Understanding Geo-spatial Information on Social Media

Li, Wen

DOI
10.4233/uuid:06c072ad-4db6-4d3b-b747-784e30d862a4

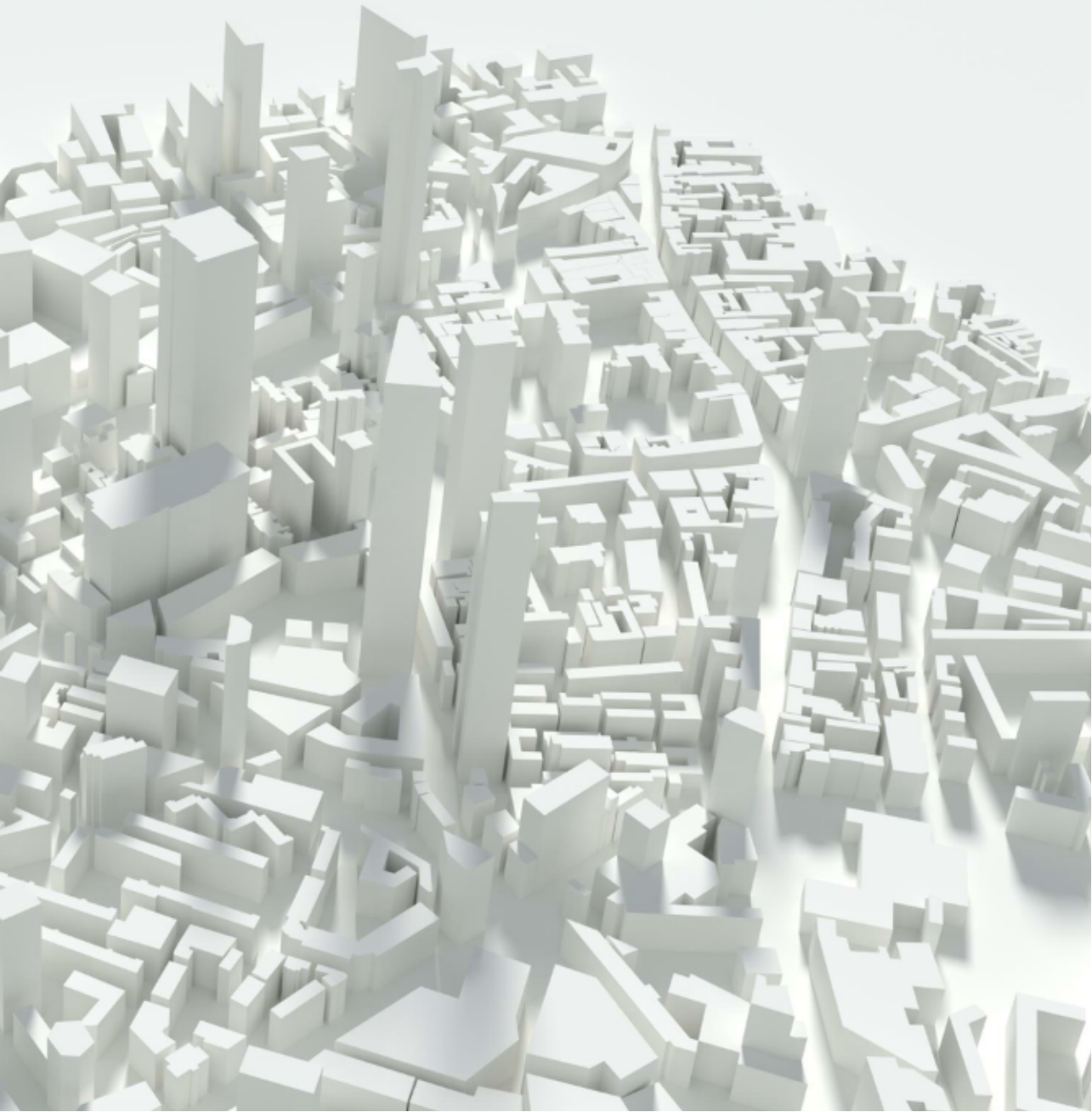
Publication date
2016

Document Version
Final published version

Citation (APA)
Li, W. (2016). *Understanding Geo-spatial Information on Social Media*. [Dissertation (TU Delft), Delft University of Technology]. SIKS. https://doi.org/10.4233/uuid:06c072ad-4db6-4d3b-b747-784e30d862a4

Wen Li

Understanding Geo-spatial information on Social Media



Understanding Geo-spatial Information on Social Media

Understanding Geo-spatial Information on Social Media

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universtiteit Delft,
op gezag van de Rector Magnificus Prof.ir. K.C.A.M. Luyben,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen
op maandag 23 mei 2016 om 10.00 uur
door

Wen LI

Master of Engineering, Xi'an Jiaotong University

geboren te Xi'an, China

Dit proefschrift is goedgekeurd door de promotor:

Prof.dr.ir. A.P. de Vries

Samenstelling promotiecommissie:

Rector Magnificus,	voorzitter
Prof.dr.ir. A.P. de Vries,	Radboud Universiteit, promotor

Onafhankelijke leden:

Prof.dr.ir. G.J.P.M. Houben,	Technische Universiteit Delft
Prof.dr. A. van de Bosch,	Radboud Universiteit
Prof.dr. H. Scholten,	Vrije Universiteit Amsterdam
Prof.dr. P.M.E. De Bra,	Technische Universiteit Eindhoven
Prof.dr. A. Hanjalic,	Technische Universiteit Delft, reservelid

Overige leden:

Dr.ir. R.A. de By,	Universiteit Twente
Dr.ir. M.-C. ten Veldhuis,	Technische Universiteit Delft

SIKS Dissertation Series No. 2016-27
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Cover designed by Wen Li & Jie Jiang
Printed by CPI Koninklijke Wöhrmann
ISBN 978-94-6186-665-3

PRINTED IN THE NETHERLANDS

to Jie

Contents

Part I

Prelude

Chapter 1

Introduction

The development of hardware and software regarding the Internet in the past decade has enormous influence on how people communicate with each other. With up-to-date smart phones, users can share all kinds of information with their peers almost in real time. The so-called online social media platforms act as portals for the shared information and make them available almost anywhere at any time on users' devices. Twitter is one of the most popular social media platforms, which focuses on short textual messages up to 140 characters. Many web applications have emerged to help users attach images and video clips to their messages within such limitation. In spite of this particular limitation, it has become successful real-time broadcasting channels for users as small as individuals talking about their lives or as big as companies and organizations announcing important changes.

The popularization of social media drives a huge volume of data through the platforms hosting it, including text messages, photos, video clips, *etc.* For example, Twitter receives more than a half billion messages per day [5] and Flickr as a photo sharing web site has a rate of 3.5 million new images uploaded daily [69]. User contributed content keeps accumulating over time and is valued as an important asset in different domains. Business companies use it to learn users' interests and purchasing habit so that they can improve users' purchasing experience or deploy more effective and efficient promotion [17]. Researchers have been granted the opportunity to observe massive crowds in real time and

learn how they interact with each other and/or with the rest of the world.

1.1 Geographical Information in Social Media

Among various kinds of information going through these social media platforms, geographical information is of particular interest, *e.g.*, it presents how users perceive their physical environment. Before the era of smart phones, geographical information on the Internet has been generally presented in the form of textual descriptions, *e.g.*, addresses, city names, post codes, area code in telephone numbers, *etc.* [27, 43]. Since the integration of positioning sensors (*e.g.*, GPS, A-GPS) in smart phones, a more precise way of describing a location is adopted by social media platforms, *i.e.*, geotags. A geotag is merely a pair of coordinates (longitude and latitude) indicating where the device was at the time when the position was inquired. When a user writes a message or takes a photo, the positioning sensor on the user's device will capture the current location and generate a geotag and then the user will choose whether or not to attach it to the message or the photo. For some applications, attaching geotags is opt-in, *e.g.*, Twitter, Flickr, while for some other applications this is compulsory, *e.g.*, Foursquare. Foursquare is a dedicated Location-based Social Network.

Twitter introduced another type of geotags for representing *Place of Interest* information in 2009[132]. Besides coordinates, these geotags contain more information, *e.g.*, names, addresses, cities in which the places are located. To distinguish them from the ones containing merely coordinates, we refer to these as POI-tags. They are more precise as the names of locations are explicitly recorded which would be difficult to derive from merely coordinates.

With geotags, it is more convenient and precise for users to express locations. For example, when a user writes a tweet (a message delivered via Twitter) and requests the application to mark (tag) it with a location, the application will enable the location sensor (usually a GPS-chip) on the user's device to find the current location and then make the returned coordinates as a geotag for the message. The message can be the user's comment on the great experience at a restaurant or a photo the user took at a place for the nice view. This message with the geotag can later be viewed by the user's friends when it is delivered by Twitter. They can easily learn the whereabouts in the tweet by checking the coordinates on any online map services, *e.g.*, Google Maps. On

top of geotags, POI-tags carry more information describing the entities at the locations more precisely, *e.g.*, names and addresses, by which they are easier to be distinguished from other co-located ones, *e.g.*, shops in a mall or a street.

The use of geographical information has become both popular and controversial in the sphere of social media. Its popularity originates from its better presentation of locations than any other forms. However, the exact precision can reveal one's daily trails to the public and become a threat to users' privacy. Both sides of the coin should be explored and investigated to make sure we fully understand how it will affect users.

In this dissertation, we focus in particular on the geographical information emerged and embedded in POI-tags on social media. We investigate the connections between location information and other sources of information on social media and how these connections can be used in real world applications.

1.2 Research Questions

Geotags bring up a new dimension of information to online social media which can reveal relationships between users' online behavior and the associated contexts. This has triggered a lot of research in different domains. One way of looking into these relationships is to find the correlation between users' words and their locations. The correlation, if there is one, may extend the use of geotags which has both positive and negative implications, *i.e.*, improving users' online experience with more intelligent services, and paying the price of potential abuse of users' privacy. For example, a user may receive more personalized recommendations of places based on the locations that other users have been to and share similar interests with the user. On the other hand, the correlation may lead to disclosing users' current location without direct evidence of where the users are. This put them or their property under threat. Many studies have been carried out by researchers on these correlations. Cheng et al. [31] and Ren et al. [124] studied the problem of predicting users' home town based on their messages on Twitter and Serdyukov et al. [135] tried to predict the origin of a collection of photos retrieved from Flickr based on tags. Their works all rely on coordinates in geotags at granularity of cities. With POI-tags, the connection between user activities and locations where these activities take place becomes more precise and clear and may enable more precise prediction of users' locations. This inspires the first research question

that we will discuss in this dissertation:

- **RQ1**: Can we predict a user's location in terms of POIs based on his/her messages on social media?

It can further be broken down to (a) What are the features can be used for modelling a POI? (b) To what extent can locations be distinguished by the models? (c) How well can the models predict the origin of the messages?

The gamification approach introduced by Foursquare[1], a dedicated location-based social network platform, makes users regularly geotag their messages, so-called check-ins, to win badges and titles. These regularly posted POI-tags, from a user, to some extent, imply the user's moving trails. For example, a user's Twitter time line, a stream of messages posted by the user, may record his/her visit to a café at 8:03, a lecture room in university at 8:45 and the library of the university at 15:34. Patterns may be observed in many similar trails like this one and the correlation between the geotags may suggest (or predict) future visits for this particular user. In general, this problem falls into trajectory mining and prediction via different sources of information. For example, Zheng et al. [176] proposed Collaborative Filtering based location recommendations based on GPS position logs. The work is based on data collected by continually sampling of way points returned from GPS devices. Geotags from social media are generally random samples of users' locations. Compared to GPS-trajectories, there are more missing parts of the observations about users' locations.

Correlations modelled from geotags have been used in research on location recommendation. For example, Kurashima et al. [74] proposed Markov Models combined with topic model for recommending routes to tourists based on photos with geotags from Flickr. Similarly, Shi et al. [137] and Clements et al. [36] recommend landmarks based on users' own interests. These works make prediction among locations that tourists are interested in and leave the question open to more general types of locations. This in turn inspires the second research question of this dissertation:

- **RQ2**: Can we predict users' future visits to POIs by only using users' visiting histories?

[1] `https://foursquare.com`

This question involves two sub-questions to answer: (a) How can we model user check-ins in both temporal and spatial dimensions? (b) How well can the models predict users' next move based on users' previous check-ins?

Correlation between user behaviour and visited locations demonstrates the interaction between users and locations, which may slip into users' memory and form the kind of knowledge that we call geo-expertise. For example, a user who often goes to different bars or pubs may know the differences between the bars in town. This knowledge can be very useful, *e.g.*, the user's friend may rely on his advice and recommendation for their graduation party. However, this kind of knowledge can hardly be interpreted in written language [39] as there might be too many different aspects and their importance may vary vastly for different people. The result from previous studies suggest to find the person who is an expert on the given topic instead of returning document containing direct answers may help solve the problem.[13, 16, 161] Different from the previous studied expertise retrieval systems, in this study, geo-expertise retrieval systems rely on non-textual information which has been barely explored before. This leads to the third research question in this dissertation:

- **RQ3**: How can we model users' knowledge about locations and build an automated retrieval system based on POI information on social media?

This question can be fulfilled by answering the three sub-questions: (a) How often, in which way, and to whom are people looking for, or giving POI advice? (b) How should a candidate's geo-expertise be determined via his/her check-in profile? (c) How well do the automated retrieval methods perform in estimating users' geo-expertise?

Though the geographical information embedded in social media is useful for sharing location information, location information has remained a scarce resource because of the lack of contributions from users. Another reason is that location information are treated as proprietary resource and scattered among different service providers who have not agreed on a standard schema for storage, communication and sharing of such information. Part of the reason is that there are few motivating applications available on the market and users are not convinced of the benefit from contributing geographical information. As a result, the insufficient data from the users may weaken the confidence of service providers making applications. This looped dependence is holding back

large scale utilization of geographical information on social media. Solving these problems may help users benefit more from using geotags.

Not only can the integration of geographical information help users better share location information but also help public management organizations to improve the understanding of their tasks. Many research works have been carried out for discovering and extracting public events from social media, *e.g.*, earthquakes [131], floods [152]. Besides, these works also demonstrated potentials in gathering minor disturbances in public space from citizens. For example, water damage is one of the most important problems for cities sitting around rivers, as storms can bring much more precipitation than what the cities' water system can handle. It would be useful to learn how storms affect the city at a finer granularity so that the improvements can be prioritized accordingly and plans can be arranged ahead. That is:

- **RQ4**: How can we extract and make use of user contributed content on social media for understanding water damage?

To approach this question, we investigate the following subquestions. (a) Can Twitter be used as a sources of information for monitoring strong precipitation? (b) What are the advantages and disadvantages of using Twitter as a resource of water damage reports comparing to the official water damage report registry?

During the course of the research presented in this dissertation, we found there are needs of efficient tools for visualizing multidimensional data. In the domain of Information Retrieval and data related science such tools are essential for data exploration, comprehension and communication, especially visualization tools. Though there are many software and libraries with charting and plotting functions, the existing ones either require a lot of coding skills for making a nice chart or do not provide sufficient or fluent human-data interaction. For this study we also require that the tool can be embedded in Web interface so that the data can be annotated for evaluation. Thus we developed our own open source tool for better data charting, a solution which is web based and requires a minimum knowledge for the users of this tool to make a usable data charting interface featuring human-data interactivity.

1.3 Main Contributions

The main contributions of this dissertation are the models of the relationships between geographical information and user behaviors published on social media. There are three aspects of such relationships explored in this dissertation, (a) the correlation between the messages sent by users and the origin of locations where the messages were sent, (b) the correlation between users' future visits and their visiting histories, and (c) the correlation between users' knowledge about different locations and their visiting histories. Diverse models are discussed and tailored in order to capture the characteristics of these correlations. Furthermore, we propose and demonstrate various techniques to make use of diverse information sources and alleviate the sparsity problem in the real data.

Throughout the dissertation we evaluate and compare the proposed models and techniques in the setting of prediction systems using the real data collected from online social media platforms, *i.e.*, Twitter and Foursquare, and demonstrate the feasibility and effectiveness of the predictive systems based on these models.

In order to facilitate data exploration, we develop and open-source a social media workbench for researchers in the community of social media, which can be easily extended to other domain problems. The workbench is a flexible yet easy-to-use tool for exploring social media data by providing interactive access to aggregated information via a Web user interface. We also make the code and data sets (prepared and anonymized according to the term of use imposed on these data sets) used in the experiments in this dissertation available online, hopefully inspiring more research carried out in this domain.

1.4 The Outline of Dissertation

In the following chapters, we address the research questions presented in the previous section. In Chapter 2, the related concepts and studies will be introduced and discussed, as well as the services and APIs used for this study. Chapter 3 is dedicated to the first research question and presents an approach to predict user locations from a single Twitter message. We continue with the problem of predicting user locations in Chapter 4, but from a different angle, in which we look into users' mobile patterns and predict their future visits

based on their visiting history (the second research question). Since users may get familiar with the places they visited and such knowledge can be helpful for others, we investigate how the knowledge (called geo-expertise) can be estimated and retrieved in Chapter 5. Besides individual users, organizations can benefit from the geographical information in social media. Thus in Chapter 6, we present a study for extracting and comprehending water damages reported in social media. To facilitate our research, an open-source and easy-to-use tool has been developed in the course of this study and is detailed in Chapter 7. Finally, we summarize our findings and discuss future routes along this study in Chapter 8.

Chapter 2

Background

The development of Web and Internet technologies allows people to communicate easily in spite of distance. More and more applications and platforms are available for users to communicate with each other, in ways which are very different from that of the pre-Internet era. Not only do the platforms convey messages from one to another, the content of communication is also stored online which the users can revisit at any later time, often publicly available. Many different types of media are used for communicating information, such as photos, video clips and geotags.

There are in general two terms to refer to this new kind of platforms. *Social Media* is one of the most commonly used terms, which is also what we use in this dissertation to refer to these platforms. This term emphasizes one of the important functions of such platforms, *i.e.*, serving as channels for user-generated content [71]. For example, Flickr[1] is qualified as a social media platform because its main purpose is to host users' photo albums. Another term often used is (*Online*) *Social Networks*, which emphasizes the function of connecting people [26], *i.e.*, to help people getting to know each other. For example, this is the purpose of Facebook on which people are connected online and the platform can also give suggestions on whom to follow based on mutual friends. Some platforms have been designed for both purposes. Twitter[2] is one

[1] `https://flickr.com`
[2] `https://twitter.com`

of such platforms. Many people use it for communicating with friends and there are also accounts dedicated to spreading information. Though each platform or application has its own focus, they all share some common characteristics:

- Users are encouraged to create content in varying types of media.
- Users are connected by online friendships via which they can acquire content created by each other.

In this dissertation we focus more on the aspect of user generated content than friendship networks. Thus we use social media to refer to the platforms that we study in this dissertation.

The recent introduction of geotagging in social media platforms became popular, via which users can express locations more precisely than before. Geotags are a pair of coordinates in spatial reference, which are usually obtained from positioning sensors such as GPS-enabled devices. Compared to geotags, textual addresses and place names are sometimes vague, context dependent and error-prone. For example, there are several supermarkets of the same brand in the city of Delft and it would be unclear to only specify the name of the supermarket, though it may be possible to infer which one from the context. It is also much easier for machines to extract the locations from the geotags in users' messages or photos rather than inferring the information from textual or visual features. Textual and visual features may be useless for inferring the location, *e.g.*, consider the case of news on Twitter or photos taken indoor [89]. With geotags, users can attach a location to their messages, photos, *etc.*, which enables them to later recall the location or communicate the location precisely to their fellows. Social media enabled by geotags also provides an opportunity for researchers to study users' accurate locations and the derived knowledge can be used for assisting users' daily lives. For example,

- recommending locations (*e.g.*, [137]),
- recommending traveling routes (*e.g.*, [74]),
- predicting social tie strength (*e.g.*, [58]),
- recommending friends (*e.g.*, [128, 60, 158]),
- improving local search (*e.g.*, [100]),

- identifying local experts (*e.g.*, Chapter 5, [33])

2.1 Geographical Signal in Social Media

The introduction of geotags to the online world is a consequence of the widespread availability of GPS-enabled devices such as smart phones and digital cameras with GPS-recorders. Users with these devices can easily locate themselves and post their locations online via geotagged tweets and photos. These geotags stored online can associate the users to the locations they have been to. For example, if a user uploads a photo with a geotag to Flickr and makes it publicly visible, the user and his/her friends can later check where the photo was taken. If users want to explore photos on Flickr, a map of photos will show up, on which they can choose to zoom in to any area and browse the photos taken in the area.

Twitter also supports that users mark tweets (messages sent to Twitter) with a geotag representing a location. For example, a user may attach a geotag to mark the location of the restaurant where he had his breakfast and tweet about the special flavoured coffee. The geotags are usually generated via the GPS-chips on users' devices which record users' current coordinates. When the users' followers look at the geotagged tweets, they will also see a small map showing where the tweets were sent.

Twitter later improved geotags by including place entities (referred to in this dissertation as Point-of-Interest tags, or POI-tags for short) which not only have the spatial information (coordinates) but also the meta information about the locations, such as the name of the place, the address. The meta information about the location allows more precise description of a location since it can disambiguate collocated places. For example, if two shops are both located in a large shopping mall, they can hardly be distinguished by the coordinates from the GPS as GPS are inaccurate for indoor use or the two shops may be at the same coordinates but on different floors. With the name and category information embedded in POI-tags, it is possible for user to refer to either shops in their messages without ambiguity.

Foursquare[3] is one of the largest Location Based Social Networks (LBSN), on which users can check-in at, comment on or leave a tip about a place. The

[3] `http://foursquare.com`

platform also encourages users to use this service by giving them badges and titles when they achieve a certain amount of activity on the platform This may be the reason it has become prosperous [134] and outdone its two competitors: BrightKite[4] and Gowalla[5]. BrightKite was one of the pioneers in the field of LBSN, but ceased operations, and only a few studies were carried out on its data, *e.g.*, [85]. Gowalla was another LBSN around the same period of time as BrightKite, upon whose data more studies were carried out [23, 168, 95, 60]. Among these three, Foursquare attracted more research on its data, examples of which can be found in [113, 150, 88, 129, 86].

Besides the coordinates, names, addresses, *etc.*, Foursquare also provides a taxonomy of locations stored in its database. The taxonomy has 9 top categories and each of them contain one or more subcategories at a lower level. For example, under the category *Food*, there are *Chinese Restaurant*, *American Restaurant*, *etc.* Some of these lower-level categories also have subcategories (in finer granularity). This meta information provides an opportunity for researchers to learn about what kinds of activity users may carry out there and/or their purpose of going there. Inferring location information from other attributes is a challenging task. For example, *De Waag* is a restaurant in Delft which used to be a weight house[6] and the name does not have a connection to its current function. Knowing the place is a restaurant can link it to dining activity, and observations of a user checking in at a significant number of POIs in some area all classified to this category may suggest evidence of the user's knowledge about dining in the corresponding neighbourhood.

These social media (social network) platforms have been developing very fast. They change their interfaces, add new features and redefine their functions all the time. This brought some issues for long-term consistency of data as some resources may be not available for later access. Carrying out the studies for this dissertation, we have been experiencing some of the changes. For example, Twitter revoked the prior white-listing of university's IP addresses, changed the ways how locations are selected and presented in their Web interface, and changed the authorization method. As to Foursquare, they redefined and restructured the category labels, *e.g.*, adding new categories, renaming existing categories. Though the already collected data would not be affected, it would

[4] `http://en.wikipedia.org/wiki/Brightkite`
[5] `http://en.wikipedia.org/wiki/Gowalla`
[6] A place where commodities are officially weighed

undermine the consistency of continuous collected data for long periods, *e.g.*, crawling data for several years. The data sets collected for the studies in this dissertation are all in short periods and we tried to avoid the influence of changes by re-crawling.

2.2 Related Works

Location is useful information in a wide variety of applications, but it has been rarely studied in the past, particularly due to the lack of recording devices or the cost of precise positioning. Uncovering location information has become feasible with the development of geotags in social media. Bhattacharya and Das [25] proposed a probabilistic model for tracking mobile phone users in the network to reduce the cost of paging. For effectively retrieving geographical information from online resources, researchers have proposed many ways of extracting and indexing and ranking pieces of information related to geography, *e.g.*, [29, 41, 4].

In the era that GPS chips are integrated to personal devices which are affordable by normal users, it has become easier for researchers to obtain large-scale mobility data from city residents to learn their patterns of mobility. For example, finding users' home locations has been studied by a number of researchers. Backstrom et al. [11] proposed to predict users' home locations via their friends in Facebook. Fink et al. [51] built models to predict blog owners' home locations based on place names mentioned in their posts. Cheng et al. [31] tried to predict Twitter users' home locations based on the local words. Mahmud and Nichols [101] and Mahmud et al. [102] improved the performance of the prediction by using temporal knowledge (time zone) and textual knowledge (city names). Flatow et al. [52] approached the problem of geotagging social media messages by modelling regions with N-grams from the textual messages. They also found that the quality of models trained by messages from different sources may vary a lot.

Besides the users' general active areas (home locations), fine-grained mobility patterns have also been explored. González et al. [59], Song et al. [140] and Lu et al. [98] respectively studied and confirmed the predictability of human mobility based on mobile phone records. Herder and Siehndel [65] showed that daily and weekly patterns can be observed clearly from the data of GeoLife GPS Trajectory Dataset. As suggested by Cho et al. [34], periodical patterns and

social ties can be useful features in the prediction of user locations. According to Kulshrestha et al. [73], people are tied to their geographical locations because of limitation of mobility, the influence from their peers, their culture background, *etc.*

The patterns of human mobility imply the possibility of characterizing user preferences via their mobility profiles. This triggered many studies on recommending locations for users based on their visiting histories. Leung et al. [80] proposed a location recommendation system based on activity (sequence of visited stay points in GPS trajectories). Clements et al. [35, 36], Popescu and Grefenstette [120] proposed models for recommending locations tailored to individual tourist taste based on geotagged photos shared on Flickr. O'Hare and Murdock [117] proposed to build smoothed language models for grid cells over the map to improve location prediction for photos. Similar techniques were also applied to online video clips with textual tags to determine location where the videos were taken [77]. Kurashima et al. [75] proposed geo-topical models for restaurants and landmarks based on data from online reviews and geotagged photos. Based on temporal information from locations, Li and Sun [82] proposed a method based on Conditional Random Fields to find location entities in the messages.

Messages on LBSNs also have explicit expression of locations which can be used for location recommendation/prediction. Berjani and Strufe [23] studied location recommendation based on Gowalla data by modelling ratings from users' visits and then applying normal Collaborative Filtering techniques. Ference et al. [50] specifically studied recommending locations when users are out-of-town. Liu et al. [95] presented a location recommendation system based on the category information. Besides these dedicated recommender systems for locations, Yuan et al. [166] proposed a comprehensive probabilistic model to incorporate various attributes related to social media messages, *e.g.*, user, location, vocabulary, time. With this model, they can predict any of the attributes from the others. To alleviate the sparsity problem in geotagged data from social media, Li and Pham [90] proposed an object function based on ranking errors with matrix factorization and a stochastic gradient descent method for learning the parameters. Hu et al. [67] found that the ratings of a business can be affected by that of its neighbour's and proposed a model to incorporate such effect for better rating prediction. A detailed survey regarding LBSNs and location recommendation can be found in [127] and [162].

With GPS-enabled devices, users can know precisely where they are and let the devices record their trajectories while they are moving. This gives researchers an opportunity to study user mobility at a finer granularity, compared to the data from mobile network providers. Yuan et al. [164] and Yuan et al. [165] proposed a system to find the fastest route in a city for taxi drivers based on the knowledge mined from GPS trajectories recorded from taxis. Veloso et al. [151] analysed the same type of data (collected from Lisbon) and explored the possible ways of raising taxi drivers' income. This type of data is also used to reveal the functions of regions (activities related to the regions) within cities [163, 172] and a system was proposed by Giannotti et al. [57] to answer general queries, such as how to predict areas of dense traffic in the near future. As to location recommendation, Zheng et al. [175] demonstrated how GPS trajectories can be used in such systems. To ground all these applications, some mining techniques have been developed for GPS trajectories. Tang et al. [145] and Emrich et al. [46] proposed a method to improve the retrieving of similar trajectories. Detailed surveys about the analysis of GPS trajectories can be found in [119, 173].

Besides building prediction models, location information on social media is used for visualizing mobility patterns in cities. Cranshaw et al. [38] demonstrated a prototype system for revealing clusters of areas based on mobility patterns from Foursquare users. Silva et al. [138] proposed a method to visualize and classify cities based on transition graph of users.

Location information can also be inferred from other sources. Buyukokkten et al. [27] studied and implemented a prototype system to recover the geographical scope of a Web resource using information from domain registries, such as zip codes, telephone area codes and IP addresses. Bennett et al. [22] modelled the spatial distribution of web site visitors and use the models to improve location-centric web page retrieval. Mei et al. [108] proposed a probabilistic model for summarizing blogs' spatio-temporal theme patterns. Wang et al. [155, 154] and Zong et al. [179] studied the location information in Web pages and queries.

