to analyze other properties, in particular more document types and from different sources. We plan to carry on further experiments using articles written not only on online newspapers, but also research reports, scholarly articles, tech articles, etc. We can also further investigate on sources and their properties in order to study the different impact they may have. The nature of a document included into a dataset might be dependent on certain dataset "meta-characteristics" such as what kind of users generated the documents, indeed they could be general users or specifically selected users specialized and skilled in a particular field; what kind of format has the source, indeed for instance a forum website contains documents with a structure and a number of words different than a blog or a social network, etc. Another interesting future work could be to run all the experiments we designed on short texts in different languages. It would allow us to study if there is a different impact of enrichment with languages particularly simple or with complex grammars and large vocabulary. The short texts analyzed in these studies have been extracted from Twitter at this stage, but in the future we plan to test other kinds of short texts, such as instant chat messages, mobile text messages, posts on Facebook, and online comments on eCommerce websites. The context and enrichment phase could affect the categorization process in different ways and diverse text types might need different settings. The user model, built exploiting enriched short texts, will be extended with multilanguage support to test cross-language similarities and we plan to run a complete evaluation on larger texts and users datasets.

Regarding user movements, we want to analyze different features to use as place representation, in order to further study the context related to each visited place and find new data to exploit to improve the place categorization process. We want to analyze user movements types, and study how the context and data enrichment could help in discovering information about what vehicles people use. Moreover, we plan to address the problem related to categorization of places co-located inside the same building, or within a small area, which have to be considered as separated entities. To follow the design of our study of short texts, we plan to continue the study of user movements with a user model also in that field, to analyze the data enrichment effectiveness carried to another level of evaluation. We will use a network-based approach to model the relationships among important places visited by users emphasizing the daily routines with weighted paths.

An interesting and recently addressed study is the connection between the two fields we analyzed. Both short texts and user movements are user generated data, and in particular they constitute the information provided by people during their daily routine, therefore the presence of semantic relationships is an issue worth to study. Exploiting the network-based approach we previously used we plan to build a uniform approach for user modeling based on a multilayer network: we want to design a two layer network, in which each layer conveys a different kind of activity, in particular texts posted on Twitter and spatio-temporal data related to users movements. Using multiple datasets, each one consisting of online activity and visited locations during common daily routines, we want to explore possible interconnections and correlations between the data represented in the two layers, and to reason about the real organization of information which represent

users. According to our knowledge, this is a novel approach: in the literature the attempts to interconnect short texts published on social networks and users movements heavily rely on the georeferenced data associated to the former, whereas we also exploit data coming from sensors of mobile devices, allowing us to model users behaviors even when they are not interacting with social media. Hence, this model allows us to study how distinct types of activity affect the network structure and statistical properties, in particular concerning the degree distribution and centrality indices of the network.

Finally, it could be also interesting to look at computation performance issues to address the problems related to embedding enrichment phases in mobile applications.

A

List of Publications

The work performed in these years by the author of this dissertation has been presented in the following publications. The papers are grouped by type and listed in reverse chronological order.

International Journals

- Dario De Nart, Dante Degl'Innocenti, and Marco Pavan. *Finding Esteemed Users with Abstract Argumentation.* ACM Transactions on Internet Technology (TOIT). [Under review]

Chapters in Books

- Donatella Gubiani and Marco Pavan. From trajectory modeling to social habits and behaviors analysis. In Maturo, A., Hošková-Mayerová, Š., Soitu, D.-T., and Kacprzyk, J., editors, *Recent Trends in Social Systems: Quantitative Theories and Quantitative Models*, pages 371–385. Springer International Publishing, Cham, 2017.

- Marco Pavan, Stefano Mizzaro and Ivan Scagnetto. Mining movement data to extract personal points of interest: A feature based approach. In Lai, C., Giuliani, A., and Semeraro, G., editors, *Information Filtering and Retrieval, DART 2014: Revised and Invited Papers*, pages 35–61. Springer International Publishing, 2017.

International Conferences

- Marco Pavan, Thebin Lee, and Ernesto William De Luca. Semantic enrichment for adaptive expert search. In *Proceedings of the 15th International Conference on Knowledge Technologies and Data-driven Business*, i-KNOW 15, pages 36:136:4, New York, NY, USA. ACM, 2015.

- Marco Pavan and Ernesto William De Luca. Semantic-based expert search in textbook research archives. In *Proceedings of the 5th International Workshop on Semantic Digital Archives*, volume 1529 of *CEUR Workshop Proceedings*, pages 18–29. CEUR, 2015.

- Stefano Mizzaro, Marco Pavan and Ivan Scagnetto. Content-based similarity of twitter users. In Hanbury, A., Kazai, G., Rauber, A., and Fuhr, N., editors, *Advances in Information Retrieval: 37th European Conference on IR Research, ECIR 2015, Vienna, Austria, March 29 - April 2, 2015. Proceedings*, pages 507–512. Springer International Publishing, Cham, 2015.

- Marco Pavan, Stefano Mizzaro, Ivan Scagnetto, and Andrea Beggiato. Finding important locations: A feature-based approach. In *Proceedings of the 16th IEEE International Conference on Mobile Data Management*, volume 1, pages 110115. IEEE, 2015.

International Workshops

- Marco Pavan, Stefano Mizzaro, Matteo Bernardon, and Ivan Scagnetto. Exploiting news to categorize tweets: Quantifying the impact of different news collections. In *Proceedings of the 1st International Workshop on Recent Trends in News Information Retrieval, NewsIR 2016*, volume 1568 of *CEUR Workshop Proceedings*, pages 54–59. CEUR, 2016. [Best Paper Award]

- Stefano Mizzaro, Marco Pavan, Ivan Scagnetto, and Martino Valenti. Short text categorization exploiting contextual enrichment and external knowledge. In *Proceedings of the First International Workshop on Social Media Retrieval and Analysis*, SoMeRA 14, pages 5762, New York, NY, USA. ACM, 2014.

- Stefano Mizzaro, Marco Pavan, Ivan Scagnetto, and Ivano Zanello. A context-aware retrieval system for mobile applications. In *Proceedings of the 4th Workshop on Context-Awareness in Retrieval and Recommendation*, CARR 14, pages 1825, New York, NY, USA. ACM, 2014.

Bibliography

[1] INEX 2013 Tweet Contextualization Track. http://inex.mmci.uni-saarland.de/tracks/qa/, 2013.

[2] Facebook check-in. (2014, Jul.) Who, what, when, and now...where. [Online]. https://www.facebook.com/notes/facebook/who-what-when-and-nowwhere/418175202130, 2014.

[3] Foursquare check-in. (2014, Jul.) About. [Online]. https://foursquare.com/about, 2014.

[4] Twitter check-in. (2014, Jul.) How to tweet with your location. [Online]. https://support.twitter.com/entries/122236-how-to-tweet-with-your-location, 2014.

[5] Python wrapper around the Twitter API. https://dev.twitter.com/rest/public, 2016.

[6] The Next Web. http://thenextweb.com /twitter/2012/01/07/interesting-fact-mos -tweets-posted-are-approximately-30-char acters-long/#gref, 2016. [Online, visited Feb-2016].

[7] Twitter REST APIs. https://dev.twitter.com/rest/public, 2016.

[8] Ahmad Abdel-Hafez and Yue Xu. A survey of user modelling in social media websites. *Computer and Information Science*, 6(4):59, 2013.

[9] Fabian Abel, Qi Gao, Geert-Jan Houben, and Ke Tao. Semantic enrichment of twitter posts for user profile construction on the social web. In *The Semanic Web: Research and Applications*, pages 375–389. Springer, 2011.

[10] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. Effects of user similarity in social media. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 703–712. ACM, 2012.

[11] Microsoft Research Asia. Geolife project. http://research.microsoft.com/en-us/downloads/b16d359d-d164-469e-9fd4-daa38f2b2e13/, Aug. 2012.

[12] Sandhya Avasthi and Avinash Dwivedi. Prediction of mobile user behavior using clustering. *Int. J. Sci. Res. Publ*, 3(2):1–5, 2013.

[13] Ramakrishna B. Bairi, Raghavendra Udupa, and Ganesh Ramakrishnan. A framework for task-specific short document expansion. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (CIKM '16)*, CIKM '16, pages 791–800, New York, NY, USA, 2016.

[14] Somnath Banerjee, Krishnan Ramanathan, and Ajay Gupta. Clustering short texts using wikipedia. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 787–788. ACM, 2007.

[15] Tanusri Bhattacharya, Lars Kulik, and James Bailey. Extracting significant places

from mobile user gps trajectories: a bearing change based approach. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*, pages 398–401. ACM, 2012.