As many researchers become interested in the geographical information in social media, there is an increasing need for a forum to compare the methods and share the data, which are usually hard to acquire or preserve due to the technical challenges or regulations. MediaEval[7] is one of the European

[7] `http://www.multimediaeval.org`

annual events for researchers to evaluate their methods on common multimedia dataset and communicate their findings. Placing Task [47, 61, 122, 123] is one of the running tasks in MediaEval. The data sets that have been used by the task include geotagged video clips and geotagged photos. The participants are expected to predict the location for each testing item in the test set by their models trained with the textual and visual information in the training set. The contextual suggestion track in TREC is another annual event for geographical information retrieval. Different from MediaEval Placing Task, this event expects participants to rank suggestions of places to go according to the profiles of suggestion receivers and the contexts which are large metropolitans in the US. Neither of the events have used data set from social media. One possible reason is that redistribution of collected data from those large social media platforms, *e.g.*, Twitter, Facebook, Foursquare, is discouraged or prohibited. For example, the participants of microblog track in TREC had to collect the data themselves from Twitter API via the ids given by the track host [115]. To the best of our knowledge, there is no public available dataset suitable for our studies, thus we had to collect our own data sets. Two close related data set can be found in Cho et al. [34]'s paper or on SNAP[8], which are respectively collected from BrightKite and Gowalla. However, these two data sets do not contain the meta information required in our study, such as category information and location names.

2.3 Methodology

Language Models

Language modelling is one of the most commonly used methods in extracting features from textual content, which is well-known in the information retrieval community [167, 40]. The general idea is to model the probability of words observed from a source, such as a document, a query, a tweet about a location. By comparing the models from different sources, one can obtain a similarity measure on any two sources, *e.g.*, a tweet and a location. We use this method to approach the problem of predicting the origins of tweets from the textual content.

[8] `https://snap.stanford.edu`

Collaborative Filtering

Collaborative filtering techniques are widely used for estimating missing values in a matrix, which are used a lot in the recommender systems community [2]. A common use case is recommending shopping items to users based on historic purchases/ratings made by a group of similar users. The ratings are represented by a matrix where each entry is a value indicating whether a user have purchased an item or how much he/she likes the item. Then collaborative filtering methods can be applied to predict the most likely values of those missing entries, and the corresponding items can then be ranked according to the predicted values.

In general Collaborative filtering methods can be categorized into the memory based methods and the model based methods. The memory based methods are in general easy to implement and perform relatively well. The model based methods use various matrix factorization techniques to model users and items and the missing values can be predicted by minimizing the difference between the prediction (production of the factors) and the reference of the existing ratings [136]. In the study of predicting location from users' previous trails (see Chapter 4), we adopt techniques based on memory based Collaborative Filtering for its simplicity and effectiveness.

2.4 Privacy Issues

Online social networks were born with privacy issues, because they provide a central repository of personal information [20]. Such information is semi opened for general access to serve its use of social networks, which may result in potential exploits. These online social network service providers allow sharing information via different media types, which triggers users' interests in using such services. However, using the function of sharing information in multimedia also makes users to give out personal information. The motivation for the providers is that the more users engaged in the system, the more they can benefit from precise targeted and long exposure advertisement. For example, Facebook allows users' to tag their friends' faces in a photo and users have limited control on whether or how their friends tag them. It may help Facebook to learn more about the users via those tagged pictures and make their advertisement system more effective. Regarding location based social networks, Minch [109] enumerated thirteen issues ranging from collection and

storage to regulation and application and Tang et al. [144] studied how users perceive visual representations of locations with respect of privacy. Heatherly et al. [63] demonstrated how privacy inference attacker can obtain personal information from a vulnerable public dataset. Particularly, Xue et al. [159] studied both prediction and protection of location inference.

Privacy is also a concern in the usage of geotags in social media. It is relatively easy to connect users' locations to their personal lives because the activity normally carried out at a location is closely related to the function of the location. For example, restaurants are for dinning, cinema are for watching films, banks are related to money issues. As demonstrated by Soper [141], there are many good applications that can be derived from analysis of human mobility, especially for local mobile search [100], or mobile advertisement [42], but it can also be exploited by criminals. Johanson [70] reported that posting one's holidays online may attract burglars' attention, as it provides evidence of that the owners are not home. Helped by advances of information technology in social media, it seems that it would be easier for criminals to exploit the information, *e.g.*, they could harvest such information by searching holiday posts in social media.

To address the concern, as suggested by Soper [141], the first step is to study the impact of the technology, *e.g.*, how likely users' locations can be inferred from their online social network behaviours. It is easy to ascertain that one is on holiday if there is a message or photo with words or geotags stating that its owner is not at home. It is not that obvious whether stating oneself is enjoying watching a game can be a clue that the owner of the message will be away from home for a couple of hours. Though resulting methods can disclose more privacy from users' public messages, it can also help service builders to provide guidance for minimizing potential privacy leaks. For example, a client application of social networks may warn you when you post messages revealing too much personal information. Policy makers may also benefit from explicit evidence of how much private information can be learned from users' public social network behaviours so that they can detail how users' privacy should be protected. Otherwise, it may lead to unfair policy for either companies or users.

Part II

Links Between Social Media and Reality

Chapter 3

From Tweets To Locations

The rise of social media makes new dimensions of information about users available in the online world. Geographical information is one of those dimensions, and has only recently become widely available. In general, this dimension of information is contributed by users who want to share their experience at a place, bookmark it or play games in the real world. In this chapter we discuss the correlation between messages posted in online social media and the places where they were posted from. Then we can answer our first research question (RQ1), *i.e.*, whether we can predict a user's location in terms of POIs based on his/her messages on social media.

3.1 Introduction

As introduced in Chapter 1, Twitter is one of the most popular social media publishing and exchanging information online. Twitter allows users to publish messages of up to 140 characters, so-called tweets [76]. Besides textual information, users of Twitter can also attach photos, videos, web pages by including (shortened) links. When viewed through the Twitter API, a tweet is

This chapter is an extension to the publication "The where in the tweet" by W. LI, P. Serdyukov, A. de Vries, C. Eickhoff and M. Larson, in Proceedings of the 20th ACM international conference on Information and knowledge management - CIKM '11

associated with various meta data including the profile of the author, the time of posting, and in certain cases, location (coordinates) where the users sent the tweet. With the assistance of modern technology (hand-held smart devices) users of Twitter can get access to Twitter at almost any time and any place.

In March 2010, Twitter extended its API to provide more accurate geographical information for tweets. As shown in Figure 3.1, Users can specify their locations by tagging a *Place of Interest* (POI) to their tweets, suggestions of which are provided in the Twitter user interface. Each of this kind of tags includes information about the place it represents, such as the name of the place, the address, the city it locates at, *etc.* This service is not yet widely used. According to the data from the microblog track of TREC 2011, there are about 0.6% tweets marked with geotags and only 0.04% is marked with POI-tags. One of the possible reasons that limits the adoption of geotags on Twitter may be related to the privacy issues [178].

Figure 3.1: Twitter interface of selecting users' current location[1]

[1]From https://blog.twitter.com/2010/twitter-places-more-context-your-tweets

In this chapter, we investigate whether and how well such POI information can be inferred from the textual and temporal information of a tweet. From the perspective of service providers, this study may support them to learn more information from users' tweets. Such inferred location information may provide accurate contexts for systems to better understand users' needs, and lead to more efficient and effective interaction between systems and users. For example, they may lead to better mobile search systems [177].

To users who are concerned more about their privacy, the results may serve to raise the awareness that users might already leak location information in their tweets through the textual and temporal information. This study may provide evidence for those privacy concerning users about how much they may expose their privacy through just their normal tweets.

The task, as we have defined, treats places not solely as points located in space, but rather as tags implying the social function of that place. People associate social functions with a place based on why they go there and what they do there. It is our consideration of this semantics that makes our POI prediction more meaningful and better interpretable than mere pairs of latitudes and longitudes. For example, it is normal to have a restaurant and a sports store collocated in the same large mall. The conventional perspective defines a place by its geo-coordinates and is inherently not possible to differentiate such collocated places. Both places may occupy the same geo-coordinates (on different floors) or nearly indistinguishable geo-coordinates (contiguous in the same building). But for humans, it is a relevant distinction whether a tweet is associated with a restaurant or with a shopping place for sports, because the tweets may be composed of words from different vocabulary for the two places. Users may be more likely to tweet about food from a restaurant than from an electronic device store. Thus in our work, we carefully avoid conflation of human-perceived places on the basis of geo-proximity.

In this chapter, specifically, our task is to rank a set of candidate POIs by their relevance to a given tweet. Our assumption is that tweets from a place usually follow a certain set of patterns, especially, in vocabulary which can be modelled by Language Models. Due to the data source, we are facing a two-fold sparsity.

1) From the perspective of tweets, the terms in a tweet, as limited by the length of 140 characters, may be not as abundant as any documents

studied in other tasks to characterize the tweet itself, such as news articles, blogs and web pages.

2) From the perspective of POIs, there might be insufficient tweets for building comprehensive models for POIs.

To alleviate the first problem, we use the temporal information embedded in tweets, *i.e.*, the timestamps when tweets are posted, as additional evidence of its origin. The intuition is that places may have different opening times, *e.g.*, bars are crowded during nights and food places peak during noons and evenings. For the second one, we use the information from web pages returned by search engines for each POI to enrich the corresponding Language Model, as web pages closely related to a place would be ranked at the top positions and they may share vocabularies with the tweets posted from the place.

3.2 Related Work

As shown in Chapter 2, since the rise of smart hand-held devices like smart phones, and GPS-enabled digital cameras, more and more location related applications and platforms become prosperous and it is much easier to record ones' geo-locations via these devices. This new source of information allow scientists to study how geographical information is related to other aspects of human activity and its potential application in Information Systems. Twitter as one of the most popular social media platform introduced its own APIs for location services to facilitate location sharing via POI-tagged tweets. With POI-tags, users can share their experience at a location, bookmark the locations, engage social games online about locations, share photos at the location, *etc.* Not only does a POI tag contain a pair of coordinates indicating its spatial position but it also includes information like the name of the place, the human readable address and the city or state it locates in. Besides POI-tags introduced by Twitter, there are other information sources that convey geographical information, such as addresses of domain name holders in DNS registration data, the addresses in bloggers' profiles, geotags embedded in photos on Flickr, users' check-ins on Foursquare, the place names in web pages.

3.2.1 Geographical Information on Web

At the early stage of the Web, only limited geographical information was available online. Buyukokkten et al. [27] proposed an interesting system using a database of phone numbers of network administrators and a post code (zip code) database to estimate the geographical location of a web site. Ding et al. [43] studied the problem of estimating a geographical scope of web resources (pages) by exploring the features like locations mentioned in the pages and the links from pages with geographical scopes. McCurley [105] discussed various features which can be explored to decide the geo-spatial context of a web site, such as information from WHOIS services and DNS services, routers via which the site is connected to the backbone networks, addresses, postal codes, telephone numbers recognized from Web pages, names of geographic entities, links between a page and other pages with geo-spatial contexts and META tags authored manually in HTML pages.

Another interesting line of research concerns geographical named entities mentioned in web pages. Amitay et al. [8] used a set of heuristics to identify geographical entities specified in a well defined gazetteer and assigned a focus (geographical scope) for a page. Leidner et al. [79] proposed two heuristic rules in disambiguation of geographical named entity (grounding place names). Li et al. [83] and Li et al. [84] studied the ambiguity of geographical named entities and proposed to construct a similarity graph of locations and names and maximize the total score of assigning locations to names. Based upon these studies, Purves et al. [121] and Arampatzis et al. [9] brought up an interesting topic about learning region boundaries from textual content on Web. They used trigger phrases to gather relationships between two geographical named entities, and separated them into two group of points, i.e., inside and outside, with which they then used an algorithm to decide the boundaries.

Compared to our work, these studies have focused on the geographical orientation of Web pages and Web site, which have abundant textual clues compared to tweets. The authors discusses various direct features regarding the geographical information that can be used for their tasks, *e.g.*, place names, addresses registered. As for our task, we focus instead on those implicit information such as vocabulary usage and time of check-in.

3.2.2 Geographical Information in Photos

Naaman et al. [111] initiated a study on the correlation between tags and locations and proposed a prototype system for retrieving photos by implicit tagging and tag suggestion. The rise of social media and introduction of APIs for geographical information stimulated a large body of research on using the geographical distribution of photos. Hays and Efros [62] explored various visual features to locate photo via k-Nearest-Neighbour (kNN) in a global settings, such as direct matches of thumbnails, colour histogram, Texton histogram, line features, Gist descriptors, Geometric context. Along the same line, Crandall et al. [37] explored classification using Support Vector Machines for both visual features and textual features for locating a photo. They found that smoothing visual features with photos taken around the same time by the same person may help locating a photo. Different from those two, Serdyukov et al. [135] proposed a Language Model based method for estimating locations where photos were taken. They built Language Models for each cell of the grid based on coordinates and match them with the Language Models built for a given photo.

In general, the tag based photo locating problem is close to that of locating tweets. We followed the general settings of Serdyukov et al. [135], *i.e.*, using language models for locations. However, different from the photo locating problem, we do not explicitly consider spatial distance in our problem and focus on locations at the level of POI-tags. This difference in levels enables us to distinguish collocated indoor locations such as shops in a mall. It also limits the use of smoothing techniques in spatial dimension as we are trying to differentiate approximated locations. Thus we approach the sparsity problem from another angle, detailed in Section 3.3.1.

3.2.3 Geographical Information on Social Media

The study of geographical signals in social media like Twitter began to intensify when Twitter introduced their geographical APIs. Cheng et al. [31] looked into the embedded geographical information in social media and proposed a method of utilizing identified local words to estimate a user's home town at city level. An interesting fact is pointed out by Hecht et al. [64] concluding that the location entered by users was not as accurate as people had thought before. They even found that those fields may not relate to any geographical location at all. Based on a Multinomial Naïve Bayes Model and a 10,000-term

vector space, they proposed a machine learning method of predicting the users' home city. Leuski and Lavrenko [81] investigated a similar topic based on the chat messages in an on-line game which resulted in a method of predicting events at given virtual locations.

Similar to the argument for the studies in photo locating, Cheng et al.'s [31] and Leuski and Lavrenko's [81] works are based on coordinates which usually suffer from the problem of coarse positioning. That is, they both modelled locations at the level of cities while our task is locating tweets at a sub-city level. We look into individual messages (tweets) which provide less information than all the tweets from a user or a conversation.

3.2.4 Language Models for Information Retrieval

Besides used for locating photos [135], language modelling has been successfully used in speech recognition, machine translation, part-of-speech tagging, and information retrieval [167]. In general, there are two schemes of using language models for the ranking of documents. One is directly based on the probability of the language models generating a given query and the other one is based on the similarity of two language models. As shown by Zhai [167], the latter is actually a generalization of the former scheme. We use the latter one to formalize the proposed method.

3.2.5 Sparsity in the Data

As pointed out by Cheng et al. [31] and Serdyukov et al. [135], occurrences of words in geotagged tweets are sparse, *i.e.*, many words only appear once in the whole corpus. In both of these works, they tried to smooth their models in the spatial dimension: the former defined local words and their weights across space and the latter combined language models from neighbouring cells. The problem is more severe in our task, because fewer tweets (only 0.04% of tweets) have been tagged with a POI and the spatial proximities between POIs do not necessarily indicate their similarities. Moreover, the distances between places varies a lot at the sub-city level, and the distances between similar places are not bounded. These problems limit the use of spatial proximity as similarity for smoothing our models.

Sahami and Heilman [130] proposed a web-kernel similarity measurement for short text snippets, which uses web search results to generate strong models for

them. They queried the commercial search engine with candidate text snippets and collected returned web pages as supplement to the models. Following this idea we propose to use web pages related to location for building richer models. Instead of querying with tweets, we query location names and collect the returned pages for building language models for the given locations.

Another aspect we looked into to dig more evidence of tweets' origin is the time dimension. Twitter associates each tweet with a timestamp which accurately records the time when the user post the tweet. Intuitively, users as humans follow activity patterns in their daily lives, *e.g.*, bars get crowded around midnight and parks are popular on weekends. On the other hand, almost all places have their own opening times and users rarely visit places outside that time. Therefore, the timestamp embedded in tweets may have useful information about where they come from.

3.3 Methods

In general, we consider predicting POI-tags of a tweet as a ranking problem, in which POIs are ranked according to their relevance to a given tweet. A series of models are built for POIs based on different sources of evidence. Based on these models, scores are assigned for each model according to a given tweet. These models are then ranked by the scores and the higher a model of POI is ranked the more likely the tweet is considered to be sent from the corresponding place.

3.3.1 Dimension of Textual Information

To model the relevance of tweets in textual dimension, a unigram Language Model is built for each POI based on the content of the tweets the POI is attached to. Then, those Language Models corresponding to a set of candidate POIs are ranked by their KL-divergences towards the Language Models built from the content of a given tweet.

Formally, for each POI, a set of tweets $\{c_i\}$ which it is attached to is collected, *i.e.*,

$$C_l = \{c_i | l_{c_i} = l\}.$$

The content of each tweet is represented by a set of random variables of terms (unigram), *i.e.*, $c_i = \{W_j\}$. Then the maximum-likelihood estimation is used

for building the Language Model θ_l for the location l.

$$P(w|\theta_l) = \frac{1}{n}\sum_{j=1}^{n}\mathbf{1}(W_j = w)$$

where $W_j \in c_i$, $c_i \in C_l$ and n is the size of the set of tweets in terms of the number of words. $\mathbf{1}(\cdot)$ is an indicator function evaluated to 1 *iff* the input condition is satisfied.

Similarly, we can build a Language Model for a tweet $P(W|\theta_c)$ and compare it with the Language Model for each POI by the KL-divergence between the two models.

$$\mathcal{S}_{\mathrm{KL}}(\theta_c\|\theta_l) = \sum_i P(w_i|\theta_c)\log\frac{P(w_i|\theta_c)}{P(w_i|\theta_l)}$$

Then the POIs are ranked by their KL-divergence to the tweet; the smaller the divergence is, the higher the POI will be ranked.

Due to the limitation of tweet length, it is hard to build strong models for those impoverished POIs using traditional text classification methods, which are typically based on domains that offer numerous documents per category. In this case, we turn to a potential rich source of evidence, *i.e.*, web pages about the location. In general a relevant page about a location contains much lengthier information about the location than short tweets such as the functionality of the location, the designated human activity at the location, the relating items at the location. For example, the pages about cinema usually have information about the cinema itself such as the facility, the films on screening, characters in the films. The vocabulary used in web pages may also match what users tweet about the locations. For another place, the topical focus may be different. As for restaurants, the related pages may render the menus, the styles of decoration and food. For example, it may be likely that users at a cinema post tweets regarding the film they have just watched or the characters in a film which is also listed on the home page of the cinema. It should also be noted that the mental setting of users for posting a tweet from a place may be different from publishing a web page about the place. As revealed by Java et al. [68], most tweets are talking about daily routines and what people are currently doing. Web pages regarding a place, on the other hand, aim at describing the place in a more objective way. The vocabulary use may be different between these two different sources of textual information. Thus it is necessary to investigate how the combination of the two sources

of information can affect the performance of predicting POIs of tweets. For example, one of the geotagged tweets we collected from *Ace Hotel New York* says:

> *Hanging out in the lobby working before my first meeting. If you're up, come say hi.*

A web page returned by the search engine about the hotel highlighted some of the words in the tweets:

> *... a fantastically detailed lobby ... provides space for impromptu meetings ...*

The correlations and differences are detailed in Section 3.3.3

Similar to modelling POIs with textual content of tweets, a Language Model is built for each location with a set of relevant web pages returned from a search engine.

$$P(w|\psi_l) = \frac{1}{n}\sum_{j=1}^{n}\mathbf{1}(W_j = w)$$

where W_j is a random variable of words in the set of pages and n is the size of the set of pages in terms of the number of words.

3.3.2 Dimension of Temporal Information

Human activity usually follows some patterns in daily lives. Accordingly, locations related to certain activities also have visiting patterns. Such patterns may be caused by the conventions and life styles, *e.g.*, the time for lunch or dinner, or by the opening time of a place such as dinning places or parks. As shown in Figure 3.2, the temporal distribution of tweets from *In & Out Burger* is very different from that of *Runyon Canyon Park* within a day, which are both located in Los Angeles. With such kind of knowledge, a tweet is more likely coming from *In & Out Burger* than the park if it is posted around 11:30 pm.

Thus the temporal information embedded in these tweets is an important source of evidence showing the origin of the tweets. On this basis, we propose

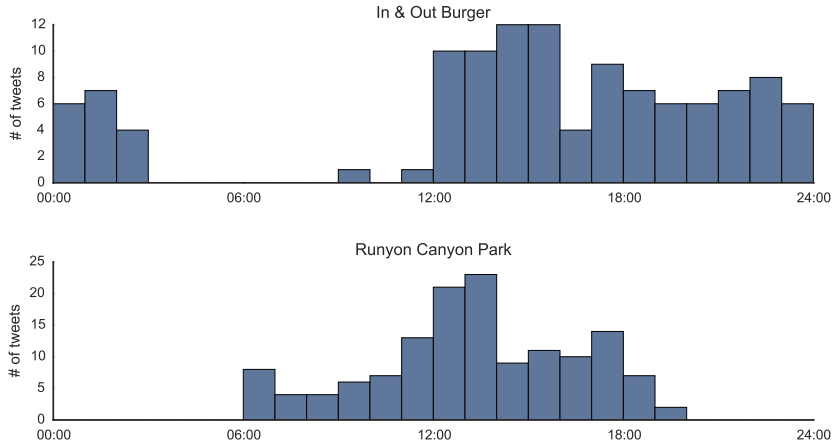


Figure 3.2: The temporal distribution of tweets from *In & Out Burger* and *Runyon Canyon Park*

a temporal model for ranking POIs, *i.e.*, $P(l|t_\delta)$, where t_δ is the relative point of time in a given cycle having a length of δ. Applying Bayes' law to the conditional probability of time given a location, we can obtain the estimation as follows:

$$P(l|t_\delta) = \frac{P(t_\delta|l)P(l)}{P(t_\delta)} \propto P(t_\delta|l).$$

Besides daily cycles, we also consider two additional scales of cycles that may influence human activity, *i.e.*, day, week and month. This is based on the observation of general activity patterns of humans[2]. Thus, a timestamp of tweet can be represented by a vector $\boldsymbol{t} = \begin{bmatrix} t_\mathrm{d} & t_\mathrm{w} & t_\mathrm{m} \end{bmatrix}$, and

$$\boldsymbol{P}(l|\boldsymbol{t}) = \begin{bmatrix} P(l|t_\mathrm{d}) & P(l|t_\mathrm{w}) & P(l|t_\mathrm{m}) \end{bmatrix},$$

where t_d, t_w, t_m are respectively the relative point of time during a day, a week, and a month. To estimate each component $P(t_*|l)$, we create a histogram out of observed tweet timestamps from a location l. That is, for $P(t_\mathrm{d}|l)$ we divid a day into hourly bins and estimate the frequency of tweets posted during each

[2]See Section 3.4.1

hour. Similarly, we estimate the frequency for each day in week cycles and for each day in month cycles.

To combine the evidence from these three scales, we use linear combination, *i.e.*, weighed sum with a parameter $\boldsymbol{\alpha}$. In this study, we use a uniform set of weights, *i.e.*, $\boldsymbol{\alpha} = \begin{bmatrix} 1/3 & 1/3 & 1/3 \end{bmatrix}$. That is, the score from time models is

$$\mathcal{S}_t(l,t) = \boldsymbol{P}(l|\boldsymbol{t})\boldsymbol{\alpha}^T.$$

Then the location can be ranked according to the score $\mathcal{S}_t(l,t)$.

3.3.3 Combining Different Sources of Evidence

As shown, we have multiple sources of evidence for ranking POIs for a given tweet, *e.g.*, textual and temporal information . Similarly to the score function for the temporal evidence, we use linear combination as a total score function of both textual and temporal models. Since KL-divergence (for textual evidence) scales differently with respect to different sources (*e.g.*, tweets, web pages), we first normalize our ranking scores with respect to their own dimensions, *i.e.*, map the scores to $[0, 1]$ and then linearly combine the scores for each POI. Generally, let $\mathbf{X}$ be the score matrix where x_{ij} is the score of POI_i given by the jth model. The normalized matrix is given by $\hat{\mathbf{X}} = \begin{bmatrix} \hat{x}_{ij} \end{bmatrix}$ where

$$\hat{x}_{ij} = \frac{x_{ij} - \min\limits_j x_{ij}}{\max\limits_j x_{ij} - \min\limits_j x_{ij}}.$$

Then the ranking score is based on a linear combination of contributions from all the component dimensions,

$$\mathcal{S} = \hat{\boldsymbol{X}}\boldsymbol{\beta}^T.$$

Here, $\boldsymbol{\beta}$ is a weight vector controlling the contribution of the different sources of evidence in final rankings. We focus our investigation on the performance that can be achieved without tuning the balance between dimensions, and thus in our experiments, all the sources of evidence are weighted equally. Then, we can rank POIs in a balanced manner taking multiple information sources into account.

3.4 Evaluation

In this section, we evaluate the proposed method based on experiments with a collection of tweets from Twitter.

3.4.1 Data Preparation

For our experiment, we need to collect a reasonably large set of tweets with POIs from Twitter's APIs[3]. To achieve this, the following strategy is used:

1) Retrieve an initial set of tweets from Twitter's stream API and filter out those without POI-tags.

2) Collect all the users who sent these tweets with POI-tags

3) Collect tweets from the users in the previous step via Twitter's REST API.

4) Collect all the POI-tags attached to these tweets.

5) Collect tweets by searching for the POI-tags gathered in the previous step via Twitter's Search API.

6) Update the data set with new incoming POI-tagged tweets and

7) Repeat Step 2)-6) to expand the dataset.

Following this strategy, we collected about 31.6 million tweets, crawling from September 2010 to May 2011. However, there are only a small proportion of tweets attached with POI-tags.

A close inspection suggests that most tweets with POI-tags originate from Foursquare[4], an online location sharing platform. Users of Foursquare can check-in at places to share their experiences or bookmark the place. Moreover, they can win titles (*e.g.*, Mayorship) or special treatment (*e.g.*, coupons) if they check-in at a place many times[5]. Some of the check-in messages are

[3] http://dev.twitter.com

[4] https://foursquare.com

[5] http://mashable.com/2010/05/17/starbucks-foursquare-mayor-specials/

re-posted to users' Twitter account and shown in Twitter with a POI-tag to the place. These tweets often follow the pattern "I'm at <place name>. https://4sq.com/XXXX", where the name of the place is embedded and the short link containing a unique code to the page about the place on Foursquare. These text snippets are usually automatically generated by Foursquare. In general, for our experiment, it would be trivial to predict the origin of the tweets having the POI names or code in their text content. Thus, we remove the links started with "https://4sq.com" and the text snippets in the pattern "I'm at <place name>.", *i.e.*, we remove the tweets containing only the auto-generated content.

In the end, we have collected for our experiments a dataset of 700,288 tweets with POI-tags from 177,817 POIs, posted by 52,488 different users. As shown in Figure 3.3, the distribution of tweets follows the power law. Only a few POIs (about 0.16%) are supported by more than 100 tweets and 93.11% of POI-tags are used less than 10 times in our data set.

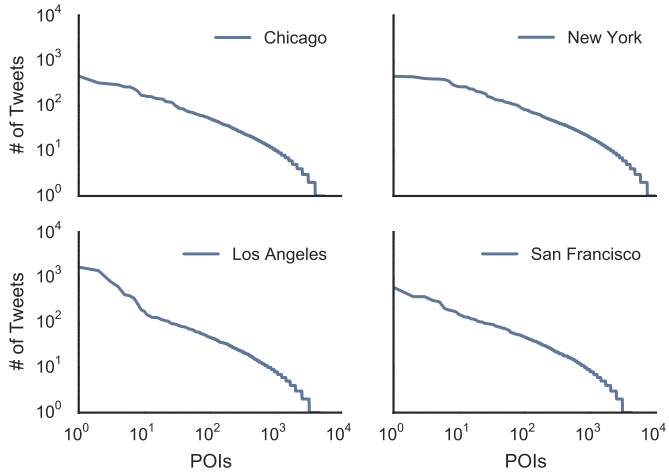


Figure 3.3: Tweet distribution among POIs log-log scale

In our experiment, we focus on four big cities in the USA, namely, *Chicago*, *Los Angeles*, *New York* and *San Francisco*, since these cities have more tweets than other cities and more diverse users and POI-tags. For each city, we select the top 10 popular POI-tags, including shops, restaurants, parks, cafeterias and clubs. These POIs are selected because they have relatively rich sources

of tweets (around 100-400) in our dataset.

3.4.2 Building Models

Before building language models for each POI-tag, we pre-process the textual content of the tweets by streaming it through a stemming tokenizer with a stop words filter from WHOOSH[6], which is used to extract terms from the text. Then, for each POI-tag, we build a language model based on the terms from the tweets with the POI-tag. In addition to textual evidence from tweets, we also build a language model for each POI based on the relevant web pages returned by a search engine. To achieve this, we query each POI name in Microsoft Bing and gather the textual content of the top 30 returned web pages with HTML tags being filtered out. Similarly, we stream the textual content of the web pages into the tokenizer and filter pipeline to build language models for each POI.

3.4.3 The Ability of Differentiation

Our first task is to find out whether POIs can be distinguished through the language models built from tweets. We split the set of tweets from each POI into two equally large subsets and build a language model for each subset. The distance between two POIs is then calculated by the KL-Divergence between the two language models, *i.e.*, the distance between the language model built from the first subset of tweets of a POI and that from the second subset. The distances are rendered into a confusion matrix for the set of POI-tags from each of the four cities. The confusion matrix for Chicago is shown, as an example, in Figure 3.4, in which the lighter a cell is, the farther the models are apart from each other in KL-divergence, *i.e.*, the more different the two models are.

Significant differences can be observed between all pairs of POIs except for those that compares to itself. This observation supports our assumption that language models are able to capture the differences between POIs. To put it in another way, the words used in tweets are more similar among that from the same POI than across POIs. From the figure, we can also observe that some pairs of language models are more like each other than other pairs, *e.g.*, *AMC River East 21* and *Century Center Cinema* which are both cinemas. For

[6] `http://whoosh.ca`

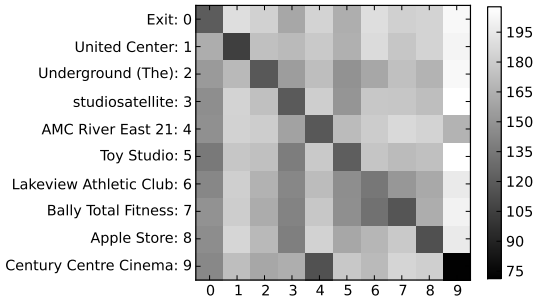


Figure 3.4: Confusion Matrix for POIs in Chicago

another example, the language models of *Lakeview Athletic Club* and *Bally Total Fitness* are closer in terms of KL-divergence and they are both fitness place. This implies that tweets from similar POIs (in terms of human activities or functionality) are also similar. This fits into the assumption that tweet vocabulary is to some extent in accordance with the place where they are.

3.4.4 Evaluation Methodology

For evaluation, we prepare the training and test data as follows. We select the top 10 places from each of the four cities to assure there are enough tweets for building the models. Then from each place in a city we randomly select a set of tweets in a size of s as the training dataset and other 10 tweets as the test dataset. For each city, we have $s \times 10$ tweets for training the model and 100 tweets for testing. We build models from the tweets in the training dataset according to Section 3.3. Then each tweet in the test dataset is queried against the model, which will produce a ranking list of the 10 POIs in the corresponding city.

To evaluate the proposed models, we use a modified precision curve. Different from typical retrieval systems which usually have multiple relevant items, in our task, there is only a single POI-tag considered relevant (*i.e.*, the reference place). The higher the reference POI is ranked, the better the ranking method performs. In our evaluation, we choose to use the frequency of the reference POI being ranked above the p-th place. Then a curve is plotted to show the relation between the frequency and the threshold p. The larger the area under the curve, the better the ranking model performs.

To verify the statistical significance of the difference between the proposed methods, we use Wilcoxon Signed-Rank test [21]. For each tweet in the test set we record and compare the ranking positions of the reference POI (*i.e.*, that is attached to the tweet) from different methods. Then we test the ranking positions with Wilcoxon Signed-Rank tests and consider those pairs of methods giving p-value < 0.05 as having statistical differences in performance.

3.4.5 Experiments

In this section, we will show the experimental performance of the proposed methods and discuss different factors that can affect the performance.

Models Trained With Sufficient Tweets

To evaluate the performance of the proposed methods, we first train the models with sufficient number of tweets ($s = 100$). The results are shown in Figure 3.5.

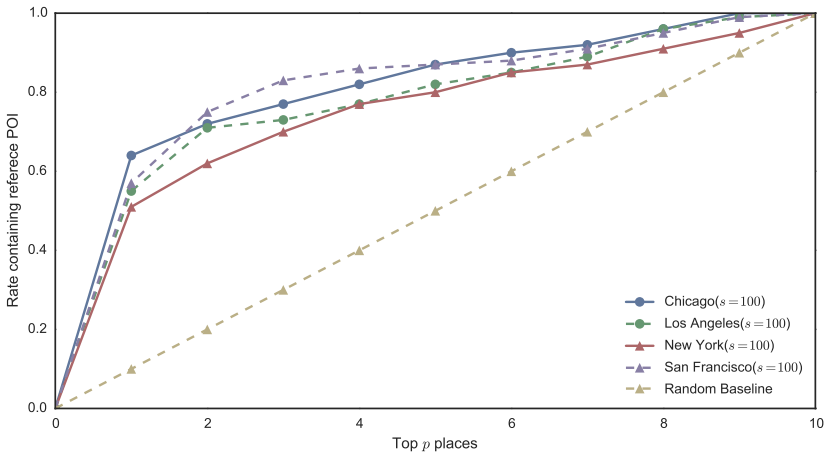


Figure 3.5: Ranking with rich set of tweets

The evaluation results are positive (41%–54% better than random baseline under P@1) in all four cities when we have enough tweets to build strong models for POI-tags.

A close inspection on the dataset suggests three types of tweets which the proposed models do not work well with:

1) The tweets whose content is short and do not carry any location-specific information, such as "Thank you", "yaaaa".

2) Words with a strong relationship to a place may not appear often enough when the data size is small, *e.g.*, "swim" usually implies fitness, however, it only appears in 3 tweets in the dataset.

3) Tweets that contain only transient terms for a location, *i.e.*, words relate to a place for a short while and then rarely come back again, *i.e.*, the title of a film.

In general, the failure of prediction happens when neither models nor the given tweets contain location related terms. To alleviate the problem, a large quantity of tweets is needed for training comprehensive models for POI-tags.

Sparsity

As mentioned before, the sparsity of the data has potentials to affect the performance of the proposed models in our task. To show how much data sparsity can affect the ranking performance, we reduce the number of tweets s used for training the models and compare how the models perform at different levels of sparsity (*i.e.*, POIs at different levels of popularity). The results of this modified setting are shown in Figure 3.6, which renders deterioration of performance when the number of tweets used for training drops. Particularly, for the case $s = 10$, the model degenerates to a random baseline (shown as the diagonal line in the chart). Therefore, we turn to our additional source of information.

Smoothing with Additional Sources of Evidence

As mentioned, to deal with the sparsity problem we proposed to use two addition sources of information for modelling POI-tags, *i.e.*, web pages returned by a search engine and the posting time of tweets.

To integrate the models enriched by web pages, we rank the candidate POIs for a given tweet by both models trained with tweets and models trained with

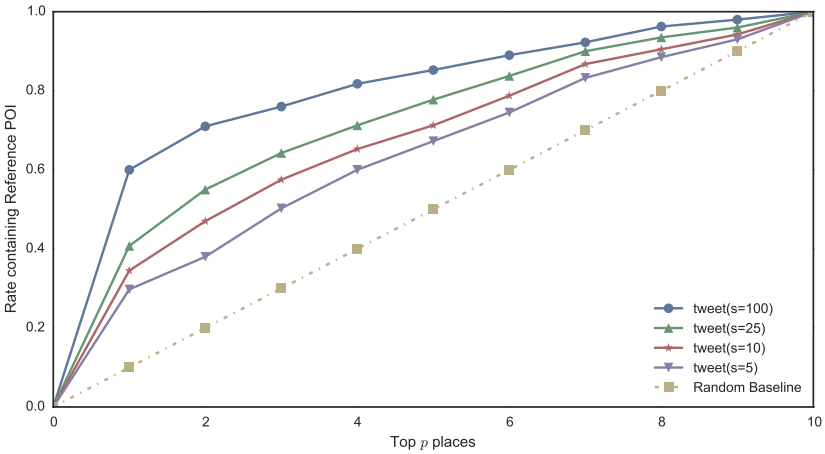


Figure 3.6: Ranking at different levels of sparsity

web pages related to POI-tags. Then the scores from the two set of models are combined (see Section 3.3.3) to generate the final ranking. The curves in Figure 3.7 shows that the evidence coming from web pages can help increase the performance (around 10% increase in P@3) of ranking when POI-tags are supported by only a few tweets ($s = 5$). On the other hand, it degrades the performance somehow in the case that POI-tags have sufficient supporting tweets ($s = 100$).

A positive example for web page based smoothing can be found with the place *Lakeview Athletic Club*[7]. It is ranked higher by the web-enriched models than by the models trained only with tweets. An opposite case is observed in the place *Nokia Theatre*[8] which is better ranked by models trained with pure tweets than by web-enriched models. An inspection on the tweets suggests that the difference may be attributed to the gap between vocabulary use in tweets and web pages or the lack of extractable textual information from web pages. Continuing with the negative example, most of the tweets from the theatre are talking about ongoing shows while the top three web pages from

[7] `http://chicagoathleticclubs.com/locations/lakeview`
[8] `http://www.nokiatheatrelalive.com/about`

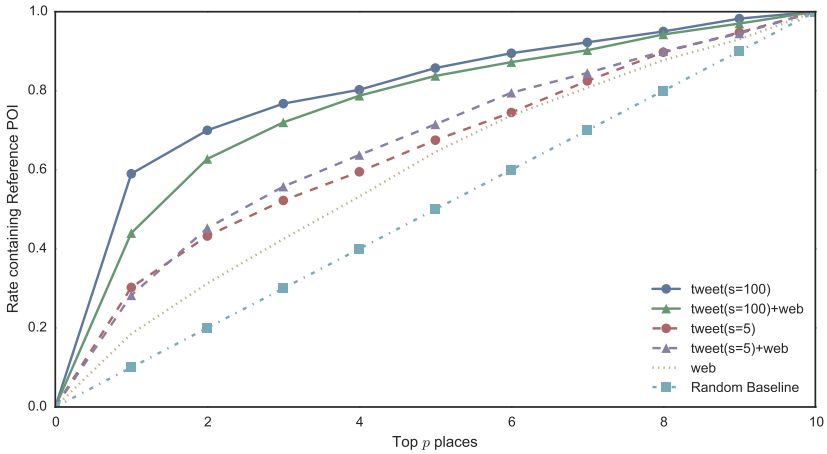


Figure 3.7: Web-enriched rankings

the search engine are as follows. The top page returned is its home page which includes only big Flash objects without much textual information extractable and the second page is about a related theatre (*Best Buy Theatre*). The third page is a Wikipedia page about the theatre. All of them do not have much textual information about the shows and thus can not match the content of the tweets.

To alleviate these problems, a filter should be used to insure the textual content in the web pages returned by search engine and its relevance to the POI-tags. Moreover, a better tuned weight parameter may help balance the diversity and consistency of using the two sources of evidence, *i.e.*, tweets and web pages.

Another source of evidence is the temporal information embedded in tweets. The experiment results of the proposed models enriched by temporal information is shown in Figure 3.8, in which the model is compared with the web-enriched models and models trained with sufficient tweets. As can be seen, the time model can further improve the performance of both language models of POIs having sufficient tweets and web-enriched models. However, the time model is also affected by sparsity problems, therefore it cannot substantially

increase ranking performance (around 10% increase in P@3). Looking into the POIs that the model does not work well with, we find that some POIs like *In & Out Burger* and *Best Buy* are always busy and cannot be characterized by the time model.

This experiment shows evidence that the proposed method can distinguish between tweets sent from the top 10 POIs of each city. However, the data samples are relatively small in the experiment set up due to data sparsity. Looking into the data in the experiment, we find that the POIs used in the experiment vary in their kinds. For example, there are two clubs, two gym centers, two cinemas, two office buildings, a stadium, and a tech shop in the data set of Chicago. In general, a POI category (*e.g.*., clubs) is associated with more tweets than a single POI in that category does, and POI categories often carry useful information about human activities. Thus, we also explore the proposed method for predicting tweet origins in terms of POI categories, which is detailed in the next section.

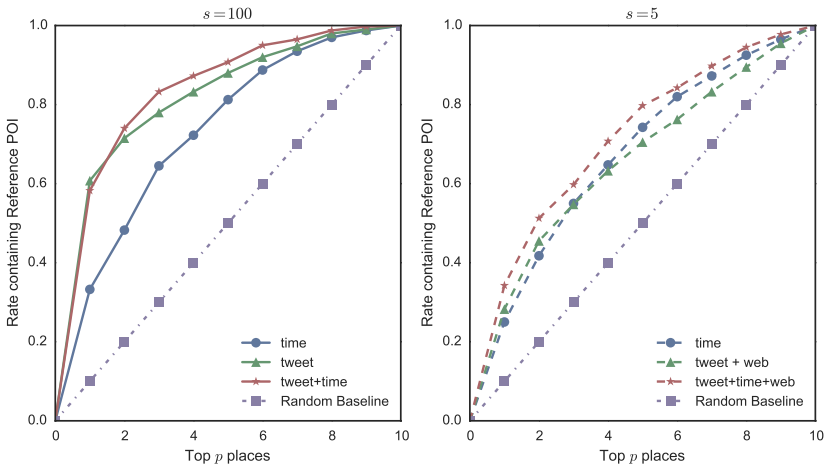


Figure 3.8: Ranking with comprehensive models

3.4.6 Predicting Location Categories

As mentioned in Section 3.3.1, each location has its own function in human society, a restaurant for dining, a cinema for watching films, *etc.* Thus it is

possible that the function of the location can be reflected from users' check-ins, via words used in the tweets and the time when the tweets are posted. To verify the conjecture, we collected the category information from Foursquare [54] for the locations in the dataset, via searching the names of the location around the coordinates with Foursquare's APIs. The returned places are with categories like *Fast Food*, *Art Museum*. However, some locations in our dataset have no corresponding category information in Foursquare's database, thus the dataset of tweets with categorized POI-tag is smaller than the location dataset. The key statistics are shown in Table 3.1.

Table 3.1: The statistics of geo-tagged tweets with location category information

Entity	Number
Tweets	382 654
Locations	52 775
Users	36 150
Location Categories	306

We use the same approach to model each *location category* and predict the category information of test tweets. As shown in Figure 3.9, the number of tweets used for building the models of location categories vastly influence the performance of prediction. Similar patterns are observed, *i.e.*, the more tweets used for building the models, the better the prediction is.

In general, though the prediction performance is better than random guess (which produces a diagonal line), the category prediction is not as good as location prediction. This may be due to that the tweets from the same location are more coherent than those from diverse locations belong to the same category. It is likely that the model of a location category is trained with tweets from a group of locations and is tested by tweets from a different group of locations, though both belong to the same location category.

3.5 Conclusion

In this chapter, we have shown the viability of applying a ranking approach to the prediction of the POIs of tweets' origin at location level within a city. Using a language modelling method, we can achieve better performance given enough tweets for building models of a POI-tag than random guess. For those POIs

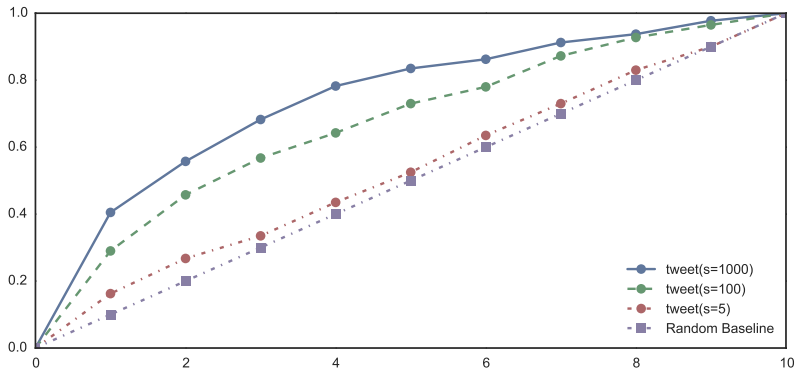


Figure 3.9: Ranking location category

not associated with sufficient tweets, we show that using web-enriched models can significantly improve ranking performance by combining the prediction from models trained with tweets and models trained with web pages. We also demonstrate the use of temporal information to improve the prediction performance. In spite of the sparsity problems, the proposed temporal models can improve the performance of prediction in all circumstances, *e.g.*, sufficient tweets, insufficient tweets, with or without web-enrichment. In general, we conclude that POI-tagged tweets have a strong correlation with their place of origin, which can be predicted from the textual and temporal information. Moreover, we also verified the conjecture that the functions of locations are also related to the tweets where they were sent from.

To answer the research questions, we use words from both tweets and web searches and temporal features to build models of POIs (RQ1a) and also demonstrate the ability of both textual feature and temporal feature to distinguish POIs (RQ1b). In the experiments we show that the proposed models can achieve better performance than the random baseline(RQ1c).

There are still challenges in this task. The web-enriched models have their own limitation if there are sufficient tweets for modelling a POI-tag. Presum-

ably, the gap between tweet vocabulary and web vocabulary regarding the same location severely affects the performance of prediction. It is challenging to decide on a good parameter for blending in the information from the web source. It is not clear what role the web source will play if the prediction is in a larger pool of POI-tags.

The users' identity, online friendships may be good resources of evidence for locating a tweet, which should be included in the future work along this line. The parameters used in combining scores can be tuned according to different places since places of different categories may have their own preference on the sources of evidence.

It should also be pointed out that the privacy disclosure can be harder to avoid in current social media era. This study demonstrates how location information can be inferred from other aspects of messages from social media. This may also lead to studies and applications that can warn users when their activity leads to potential privacy leaking.

Chapter 4

From Past Locations To The Future

Previously, we have shown empirically that the content of tweets can be exploited to predict locations or their associated properties (categories). In this chapter we look into the problem of predicting the origins of tweets from a different angle, that is, trails. Then we can answer the second research question (RQ2), *i.e.*, whether we can predict users' future visits to POIs by only using users' visiting histories.

4.1 Trails as Activity Patterns

As an intuition, everyday, a city resident need to visit several places to accomplish daily routines, *e.g.*, work, food, clothes and entertainment. However, the sequence of visiting all these places usually depends on their own schedule of a day. For example, a student may have different types of places to go to on working days, such as the lecture rooms, the canteen and the apartment. In the weekend, a trail through the city centre, a cinema and followed by a restaurant,

This work is an extension of the publication "Want a coffee? Predicting Users' Trails" by W. Li, C. Eickhoff and A. de Vries, in Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '12

is more likely. Since sharing location has become a trend, many users post about the locations they visit on social networks, like Twitter, Foursquare[1], Facebook[2]. This grants researchers an opportunity to study the mobility of city residents though the public data provided by online social networks. The knowledge derived from that data may lead to applications that are more context-aware.

Twitter, as one of the most popular online social network platforms, provides APIs for geo-information, via which users can attach a location to their posting messages. In general there are two levels of information regarding the location provided by two types of tags. Low level location information is encoded in geotags which are purely the coordinates which can only represent physical locations. High level information is encoded in *Point of Interest* (POI-tags) which are composed of more human friendly information like names of places, addresses, etc. Geotags are usually attached automatically by the devices if users approve it, while POI-tags are manually selected from a list of suggestions based on the user's current location. The geo-information APIs and the presentation of geotags and POI-tags have been adjusted a couple of times which may be due to the concern of privacy.

Besides general online social networks like Twitter, location based social networks focus specifically on the function of sharing user locations. Foursquare as one of the leading location based social networks includes category information in their POI-tags which can be accessed though their public APIs. The category information embedded in Foursquare's POI-tags reveals the connections between the location and users' activities. For example, a location labelled as a restaurant is related to the activity of dining while shops are related to buying and selling stuff. While a clue of what the place is for may have been embedded in names, a lot of places are named very distinctively so that consumers are more likely to remember them. Understanding a place's role can be more easily achieved through manual labels, provided by Foursquare.

A reasonable prediction of user locations in terms of activity (estimated from the POI category) would support accurate context modelling for context-aware applications. One example of such applications would be precise targeted advertisement. Such systems can display advertisement more relevant to users' activity which may be more persuasive rather than listing all kinds of places

[1] `https://foursquare.com`
[2] `https://facebook.com`

around the place. It would even be better if the system could know ahead about what users are going to do in the near future and response accordingly. For example, it would be more useful to show advertisements about bars when knowing a user is going for a drink after work rather than showing bars when the user is already in a bar.

In this chapter, we will look into how to predict users' next locations in terms of categories based on their previous visits. Geotagged tweets are sparse in the Twitter sphere which are only 4.3% and POI-tagged tweets are even sparser which contribute to 0.03% according to the data set provided by TREC 2011 [115]. Besides, even the active users of POI-tags do not check-in at every place they visit, which leaves "holes" in their trails and makes the trails even more sparse.

4.2 Trail Prediction

Consistent with the previous chapter, we refer to an online message with a POI-tag as a check-in, formally represented as $\langle l, t\rangle$. In this chapter, we focus on the location types rather than individual locations, therefore, l represents the category of the location indicated in a POI-tag; t is the timestamp when the check-in is made. A user's trail is defined as the set of check-ins made by the user u, *i.e.*, $\{\langle l, t\rangle | \langle l, t\rangle \in C_u\}$. For example, a trail from a university student can be represented as $\{\langle \text{University}, 8:00\rangle, \langle \text{Food}, 12:43\rangle, \langle \text{University}, 13:30\rangle\}$. With this, we make an estimation on how likely a message tagged with a library POI will appear in this trail (check-in set) during the evening. This is based on the knowledge that a typical trail of students usually has a check-in at a university library during the evening.

Formally, based on a set of location categories L and a trail $\langle l_i, t_i\rangle$ where $l_i \in L$ and t_i is a timestamp, we would like to rank all location categories in L according to the likelihood they will be observed at a give timestamp queried. The problem of trail prediction is then defined as a prediction problem, where given a trail of a user C_u and a query time, we predict the category of next location that the user will visit.

4.3 Related Work

Before the rise of smart phones and social media, to study human mobility, subjects either wore GPS-loggers to record their trajectories [174, 110] or were tracked though cellular networks [25, 10, 107]. For cellular networks, it is important to track users' location in terms of base stations as they need to be paged for incoming calls. The first study on predicting mobile users' behaviours to reduce the paging traffic in cellular networks was carried out by Bhattacharya and Das [25] and they proposed Markov Chains based models for the problem. As for recorded GPS-trajectories, Ashbrook and Starner [10] showed that second-order Markov Chains have the best precision among nth-order Markov Chains.

Instead of physical locations, the environment that users are surrounded with (also known as contexts) usually provides much richer information and many researchers have tried to infer context information based on coordinates. Meeuwissen et al. [107] showed that in the setting of inferring and predicting mobile users' context based on cellular network position trajectories, a prediction by partially matching the trails is better in continual updating datasets. By analysing recorded trajectories, the transportation type (walk/bus/car/bike) can be distinguished in each segment of trajectories (see [174]). Monreale et al. [110] tried to predict future visits for cars from recorded GPS-trajectories by building a prefix tree on the collected trajectories. Lv et al. [99] studied temporal features of trajectories (*e.g.*, dwell time, visiting frequency) which were then used for estimating the semantic categories of locations. Based on GPS-trajectories, Zheng et al. [175] try to recommend locations or activities to users based on users' historical visits and correlations between locations. Later, by abstracting *stay regions* from trajectories, they proposed methods to recommend locations and activities to users based on their current location [171]. As GPS trajectories may be sampled at a low rate which may introduce uncertainty, Zheng et al. [170] proposed a History Route Inference System to reduce such uncertainty. These studies based on fully recorded trajectories which covered all the locations the subjects had been to. Trails embedded in social networks usually reflect only part of the trajectories, or, in other words, these can be seen as trajectories with many missing values.

As Location-based Social Networks have become popular, user contributed location information not only enables other users to find out interesting locations but also provides a huge volume of data about locations and user mobility.

Such data triggered a large number of studies on location recommendation and prediction. Flickr is a popular online photo sharing service where many photos are shared with geotags (geographical coordinates). By clustering those geotagged photos into landmarks, it is possible to make personalized recommendation of landmarks to users [137, 36].

Since Twitter started supporting geotags, researchers have been able to locate users' home towns by the vocabulary used in geotagged tweets [31]. Mahmud and Nichols [101] later investigated different features such as place names and hashtags for the same problem. Bao et al. [18] used Foursquare tips for recommending locations based on matching users' preferences and experts' preferences. As Foursquare's check-ins and Twitter's POI-tags carry more accurate information about the users' locations, more knowledge can be learned from users' mobility, such as cyclic visiting patterns [32, 113], regional visiting patterns [38]. The influence between users can also be used for recommending new locations [56]. Similarly, Sadilek et al. [128] studied the relation between friendships and locations and tried to estimate users' locations based on their friends' locations. All these location recommender systems focused on the exact locations and the time range for the prediction is very large and loosely limited, *e.g.*, in months or even years. In our trail prediction, we focus on the category information and the predicted visits are in a short time frame, *e.g.*, in hours.

A closely related work is carried out by Kurashima et al. [74] who combined Markov models with a topical model based on geotagged photos on Flickr for recommending the landmark visited next. Cheng et al. [30] studied the same problem and they integrated visual clues from photos in their personalized location recommendations, *i.e.*, they classify the users by the facial attributes extracted from the photos. Their tasks are in general similar to ours, which is predicting users' future activities. However, we focus on the category of locations within a city while they focus on individual locations. The photos on Flickr of which a trajectory is composed of usually falls into a small set of categories of locations which are more popular for tourists. The check-ins on the other hand, are more diverse in categories of locations, ranging from restaurants to universities.

4.4 Methodology

As mentioned in the example in the previous section, student trails may share a lot in common. The evidence presented by Gao et al. [56] suggests that users have more similar interests within their local circles than to arbitrary users. Similarities between users can be an important factor in predicting user mobility.

In both Noulas et al.'s work [113] and Chapter 3, check-in behaviour shows a correlation with time. For example, a lot of check-ins originate in café locations in the morning, while supermarkets dominate the early evening. Noulas et al. [113] also illustrated the difference in check-in distribution between weekdays and weekends. The temporal information is also an important factor to include in predicting user mobility.

Users' daily trails may also depend on many other factors, such as the days of the week, local culture, their own social circles, age groups and gender. The variety of factors that may be related to user mobility patterns motivates us to focus on a data-driven approach and model the similarity between users purely based on their life style reflected in their trails.

In this section, two different approaches to the problem are discussed. In general, a user's trail can be seen as a sequence of transitions between different location categories (activities), which can be modelled by Markov Chains. We propose a simple method based on Markov Chains to predict users' trails. From another perspective, as users' trails reflect their life styles, the trails can be seen as users' interests of going somewhere or performing an activity at a certain time. We propose therefore a second approach to the problem with a Collaborative Filtering based method, as Collaborative Filtering has been shown effective for predicting user interests in items [136]. In this study, the performance of both approaches are evaluated empirically on real data collected from Twitter and Foursquare.

4.4.1 Markov Chain Models

As shown by Kurashima et al. [74], Markov Chain Models can be used for recommending tourism routes based on geotagged Flickr photos, which in general is also a trail predicting problem, *i.e.*, predicting the trail a user would follow during his/her tour. In our trail prediction problem, the category of locations a user checked-in at can be seen as a state. Check-in at another

category of location infers a transition to another state. Formally, let a trail be a sequence of location categories represented by random variables X_i indicating the location category. Predicting the next location category at which a user would check-in will then be based on the following conditional probability.

$$\mathcal{M}_{\text{MC}} = P(x_i|x_{i-1}, x_{i-2}, \ldots, x_0)$$

By applying the first-order time invariant principle (Markov Chain Model), we have:

$$\mathcal{M}_{\text{MC}} = P(x_i|x_{i-1}, x_{i-2}, \ldots, x_0) = P(x_i|x_{i-1})$$

Then location categories are ranked according to the conditional probability $P(x_i|x_{i-1})$. In this model, the timestamps in a trail of check-ins are simplified to the chronological order of the states of location categories visited.

To estimate $P(x_i|x_{i-1})$, we use simply the Maximum-likelihood estimation:

$$P(x_i|x_{i-1}) = \frac{|\bigcup_{u \in U}\{\langle c_j, c_k\rangle | \langle c_j, c_k\rangle \in R_u, l_{c_j} = x_{i-1}, l_{c_k} = x_i\}|}{|\bigcup_{u \in U}\{\langle c_j, c_k\rangle | \langle c_j, c_k\rangle \in R_u, l_{c_j} = x_{i-1}\}|}$$

where $R_u = \{\langle c_j, c_k\rangle | c_j, c_k \in C_u, \nexists c' \in C_u : t_{c_j} \leq t_{c'} \leq t_{c_k}\}$ includes all pairs of consecutive check-ins in the user's trail, c_j and c_k are check-ins, u indicate a user, t_c and l_c are respectively the timestamp and location category of a check-in.