[16] D Brickley, L Miller, T Inkster, Y Zeng, Y Wang, D Damljanovic, Z Huang, S Kinsella, J Breslin, and B Ferris. The weighted interests vocabulary 0.5. *Namespace document, Sourceforge (September 2010)*, 2010.

[17] Dan Brickley and Libby Miller. Foaf vocabulary specification 0.91. namespace document, foaf project, 2007.

[18] Maike Buchin, Anne Driemel, Marc van Kreveld, and Vera Sacristán. Segmenting trajectories: A framework and algorithms using spatiotemporal criteria. *Journal of Spatial Information Science*, 2011(3):33–63, 2011.

[19] William M Campbell, Charlie K Dagli, and Clifford J Weinstein. Social network analysis with content and graphs. *Lincoln Laboratory Journal*, 20(1):61–81, 2013.

[20] Mario Cataldi, Luigi Di Caro, and Claudio Schifanella. Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining*, page 4. ACM, 2010.

[21] Chen Cheng, Haiqin Yang, Michael R Lyu, and Irwin King. Where you like to go next: Successive point-of-interest recommendation. In *IJCAI*, volume 13, pages 2605–2611, 2013.

[22] Yohan Chon, Yunjong Kim, and Hojung Cha. Autonomous place naming system using opportunistic crowdsensing and knowledge from crowdsourcing. In *Information Processing in Sensor Networks (IPSN), 2013 ACM/IEEE International Conference on*, pages 19–30. IEEE, 2013.

[23] Markus Christen, Thomas Niederberger, Thomas Ott, Suleiman Aryobsei, and Reto Hofstetter. Micro-text classification between small and big data. *Nonlinear Theory and Its Applications, IEICE*, 6(4):556–569, 2015.

[24] Joana Costa, Catarina Silva, Mário Antunes, and Bernardete Ribeiro. Defining semantic meta-hashtags for twitter classification. In *International Conference on Adaptive and Natural Computing Algorithms*, pages 226–235. Springer, 2013.

[25] Zichao Dai, Aixin Sun, and Xu-Ying Liu. Crest: Cluster-based representation enrichment for short text classification. In *Advances in Knowledge Discovery and Data Mining*, pages 256–267. Springer, 2013.

[26] Stefano De Sabbata, Stefano Mizzaro, and Luca Vassena. Spacerank: Using pagerank to estimate location importance. In *Proceedings of ECAI08 Workshop on Mining Social Data (MSoDa08)*, pages 1–5, 2008.

[27] Stefano De Sabbata, Stefano Mizzaro, and Luca Vassena. Where do you roll today? trajectory prediction by spacerank and physics models. In *Location Based Services and TeleCartography II*, pages 63–78. Springer, 2009.

[28] Inoshika Dilrukshi, Kasun De Zoysa, and Amitha Caldera. Twitter news classification using SVM. In *Proc. of ICCSE'13*, pages 287–291. IEEE, 2013.

[29] Miles Efron, Peter Organisciak, and Katrina Fenlon. Improving retrieval of short texts through document expansion. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 911–920, New York, NY, USA, 2012. ACM.

[30] Huiji Gao, Jiliang Tang, Xia Hu, and Huan Liu. Content-aware point of interest recommendation on location-based social networks. In *AAAI*, pages 1721–1727.

Citeseer, 2015.

[31] Longwen Gao, Shuigeng Zhou, and Jihong Guan. Effectively classifying short texts by structured sparse representation with dictionary filtering. *Information Sciences*, 323:130–142, 2015.

[32] Donatella Gubiani and Marco Pavan. *From Trajectory Modeling to Social Habits and Behaviors Analysis*, pages 371–385. Springer International Publishing, Cham, 2017.

[33] Weiwei Guo, Hao Li, Heng Ji, and Mona T Diab. Linking tweets to news: A framework to enrich short text data in social media. In *ACL (1)*, pages 239–249. Citeseer, 2013.

[34] Jungkyu Han and Hayato Yamana. Why people go to unfamiliar areas?: Analysis of mobility pattern based on users' familiarity. In *Proceedings of the 17th International Conference on Information Integration and Web-based Applications & Services*, iiWAS '15, pages 28:1–28:10, New York, NY, USA, 2015. ACM.

[35] Chung-Wei Hang, Pradeep K Murukannaiah, and Munindar P Singh. Platys: User-centric place recognition. In *AAAI Workshop on Activity Context-Aware Systems*, 2013.