4.4.2 Collaborative Filtering

Collaborative filtering is a successful technique developed in recommender systems [93], where users' interests on un-purchased/un-rated items are estimated. General location recommender systems have applied Collaborative Filtering algorithms, see *e.g.*, [137, 36, 18]. Trail prediction can be seen as finding out users' interests in visiting certain categories of locations at certain times. This analogy inspired us to use Collaborative Filtering for the trail prediction problem. Collaborative Filtering is suitable in cases with missing values, a problem that also presents in our data due to the hidden part in trails because users may not check-in at every place they visit.

Interest Profiles

Collaborative Filtering models user interests based on their ratings on different items. The items that the algorithm predicts the users will rate highly but

have not yet bought/watched/read/*etc.*, are then recommended to the users. In trail prediction problem, instead of ratings towards an item, we consider check-ins as ratings towards visiting a location of a certain category at a certain time. Predicting a user's future trail can then be interpreted as predicting the user's interest in visiting a location of some category in some future time window.

Many different approaches have been proposed for improving the performance of general and task-specific Collaborative Filtering based methods. In our experiments, we base our methods on the class of memory-based Collaborative Filtering, which measures similarities between neighboring data points, and predict rating on a given item by aggregating ratings from similar data points (users or items) [125, 133]. While alternatives may be considered in future research, memory-based Collaborative Filtering has the advantages of good performance and ease of implementation, providing a suitable baseline to explore.

As shown in Section 4.2, the trail is composed of check-ins which is a set of points in both location and time dimension. A method to converting point sets to vectors is needed to define a similarity measure on two trails in memory-based Collaborative Filtering methods. In the experiment, we discretize time dimension into slots for vectorizing trails.

Formally, each user's check-in profile is represented by a matrix $\boldsymbol{r}_{|L|\times m}$, in which the value of each entry $r_{l,t}$ represents the likelihood that the user is at a location with associated category l at time slot t. Thus, a check-in at a location will indicate that the user is at the location and the time slot of the location category will be increased by 1.

For a given trail $\boldsymbol{r}$, the top n similar trails $\tilde{\boldsymbol{R}}$ are selected and aggregated with the following formula. This estimates the user's interests towards a location at a certain point of time, when going along the given trail:

$$\hat{\boldsymbol{r}} = \frac{1}{\sum_{i=1}^{n} f_{\text{sim}}(\boldsymbol{r}_i, \boldsymbol{r})} \sum_{i=1}^{n} f_{\text{sim}}(\tilde{\boldsymbol{r}}_i, \boldsymbol{r}) \boldsymbol{r}_i$$

where $f_{\text{sim}}(\cdot, \cdot)$ is a function for measuring the similarity of two vectors. In this study the commonly used cosine function has been taken as the similarity

function.

$$f_{\text{sim}}(\mathbf{x}, \mathbf{y}) = \cos(\mathbf{x}, \mathbf{y}) = \frac{\sum x_{ij} \cdot y_{ij}}{\sqrt{\sum x_{ij}^2}\sqrt{\sum y_{ij}^2}}$$

The estimation $\hat{r}_{l,t}$ is used for ranking the location category l as a prediction at the time point t, which is noted as

$$\mathcal{M}_{CF} = \hat{r}_{l,t}.$$

Trail Smoothing

Check-ins are only snapshots (sample points) of users' trajectories and they are too sparse for training models. To address this problem, we use smoothing techniques to propagate the positive evidence of users' presence to the neighbouring time slots, which can be considered as reconstructing actual trails from snapshots. The intuition is that when a user checks in at a location at some point of time, the user does not merely stay at the location for a very short time. On the contrary, the user usually stays there for longer time, which reflects the user's interest in the place where the user checks-in. Inspired by such a heuristic, we assume that users usually stay at the location around the time point they check-in at for, *e.g.*, one hour.

To model how likely a user stays at the location at a point of time that is different from the time point when the check-in is made at the location, we use a bell curve function to approximate the likelihood. The bell curve function is defined as follows:

$$\mathcal{K}(t) = \mathrm{e}^{-\frac{t^2}{h}},$$

where h is a parameter to control the span of the curve.

There are four reasons motivating us to choose this specific function. The first is that the function gradually goes to a lower value when t goes off the centre, which represents that the likelihood a user stays at the location diminishes when no further check-ins are observed at the location. The second is that the function goes down slowly when t is leaving the centre and goes down faster when it is further away from the centre, and then gradually approaches to zero. This indicates that there is a time range around the check-in, within which the user is very likely to stay and the likelihood will steeply drop when t is beyond the range and the range can be controlled by h which can be considered as a bandwidth parameter. The third reason is that the maximum

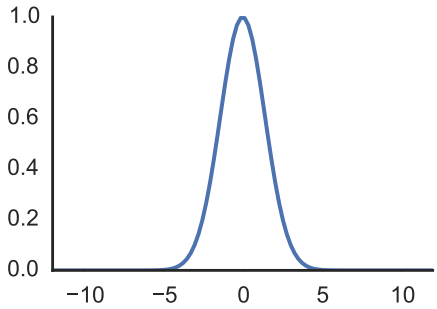


Figure 4.1: The bell curve function ($h = 2$)

value 1 is achieved at $t = 0$, which indicates that the user is absolutely at the location when a check-in made by the user is observed. The fourth is that the function is easy to achieve in a simple mathematical formula. As can be seen in Figure 4.1, a check-in observed at a point of time will bring up the likelihood of the user being at the same location around the time.

For multiple check-ins at locations belonging to the same category, we aggregate the bell curves for each check-in:

$$r_{l,t} = \sum_{i=1}^{n} \mathbf{1}(l = l_i)\mathcal{K}(t - t_i) \tag{4.1}$$

By this smoothing step, we can to some extend alleviate the problem of sparsity in the data set and we refer to the collaborative filtering method with smoothing as $\mathcal{M}_{CF-K}$.

Figure 4.2 shows a smoothed trail, which contains 6 check-ins at 6 different locations respectively belonging to 5 categories. The two horizontal axes respectively represent the time dimension and location category dimension of the trail and the perpendicular axis represents the likelihood of user being at the location at the time. For example, the rising surface at *bar* and around 00:00* means that the user was very likely in a bar around those time slots. In the experiments, the smoothing is only applied to time dimension. A smoothing method for category will be explored in the future work, *e.g.*, a method based on topical similarity between categories.

* indicates the next day

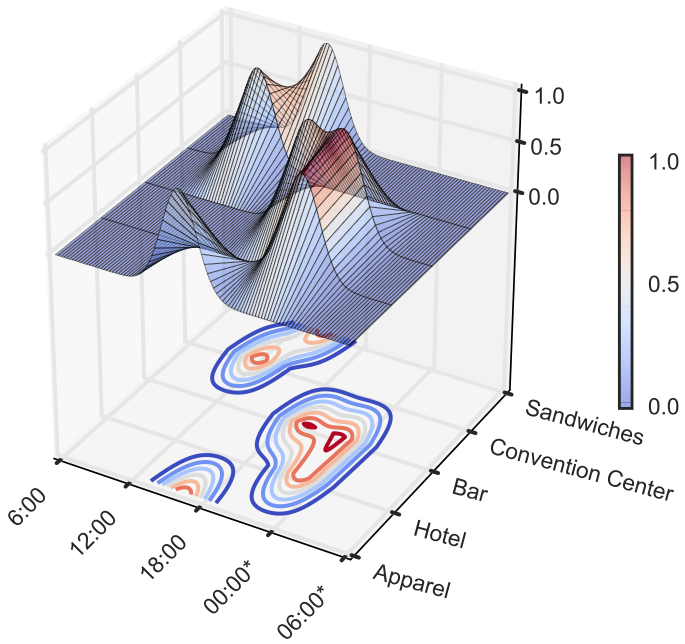


Figure 4.2: A kernel density estimation of trail ($h = 2$ hours)

4.4.3 Baselines

To show the effectiveness of our proposed models, we construct three baselines for the problem. For classification problems, an important baseline to consider is simply to use the major class ($\mathcal{M}_M$) as the prediction. In general, predicting the next location category can be seen as a classification problem where the category of the location that users visit most often can be a good baseline. Inspired by the general check-in patterns of a day, we devise a simple baseline by incorporating temporal information, *i.e.*, using the major class at a given time as the prediction ($\mathcal{M}_{TM}$). Besides, by manually inspecting a subset of the data, we observed that there are some consecutive posting of tweets from

the same location. This motivates a baseline ($\mathcal{M}_L$) of always use the last seen location categories as the prediction of the future check-in.

4.5 Data

As explained in Section 4.1, Foursquare has a rich source of location information including category information which largely reflects users' activities at various places. It allows users to post their current locations and share them with friends. Such posts are usually referred to as *check-in*s. Due to privacy concerns, Foursquare only allows users' friends to view their check-ins and, thus, it is not possible to get access to sufficient data from Foursquare. To this end, we turn to POI-tags in tweets which can be accessed via Twitter's APIs. As an observation, a large proportion of POI-tagged tweets are from Foursquare due to its automatic tweeting function, by enabling which users can instruct Foursquare to post their check-ins on their Twitter account. These POI-tagged tweets from Foursquare carry the information about the geographical entities from Foursquare, though the two location databases may use different identifiers for the same location. Thus we collect a set of POI-tagged tweets and then try to match the geographical entities from both sides so that we can map Twitter's POI-tags to Foursquare's location categories. Using the data collecting strategy in Section 3.4.1, we collected about 1.2 million POI-tagged tweets from 62 thousands users posted from 236 thousands locations.

To collect the category information for all the locations represented by these collected POI-tags, we search each location via Foursquare's API[3] by its name and coordinates, with a filter indicating that only places within 100 meters around the provided coordinates should be returned. Because the same location may appear with slightly different names and coordinates in the two data sources, a simple matching algorithm is used to match them up. Based on a linear combination of textual distance of the names (Levenshtein distance) and spatial distance (Euclid distance of geo-coordinates), we select the closest location entity in Foursquare for each Twitter POI-tag. After matching up the locations there are still 14.3% locations from Twitter that cannot be labelled with a category which leaves 8.6% of the tweets with POI-tags uncategorized. The missing categories arise due to locations without a Foursquare category or locations that cannot be matched to locations on Foursquare. We removed

[3] `https://developer.foursquare.com/docs/venues/search`

from our experiments the POI-tags that cannot be matched to any categories and the tweets to which they attach. Because we are only interested in users who actively use geotags for their tweets, we only keep the tweets whose owners have more than 5 POI-tagged tweets. This results in a dataset of 326 782 POI-tagged tweets posted by 11 087 users from 24 632 locations.

Category information on Foursquare is organized in a hierarchy of in total 400 categories and 9 of them are the top categories which the rest belongs to. Table 4.1 lists the 9 top categories with some examples of direct subcategories and the number of direct subcategories. As shown, the categories reflect the function aspect of locations, *i.e.*, a check-in is able to be categorized according to the potential activities that could be performed at the location. Within each top category, the number of subcategories varies. For example, the category of *Food* has *Sandwiches* and *Chinese* as subcategories. Some of the subcategories are further divided. We consider two settings regarding the categories used in the experiments. The first setting uses categories without the hierarchy, *i.e.*, we use the category originally associated with each location which results in 400 candidate categories for prediction. In the second setting, we replace the category with the top category it belongs to in the hierarchy for each location which results in 9 top categories as candidates for prediction.

As shown in Chapter 3, most of the POI-tags are locations in densely populated areas, usually big cities. We focus our analysis on the four major cities which contribute the most POI-tagged tweets, *i.e.*, New York (NY), Chicago (CH), San Francisco (SF), Los Angeles (LA). Table 4.2 shows the breakdown of the numbers for each of the 4 cities. Similar to the previous studies (Chapter 3), New York has the largest number of active geotag users on Twitter, which contributes the largest volume of geotagged tweets.

Figure 4.3 shows the distribution of the collected check-ins, using different colors for each top category of locations attached to the check-ins. It can be seen that the check-ins depict the outline of the city, especially for the down town areas, *e.g.*, Manhattan area in New York. For places in the category of *Travel & Transport* (marked in purple), most of them are near the down town area. The remote cluster to the south west of Los Angeles is the *International Airport* where many facilities are marked as the category. On the other hand, places in the category of *Shops & Services* spread more uniformly. There are also small clusters marked in blue which are places in the category of *Colleges & University*, *e.g.*, the ones around the coordinates (87.60°W, 41.80°N) in

Table 4.1: The top categories from Foursquare

Top Category	Example of subcategories	Number of subcategories
Arts & Entertainment	Stadium, Museum	18
College & University	College Library, College Classroom	23
Food	BBQ Joint, Ice Cream Shop	86
Great Outdoors	Bridge, Park	31
Nightlife Spot	Bar, Pub	19
Professional & Other Places	Post Office, Factory	18
Residence	Home (Private), Residential Building	3
Shop & Service	Car Wash, Clothing Stores, Spa or Massage	56
Travel & Transport	Airport, Bus Station, Hotel	17

Table 4.2: The statistics of the data set for experiments

City	Tweets	POIs	Users
New York	148992	10200	5537
Chicago	70997	5574	2113
Los Angeles	52195	4432	2220
San Francisco	54594	4426	2305

Chicago are places in *University of Chicago*.

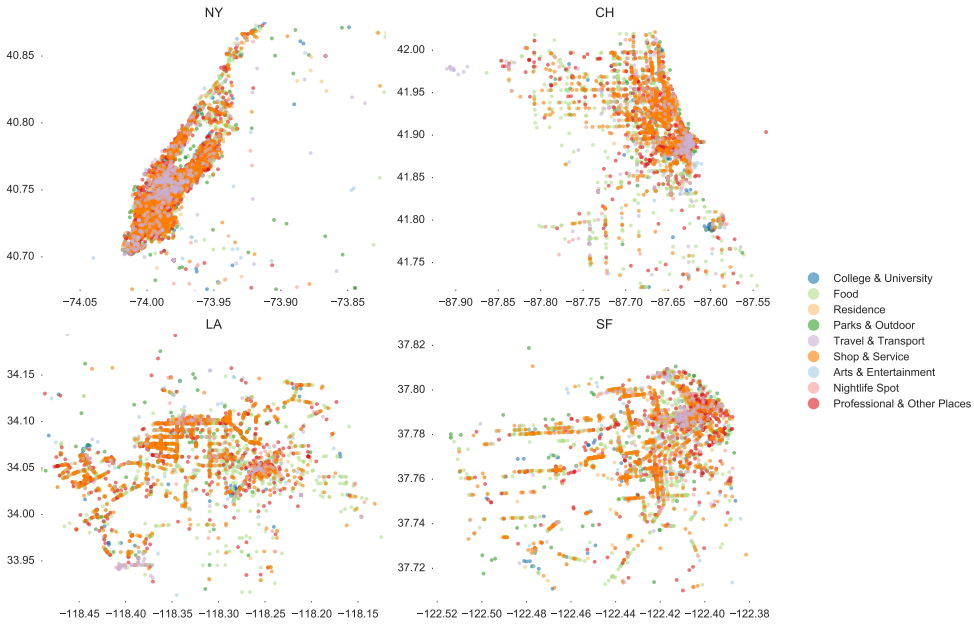


Figure 4.3: POI distributions in the four cities

Similar to the datasets reported by Cheng et al. [32] and Ye et al. [160], our data set renders users' cyclic patterns of check-in behaviours, e.g., day cycles and week cycles. Particularly, as shown in Figure 4.4 check-in volume starts to rise around 5:00 and slopes down after midnight. Two spikes correspond to check-ins at places for *Food* around 12:30 and 20:00 and places of *Nightlife Spot* become crowded after 18:00. In general, places for *Food* are more popular

for check-ins than other categories.

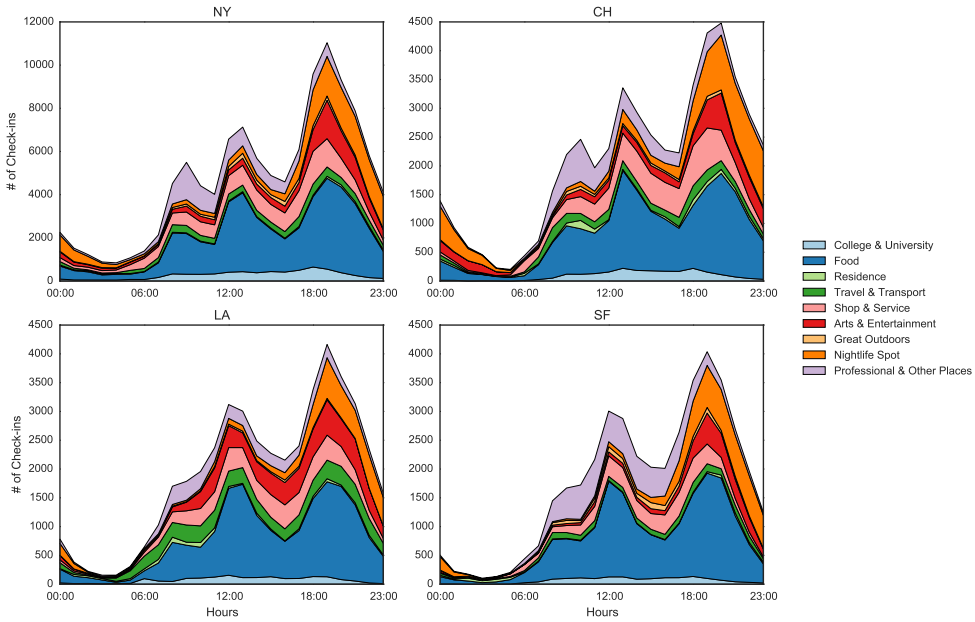


Figure 4.4: Check-in distribution over local time in working days

On weekend days, shown in Figure 4.5, we observe slight differences from check-in behaviour of working days. The rise of check-in volume is 2 hours later at 8:00, and lunch time in general moves to 13:00. Places in the category of *Shops & Services* are visited more often in the morning than in the afternoons during weekends. Visits at this type of places are more evenly distributed on working days, slightly biased towards evenings. The places in the category of *Nightlife Spot* receive more check-ins in the evenings of working days than that in the mornings and afternoons of working days, compared to the smooth distribution of check-ins during weekends.

As can be seen in Figure 4.6, Day-wise weekly cycles do not show as clear patterns as hour-wise daily cycles. In general, we observe more visits to *Food* places and *Nightlife Spot*s in the weekends. The reduction of check-ins on Sundays can be explained by the tradition of resting on that day.

Daily cyclic patterns of check-in behaviour inspires us to consider trails in a

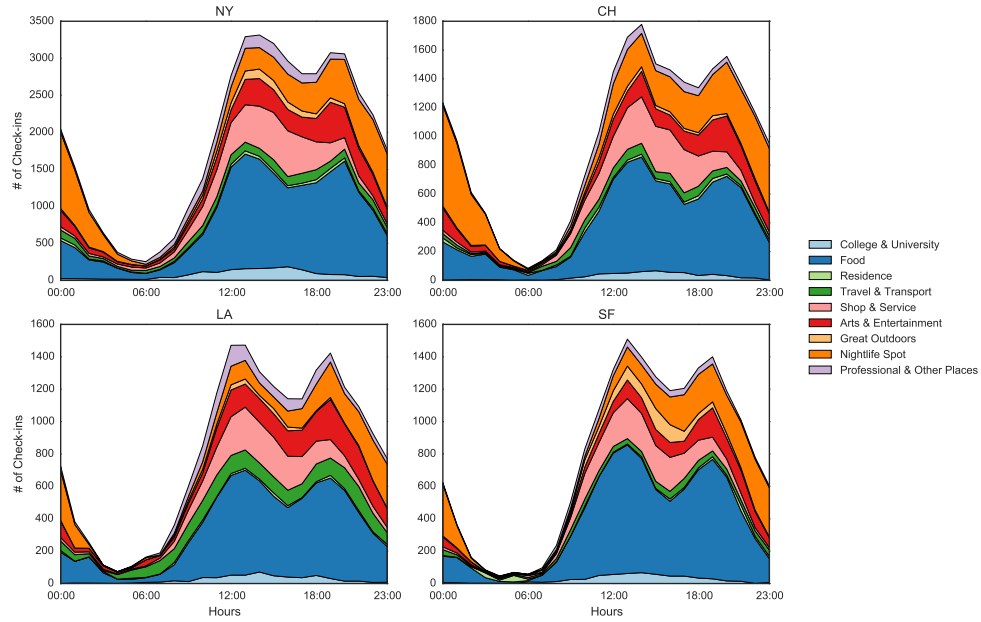


Figure 4.5: Check-in distribution over local time in weekends

length of day. As shown in Figure 4.7, the number of check-ins in daily trails follows the power law, i.e., there is a large proportion of daily trails containing very few check-ins.

Trails represented by check-ins have several intrinsic differences from other types of trails/trajectories: (1) The trails of check-ins are sparse and random snapshots of users' daily mobility. Unlike GPS recorded trajectories, check-ins are totally volunteered records of a user being at different locations. There is no obligation for users of location based social networks to report their positions. This means that the collected data reflects only a part of users' actual trails and there are many missing data between users' check-ins. (2) Triggers causing users to check-in may be more diverse than those to take geotagged photos used in travel route recommendation. Photos tend to be taken for bringing back memories later on, forming a natural representation of the value of going there. Usually the more users post a photo from a location, the more that place is liked. This embedded signal of worthiness and liking can easily be transferred

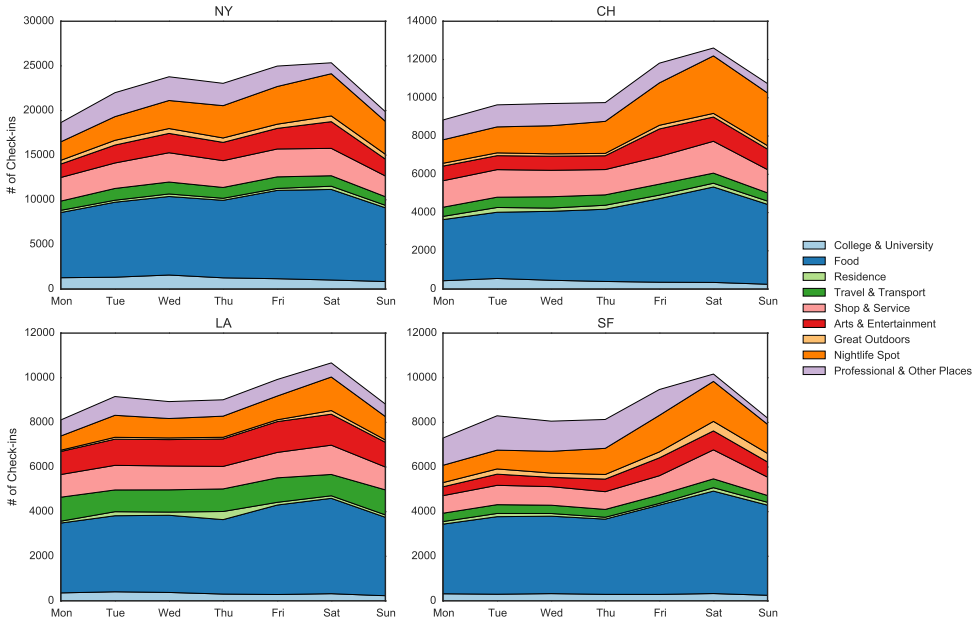


Figure 4.6: Check-in distribution over week days

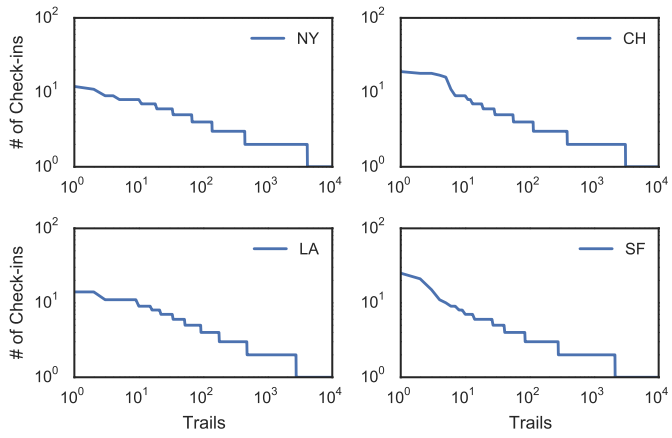


Figure 4.7: Trail length distribution

to other tourists and that is the signal that recommendation is based upon. Due to the gamification strategy adopted by Foursquare, however, users may compete for virtual titles and badges by checking-in at otherwise irrelevant places. They may also check-in because they like the place and want to share it with their friends; and they may simply have a conversation with each message being a check-in, *i.e.*, chatting on Foursquare. These different characteristics of check-in data from other sources of user trajectory information bring challenges to this study.

4.6 Evaluation

To evaluate the proposed methods for predicting users' locations, we conduct two series of experiments. The first series of experiments is based on the top 9 category labels defined by Foursquare's location category hierarchy and each of the 9 labels is propagated to POI-tags whose category labels belong to it in the hierarchy. The second series of experiments is based on the category labels assigned by Foursquare which are all subcategories to the top 9 categories. Both series of experiments are carried out for the four major US cities, *i.e.*, New York (NY), Chicago (CH), Los Angeles (LA), San Francisco (SF), which is due to the reason that, as mentioned in Section 4.5, the locations in these cities are popular among the users in the data we collected. The experiments are conducted using 10-fold cross validation. The folds are cut along the user dimension of the data, which means there is no user overlap between the folds, to make sure that the experiment results can be generalized to different user sets. It reveals the commonality of check-in behaviour between different users and also demonstrates the performance of the proposed methods under minimum privacy intrusion. Similar to the previous chapter, Mean Reciprocal Rank (MRR) is used for evaluating our methods since there is only one relevant location category for each testing trail in the ground truth.

4.6.1 Trail Set Preparation for CF

For the Markov Chain based method and baselines, training and testing trails can be simplified as pairs of consecutive check-ins posted by users in training set and testing set respectively. For the proposed collaborative filtering based methods, trails are composed of check-ins over a certain period. As observed in Section 4.5, users' check-in behaviours are strongly related to day cycles. Thus we focus on the trails of 24 hours long in the evaluation. Both the training trail

set and testing trail set are dynamically composed according to the reference check-in. That is, given a reference check-in from the testing set, a testing trail is composed by grouping the check-ins within a certain length of period (*e.g.*, 24 hours) before the timestamp of the reference check-in. Check-ins in the training set are grouped into trails so that they are aligned with the given testing trail, *i.e.*, both training trails and the given testing trail start and end at the same time within a day. This is based on the consideration that users' activities are largely determined by the time of day. For example, given a 24-hour testing trail with the reference check-in at 10:21 Feb 22, 2011, the check-ins in the training set are grouped into trails such that each trail starts at 10:22 and ends at 10:21 the next day.

Because of the dynamic grouping, it would be inefficient to prepare the training set for each testing trail. Thus, we develop the following strategy for training set preparation. The check-ins are first grouped into trails of 24 hours, from midnight to midnight. Then all consecutive pairs of these trails are concatenated to form double trails, *i.e.*, the trails of doubled lengths. Each of the double trails has a half overlapped with the preceding double trail and another half overlapped with the succeeding one. With these double trails, we can dynamically extract aligned training trails, *i.e.*, for each testing trail, each aligned training trail can be extracted from one of these double-length trails.

4.6.2 Experimental Results

In this section, we present the results of the methods introduced in the previous sections, *i.e.*, using the major class as prediction ($\mathcal{M}_M$), using temporal major class as prediction ($\mathcal{M}_{TM}$), using the last visit as prediction ($\mathcal{M}_L$), Markov Chain based method ($\mathcal{M}_{MC}$) and Collaborative Filtering based methods ($\mathcal{M}_{CF}$ without smoothing and $\mathcal{M}_{CF-K}$ with smoothing). Based on the trails extracted respectively from both training set and testing set, we evaluated the proposed methods as well as the baselines with Mean Reciprocal Ranking (MRR). The reason of using MRR as the evaluation method is that there is only one correct (relevant) category in the resulting rank lists. MRR has been used in similar settings, for example in TREC tracks for the evaluation of question answering and contextual suggestions. The MRR scores are listed in Table 4.3 and Table 4.4 respectively for trails at two levels of categories. The parameters used for $\mathcal{M}_{CF-K}$ are $h = 0.5$ hour (bandwidth), $n = all$ (number of neighbours) in experiments at the top-level of location categories

and $h = 4$ hour, $n = all$ in experiments at the lower-level of location categories. It can be seen that $\mathcal{M}_{CF-K}$ performs the best among all the methods at both levels of the category hierarchy. The collaborative filtering method without smoothing performs the worst, because the sparsity dramatically increases when converting the trails into matrix presentation. Such a big difference in performance demonstrates the necessities of the smoothing technique for the collaborative filtering method.

Table 4.3: The MRR scores on 9-top-category trails

Model	NY	CH	LA	SF
$\mathcal{M}_L$	0.4932	0.5425	0.5209	0.5552
$\mathcal{M}_M$	0.5481	0.5437	0.5928	0.5751
$\mathcal{M}_{TM}$	0.5705	0.5891	0.5987	0.6066
$\mathcal{M}_{MC}$	0.5647	0.6049	0.6100	0.6564
$\mathcal{M}_{CF}$	0.5227	0.5217	0.4481	0.5519
$\mathcal{M}_{CF-K}$	**0.5985**	**0.6398**	**0.6162**	**0.6747**

Table 4.4: The MRR scores on 400-category trails

Model	NY	CH	LA	SF
$\mathcal{M}_L$	0.1593	0.2349	0.1723	0.2343
$\mathcal{M}_M$	0.1469	0.1862	0.1425	0.1933
$\mathcal{M}_{TM}$	0.1724	0.2172	0.1526	0.2244
$\mathcal{M}_{MC}$	0.2270	0.3025	0.2279	0.3139
$\mathcal{M}_{CF}$	0.0863	0.1354	0.0755	0.1529
$\mathcal{M}_{CF-K}$	**0.2432**	**0.3124**	**0.2304**	**0.3308**

Markov Chain based method ($\mathcal{M}_{MC}$) performs the second best among these methods except in New York and it is more computationally efficient than the Collaborative Filtering based methods. $\mathcal{M}_{TM}$ outperforms $\mathcal{M}_{MC}$ with the New York data set at the top-level location categories and is the third best in the other cities. On the other hand, $\mathcal{M}_L$ performs better than $\mathcal{M}_{TM}$ at the lower-level location categories except for New York. This suggests that there are quite some trails in which the reference check-in is at the same location category as the preceding check-in, *i.e.*, there are quite some check-ins belonging to the same location category.

The differences in MRR scores for the evaluated methods are all statistically significant (tested by Wilcoxon Signed Rank Test $p < 0.05$), except for the difference between $\mathcal{M}_{MC}$ and $\mathcal{M}_{CF-K}$ on the check-ins from Los Angeles at the lower-level categories.

4.6.3 Discussion

In this section, we will discuss how the parameters can affect the performance of $\mathcal{M}_{CF-K}$.

Making prediction of a location category given a trail using memory-based Collaborative Filtering relies on the neighbours of the trail to provide an estimation of the user's interest. In the trail prediction problem, the number of neighbours can affect the estimation. As shown in Figure 4.8 and Figure 4.9, a small number of neighbours may not produce accurate estimation. With the number of neighbours going up, the performance of $\mathcal{M}_{CF-K}$ increase as well. This is due to the weighted combination of neighbours in which dissimilar trails will contribute little to the estimation of future trails. However, the performance becomes stable after n reaches 2000.

As explained in Section 4.4.2, the bandwidth parameter h can be interpreted as an approximation of the average length of the time span users stay at a place. Figure 4.10 shows the performance of $\mathcal{M}_{CF-K}$ predicting the top-level location categories and Figure 4.11 for the lower-level location categories. The parameter resulting the best performance is 0.5 hours. Since the parameter is roughly the width of one side bell curve, $h = 0.5$ hour resembles an average time span of 1 hour that users are assumed to stay at a location. As for the lower level categories, the best parameters is around 4 hours, which is much longer than that of the top level categories. An explanation is that the data set for the lower level categories is more sparse than that of the top level categories and the temporal factors are less important than the evidence of having been to a place. CF-K performs better with stronger smoothing parameters, as they can differentiate the case where two users having never been to places under the same category and the case where two users having been to the same place but at different points of time.

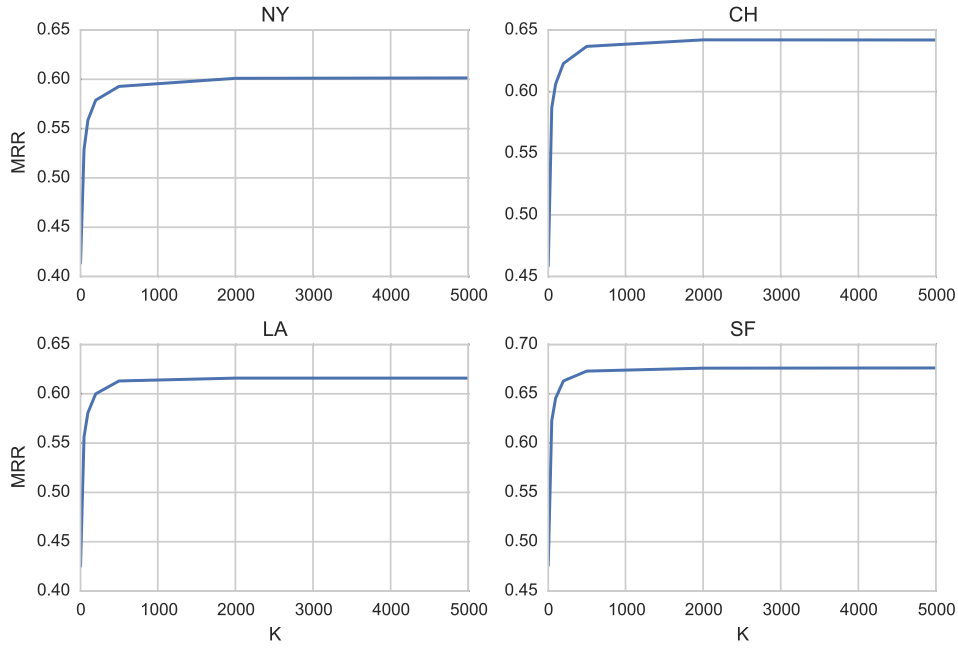


Figure 4.8: The performance of $\mathcal{M}_{CF-K}$ with different number of neighbours for estimation at the top level of categories

4.7 Conclusion

In this chapter, we have investigated the problem of trail prediction from the perspective of location categories. Different from previous studies on location recommendation and travel route recommendation, we focus on the category information of the locations posted on Twitter. The location categories in general resemble the functions of locations and are strongly connected with human activities.

In general there are two major challenges in solving the problem. The first challenge is that there is no public dataset that connects all users' check-ins, locations and category information. Thus to prepare the dataset, we matched two sources of information, *i.e.*, POI-tags from Twitter and category information from Foursquare, and released the ids for public use. The second

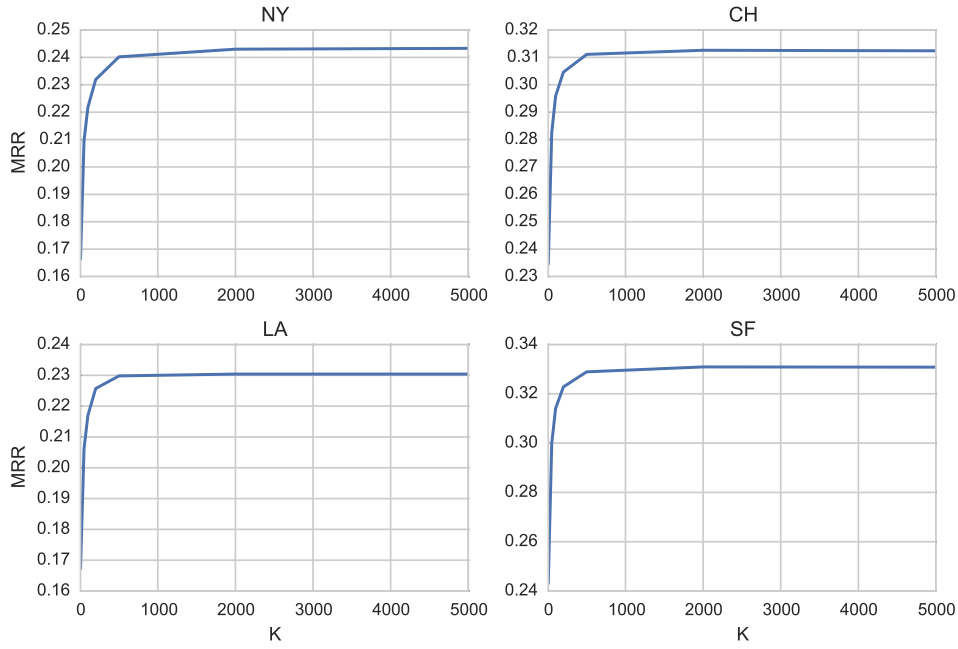


Figure 4.9: The performance of $\mathcal{M}_{CF-K}$ with different number of neighbours for estimation at the lower level of categories

challenge is that the data is very sparse considering that the problem we try to solve involves both temporal and spatial dimensions. We resolve this problem by introducing a smoothing technique which tries to approximate users' real trails. Then we transform the problem of trail prediction into rating prediction in recommender systems by considering each pair of location category and point of time as an item for rating (RQ2a). By applying memory-based collaborative filtering, we can predict users' rating on all the location categories at a given time.

To demonstrate the effectiveness of the proposed methods (RQ2b), we include several baseline methods which are inspired by the investigation over the dataset. These include using the majority class as Prediction ($\mathcal{M}_M$), using the temporal majority class as prediction ($\mathcal{M}_{TM}$), using the Last check-in as prediction ($\mathcal{M}_L$). Besides these, we also include a Markov Chain model, as it

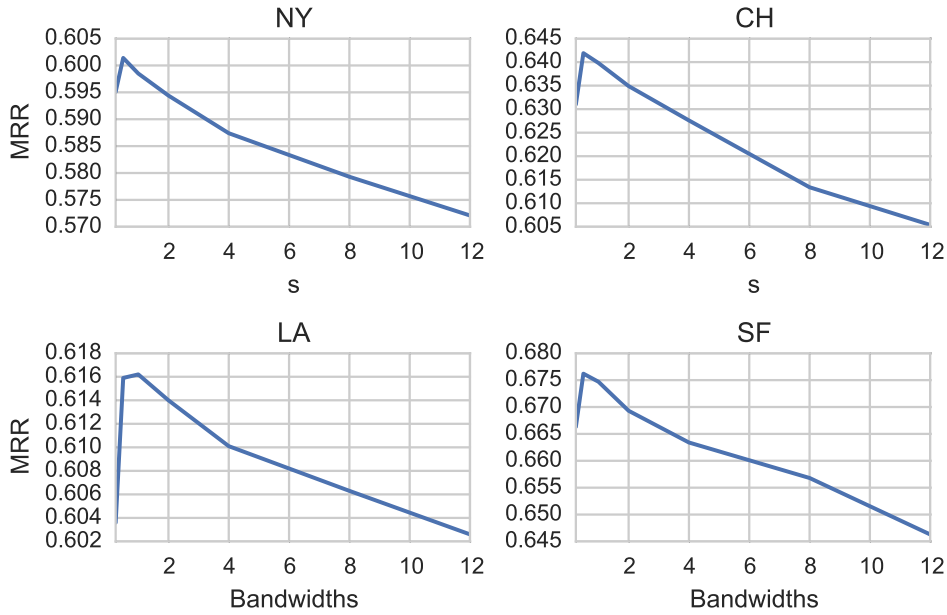


Figure 4.10: The performance of CF-K with different bandwidths for the top level categories

is reported to be useful in predicting users' travel routes.

By evaluating the studied methods on the collected data, we find that the proposed methods perform the best in all the four cities at both levels of location categories. The Markov Chain based method is competitive in this problem, which also has the merit of efficiency. The experiments show that the lower-level location categories are more sparse than the top-level location categories and require a large bandwidth for smoothing. The number of neighbours used for estimating users' future visits are insensitive when it is large enough ($n > 2000$).

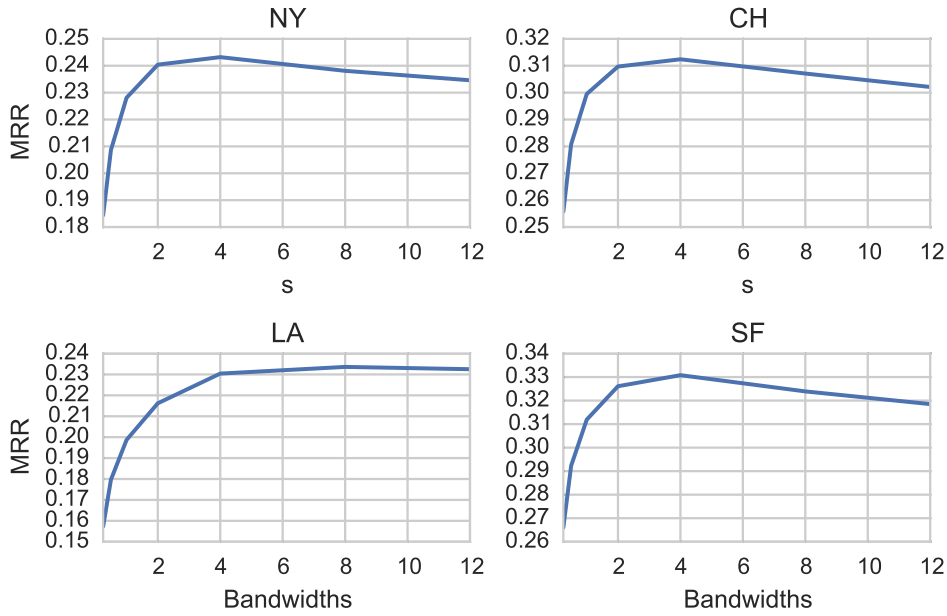


Figure 4.11: The performance of CF-K with different bandwidths for the top level categories

Chapter 5

From Location To Geo-expertise

When a user checks-in at a place via a location-based social network, the action of *checking-in* is not merely a notation of the user sharing the information of his/her being there. The physical attendance at the location also suggests that the user gets familiar with the location and its environment, at least to some extend. In this chapter, we consider check-ins as links to relate the locations users have visited to their knowledge about the locations. With this interpretation of check-ins, we present ways of measuring users' expertise (*i.e.*, geo-expertise) and evaluate the performance of proposed methods. Then we can answer the third research question (RQ3), *i.e.*, how we can model users' knowledge about locations and build an automated retrieval system based on POI information on social media.

Part of this work was published as "Geo-spatial Domain Expertise in Microblogs" by W. Li, C. Eickhoff and A. de Vries, in Advances in Information Retrieval - ECIR '14, pp. 487–492, and part was published as "Probabilistic Local Expert Retrieval" by W. Li, A. de Vries and C. Eickhoff, in Advances in Information Retrieval - ECIR '16, pp. 227–239

5.1 Introduction

When a person comes to a new city, he/she is confronted with a lot of challenges regarding the new environment, especially when he/she wants to find the right places for diverse occasions, such as birthday parties, family get-together. Some people enjoy wandering around the city and discovering new places by themselves, even though they may not always be prepared for all occasions. Not everyone can afford that or think it is worth doing so. Even those who grew up in the city may not be aware of where to find some places/facilities they need. Besides, exploring by oneself is an elaborate and costly task, involving a lot of time and effort, as cities can be large and may change over time. This poses us a problem of fulfilling users' needs of knowledge regarding locations in a city.

5.1.1 Geo-expertise: A definition

In this study we refer to the knowledge facilitating ones' lives or staying in a city or a town as geo-expertise. One example application of such knowledge is illustrated in the following scenario.

> *Mike moves to a city for his new job and he does not know much about the city. One day, his friend comes to visit him and Mike would like to invite him to a nice dinner. Although he frequents a few places, he wants something special in his new neighbourhood for this occasion. How can he get informed?*

5.1.2 Do We Have A Solution Already?

The ultimate goal of geo-expertise retrieval is to give users advice on places to go. This in general is aligned with the goal of location recommender systems, *i.e.*, recommending the locations that a user may be interested in paying a visit. For example, Shi et al. [137] and Clements et al. [36] proposed systems for recommending landmarks based on users' past visits. Some studies also looked into recommending groups of locations, *e.g.*, Kurashima et al. [74] and Lu et al. [97] proposed systems that can help users plan their travels based on geotagged photos. However, these systems have several drawbacks in fulfilling users' needs of geo-expertise. They usually do not consider users' specific needs in each case and only cover a few types of locations (*e.g.*, restaurants, tourist

attractions). Their recommendations are returned without explanation, due to the complex latent factors employed in algorithms. They cannot answer users' needs in an elaborative and interactive way which hinder the development of query expression.

Though users can turn to review websites such as TripAdvisor[1] and Yelp[2] or general purposed search engines for information about locations, there is few automated system for users to query for location information. The lack of automated systems for answering users' needs of geo-expertise is because of relevant information (*e.g.*, that is necessary for assessing relevance of a venue) may just not be freely accessible in digital format. The information is neither considered important for documentation nor easy to document in a written language [39]. For example, it might be either trivial or hard to explicitly document the feel and look in a restaurant or the kind of regular visitors to a pub in written language. Instead of asking the computer to find every piece of text about a location which may still only cover a few aspects of it, users may alternatively turn to a person who is familiar with the locations the user queries about. In this case, it is essential to determine which person is a promising candidate for answering questions about a location.

In the example scenario, if Mike happens to know that one of his colleagues has been to a lot of local restaurants, that colleague would be a good candidate to answer Mike's information need. Such knowledge about whether one's friends have been to a place before become much easier to obtain with the development of online social networks. It also brings the opportunity of automated searching for the right person who could answer a user's questions.

The expansion of online social circles and increase of short-/long-distance travelling makes it hard for one to catch up with all his fellows' recent discoveries in real life, *e.g.*, via chatting. On the other hand, online social networks enable automated ways of analysing the life traces of users' friends. In this study, we use these traces to recommend the most appropriate friends for fulfilling the needs of geo-expertise. That is, an automated system can be employed for processing all friends' message feeds and rank them based on their knowledge for answering questions about queried places.

The task of expertise retrieval has first been addressed in the domain of

[1] `http://www.tripadvisor.com`
[2] `http://www.yelp.com`

enterprises managing and optimizing human resources (detailed in [16, 161]). Early expert retrieval systems require experts to manually fill out questionnaires about their expertise to create so-called expert profiles. Later, automated systems were employed for building and updating such profiles and probabilistic models were introduced for estimating candidates' expertise based on the documents they authored, *e.g.*, [16, 48, 49, 28, 153]. These studies focused on expertise regarding general topics, while ours focuses on location related knowledge. Cheng et al. [33] also proposed finding local experts as a retrieval problem, for which they combined models for local authority and models for topical authority to rank candidates based on data collected from Twitter API. In their settings, queries are keywords with a location to specify the spatial proximity. The needed expertise is not limited to the location related in their system, *e.g.*, "technology in New York". In our settings all queries are interpreted in locations (*e.g.*, a specific restaurant or a type of restaurants) to which users are interested in paying a visit. And the candidates returned to the query must have been to the locations at least once.

Another type of closely related systems to expertise retrieval is found on online questions & answers platforms, such as Quora[3], Stackoverflow[4], Yahoo Answers[5]. These platforms rely on effective methods to route questions to the most suitable answerers, since it is not practical to let answerers go through all questions on the platform to find those he/she can answer. These platforms also provide researchers with an opportunity to access to public data about users' expertise. In particular, the data includes evidence for evaluation, *i.e.*, whether a candidate gives a satisfactory answer to the question he/she has been asked. Based on this kind of data, Liu et al. [94] proposed to use Language Models to profile candidates, Zhang et al. [169] used heuristic features from asker-answerer networks to rank candidates, and Horowitz and Kamvar [66] used probabilistic models similar to the approach by Balog et al. [16] in their social search engine Aardvark. Aardvark as a Q&A platform based on instant messaging system implemented a LocationSensitiveClassifier to decide whether a question needs experts from a certain spatial area. Though they did not detail the algorithm of the classifier [66], the example in the paper renders it as place entity recognition and matching. The studies based on data from Q&A platforms also focused on textual and social network features but did not

[3] http://www.quora.com
[4] http://stackoverflow.com
[5] https://answers.yahoo.com

explore candidates' visiting history which we consider as an important factor in geo-expertise retrieval tasks.

In social networks, topic influencers, whose words are valued more than others and widely propagated can be considered as a kind of experts. For example, such influencers draw a lot of attention to certain products and may affect users purchasing those products. This indicates implicit interaction between experts (influencers) and knowledge seekers (buyers). Identifying those influencers may help users find out more about a topic. There are several studies in this direction, such as [143, 156, 118, 157, 116]

In general, the above discussed studies can be categorized in three groups. The first group focuses on text-based features and tries to build formal probabilistic models from the text authored by candidate experts. Fang and Zhai [48], Fang et al. [49] tried to formalize the models in a probabilistic way based on topic generation models and candidate generation models via document retrieval. Balog and de Rijke [15] proposed non-local features (*e.g.*, IDF, query expansions) that may help better estimate the prior. Liu et al. [94] considered a candidate's profile on previous questions as a document and convert the expertise retrieval problem to a general information retrieval problem, *i.e.*, retrieving the answerer whose profiles best matches the questions. Balog et al. [16] reviewed studies regarding modelling the association between topics and candidates, which also involves natural language processing, entity recognition, entity disambiguation, *etc.* Wagner et al. [153] proposed a LDA based method for identifying experts on a topic among Twitter users and suggested using external resources other than tweet content for identifying candidates' expertise in order to overcome the high amounts of noise pertaining to the domain. Some other works also tried to approach the problem through graph models, which consider the email conversation between candidates [28] or the co-occurrence of candidates in documents [14]. The studied methods in this group are based on solid probabilistic models which are proved to be successful. They inspired us to use probabilistic models for modelling geo-expertise. However, geo-expertise and location information are not always presented in textual format. For example, the types of locations are rarely mentioned in the textual content of the messages that relate to the location, and location names do not always reveal their functions (*e.g.*, Exit is a nightclub) or only indirectly (*e.g.*, NY Pizza is in the category of Italian food). We approach the problem by building models based on candidates' check-in profiles which explicitly links a location (with category information) to the visitors' knowledge.

The second group of studies focuses on heuristic and non-textual features combined with machine learning techniques. Zhang et al. [169] proposed a PageRank like method for expertise ranking in asker-answerer relationship networks in community-based Q&A and their method shows promising performance on a set of more than 100 users from Java Developer Forum. Agarwal et al. [3] proposed a method of scoring bloggers' influence based on heuristic indicators (*i.e.*, in-link, out-link, number of comments and post length) and evaluated them by the data crawled from Digg[6]. Weng et al. [156] combined knowledge from topics (distilled by LDA) and social networks to produce a topic-specific network and used random walk methods to find topic-specific influential users (experts on a topic). Lehmann and Castillo [78] studied how to identify news story curators on Twitter based on random forest models with the heuristic non-textual features (*e.g.*, followers, lists, retweets) selected by an information-gain based method. Tinati et al. [147] created a more finer categorization of users, namely, idea starter, amplifier, curator, commentator, viewer, which define types of human experts in topics. Smirnova [139] extends author-topic based models with the friendship on social networks for retrieving experts which is shown to outperform the ones without the extension. Sun and Ng [143] proposed a method of identifying influential users on Twitter by creating a post-graph out of influential posts on a given topic, transforming it into user graph and then measuring users' influence by both graphs. Pal and Counts [118] used 17 social features (in 4 groups), *e.g.*, number of tweets, number of hastags, to identify authoritative users via Gaussian Mixture Models. Overbey et al. [116] proposed a method for identifying influential users (based on graphs of retweeting) in a particular event, *i.e.*, Egyptian Revolution, and analysed the method against the tweets collected during the time of the event whose authors are tagged as living in the country. A closely-related project, named Aardvark[7], routes questions between Instant Messenger (IM) users, as Horowitz and Kamvar [66] explained, by employing a classifier for recognizing location entities embedded in the questions. For example, it can prompt to the users living in Utrecht (*e.g.*, by looking at the candidate's profile) a question regarding cinemas in Utrecht in the Netherlands.

Bao et al. [18] proposed a method to find local experts having knowledge about a category of locations. They applied Hyperlink Inference Topic Search

[6] `http://digg.com`

[7] Aardvark was ceased by Google in September, 2011. `http://googleblog.blogspot.nl/2011/09/fall-spring-clean.html`

algorithm on the user-location matrix. The algorithm will eliminate users who only have check-ins at less-known places. However, the algorithm only served as a feature for selecting user profiles for location recommendation. Though the feature shows a positive effect on the final recommendation performance and efficiency, it is not clearly shown whether the candidates labelled with a high score are really *experts* as perceived by normal people. Thus we include their algorithm as a baseline in this study which is detailed in Section 5.3.6.

The studied methods in this group inspired us to explore the relevant features to solve the problem of geo-expertise retrieval, such as number and recency of check-ins. However, we did not include social links explicitly in our proposed methods due to the sparsity in the data we collected. We leave it for future work and a possible direction is to explore the usage of social links as a candidate prior.

The third group tries to modelling candidates' behaviour directly based on external evidence of performing expertise. Bar-Haim et al. [19] tried to identify stock experts on Twitter by evaluating their expertise according to stock market events and their tweeted buys and sells. Whiting et al. [157] suggested to use changes of Wikipedia pages as clues to real-time event topics, then retrieve tweets containing (stable and temporally important) terms and consider the authors (with more followers) of these tweets as influential Twitter users. These studies rely on particular sources of direct evidence which is not easy to replicate on the data from other domains. Though, they inspired us to send out evaluation questionnaires to the retrieved candidates from our system and let them to directly evaluate themselves with respect to the geo-expertise they were queried about.

5.1.3 Check-ins As Evidence Of Expertise

In the previous chapter, we show that modern city inhabitants usually go to several places on a daily basis, *e.g.*, places for food and for entertainment. The repeated attendances at a place may reinforce their impression about the place and the physical environment around it. Such impression is likely to turn into a long-term memory and transform into the knowledge about the place. Mainstream online social networks enable users to share their locations by check-ins, which we consider as a suitable indicator of geo-expertise, since they provide direct evidence of users' physical attendance of the place.

As there have been few previous studies specifically focusing on the domain of geo-expertise retrieval, we first carry out a study to establish the real-world needs of a solution to the problem (RQ3a), for which we designed and handed out a survey to general users of the Internet and ask their opinion on the type of information needs. The outcome of the survey strongly suggests, among average users, the needs of a geo-expertise retrieval system. Based on the answers to the survey, we propose and investigate four probabilistic model based methods of estimating geo-expertise from candidate profiles (RQ3b). To evaluate the proposed methods, we select a set of topics at random, rank candidates according to these topics, pool the top k candidates retrieved for each query, collect annotation of pooled topic-candidate pairs, and evaluated the proposed methods based on the annotation (RQ3c). The following sections detail the answers to these questions.

5.1.4 The problem of geo-expertise retrieval

In the previous section, we define geo-expertise as knowledge regarding locations. Here, a topic is either a specific location or a category of locations within a geographical scope, *e.g.*, the Blue Ribbon Fried Chicken in New York and Chinese Restaurants in Los Angeles. The former ones are referred to as *POI topics*, which relate to knowledge regarding a particular location. The knowledge to potential inquiries about that location may include the opening time, the special menu, the service quality, *etc.* The latter ones are referred to as *category topics*, which relate to knowledge regarding the locations in a specific category as a whole, such as theme or decoration difference between locations in the given category. The successful candidate to the given topic should be able to recommend a location for a give occasion (*e.g.*, birthday, dating) considering some given constraints (*e.g.*, parking spaces, facility availability). A candidate ranked at a top position by the system should be able to answer more questions than those lower ranked candidates.

5.2 Understanding Geo-Spatial Expertise

In order to have a better understanding of users' needs of geo-expertise retrieval systems (RQ3a), we handed out a survey via CrowdFlower[8]. The survey is composed of questions of two types, namely, questions related to demographics

[8] `http://crowdflower.com`

and questions related to geo-expertise seeking experiences. A number of options were offered for each question, as well as the possibility to give free-text answers if no given option was deemed applicable. Participants were also invited to select more than one option if applicable.

The participants were informed about our privacy confidentiality policy and asked to be honest with all the questions. A total of 199 forms were received within one week. In the following, we discuss the main findings and implications derived from the survey.

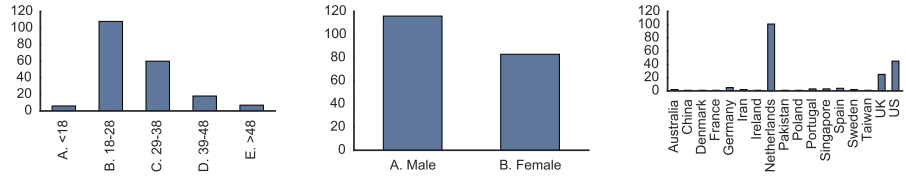
5.2.1 Participant Demographics

To review the representativeness of the participants in the survey, we include questions about the demographics at the beginning of our questionnaires. The questions and options are listed as follows.

- **Age**: Under 18, 18–28, 29–38, 39–48, Over 48
- **Gender**: Male, Female
- **Country**: (Free text)
- **Experience of Social Network**: Twitter, Facebook, LinkedIn, Foursquare, Google Plus, Path, Other
- **Usage of Social Network**: Several times a day, Once a day, Once the other day, Once per week, Once per month, I seldom use online social networks, Other
- **Sharing Locations**: Not at all, Once in a while, Regularly, Frequently

Figure 5.1 shows the general demographics of our participants in terms of (a) age, (b) gender and (c) country. As can be seen, most of our participants are between 18 and 38 years old and both genders are well represented. As of country dimension, we see most of them are from the Netherlands and the second largest group is from the US. Though the participants are not evenly distributed, we consider they can reassemble to some extend the Internet users.