[36] Jeffrey Hightower, Sunny Consolvo, Anthony LaMarca, Ian Smith, and Jeff Hughes. Learning and recognizing the places we go. In *International Conference on Ubiquitous Computing*, pages 159–176. Springer, 2005.

[37] Xiaohua Hu, Xiaodan Zhang, Caimei Lu, Eun K Park, and Xiaohua Zhou. Exploiting wikipedia as external knowledge for document clustering. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 389–396. ACM, 2009.

[38] Yin-Fu Huang and Chen-Ting Huang. Mining domain information from social contents based on news categories. In *Proc. of IDEAS'15*, pages 186–191. ACM, 2015.

[39] Pan Hui and Jon Crowcroft. Human mobility models and opportunistic communications system design. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 366(1872):2005–2016, 2008.

[40] Sibren Isaacman, Richard Becker, Ramón Cáceres, Stephen Kobourov, Margaret Martonosi, James Rowland, and Alexander Varshavsky. Identifying important places in peoples lives from cellular network data. In *International Conference on Pervasive Computing*, pages 133–151. Springer, 2011.

[41] Jungwook Jun, Randall Guensler, and Jennifer Ogle. Smoothing methods to minimize impact of global positioning system random error on travel distance, speed, and acceleration profile estimates. *Transportation Research Record: Journal of the Transportation Research Board*, 1972:141–150, 2006.

[42] Jong Hee Kang, William Welbourne, Benjamin Stewart, and Gaetano Borriello. Extracting places from traces of locations. In *Proceedings of the 2nd ACM international workshop on Wireless mobile applications and services on WLAN hotspots*, pages 110–118. ACM, 2004.

[43] Dmytro Karamshuk, Chiara Boldrini, Marco Conti, and Andrea Passarella. Human mobility models for opportunistic networks. *IEEE Communications Magazine*, 49(12):157–165, 2011.

[44] Georgios Kellaris, Nikos Pelekis, and Yannis Theodoridis. Map-matched trajectory

compression. *Journal of Systems and Software*, 86(6):1566–1579, 2013.

[45] K. L. Kwok. A neural network for probabilistic information retrieval. *SIGIR Forum*, 23(SI):21–30, May 1989.

[46] Nicholas D Lane, Ye Xu, Hong Lu, Andrew T Campbell, Tanzeem Choudhury, and Shane B Eisenman. Exploiting social networks for large-scale human behavior modeling. *IEEE Pervasive Computing*, 10(4):45–53, 2011.

[47] T Durga Laxmi, R Baby Akila, KS Ravichandran, and B Santhi. Study of user behavior pattern in mobile environment. *Research Journal of Applied Sciences, Engineering and Technology*, 4(23):5021–5026, 2012.

[48] Chenliang Li, Aixin Sun, and Anwitaman Datta. Tsdw: Two-stage word sense disambiguation using wikipedia. *Journal of the American Society for Information Science and Technology*, 64(6):1203–1223, 2013.

[49] Chenliang Li, Haoran Wang, Zhiqian Zhang, Aixin Sun, and Zongyang Ma. Topic modeling for short texts with auxiliary word embeddings. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 165–174. ACM, 2016.

[50] Quannan Li, Yu Zheng, Xing Xie, Yukun Chen, Wenyu Liu, and Wei-Ying Ma. Mining user similarity based on location history. In *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*, page 34. ACM, 2008.

[51] Xin Li, Mengyue Wang, and T-P Liang. A multi-theoretical kernel-based approach to social network-based recommendation. *Decision Support Systems*, 65:95–104, 2014.

[52] Miao Lin and Wen-Jing Hsu. Mining gps data for mobility patterns: A survey. *Pervasive and Mobile Computing*, 12:1–16, 2014.

[53] Haifeng Liu, Zheng Hu, Ahmad Mian, Hui Tian, and Xuzhen Zhu. A new user similarity model to improve the accuracy of collaborative filtering. *Knowledge-Based Systems*, 56:156–166, 2014.

[54] Xin Liu, Yong Liu, Karl Aberer, and Chunyan Miao. Personalized point-of-interest recommendation by mining users' preference transition. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 733–738. ACM, 2013.

[55] Chunliang Lu, Wai Lam, and Yingxiao Zhang. Twitter user modeling and tweets recommendation based on wikipedia concept graph. In *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[56] Mingqi Lv, Ling Chen, and Gencai Chen. Discovering personally semantic places from gps trajectories. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1552–1556. ACM, 2012.