Besides the questions about basic demographics, we also include questions towards participants' experience on online social networks. As shown in



(a) Age (b) Gender (c) Country

Figure 5.1: Demographic

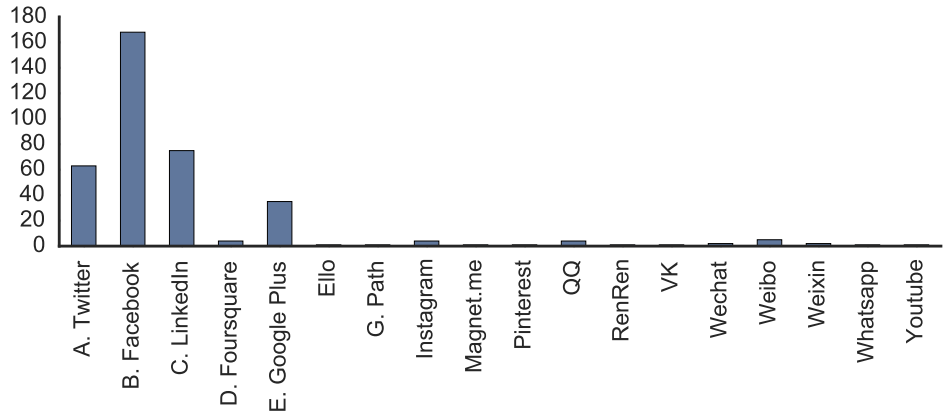


Figure 5.2: Usage of online social network platforms

Figure 5.2, Facebook has the most users in our participants and Twitter follows.[9] This pattern matches the outline given in the report by Duggan and Smith [44].

Many of the participants visit online social networks frequently (Figure 5.3). As for sharing location, very few of them actively post information about their locations, though many of them claim they have experiences of using such features (Figure 5.4).

[9]The labels starting with A., B., *etc.* are the provided choices for the question, where others are from the answers of free text input.

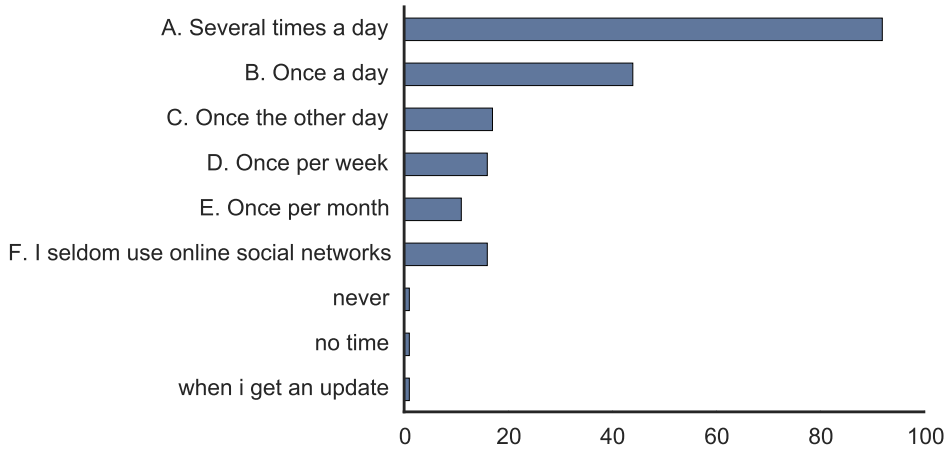


Figure 5.3: Social network engagement

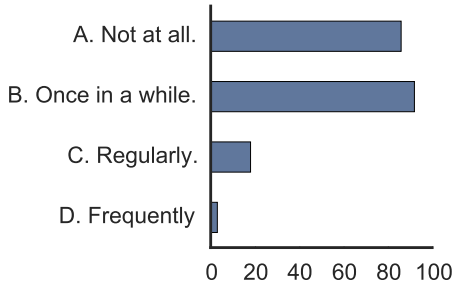


Figure 5.4: Sharing locations on social networks

5.2.2 Perception of Online Geo-expertise Seeking

To answer whether there are general needs of geo-expertise, we investigate how this types of tasks are perceived by online users. To this end, we included the following questions to outline participants' real life experience of seeking advice about locations.

1. How often do you research places for specific occasions?

2. How often do you ask your friends, family, colleagues or any other people for advice about a place to go?

3. Attitude towards looking for help: Asking friends (online or offline), colleague, family, using review sites.

4. Preference of means: Send online message (Twitter, IM, Facebook Message), Emails/Telephone/SMS, Face to face or in person.

5. Preference of candidates: Family, colleagues, Friends, Online Friends (not known offline), Blogs

We summarize the key findings in the survey as follows. As shown in Figure 5.5, looking for places is a common demand which accounts for 89% of the participants. One of the participants reported that he/she would like to go where his/her friends go. This leads to an interesting question to investigate in the future, *i.e.*, what alternative method do people use for finding places to go besides rational decision based on the information acquired about the places.

According to Figure 5.6, up to 84% of the participants see the usefulness of others' advice on a place to go, *i.e.*, geo-expertise; only 16% participants would like to research the problem by their own. These findings confirm our assumption about the potential benefits of geo-expertise retrieval systems. When participants consider the advice given by others, trust is their main concern. Figure 5.7 shows a clear preference among the participants favouring family and friends over on-line contacts or even unknown review websites. When they decide to ask for help on geo-expertise, the most favourite way is to talk in person, followed by telecommunication/Email and online social networks (Figure 5.8). From Figure 5.9, we can observe that telling friends about needs of geo-expertise is preferred over the means of searching for geo-expertise online, which is followed by asking online friends and posting questions on forums/Q&A. There are some participants who prefer *Googling it* when they have needs of geo-expertise. Combined with other similar responses (*e.g.*, search online), those who prefer researching the problem themselves account for about 8% of the participant population.

To sum up, there is a demand of geo-expertise retrieval system according to the survey responses. Participants would trust the advice from their family or friends, *i.e.*, those who are close to them, other than any strangers online

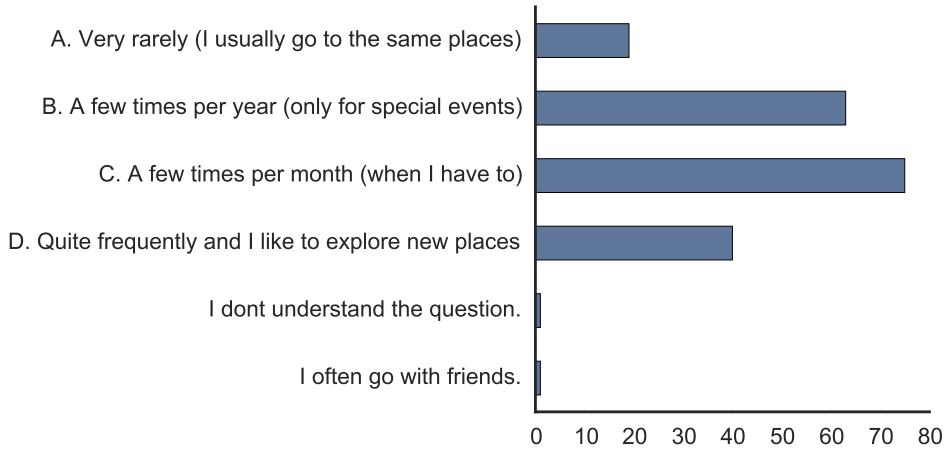


Figure 5.5: Demand of location knowledge

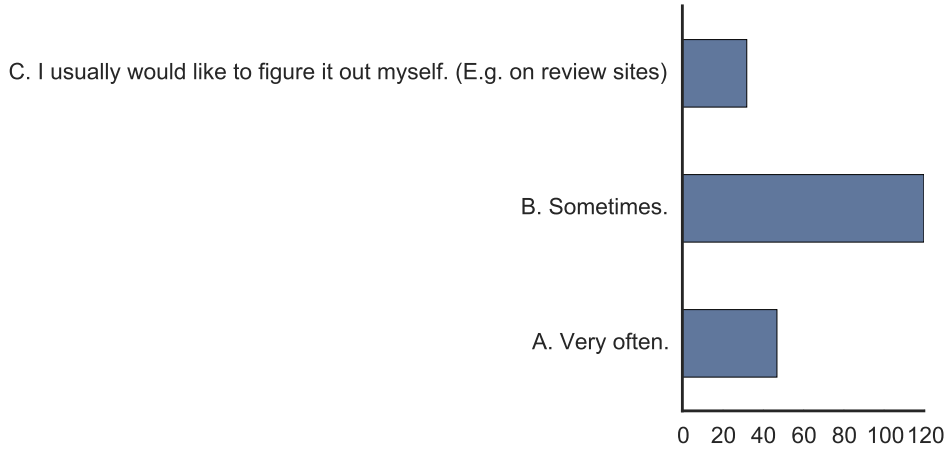


Figure 5.6: Demand of advice from others

or professional review writers. This suggests that we should consider social circles in our proposed geo-expertise retrieval system. However, due to the

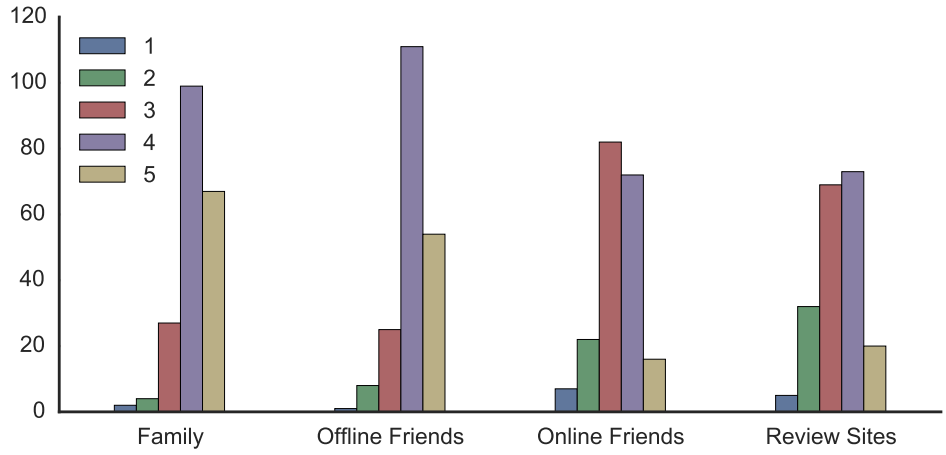


Figure 5.7: Trust in different sources of geo-expertise

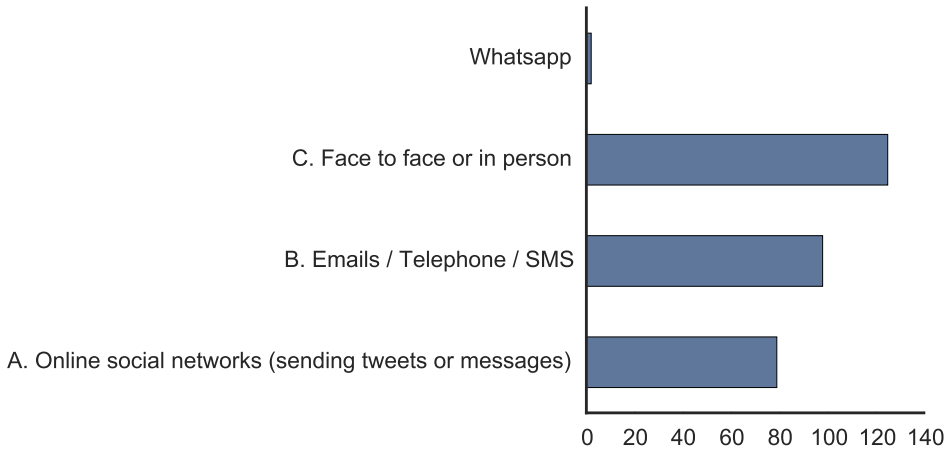


Figure 5.8: Channel preference for geo-expertise

sparsity in the data we collected, we decide to focus on candidates' visiting histories rather than how close they are with the geo-expertise seekers.

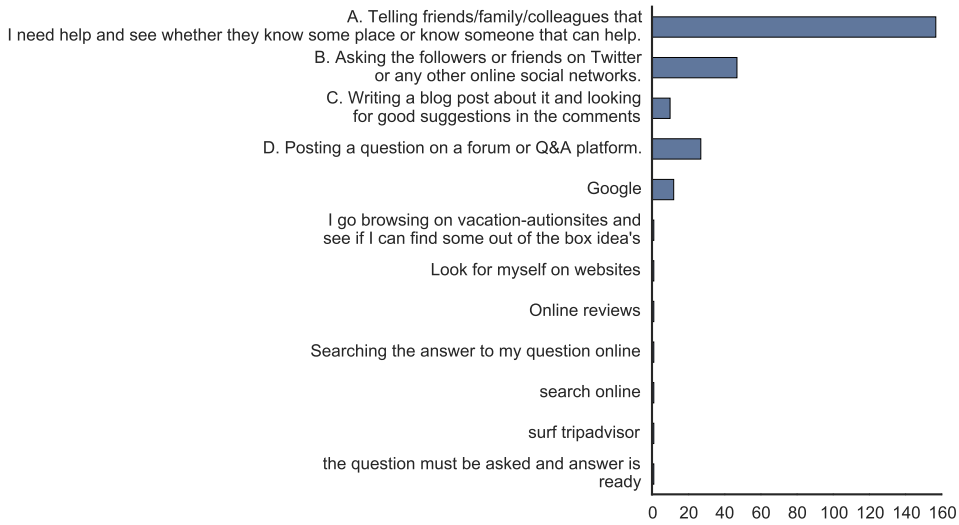


Figure 5.9: Means of searching for geo-expertise

5.3 Methodology

Geotags have been commonly used in social media to enable users sharing their locations. Geotagged tweets also provide evidence of users' physical presence at a place. For example, a tweet, "*I really love sandwiches here*", with a POI-tag *Blue Barn Gourmet* shows that the user has been to the place *Blue Barn Gourment* and the user has had sandwiches there, and experienced the food and environment at the place. Such experience or knowledge may be useful for others, for example, for those who are visiting the area and planning to have a brief business lunch there. Though the check-ins are single time points of users' actual trails in the physical world, they reveal part of their visits. The interaction between users and locations in the form of check-ins suggests however that these users have paid considerable attention to those locations and were willing to make effort to check-in at the location. Such attention implies that users perceive more about their physical surroundings which means they notice more details about the location than those locations that they have no interests in checking-in at.

Technically, it is possible for one to fake his/her check-ins, but we presume that such check-ins would be rare on social network platforms. First, faking one's location on a smart phone is beyond normal users' knowledge. Second, the purpose of using a social network platform is to communicate and share interests between friends, and faking one's location is not going to be highly appreciated in friend circles.

The only issue is that companies and organizations also use social networks for advertisements and promotions. Some of them may also bind a location to their messages and these accounts probably will not provide useful response to any users who want to know about the locations. By calculating the speed of moving from one check-in location to another, we eliminate those accounts having a pair of check-ins implying a speed of over 700 kph. This threshold is chosen because it is roughly the speed of aircraft and any higher speed between two check-ins means it is not likely to be a normal user account. In this way some promotion/branding accounts for large companies and having check-ins all over the country will be eliminated. There may be accounts from small local companies in the dataset, and we presume they are managed by the ones live around the place and can be potential sources of location advice. For example, users may consult the account representing a local restaurant for the information about it.

5.3.1 Motivation

In our aforementioned survey, we include a poll for the preferences regarding intuitive methods for geo-expertise retrieval and see how the participants perceive the task. The following criteria are listed in a survey question. Participants are asked to assign a score for each of them ranging from 1 (not important) to 5 (very important).

- **Within-topic Activity**: The total number of visits at a given place (or category) should be above a certain number.
- **Number of Visited Place**: The total number of places visited should be above a certain number.
- **Within-topic Recency**: There should be recent visits at a given place. Otherwise the candidate may not know much about the current status of the place.

- **Within-topic Diversity**: There should be visits at many different places in a category if a given topic is about a category of locations. Otherwise the candidate may only know things about a specific location rather than the entire category to which the location belongs.

- **Visits in Other Cities**: There should be check-ins at places in other cities for a given category.

- **Spatial Distribution**: The visited places should cover most part of the city. Otherwise the user may only know places in a small area.

The result of this poll is shown in Figure 5.10. The top three criteria matches the methods we proposed in Section 5.3, which are respectively *Within-topic Recency*, *Within-topic Activity*, *Within-topic Diversity*. It should be noted that the participants of this survey also annotated the ground truth for the experiment afterwards.

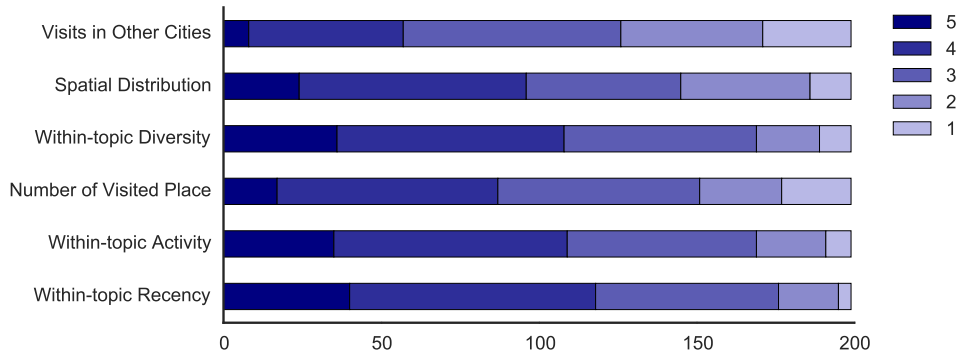


Figure 5.10: Criteria for determining geo-expertise

In the open-ended feedback on our survey questionnaires, the participants also provided us with some interesting ideas that worth exploring in future work. One of the participants would like to consider the average frequency of a candidate visiting the location about the given topic. Candidates regularly visiting a place may know better about the location than those who recently have paid several visits to the location.

Another participant suggested to higher rank those whose visits are generally more geo-spatially widely spread. It is presumed that if a candidate has been to many different countries or regions, they will know more about locations, as they may visit some locations without checking-in at them. An interesting direction along this is to consider about users' regional/cultural knowledge. A user would give good advice about Italian restaurant if he has been to Italy. However, these semantic links are beyond the scope of this study.

There is also a suggestion of including in consideration the ratio between relevant posts and irrelevant posts in one's Twitter timeline. This involves inspection on textual features of candidates' profiles. It is a method of identifying users' topical interests. However, in this study we focus on the temporal and spatial domain of check-ins and may explore this direction in the future.

Several of the participants mentioned that the shared interests and friendships should be considered between the query issuer and the candidate experts. This is based on the same reasoning on the trust and common interests between friends and/or families as discussed in the previous section. It is an interesting direction to explore in the future for personalized geo-expertise retrieval.

In this study, we focus on only the candidates' check-ins profiles and propose to model their knowledge by considering three aspects, *i.e.*, the knowledge about locations, the knowledge about differences between locations and the recency of knowledge. The following sections explain how these intuitions are expressed computationally.

5.3.2 Within-Topic Activity

The first method we propose considers only the candidates' check-in activities. We take a co-occurrence modelling approach, inspired by expert finding via text documents [48]. To be specific, we rank a candidate u by the probability of him/her having geo-expertise on the given topic q, *i.e.*, $P(u|q)$. We estimate the conditional probability by aggregating over users' check-ins at all possible locations (l), that is

$$P(u|q) = \sum_{l} P(u|l,q)P(l|q).$$

Assuming conditional independence of the candidate u and the query topic q given the location l, *i.e.*, $P(u|l,q) = P(u|l)$, we obtain

$$P(u|q) = \sum_{l} P(u|l)P(l|q).$$

As for $P(u|l)$, we apply the Bayesian Rule which gives

$$P(u|l) \propto P(l|u)P(u)$$

where we assume a uniform prior for $P(l)$.

Putting these together, we obtain

$$P(u|q) \stackrel{\text{rank}}{=\!=} \sum_{l} P(l|q)P(l|u)P(u). \tag{5.1}$$

$P(u)$ is the query-independent expertise and we estimate it by the number of check-ins a user posted online, *i.e.*,

$$P(u) = \frac{N_u}{N}$$

where N_u and N are respectively the number of check-ins posted by the candidate u and the number of check-ins in total. Intuitively, the more a candidate posts his/her check-ins, the more we trust the model built from his/her check-in profiles. The conditional probability $P(l|q)$ captures users' query intent, *i.e.*, the possible locations users want to know via issuing the query. In our setting of geo-expertise, the query is a location or a type of locations. So the conditional probability can be estimated by

$$P(l|q) = \begin{cases} \dfrac{1}{|L_q|} & \text{if } l \in L_q, \\ 0 & \text{otherwise,} \end{cases}$$

where L_q is the set of locations matching the query. To estimate $P(l|u)$, we use

$$P(l|u) = \frac{N_{l,u}}{N_u},$$

where $N_{l,u}$ is the number of check-ins the candidate u made at the location l. Then the scoring function can be derived from simplifying Eq. 5.1, that is

$$\mathcal{S}_n(u,q) = \frac{1}{|L_q| \cdot N} \sum_{l \in L_q} N_{l,u} \stackrel{\text{rank}}{=\!=} \sum_{l \in L_q} N_{l,u} \ .$$

Intuitively, the more check-ins a candidate has at the queried location(s) in L_q the more likely he/she is interested in the locations and knows more about them.

5.3.3 Within-Topic Diversity

Our second method uses the language model referred to as *Model 1* in [16], that is

$$P(\theta_u|q) \propto P(q|\theta_u)P(\theta_u).$$

where θ_u is a Language Model of a candidate based on his/her check-in profile. To estimate $P(q|\theta_u)$, we assume the independence between the locations representing the underlying information needs through the given query, that is

$$P(q|\theta_u) = \prod_{l \in L_q} P(l|\theta_u) = \prod_{l \in L_q} \frac{N_{l,u}}{N_u}. \tag{5.2}$$

For the prior $P(\theta_u)$, we use

$$P(u) = \frac{N_u^{|L_q|}}{\sum_{u' \in U} N_{u'}^{|L_q|}},$$

so it will simplify the scoring function, that is

$$P(q|u) = \frac{N_u^{|L_q|}}{\sum_{u' \in U} N_{u'}^{|L_q|}} \prod_{l \in L_q} \frac{N_{l,u}}{N_u}.$$

By applying the logarithm (to avoid underflow in computation), we obtain

$$\log P(q|u) = \log \frac{1}{\sum_{u' \in U} N_{u'}^{|L_q|}} + \sum_{l \in L_q} \log N_{l,u} \overset{\text{rank}}{=\!=} \sum_{l \in l_q} \log N_{l,u}.$$

The following smoothed version is used as the scoring function of this method to differentiate the profiles containing visits to different numbers of locations but each location has been visited only once.

$$\mathcal{S}_d(u, q) = \sum_{l \in l_q} \log(N_{l,u} + 1).$$

The above scoring function indicates that check-ins at diverse locations (within the queried location sets) will increase the score more than check-ins at the same location. It means that a candidate will gain a higher score on geo-expertise if he/she makes check-ins at a variety of locations. This fits into the intuition that candidates with experience at variety of locations may know more about the type of queried locations, and may give more comprehensive information to satisfy the users' needs. For example, if one wants to ask about Italian restaurants, those who have been to a lot of Italian restaurants in town will be more suitable candidates than those who have only been to the same restaurant a lot. The former may be able to tell the differences and provide more useful advice in choosing an Italian restaurant. On the contrary, the latter, who have a lot of check-ins at a single Italian restaurant may not be familiar with other restaurants.

The prior $P(u)$ is selected for the score function so that the candidate-dependent denominator in the conditional probability $P(l|u)$ will be cancelled when combined with the prior. This is because Language models represent users' topical focus rather than users' knowledge, *i.e.*, they are biased towards the shorter profile, when two profiles have the same amount of relevant check-ins. Check-ins are positive evidence of candidates knowing about a location and knowing about another location should not affect the knowledge acquired from the former one. For example, if a candidate has been to two places A and B each for n times while another candidate only has been to A for n times, it is not reasonable to assume that the latter candidate having more knowledge about the place A than that of the former candidate, even if the latter has focused on the place A more.

5.3.4 Within-Topic Recency

Experts are humans and rely on their memories to support their expertise. Therefore, we should take into account the fact that 1) people forget the knowledge they once gained and have not been familiarized with it for a while, and 2) the world is changing as time goes by, *e.g.*, restaurants may have new chefs, old buildings may have been replaced. The longer it is since the creation of the memory, the more likely the memory becomes inaccurate.

To incorporate such effects, we model the candidates' memory by $P(c|u)$, which indicates the probability that candidate u can recall his/her visit represented by the check-in c. As suggested in the domain of psychology, human

memory can be assumed to decay exponentially [96]. For simplicity, in this study we use a straightforward exponential decay function to represent the retention of individual check-ins, by which we obtain:

$$P_t(c|u) = \frac{e^{-\lambda(t-t_c)}}{\sum_{c \in C_u} e^{-\lambda(t-t_c)}},$$

where t is the time of query and t_c is the time when the user posted the check-in. Similarly, we define a prior for each candidate as follows

$$P_t(u) = \frac{\sum_{c \in C_u} e^{-\lambda(t-t_c)}}{\sum_{c \in C} e^{-\lambda(t-t_c)}}$$

The decay of the weight on check-ins models our belief on how up-to-date the information is, while the prior reflects the average recency of knowledge borne by the whole community on the social network. Then for estimating the candidates' expertise, we weight each check-in according to its recency, *i.e.*, we marginalize the user's old check-ins.

$$P_t(l|u) = \sum_{c \in C_u} P(l|c) P_t(c|u) = \sum_{c \in C_u} \frac{\mathbf{1}(l_c = l) e^{-\lambda(t-t_c)}}{e^{-\lambda(t-t_c)}}, \tag{5.3}$$

where $\mathbf{1}(\cdot)$ is an indicator function, which equals 1 *if and only if* the condition in the parentheses evaluates to true. Given these two estimations, we obtain

$$\begin{aligned}
\mathcal{S}_r(u, q) &= P_t(u|q) \\
&= \sum_{l \in L} P_t(u|l) P(l|q) \\
&\overset{\text{rank}}{=\!=} \sum_{l \in L} P_t(l|u) P_t(u) P(l|q) \\
&\overset{\text{rank}}{=\!=} \sum_{l \in L_q} \sum_{c: l_c = l, c \in C_u} e^{-\lambda(t-t_c)}.
\end{aligned}$$

As can be seen $\mathcal{S}_r$ down-weights the older check-ins' contribution to candidates' expertise due to the fact that they may be vaguely memorized and become unreliable. Parameter λ that controls the amount of decay is fixed to $\frac{1}{150}$ at a granularity of days and we leave the fine tuning of this parameter for future work.

5.3.5 Combining Recency and Diversity

Diversity and recency of check-ins can be both important factors in estimating one's expertise about locations. Thus, we propose a combination of the two features introduced in Section 5.3.3 and 5.3.4. In Eq. 5.2, the conditional probability can be transformed into

$$\begin{aligned} P(q|u) &= \sum_{c \in C_u} P(l|c,u)P(c|u) \\ &= \sum_{c \in C_u} P(l|c)P(c|u) \\ &= \sum_{c \in C_u} \mathbf{1}(l_c = l)P(c|u), \end{aligned}$$

in which we assume the candidate and location are conditionally independent given a check-in (*i.e.*, applied on the second equal sign). Then we estimate the conditional probability $P(c|u)$ with Eq. 5.3.

$$P(c|u) = \frac{e^{-\lambda(t-t_c)}}{\sum_{c \in C_u} e^{-\lambda(t-t_c)}}.$$

Thus,

$$P(q|u) = \prod_{l \in L_q} \sum_{c \in C_u} \frac{\mathbf{1}(l_c = l)e^{-\lambda(t-t_c)}}{\sum_{c \in C_u} e^{-\lambda(t-t_c)}}.$$

Similar to the prior probability used in the diversity method,

$$P(u) = \frac{(\sum_{c \in C_u} e^{(-\lambda(t-t_c))})^{|L_q|}}{(\sum_{c \in C} e^{-\lambda(t-t_u)})^{|L_q|}}.$$

By replacing the counter parts with these into Eq. 5.2 and applying logarithm on both sides of the equation we obtain

$$\begin{aligned}
\mathcal{S}_d(u,q) &= \log \frac{(\sum_{c\in C_u} e^{-\lambda(t-t_c)})^{|L_q|}}{(\sum_{c\in C} e^{-\lambda(t-t_u)})^{|L_q|}} \prod_{l\in L_q} \frac{\sum_{c\in C_u} \mathbf{1}(l_u = l) e^{-\lambda(t-t_c)}}{\sum_{c\in C_u} e^{-\lambda(t-t_c)}} \\
&= \log \frac{(\sum_{c\in C_u} e^{-\lambda(t-t_c)})^{|L_q|}}{(\sum_{c\in C} e^{-\lambda(t-t_u)})^{|L_q|}} \cdot \frac{\prod_{l\in L_q} (\sum_{c\in C_u} \mathbf{1}(l_u = l) e^{-\lambda(t-t_c)})}{(\sum_{c\in C_u} e^{-\lambda(t-t_c)})^{|L_q|}} \\
&= \log \frac{1}{(\sum_{c\in C} e^{-\lambda(t-t_c)})^{|L_q|}} + \sum_{l\in L_q} \log \sum_{c\in C_u} \mathbf{1}(l = l_c) e^{-\lambda(t-t_c)} \\
&\overset{\text{rank}}{=\!=} \sum_{l\in L_q} \log \sum_{c\in C_u} \mathbf{1}(l = l_c) e^{-\lambda(t-t_c)}.
\end{aligned}$$

The decaying parameter λ is set to the same value as the one used in the Within-Topic Recency method.

5.3.6 Iterative Inference Model (Hub)

Bao et al. [18] proposed a model for estimating one's knowledge about locations based on Hyperlink-Induced Topic Search, an approach originally designed for link analysis on Web pages [72]. The model defines two properties for users and locations respectively, *i.e.*, hub scores for users and authority scores for locations. The hub score indicates how well a user can serve as an information source about a place and the authority score presents how popular a place is. We implement a normalized version of the algorithm and focus on hub scores for users (candidates) which is used as an estimation of users' expertise and are calculated as $\mathcal{S}_h(u,l) = \boldsymbol{h}_{u,l}^{(n+1)}$, where

$$\boldsymbol{h}^{(n+1)} = \frac{\mathbf{M}^T\mathbf{M} \cdot \boldsymbol{h}^{(n)}}{\parallel \mathbf{M}^T\mathbf{M} \cdot \boldsymbol{h}^{(n)} \parallel}.$$

and $\parallel \cdot \parallel$ is the norm of a vector.

5.3.7 Candidate Profiling

As a mainstream location-based social network, Foursquare attempts to increase user engagement by encouraging users, through gamification, to check-in at a location far more times than they actually need to. For example, users who

have the most check-ins at a place can win the mayor title of the place. To investigate the effect of this twisted relation between check-ins and visits, we define a different type of candidate profile, *i.e.*, Active-day Profile (referred to as +A while +C is used to refer to original check-in profiles). It is a subset of a user's check-ins which is defined as :

$$C_u^{+A} = \{c | c \in C_u, \nexists c' \in C_u : l_c = l_{c'}, t_c < t_{c'} < \lceil t_c \rceil_D\}$$

where $\lceil \cdot \rceil_D$ is a ceiling function towards midnights. Informally, the Active Day profile contains only the last check-in within each day at each place, reducing the influence of multiple check-ins at the same place.

5.4 Evaluation

To obtain a quantitative evaluation of the proposed methods as we raised in RQ3c, we conduct experiments on a configurable system with the implementation of the proposed methods and baselines. Each configuration combines a method and a profile type and we run experiments on all possible combinations. The system accepts a topic which is composed of a location or type of locations and a scope of city, and returns a list of related candidate experts according to the topic. The rank of candidates is determined by the system according to their geo-expertise on the topic estimated by the methods mentioned in the previous sections. The lists of returned candidates from different configurations are then pooled (see Section 5.4.2) and annotated (see Section 5.4.4). Based on the annotation on all candidate-topic pairs, the proposed methods and baselines are evaluated for their performance in retrieving geo-experts.

The data is collected from Twitter[10] and Foursquare[11] through their public APIs. We filter out the POI-tagged messages from Twitter and match the locations with that from Foursquare, so we can have each POI-tag associated with a category defined by Foursquare (see Section 5.4.1). Then we prepare a set of topics based on the data collected which is detailed in Section 5.4.2.

Two sources of annotation are used in the evaluation. One is from the pooled candidates via Twitter, *i.e.*, we directly ask the experts retrieved by the system whether they consider themselves having knowledge about the topic. (see Section 5.4.3) The other is from a third-party annotation, *i.e.*, online or

[10] `https://dev.twitter.com`
[11] `https://developer.foursquare.com`

offline recruited participants are asked to annotate the retrieved candidates according to the topic. (see Section 5.4.4)

With the annotated topic-candidate pairs, we measure the performance of the proposed systems by P@1, P@5 and MAP, which are widely used in different studies regarding information retrieval. A random baseline is also included in the evaluation to demonstrate the effectiveness of the proposed methods.

In principle, a better solution of evaluating the methods would be based on the outcome of the interaction between users and candidates [126]. However, in expertise seeking tasks, the real interactions between users and candidates are hard to capture, as they may communicate through other channels, *e.g.*, offline meetings [16]. Though, in an online environment, it is likely that the expertise seekers and experts will have conversation online, *e.g.*, via Twitter. It would either take too long or be complicated to collect meaningful data points for evaluation due to the small group of initial users, potential misunderstanding of the purpose of the system and no real stimulation of using it.

5.4.1 Dataset

We reuse the dataset collected for previous studies (described in Chapter 4), namely a collection of POI-tagged tweets from the four cities, New York, Chicago, Los Angeles, San Francisco. Since evaluating one's geo-expertise requires a full profile of user check-ins, which was not the intent of the previous dataset, we extended it to include users' full check-in profiles. Besides, in the experiments, we focus on the active users of POI-tagged tweets, *i.e.*, those who have posted a fairly number of POI-tagged tweets. We presume they hold a positive attitude towards sharing their locations and also knowledge about locations on social media. Thus, the users who have more than 5 POI-tagged tweets in the previous dataset are collected and we queried their full check-in profiles via Twitter's public API to form the dataset used in this study. This results in a dataset containing 1.3M check-ins from 8K users.

As shown in Figure 5.11, the check-in distribution among users does not follow the power law, because this dataset biases towards active users. Table 5.1 summarises the statistics of users and their check-ins collected from each of the four cities.

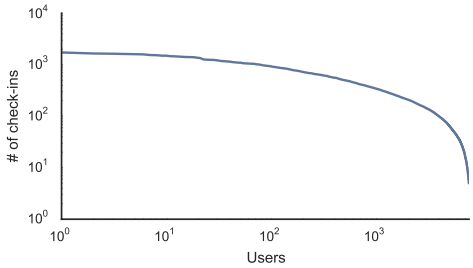


Figure 5.11: Check-in distribution

Table 5.1: Users and check-ins for each of the four city

Region	Check-ins	POIs	Users
San Francisco	113 613	6 349	2 877
New York	391 509	24 582	4 881
Los Angeles	202 770	18 740	3 508
Chicago	163 722	12 157	2 474

5.4.2 Topics

To prepare a set of topics for evaluation, we use stratified sampling to put together a seed set of location categories and POIs, based on their popularity in the dataset. Specifically, two strata are composed respectively for popular POIs (top 10%) and less popular POIs (90%). POIs are selected at random using a uniform distribution per stratum, and the number of samples taken is in accordance with the size of the stratum. As for category topics, we include all 9 top categories (*e.g.*, Food) and apply 10%:90% stratified sampling to the categories at the lower levels (*e.g.*, Chinese Food, Mexico Food). This results in a seed set of total 275 topics for all 4 cities. We remove POIs/categories which have less than 5 visitors and whose names are obscure (Building, Home Private, Field, Professional & Other, Residence). Finally, we obtain 95 topics from 4 cities in total, among which there are 71 category topics and 24 POI topics.

Pooling

There are 11 system configurations for evaluation, as shown in Table 5.2. We build a pool of all the candidates that are ranked in the top 5 positions in any of the returned rank lists from any configuration of the system. This pool contains 1588 topic-candidate pairs in total.

Table 5.2: Systems to be evaluated

Abrreviation	Method	Profile Type
WTA+C	Within-topic Activity	Checkin Profile
WTD+C	Within-topic Diversity	Checkin Profile
WTR+C	Within-topic Recency	Checkin Profile
WTRD+C	Within-topic Diversity + Recency	Checkin Profile
Hub+C	Hub score	Checkin Profile
WTA+A	Within-topic Activity	Active-day Profile
WTD+A	Within-topic Diversity	Active-day Profile
WTR+A	Within-topic Recency	Active-day Profile
WTRD+A	Within-topic Diversity + Recency	Active-day Profile
Hub+A	Hub score	Active-day Profile
Rand	Random	–

5.4.3 Self-Evaluation Through Social Channel

A nice feature about online social networks is that people can connects easily with each other. This gives us an opportunity to evaluate our system from a unique angle, *i.e.*, by simply asking the candidates. We presented each of the pooled candidates with our model's predictions of their individual expertise and asked them to judge their actual knowledge about the topic on a 5-point scale from 1 ("I do not know about this") to 5 ("I am an expert").

A group of 10 candidates volunteered to work with us. Six participants indicated high expertise (Grades 4–5) in the topics predicted by our method. Another three reported reasonable competency (Grade 3) towards the predicted topics. Only a single participant indicated mild expertise (Grade 2) towards the predictions.

This method is proven harder than recruiting annotators from elsewhere, as candidates might resist answering requests from us as we are totally stranger

to them.

5.4.4 Third-Party Evaluation

Crowdsourcing is an approach of distributing small tasks that require human intelligence. The task workers are paid small amounts of money in return. In the information retrieval community, crowdsourcing is popular for assessing ranking list returned by studied systems [7, 6]. There are two main reasons for IR scientists using crowdsourcing platforms to collect ground truth, which are: 1) it is easy to hire a group of workers, 2) Population from which the workers are sampled may be more representative than the usual approach to hire assessors close to the science community, which may reduce bias. In this way, researchers can evaluate their systems by the data annotated by humans and compare the scores between different systems. We also adopt this way for evaluating our system.

We setup a job on CrowdFlower and direct workers to our annotation system. They were asked to read the instruction and our policy of privacy and answer a survey as described in the previous section. Then they work on a batch of ten topic-candidate pairs per iteration. In the end, they are provided with a code that can be redeemed on CrowdFlower to receive their credits.

However, crowdsourcing ground truth annotation faces challenges in quality control and the potential for workers to perform badly due to cheating or lack of task comprehension, leading to managing effort and cost overhead[45]. Thus we designed and implemented an interactive annotation system for this task, which is detailed in the following section. The annotation system detects annotators' error-prone behaviour and issue a warning to explain what is wrong. To be specific, it tracks annotators' mouse movement and clicks and requires a minimum number of mouse clicks and mouse travel distance. When an error-prone submission is detected, the annotator will receive a full screen message reminding them of the need to inspect a candidate's profile prior to submitting the assessment.

Annotation System

As mentioned in the previous section, we implemented an annotation system for collecting ground truth. As suggested by Liao et al. [91], insufficient information may lead to low quality judgements from assessors. Our annotation interface

is designed to allow annotators to explore candidate check-in profiles as shown in Figure 5.12. With the interface we try to mitigate the chances of leading the participants to bias towards our desired assessment. It is designed to allow them to follow their own way of making judgements based on the displayed evidence. The variety of charts can make sure participants are able to inspect as many perspective as they consider important. The interactiveness give them the freedom of investigating candidate profiles in any scales they want. Here are the list of key features of the annotation interface:

- render statistics rather than raw data,
- depict as many aspects as possible regarding candidates' check-in profiles,
- arrange charts in a compact yet tidy way to reduce confusion,
- enable interactivity between annotators and the interface to facilitate flexible inspection, and
- to provide error-prone behaviour detection and warning.

The implementation (detailed in Chapter 7) is based on Django, a web application framework based on Python, and hosted on Google App Engine, in which the back end serves the data and the front end aggregates the statistics and makes charts. To achieve high interactivity, the front end weaves together several JavaScript libraries including *crossfilter*, *dc.js*, *d3.js*, *gmap*. The back end includes a layer wrapping over Google Apple Engine's DB Datastore[12] to provide data used by the front end, *e.g.*, the candidates, topics, check-ins and annotations. The back end also manages the flow of annotation, *e.g.*, routing between information pages and annotation pages, tracking annotators' contribution.

Each retrieved candidate expert is rendered in a page from a number of different perspectives based on their check-in profiles. Annotators are asked to review the check-ins by interacting with the charts displayed on the page. They are then asked to assign a score indicating whether they think the candidate is capable of answering questions, if being asked, about the locations or the types of locations. "5" indicates that the candidate knows the topic very well and "1" indicates that the candidate is not a suitable person to answer questions

[12]https://developers.google.com/appengine/docs/python/datastore

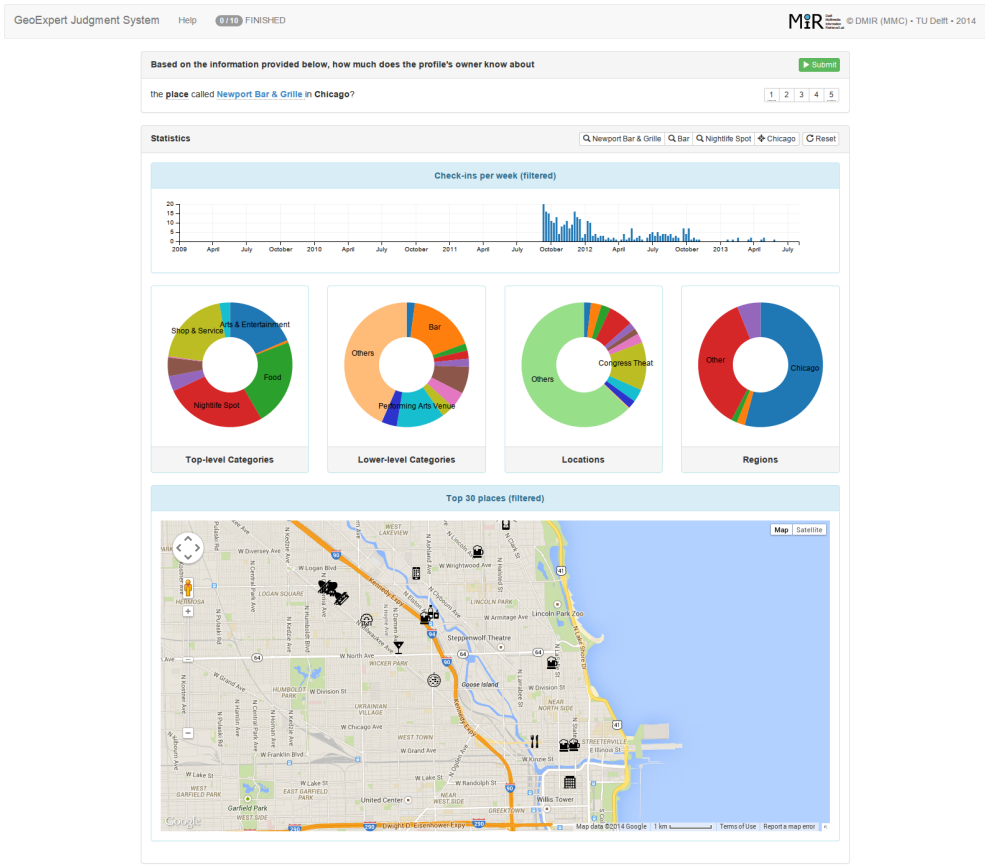


Figure 5.12: The interface of annotation system for geo-expertise

about the topic. Before annotators start, a brief introduction is presented to help them learn how to use the interface and what kind of interactions are supported (Figure 5.13).

Annotation Analysis

We have carried out three runs of annotation. The first run (CF) has been carried out on CrowdFlower where each participant is paid 0.5 USD per task, since each task contains 10 topic-candidate pairs to annotate and may take

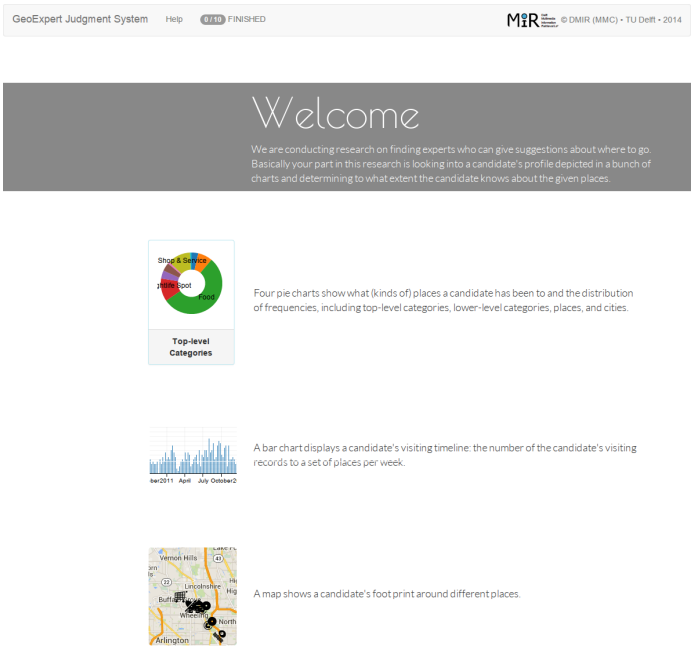


Figure 5.13: The instruction presented before annotation tasks

about 30 minutes to finish[13]. The annotators for the second (U1) and third (U2) runs are university students and staff[14] from Delft University[15], who volunteered to participate in the experiments. The general statistics of all the three batches are shown in Table 5.3. The first two runs cover the whole pool of topic-candidate pairs while the third one covers only about a half of the pool. The third one is reserved for assuring the agreement between annotators.

The raw scores between 1 and 5 assigned by the annotators are converted to binary annotations by a threshold at 3, *i.e.*, topic-candidate pairs that assigned with a score of 4 or 5 indicate that the candidates are experts on the topic, while 3 or lower indicates non-experts. Then we evaluate the agreement between the binary versions of the three annotation runs via Cohen' Kappa

[13] `http://crowdresearch.org/blog/?p=9039`

[14] Including bachelor students, master students, PhD students, postdocs, and assistant professors

[15] Including faculties of TBM, CiTG, 3mE, OTB, BK, TNW, LR, IO, EWI

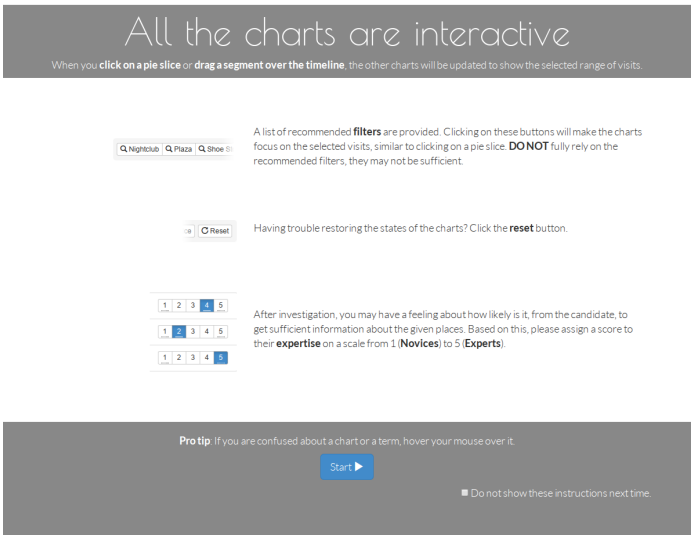


Figure 5.14: The instruction presented before annotation tasks (continued)

Table 5.3: Crowdsourcing Annotation Runs

Runs	Sources	Judges	Tasks	Topics	Candidates
CF	CrowdFlower	86	1588	95	1121
U1	University	116	1588	95	1121
U2	University	105	749	95	616

[53]. As shown in Table 5.4, the three runs only achieve *slight agreement*. This may be caused by the uncertainty of the criteria used by different annotators. Many factors can be important in judging one's geo-expertise and annotators may not agree on one single criterion for determining geo-expertise. We further investigate how much the annotators agree on definite cases, *i.e.*, the topic-candidate pairs that are assigned with score 1 or 5, the result of which is shown in Table 5.5. In these cases, Run U1 and U2 have *moderate agreement*. The increase of agreement may be due to the reason that annotators hesitate less when assessing the candidates with very strong or very weak check-in profiles. However, when the profile owners are of neither cases, the annotators may struggle in choosing criterion and this may lead to disagreement between the annotators.

Table 5.4: The agreement between annotation runs

Kappa	CF	U1	U2
CF	1.0000	0.1223	0.1762
U1		1.0000	0.1912
U2			1.0000

Table 5.5: The agreement between annotation runs based on definite topic-candidate pairs (extreme values on the rating scale)

Kappa	CF	U1	U2
CF	1.0000	0.4271	0.3734
U1		1.0000	0.4887
U2			1.0000

Table 5.6: The agreement between annotation runs based on definite topic-candidate pairs (extreme values on the rating scale) from the random baseline

	CFvsU1	CFvsU2	U1vsU2
WTA	0.4700	0.2702	0.3725
WTD	0.6370	0.2935	0.5000
WTR	0.5006	0.2666	0.3178
WTRD	0.4391	0.1447	0.3314
Rand	-0.0270	0.1886	0.1983
Hub	0.5500	0.5333	0.5000

A closer look at the definite topic-candidate pairs shows that the annotators hardly agreed on the ones generated by the random baseline systems, as shown in Table 5.6. As for the proposed methods and the Hub-Authority based method, we found annotators achieve fair and moderate agreement on the definite topic-candidate pairs. To sum, annotators are inclined to agree when they have strong opinions on whether or not a candidate holds the geo-expertise in the query.

5.4.5 Quantitative Evaluation

We carry out separate evaluations on both sets of annotations (CrwodFlower recruited workers and the university staff and students) in order to avoid having to make an arbitrary choice on merging different opinions from different annotators. `trec_eval` [16] is used for the evaluation. As mentioned before, the annotations are converted into binary values, in which topic-candidate (topic-document) pairs assigned with score 4 or 5 are considered relevant and those with score 1, 2, or 3 are considered as irrelevant. Furthermore, we test the statistical significance of differences between the evaluation scores using the Wilcoxon Signed Rank Test ($\alpha < 0.05$).

Evaluation Based on CrowdFlower Annotation

As shown in Table 5.7, under P@1 and P@5, the configuration of WTD+A performs the best, and under MAP the configuration of WTA+C performs the best. The configurations of these two methods focus on the experience the candidates have cumulated via check-ins, which is valued more by the annotators in this experiment. Though, many survey participants indicated that recency is an important factor to consider, the scores of the corresponding systems (WTR and WTRD) are not as high as the systems with WTA and WTD. This may be due to the fact that the candidates who are new in the location based social networks have only a few recent check-ins regarding the given topic and are not considered as an expert, while some other candidates who stopped checking-in recently but have a lot of diverse check-ins in the past are considered as experts. In general, all the proposed methods with both types of profiles significantly outperform the random baseline. This suggests that the proposed methods are effective in retrieving geo-expertise. However, the annotators from crowdsourcing channel do not show a very clear preference between the proposed methods and we do not observe uniform preferences between the two profile types. This may suggest that the two types of profiles do not diverge much and the check-in gamification does not have an observable influence on assessing candidates' geo-expertise.

[16] `http://trec.nist.gov/trec_eval/`

Table 5.7: The evaluation results based on CF annotation

Method	Profile	MAP	P@1	P@5
WTA	+A	0.2750	0.4211	0.3979
	+C	**0.2771**	0.4316	0.3895
WTD	+A	0.2340	**0.4789**	**0.4197**
	+C	0.2280	0.4507	0.4169
WTR	+A	0.2442	0.3789	0.3747
	+C	0.2508	0.4211	0.3768
WTRD	+A	0.2434	0.4316	0.3684
	+C	0.2491	0.4421	0.3726
Hub	+A	0.2327	0.4507	0.4113
	+C	0.2363	0.4366	0.4028
Rand	–	0.1343	0.2316	0.2063

Evaluation Based on University Annotation

Similarly, we conduct the evaluation based on the university annotation (U1) and the results are shown in Table 5.8. The annotators also prefer the configuration with WTD under P@1 and P@5 and the configuration with WTA under MAP, though no significant differences between the configurations with these two methods are observed. The proposed methods and the hub-authority based method are all significantly better than the random baseline. Different from the evaluation based on the CF annotation, there is a clear preference of the proposed methods (ones with WTD, WTA, or WTRD) to the configuration with the hub-authority based method, *i.e.*, the difference between them are significant under P@5 and MAP while not under P@1. As among the configurations with the three methods, WTD, WTA, WTRD, we do not observe significant differences. As for the two types of profiles, the differences between them in the configurations with the same method are not significant.

Summary

With the group of annotators from the University, we can observe clear preference towards the proposed methods over Bao et al.'s [18] method (Hub) and the random baseline. With more diverse annotation from the crowdsourcing channel, the advantages of the proposed methods become vague and insignificant. The uncertainty of the criteria, which is introduced by allowing the

Table 5.8: The evaluation results based on U1 annotation

Method	Profile	MAP	P@1	P@5
WTA	+A	**0.3218**	0.5579	0.4463
	+C	0.3147	0.5579	0.4337
WTD	+A	0.2878	0.5775	0.4817
	+C	0.2908	**0.5915**	**0.4845**
WTR	+A	0.2814	0.5263	0.4042
	+C	0.2824	0.5368	0.4063
WTRD	+A	0.2862	0.5368	0.4042
	+C	0.2919	0.5368	0.4168
Hub	+A	0.2041	0.5211	0.3859
	+C	0.2045	0.5493	0.3831
Rand	–	0.1041	0.1579	0.1600

annotators freely choose their own criteria, leads to diverge of judgements and no significant differences in the evaluation.

The profile types, which are introduced due to the concern of check-in gamification, do not improve the performance. This suggests that the gamification of check-in behaviours may not affect much on assessing the candidates' geo-expertise. Regarding the experiment settings, crowdsourcing shows its advantage in efficient recruiting, while controlled groups shows better annotation agreement.

5.5 Conclusion

In this chapter, we presented a novel type of domain expertise regarding location knowledge which focuses on *Points of Interest* (POI) and their categories. We conducted a qualitative user study and investigated, via online questionnaire based survey, the way in which Twitter and other communication channels are used for searching, receiving and giving location-related advice and recommendations. Doing so, we found the need of geo-expertise from the participants in our user study. As reported, there is a general needs of geo-expertise and on-line communication with close friends and families or even the wider social network are the major channels for obtaining advice (RQ3a). We also presented survey participants with examples of Twitter streams and asked

them to judge the expertise of the showcased user towards a number of topics, as well as, to explain which criteria influenced their decision. Within-topic coverage and diversity turned out to be the most frequently named features.

On the basis of these qualitative insights, we designed three automatic retrieval methods based on probabilistic models for ranking domain expertise (RQ3b). We evaluated the methods based on the annotations from both the crowdsourcing channel (CrowdFlower) and the university staff and students. By carefully designing and implementing the annotation system, we keep the balance between too little information that may lead annotators to biased decision and too much information that overloads annotators with details. With the annotation collected, we quantitatively evaluate the proposed methods.

Though the overall interrater agreement is not high between the crowdsourcing annotators and the university annotators, they agree more (moderate agreement) on definite cases that they give either "1" or "5" scores. This may imply that the middle-class in the candidate population causes the disagreement, which is a consequence of the open setting in the annotation system where various criteria are allowed. As shown in the evaluation (RQ3c), the annotators from the university have a more clear preference of the methods which are in favour of candidates with a more diverse check-in profile. On the other hand, there is no significant difference between the methods based on active-day profiles and check-in profiles, the former of which is presumed to be a better representation of users' location knowledge, since the gamification of check-ins may twist users towards non-engaged presences at a location.

The evaluation shows promising results of the proposed methods and encourages future quantitative confirmation in a Web-scale setting. A dedicated geo-expertise retrieval system can be implemented based on Twitter/Foursquare API, in which users can authorize the application to analyse their check-in profile and search in their friend circles for geo-expertise. We can then ask users to assign a score to their geo-expertise and/or their fellows' geo-expertise retrieved by the system. In this way, different methods can be evaluated and compared under real scenarios.

Part III

Applications

Chapter 6

Geo-tagged Tweeting for Water Damage

Environmental monitoring is an import aspect of modern city management and various devices are deployed for effective data collecting. Users of online social networks post various kinds of messages, from personal feelings, selfies (self-portrait photos[1]) to announcements, news. It is also likely that users post messages regarding topics that city planner and management team care about. The potential advantage of this new channel is that the social network users can be seen as a sensor network which is readily deployed and covers most of urban areas. The question is whether this new channel is potentially supplement to the current deployed ICT infrastructures for environmental monitoring and management. In this chapter, we investigate the fourth research question (RQ4), *i.e.*, whether and how we can extract and make use of user contributed content on social media for understanding water damage? We present a case study of comparing water damage related tweets with the registered damage reports from municipalities. We take two large cities in the Netherlands as examples, namely, Amsterdam and Rotterdam, and investigate the correlation between messages from Twitter and two incidents of extreme weather hit the two cities respectively.

[1] http://en.wikipedia.org/wiki/Selfie

6.1 Introduction

Water damages such as floods, broken pipes become an essential problem for modern cities, especially under climate changing. A practical way is to mitigate the problem by running risk management on this issue which requires to incorporate fine-grained and real-time data sources. As a case study, we focus on two cities in the Netherlands, *i.e.*, Amsterdam and Rotterdam, which are the top two largest cities in the Netherlands.

Both municipalities run damage reporting systems which register residents' reports on public facility issues such as floods, ponds, *etc.* However, the registration of such reports takes a lot of man-hours and scattered small issues may not be investigated due to the lack of motivation to report or busy lines of the systems. Social networks accumulate a large volume of information regarding individual users' experience, which may be a potential resource for gathering information about water damage in urban areas.