[57] Walid Magdy, Hassan Sajjad, Tarek El-Ganainy, and Fabrizio Sebastiani. Distant supervision for tweet classification using youtube labels. In *Ninth International AAAI Conference on Web and Social Media*, 2015.

[58] Gerasimos Marketos, Elias Frentzos, Irene Ntoutsi, Nikos Pelekis, Alessandra Raffaetà, and Yannis Theodoridis. Building real-world trajectory warehouses. In *Proceedings of the seventh ACM international workshop on data engineering for wireless and mobile access*, pages 8–15. ACM, 2008.

[59] Graham McDonald, Romain Deveaud, Richard Mccreadie, Craig Macdonald, and

Iadh Ounis. Tweet enrichment for effective dimensions classification in online reputation management, 2015.

[60] Matúš Medo. Network-based information filtering algorithms: ranking and recommendation. In *Dynamics On and Of Complex Networks, Volume 2*, pages 315–334. Springer, 2013.

[61] Wang Meng, Lin Lanfen, Wang Jing, Yu Penghua, Liu Jiaolong, and Xie Fei. Improving short text classification using public search engines. In *International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making*, pages 157–166. Springer, 2013.

[62] Stefano Mizzaro, Marco Pavan, and Ivan Scagnetto. *Content-Based Similarity of Twitter Users*, pages 507–512. Springer International Publishing, Cham, 2015.

[63] Stefano Mizzaro, Marco Pavan, Ivan Scagnetto, and Martino Valenti. Short text categorization exploiting contextual enrichment and external knowledge. In *Proceedings of the First International Workshop on Social Media Retrieval and Analysis*, SoMeRA '14, pages 57–62, New York, NY, USA, 2014. ACM.

[64] Krishna K Mohbey and GS Thakur. User movement behavior analysis in mobile service environment. *British Journal of Mathematics & Computer Science*, 3(4):822–834, 2013.

[65] Arturo Montejo-Ráez, Eugenio Martínez-Cámara, M Teresa Martín-Valdivia, and L Alfonso Ureña-López. A knowledge-based approach for polarity classification in twitter. *Journal of the Association for Information Science and Technology*, 65(2):414–425, 2014.

[66] Raul Montoliu, Jan Blom, and Daniel Gatica-Perez. Discovering places of interest in everyday life from smartphone data. *Multimedia Tools and Applications*, 62(1):179–207, 2013.

[67] Raul Montoliu and Daniel Gatica-Perez. Discovering human places of interest from multimodal mobile phone data. In *Proceedings of the 9th International Conference on Mobile and Ubiquitous Multimedia*, page 12. ACM, 2010.

[68] Jonathan Muckell, Paul W Olsen Jr, Jeong-Hyon Hwang, Catherine T Lawson, and SS Ravi. Compression of trajectory data: a comprehensive evaluation and new approach. *GeoInformatica*, 18(3):435–460, 2014.

[69] Parma Nand, Rivindu Perera, and Gisela Klette. A tweet classification model based on dynamic and static component topic vectors. In *Australasian Joint Conference on Artificial Intelligence*, pages 424–430. Springer, 2015.

[70] Anastasios Noulas, Salvatore Scellato, Cecilia Mascolo, and Massimiliano Pontil. An empirical study of geographic user activity patterns in foursquare. *ICwSM*, 11:70–573, 2011.

[71] Fabrizio Orlandi, John Breslin, and Alexandre Passant. Aggregated, interoperable and multi-domain user profiles for the social web. In *Proceedings of the 8th International Conference on Semantic Systems*, pages 41–48. ACM, 2012.

[72] Christine Parent, Stefano Spaccapietra, Chiara Renso, Gennady Andrienko, Natalia Andrienko, Vania Bogorny, Maria Luisa Damiani, Aris Gkoulalas-Divanis, Jose Macedo, Nikos Pelekis, et al. Semantic trajectories modeling and analysis. *ACM Computing Surveys (CSUR)*, 45(4):42, 2013.

[73] Marco Pavan, Thebin Lee, and Ernesto William De Luca. Semantic enrichment for adaptive expert search. In *Proceedings of the 15th International Conference on*

Knowledge Technologies and Data-driven Business, i-KNOW '15, pages 36:1–36:4, New York, NY, USA, 2015. ACM.

[74] Marco Pavan and Ernesto William De Luca. Semantic-based expert search in textbook research archives. In *SDA@TPDL*, 2015.

[75] Marco Pavan, Stefano Mizzaro, Matteo Bernardon, and Ivan Scagnetto. Exploiting news to categorize tweets: Quantifying the impact of different news collections. In *Proceedings of the First International Workshop on Recent Trends in News Information Retrieval*. CEUR, March 2016. Co-located with 38^th European Conference on Information Retrieval (ECIR 2016).

[76] Marco Pavan, Stefano Mizzaro, and Ivan Scagnetto. *Mining Movement Data to Extract Personal Points of Interest: A Feature Based Approach*, pages 35–61. Springer International Publishing, Cham, 2017.

[77] Marco Pavan, Stefano Mizzaro, Ivan Scagnetto, and Andrea Beggiato. Finding important locations: A feature-based approach. In *2015 16th IEEE International Conference on Mobile Data Management*, volume 1, pages 110–115. IEEE, 2015.

[78] Mohammed A Quddus, Washington Y Ochieng, and Robert B Noland. Current map-matching algorithms for transport applications: State-of-the art and future research directions. *Transportation research part c: Emerging technologies*, 15(5):312–328, 2007.

[79] Kai-Florian Richter, Falko Schmid, and Patrick Laube. Semantic trajectory compression: Representing urban movement in a nutshell. *Journal of Spatial Information Science*, 2012(4):3–30, 2012.

[80] Kevin Dela Rosa, Rushin Shah, Bo Lin, Anatole Gershman, and Robert Frederking. Topical clustering of tweets. *Proceedings of the ACM SIGIR: SWSM*, 2011.

[81] Mehran Sahami and Timothy D Heilman. A web-based kernel function for measuring the similarity of short text snippets. In *Proceedings of the 15th international conference on World Wide Web*, pages 377–386. AcM, 2006.

[82] L. Sang, F. Xie, X. Liu, and X. Wu. Wefest: Word embedding feature extension for short text classification. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pages 677–683, Dec 2016.

[83] Salvatore Scellato, Mirco Musolesi, Cecilia Mascolo, Vito Latora, and Andrew T Campbell. Nextplace: a spatio-temporal prediction framework for pervasive systems. In *International Conference on Pervasive Computing*, pages 152–169. Springer, 2011.

[84] Bill N Schilit, Anthony LaMarca, Gaetano Borriello, William G Griswold, David McDonald, Edward Lazowska, Anand Balachandran, Jason Hong, and Vaughn Iverson. Challenge: Ubiquitous location-aware computing and the place lab initiative. In *Proceedings of the 1st ACM international workshop on Wireless mobile applications and services on WLAN hotspots*, pages 29–35. ACM, 2003.

[85] Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas. Short text classification in twitter to improve information filtering. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 841–842. ACM, 2010.

[86] Jiliang Tang, Xufei Wang, Huiji Gao, Xia Hu, and Huan Liu. Enriching short text representation in microblog for clustering. *Frontiers of Computer Science*,

6(1):88–101, 2012.

[87] Ke Tao, Fabian Abel, Qi Gao, and Geert-Jan Houben. Tums: Twitter-based user modeling service. In Raúl García-Castro, Dieter Fensel, and Grigoris Antoniou, editors, *The Semantic Web: ESWC 2011 Workshops: ESWC 2011 Workshops, Heraklion, Greece, May 29-30, 2011, Revised Selected Papers*, pages 269–283, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[88] Muhammad Umair, Wan Seok Kim, Byoung Chul Choi, and Sung Young Jung. Discovering personal places from location traces. In *16th International Conference on Advanced Communication Technology*, pages 709–713. IEEE, 2014.

[89] Nagendra R Velaga, Mohammed A Quddus, and Abigail L Bristow. Detecting and correcting map-matching errors in location-based intelligent transport systems. In *12th world conference on transport research, Lisbon, Portugal*, pages 11–15, 2010.

[90] Dung D. Vu, Hien To, Won-Yong Shin, and Cyrus Shahabi. Geosocialbound: An efficient framework for estimating social poi boundaries using spatio–textual information. In *Proceedings of the Third International ACM SIGMOD Workshop on Managing and Mining Enriched Geo-Spatial Data*, GeoRich '16, pages 3:1–3:6, New York, NY, USA, 2016. ACM.

[91] Fang Wang, Zhongyuan Wang, Zhoujun Li, and Ji-Rong Wen. Concept-based short text classification and ranking. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1069–1078. ACM, 2014.

[92] Peng Wang, Heng Zhang, Bo Xu, Chenglin Liu, and Hongwei Hao. Short text feature enrichment using link analysis on topic-keyword graph. In *Natural Language Processing and Chinese Computing*, pages 79–90. Springer, 2014.

[93] Ross Wilkinson and Philip Hingston. Using the cosine measure in a neural network for document retrieval. In *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '91, pages 202–210, New York, NY, USA, 1991. ACM.

[94] Xiangye Xiao, Yu Zheng, Qiong Luo, and Xing Xie. Finding similar users using category-based location history. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 442–445. ACM, 2010.

[95] Xiangye Xiao, Yu Zheng, Qiong Luo, and Xing Xie. Inferring social ties between users with human location history. *Journal of Ambient Intelligence and Humanized Computing*, 5(1):3–19, 2014.

[96] Guandong Xu, Zongda Wu, Guiling Li, and Enhong Chen. Improving contextual advertising matching by using wikipedia thesaurus knowledge. *Knowledge and Information Systems*, 43(3):599–631, 2015.

[97] Hiroki Yamakawa, Jing Peng, and Anna Feldman. Semantic enrichment of text representation with wikipedia for text classification. In *Proc. of SMC'10*, pages 4333–4340. IEEE, 2010.

[98] Qiang Yan, Lianren Wu, and Lan Zheng. Social network based microblog user behavior analysis. *Physica A: Statistical Mechanics and Its Applications*, 392(7):1712–1723, 2013.

[99] Zhixian Yan, Dipanjan Chakraborty, Christine Parent, Stefano Spaccapietra, and Karl Aberer. Semantic trajectories: Mobility data computation and annotation.

ACM Transactions on Intelligent Systems and Technology (TIST), 4(3):49, 2013.

[100] Majid Yazdani and Andrei Popescu-Belis. Computing text semantic relatedness using the contents and links of a hypertext encyclopedia. *Artificial Intelligence*, 194:176–202, 2013.

[101] Zheng Yu, Haixun Wang, Xuemin Lin, and Min Wang. Understanding short texts through semantic enrichment and hashing. *IEEE Transactions on Knowledge and Data Engineering*, 28(2):566–579, 2016.

[102] Shitao Zhang, Xiaoming Jin, Dou Shen, Bin Cao, Xuetao Ding, and Xiaochen Zhang. Short text classification by detecting information path. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 727–732. ACM, 2013.

[103] Yang Zhang, Yao Wu, and Qing Yang. Community discovery in twitter based on user interests. *Journal of Computational Information Systems*, 8(3):991–1000, 2012.

[104] Kaiqi Zhao, Gao Cong, and Aixin Sun. Annotating points of interest with geotagged tweets. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pages 417–426, New York, NY, USA, 2016. ACM.

[105] Yukun Zhao, Shangsong Liang, Zhaochun Ren, Jun Ma, Emine Yilmaz, and Maarten de Rijke. Explainable user clustering in short text streams. In *SIGIR, ACM*, 2016.

[106] Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. Urban computing: concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(3):38, 2014.

[107] Yu Zheng, Quannan Li, Yukun Chen, Xing Xie, and Wei-Ying Ma. Understanding mobility based on gps data. In *Proceedings of the 10th international conference on Ubiquitous computing*, pages 312–321. ACM, 2008.

[108] Yu Zheng and Xing Xie. Learning travel recommendations from user-generated gps traces. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(1):2, 2011.

[109] Yu Zheng, Xing Xie, and Wei-Ying Ma. Geolife: A collaborative social networking service among user, location and trajectory. *IEEE Data Eng. Bull.*, 33(2):32–39, 2010.

[110] Yu Zheng, Lizhu Zhang, Xing Xie, and Wei-Ying Ma. Mining interesting locations and travel sequences from gps trajectories. In *Proceedings of the 18th international conference on World wide web*, pages 791–800. ACM, 2009.

[111] Yu Zheng and Xiaofang Zhou. *Computing with spatial trajectories*. Springer Science & Business Media, 2011.

[112] Deyu Zhou, Liangyu Chen, and Yulan He. An unsupervised framework of exploring events on twitter: Filtering, extraction and categorization. In *Proc. of AAAI'15*, 2015.