6.2 Related Works

In crisis management, many researchers have looked into social media sources for real-time information regarding incidents. For example, Sakaki et al. [131] proposed a set of temporal and spatial models based on probabilistic models for detecting disasters on Twitter, *e.g.*, earthquakes, typhoons, which have national impacts. Starbird et al. [142], Vieweg et al. [152] analyse the textual features of tweets related to two specific incidents, the Oklahoma Grassfires and the Red River Valley Floods. Abel et al. [1] demonstrated a system for enriching messages from emergency broadcasting services with tweets related to them. Maxwell et al. [104] demonstrated a system for detection of crisis from real time tweet stream and synthesized different sources of information on one monitor interface, such as videos, maps, and text messages. Different from these studies, we focus on the spatial and temporal analysis of the tweets related to strong precipitation within cities. We compare different sources of information, investigate the correlations between them in terms of space and time for both short-term and long-term analysis.

SHINE[2] is a research project in Delft University of Technology which aims at utilizing heterogeneous information networks for better urban management.

[2] `http://shine.tudelft.nl`

Heavy rainfall is one of the studied cases demonstrating how different types of resources can be combined for better understanding and decision making. For example, Gaitan et al. [55] combined information from radar measurements, laser precipitation monitors, automatic weather stations and damage reports to find vulnerabilities in the city of Rotterdam. The study in this chapter evaluates the value of social media in this context.

6.3 Data

In this case study, we rely on two main sources of data, *i.e.*, the records of damage reports from the municipalities and tweets regarding rainfalls and water damage in the cities. In order to investigate the spatial and temporal distribution of these two data sets, we compare the data sources with official census data and weather records. In this section we briefly introduce the selected incidents and the four sources of data used in our analysis.

6.3.1 Storm Incidents

The fast-grown user base of social media enables researchers to observe human behaviours at a large scale. We presume that the strong precipitation would affect users of social media and the reaction may be recorded in their messages on social media such as Twitter. Then these message can be used as a complementary information sources for city management.

Due to the difficulty of collecting tweets related to water damage in general, we focus on two selected incidents of storms which hit the two cities in the past years. The first incident we look into is a storm hitting Amsterdam on July 14, 2012 [112, 146]. The severe precipitation caused many areas being flooded and a tunnel being closed for a short time. A lot of reports of flooded basements were recorded by the municipality of Amsterdam. We choose a time window from July 11 to 17, 2012 for investigation.

As for Rotterdam, we investigate the storm on October 13, 2013, which caused a lot of damage to Rotterdam and nearby areas around the city [106, 114]. Similarly, the precipitation flooded many areas, temporarily stopped the metro traffic and many reports were recorded by the municipality during those days. For this incident, we choose a time window from October 11 to 17, 2013 as our investigation range.

6.3.2 Damage Reports

In the cities of Rotterdam and Amsterdam, residents can report water related problems in public space, such as floods, sewage problems, to municipalities via online forms[3] or phone calls. Usually, a report contains a series number, to identify the time of registration, the street name, the house number (for indicating the location of the incident), *etc.* Available records from the Rotterdam municipality range from 2004 to 2013, and the records from Amsterdam range from 2012 to 2014.

The textual addresses in the records can not be directly used in spatial analysis. To translate the addresses to pairs of coordinates in space, we use the geocoding service provided by PDOK[4]. The geocoding service allows free text input and returns coordinates in EPSG:28992, which is a standardized official projection for maps of the Netherlands[5].

However, not all the records can be processed correctly by the online geocoding service. For example, some of the records have no house number or misspelt street names. For records without house number or with a house number of zero, we queried with just the street names, in response to which the service will return the geometric center of the street. We rely on the geocoding service for handling misspellings of street names. PDOK is used as the primary geocoding service and for addresses not geocodable by this service we use a back up from Google Geocoding API[6]. For example, *Hudsonstraat* xxxx[7], *Rotterdam* is a valid address that is not correctly geocoded by the services provided by PDOK but can be geocoded by Google's service. Even though, there are still records (2.6% for Amsterdam, 0.09% for Rotterdam) that can not be associated with a location. Some of them basically have no valid addresses registered. There are also misformatted timestamps registered in the records. For these records, we assigned them with the time of the proceeding records (via their IDs).

In this study, we only look into the effective records, *i.e.*, having both valid

[3] http://www.gis.rotterdam.nl/msb.meldingenopkaart/MeldingenOpKaart.aspx

[4] https://www.pdok.nl/nl/producten/pdok-services/uitleg-over-services

[5] http://epsg.io/28992

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

[7] The house number is anonymized

timestamps and geocoded addresses. Table 6.1 shows the basic statistics about the records provided by both municipalities.

Table 6.1: The statistics of the registered records about water problems from Amsterdam and Rotterdam

	Amsterdam	Rotterdam
Total Records	6 181	46 355
First Date	01-JAN-2012	01-JAN-2004
Last Date	10-JUN-2012	31-DEC-2013
Unique Addresses	4 650	21 424
Geocoded via PDOK	4 207	19 928
Geocoded via Google	336	1 491
Effective Records	6 023	38 657

6.3.3 Tweets Related to Water Problems

Twitter as one of the most popular social media, accumulates half a billion tweets per day. It is neither practical nor feasible to retrieve all the tweets for analysis. In this case study we focus on two incidents and collect tweets relevant to these two incidents. To achieve this, we use several search services for potentially relevant tweets, filter out those qualified for analysis and then have them annotated manually.

Table 6.2: Keywords related to heavy rains and water damages

Keywords		
blank	hoost	overlast
overstroming	overstroomd	plas
regen	riool	sloot
verstopt	water	giet
kolk	neerslag	onweer
wateroverlast		

As a starting point, a list of keywords regarding rainfalls and water damage is compiled by two native Dutch speaker, as shown in Table 6.2. Then for each keyword in the list, we respectively queried three tweet search services, *i.e.*, Twitter Search, Topsy, Twiqs.nl [148].

Twitter provides a search API for accessing tweets by keywords, allowing to specify the location by the parameter *geocode*. However, the API can not be used in this case study since it only indexes tweets in last few days up to the time of query (6 to 9 days according to the official document[8]). The studied incidents are both much earlier beyond the capacity of the index.

The web interface of Twitter returns older tweets[9]. So we scrape the data from the web interfaces via a simple script querying HTML pages from Twitter. The queries used for collecting tweets are in the format: ⟨*keyword*⟩ *nearby*:"⟨*city name*⟩ *Netherlands*" *within:15mi since*:⟨*begin*⟩ *until*:⟨*end*⟩. It means tweets containing the keywords, have a related geotag locating within 15 miles from the city centre, and have been posted during the given time period. However, the returned tweets do not all satisfy the given constraints on the location, as Twitter also tries to return tweets from users who have the queried locations in their profile[10].

Besides the official Search API, we also used two other services. Topsy[11] is a search engine for tweets that claimed to have a full archive of tweets since 2006. Optional constraints on various dimensions are provided on the interface, such as locations and keywords. We use another script for automatically querying Topsy with the keywords. As Topsy stripes off geotag information from the returned tweets, we have to consult Twitter REST APIs for the geotags of the returned tweet IDs.

We also use the tweet search service from Twiqs.nl[12] for collecting tweets related to water problems. It is a dedicated service for indexing Dutch tweets streamed by sample endpoint of Twitter Stream API, which carries about 1% of the whole tweet stream going through Twitter. It has an option to harvest IDs of geotagged tweets posted within a give period. Then we retrieved the full tweet objects with these tweet IDs via Twitter REST APIs.

We combined tweets from different sources for each incident and semi-automatically labelled all relevant tweets with the following steps: i) Applying a keyword filter on the set of data with the keyword list. ii) Applying a

[8] https://dev.twitter.com/rest/public/search

[9] https://blog.twitter.com/2013/now-showing-older-tweets-in-search-results

[10] https://dev.twitter.com/rest/public/search

[11] http://topsy.com

[12] http://twiqs.nl

spatial-temporal filter to keep only the tweets that are posted during the time and in either cities.[13] iii) Annotating the relevance of each tweet (by a native Dutch speaker). By the three steps, we have collected a set of geotagged tweets related to water problems. During Step iii, our annotator noticed that there are tweets posted from channels dedicating to broadcasting emergencies, as both cities have official web sites for real time updates of received reports of emergencies. These tweets are categorized as auto tweets by the annotator. A brief summary of statistics of the tweets regarding these steps are listed in Table 6.3.

Table 6.3: The statistics of collected tweets regarding the two incidents respectively

	Amsterdam	Rotterdam
Twitter Search	40	30
Topsy	62	212
Twiqs.nl	55	222
Unique Tweets	99	288
Relevant Tweets	46	263
Auto Tweets	3	186

Significant difference in quantity is observed between the two sets of tweets for Amsterdam and Rotterdam. Our explanation to the differences is that the set of tweets for Amsterdam are one year earlier than that for Rotterdam. Twitter has a policy to protect users' privacy that forbid any services to store the tweets that users command to delete and the longer ago the tweet is the more likely they have been removed. Another factor could be the rise of popularity of both Twitter and location sharing. That is, more people used Twitter and especially geotags in 2013 than 2012.

For short, we use the term *tweets* to refer to the relevant tweets unless otherwise specified in the rest of this chapter.

6.3.4 Residential Population

Since the water damage reports are filed by residents in the cities, it is likely that the distribution of damage reports correlates with the population density

[13]The spatial filter is based on the boundaries provided by PDOK

distribution. Intuitively, the less the residents live in an area, the less likely an incident will be discovered and reported. To investigate such correlation we retrieved from PDOK services the resident population distributions for both cities. The data is provided as a shape file storing the numbers of residents registered (in 2013) in grids of 100 meters by 100 meters cells, as shown in Figure 6.1 and Figure 6.2. The colored squares representing the density of the residents in log scale from light green (5–10 residents) to red (500+ residents). The orange dotted lines in the figures show the boundaries of municipalities.

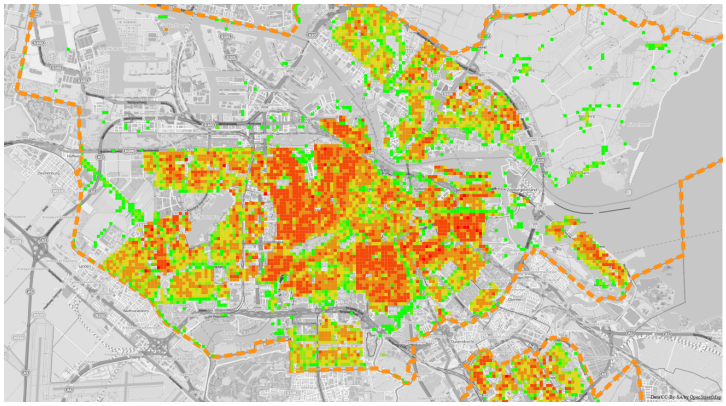


Figure 6.1: The population distribution of Amsterdam, Netherlands

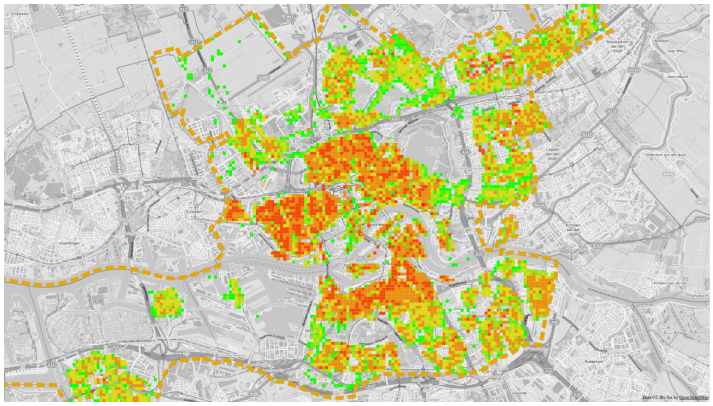


Figure 6.2: The population distribution of Rotterdam, Netherlands

As can be seen in both municipalities, there are large areas with nearly no

residents. For example, the north west part of Amsterdam and the south east part of Rotterdam are less populated because they are port areas. It should also be noted that Rotterdam city also includes the bank of the river until the sea which is not fully shown in the figure.

6.3.5 Precipitation

Water damage is usually considered relevant to the incidents of strong precipitation. To investigate the strength of the correlation between water damage reported and the precipitation in the cities, we used archived weather data from KNMI[14], which provides hourly precipitation records. Since there is no weather station in the municipality of Amsterdam, we use the data from the nearest weather station (*i.e.*, Schiphol) as an approximation of the data from the city. As for Rotterdam, we use the data from the weather station in Rotterdam.

6.4 Analysis

In this section, we demonstrate some analysis with the data obtained from different sources including the tweets we collected.

6.4.1 Long-Term Analysis

Strong precipitation is usually a significant cause of water damage in delta cities like Amsterdam and Rotterdam. Based on the data from the records of reported water damage and the weather station, we investigate the correlation between them. In this case study we use Pearson correlation:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

Though water damage in cities are more likely caused by short and strong precipitation, the correlation at monthly scale can give better understanding of how the trend of climate change will affect the facilitate currently used in urban areas. Figure 6.3 and Figure 6.4 show comparisons between the monthly precipitation and the frequency of reports. The correlations are respectively

[14] http://www.knmi.nl/klimatologie/daggegevens/download.html

0.8443 (Amsterdam) and 0.6951 (Rotterdam). The figures confirm strong correlation between the reported water problems and precipitation in a month granularity.

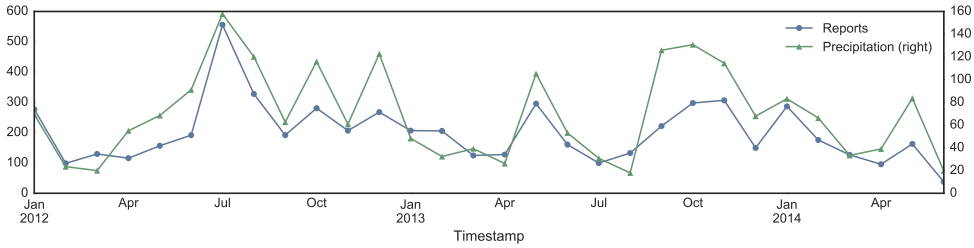


Figure 6.3: Comparison between the precipitation and reports of water problems in Amsterdam

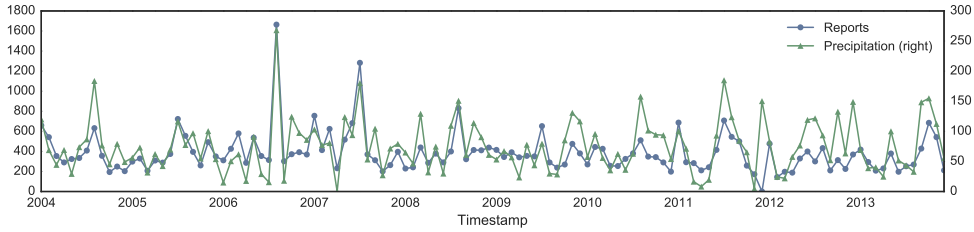
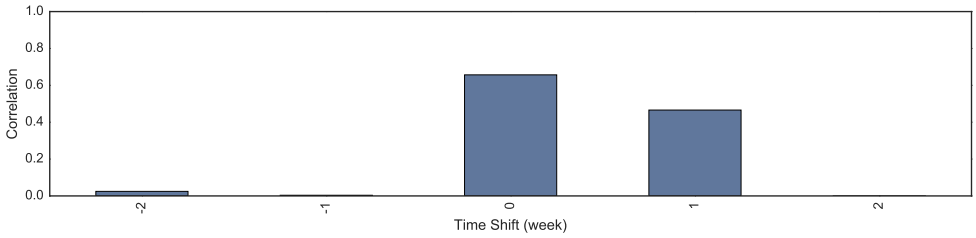
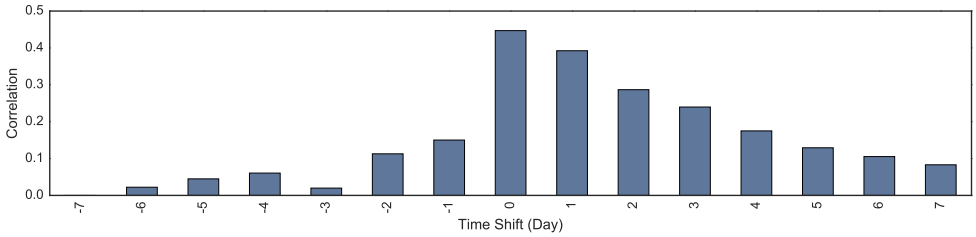


Figure 6.4: Comparison between the precipitation and reports of water problems in Rotterdam

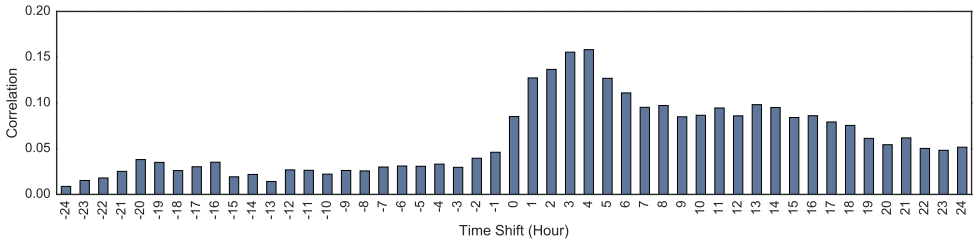
We also tried 3 other comparison of smaller granularities, *i.e.*, weekly, daily and hourly and shifting the data in temporal dimension in order to compensate the time gap between the precipitation and the damage reports. There are two reasons we include the shifting in our comparison. First, the weather station is away from the urban area which may introduce time difference when use the data from it as an estimation of the precipitation in the urban area. Second, there may be a general delay of reporting due to the reluctant of action (reporting) when the water damage happens to public facility, especially, when the system is not easy to access (*e.g.*, looking for the right phone number or website). As shown in Figure 6.5 and Figure 6.6, the general trend is that the



(a) Weekly



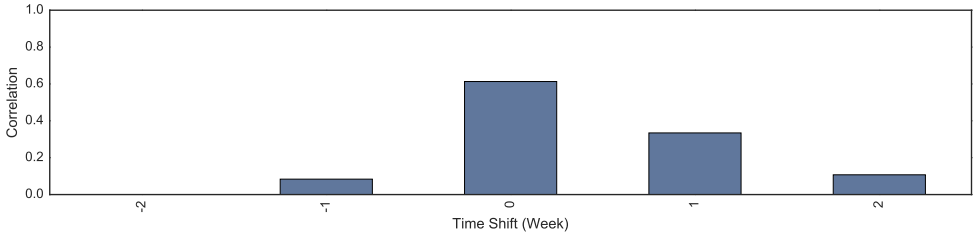
(b) Daily



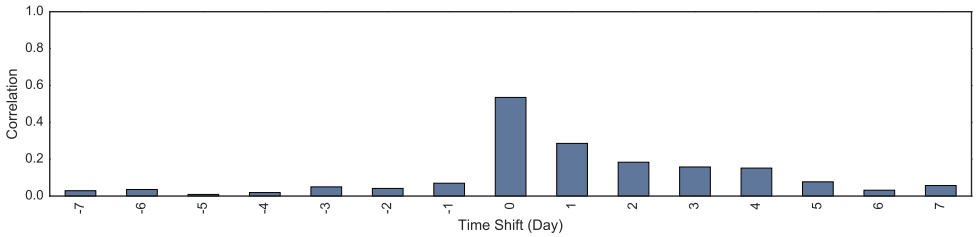
(c) Hourly

Figure 6.5: Correlation between the time-shifted precipitation and the reports in Amsterdam

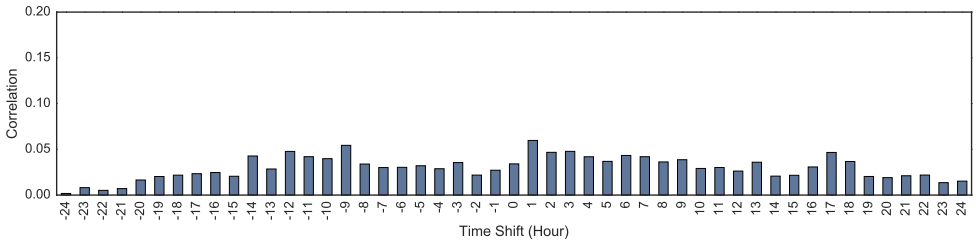
correlation decrease when the granularity of comparison gets smaller. This may be due to the uncertainty of when the damages were observed and reported and the Pearson correlation is not good at capturing this kind of uncertainty. An interesting finding is that at hour granularity, the correlation goes to a peak (0.158) if applying a 4-hour shift, though it is very weak. This particular peak may worth further investigation.



(a) Weekly



(b) Daily



(c) Hourly

Figure 6.6: Correlation between the time-shifted precipitation and the reports in Rotterdam

In order to estimate the correlation between reported issues (from either damage report channel or Twitter) and population, we use Kernel Density Estimation to associate a spatial point with a probability that the reported issue happened there. Then we calculate correlation with Pearson correlation between the estimated spatial distribution of either reports or tweets and the population density in both cities.

For the KDE we use a Gaussian kernel with a factor l estimated via Silverman's rules[15], *i.e.*,

$$\hat{f}(\mathbf{x}) = \frac{1}{2\pi n\sqrt{|l^2\boldsymbol{\sigma}^2|}} \sum_{\mathbf{x}_i \in \mathbf{X}} \mathrm{e}^{-\frac{1}{2}(\mathbf{x}_i-\mathbf{x})\boldsymbol{\sigma}^{-2}l^{-2}(\mathbf{x}_i-\mathbf{x})}$$

where $\mathbf{X}$ is a set of vectors representing geo-spatial points, $l = \left(\frac{n(d+2)}{4}\right)^{-\frac{1}{d+4}}$, and $\boldsymbol{\sigma} = \mathrm{cov}(\mathbf{X}, \mathbf{X})$.

Figure 6.7 and Figure 6.8 show comparisons between the resident distribution and the density of reports. The density of reports is estimated by Kernel Density Estimation with a bandwidth manually selected. The correlations between the distributions of reports and population density are respectively 0.4893 (Amsterdam) and 0.5070 (Rotterdam). The correlation between the distribution of reports and resident population density is moderate, weaker than the correlation with precipitation data from the weather station.

6.4.2 Short-term analysis

As mentioned in Section 6.3.1, two time periods have been selected for Amsterdam and Rotterdam respectively, during which severe storms hit the two cities. In this section, we demonstrate how the perception of these two incidents is represented on social media, by comparing them to the reports of water damage recorded officially.

Similarly to the long-term analysis, we compare the temporal (hourly) distribution of tweets mentioning rain and water issues with the precipitation recorded during the same time. As shown in Figure 6.9 and Figure 6.10, tweets bursts are closely aligned in time with the precipitation bursts recorded at

[15]Implemented in `http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.gaussian_kde.html`

Figure 6.7: The distribution of reports from Amsterdam 2012–2014 (bw=0.0005)

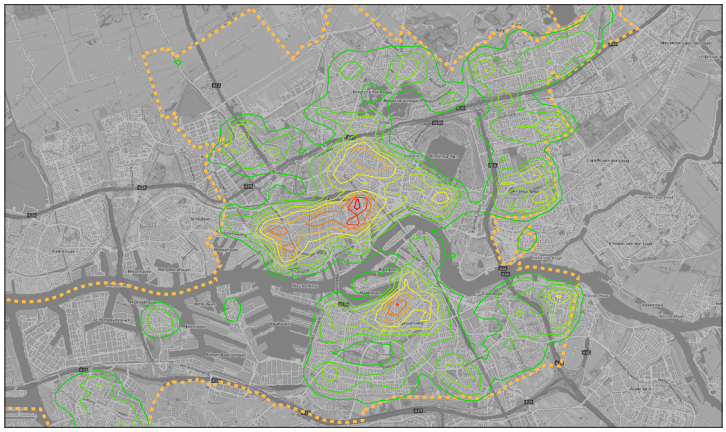


Figure 6.8: The distribution of reports from Rotterdam 2004–2014 (bw=0.05)

weather stations. The bursts in the time series of damage reports are some time behind the strong precipitation.

As shown in Table 6.4 and Table 6.5, the time series from tweets are closer correlated to precipitation than the time series of damage reports in both cities and the correlations are respectively 0.4527 (tweets) and 0.0082 (reports) in

Amsterdam and 0.3067 (tweets) and -0.0553 (reports) in Rotterdam.

Table 6.4: The correlation between precipitation, tweets and reports in Amsterdam

	Precipitation	Tweets	Reports
Precipitation	1.0000	0.4527	0.0082
Tweets	0.4527	1.0000	0.0784
Reports	0.0082	0.0784	1.0000
Tweets (-1h)	0.4947	0.3727	0.0426
Reports (-3h)	0.3574	0.1728	0.2769

Table 6.5: The correlation between precipitation, tweets and reports in Rotterdam

	Precipitation	Tweets	Reports
Precipitation	1.0000	0.3067	-0.0553
Tweets	0.3067	1.0000	0.1092
Reports	-0.0553	0.1092	1.0000
Reports (-1h)	0.0544	0.1387	0.4776

We tried to align the time series of damage report from both social channel and official channel with precipitation. It is found that for Amsterdam the correlation between them reaches a maximum of 0.3573 when the time series of reports is moved backward in time of 3 hours. As for Rotterdam, the maximum correlation of 0.0544 is achieved when moving the report time series backward in a hour. By aligning the time series of tweets, we achieve in Amsterdam a maximum correlation of 0.4947 when moving the time series backwards in time of a hour. In Rotterdam, the time series of tweets achieves the maximum without moving. In general, the correlation between the tweets and the precipitation are higher than that of the reports.

We also compare the correlation at day granularity, and observe higher correlation between the tweets and the precipitation in both cities, 0.7214 for Amsterdam and 0.7591 for Rotterdam, than that of reports, 0.4516 for Amsterdam and -0.0526 for Rotterdam. The extreme low value for the correlation may due to the delay of access to reporting issues, as when we shift the report back in one day we have a much hight correlation, *i.e.*, 0.7702. This

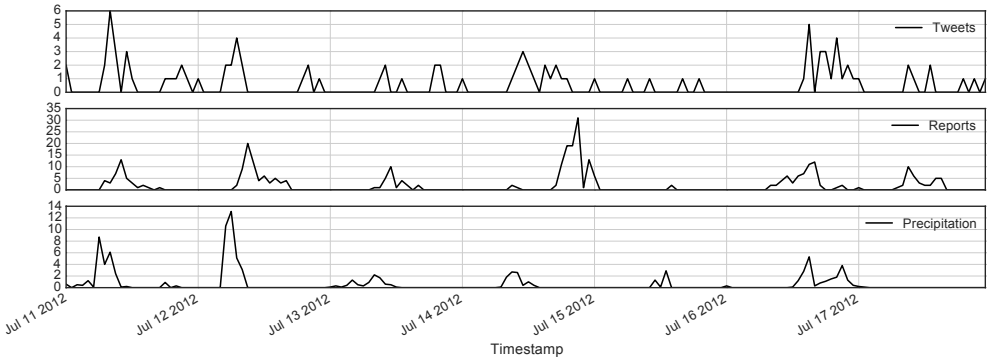


Figure 6.9: The temporal (hourly) distribution of tweets, reports and precipitation in Amsterdam

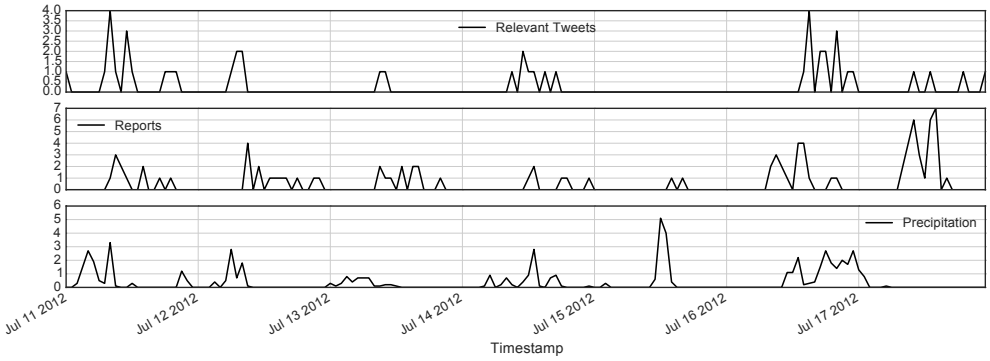


Figure 6.10: The temporal (hourly) distribution of tweets, reports and precipitation in Rotterdam

observation may imply that the Twitter users are more focus on the heavy rain, as the correlation between the reports and the tweets are in the middle between themselves and the precipitation in both cities. Respectively, they are 0.6883 for Amsterdam and 0.3178[16] for Rotterdam.

We further investigate the spatial correlations between the tweet messages concerning water problems, water damage reports and the resident population

[16]Based on the tweets and the reports on the same days

density.

Figure 6.11 and Figure 6.12 show the spatial distributions of the reports and the tweets during the incidents.

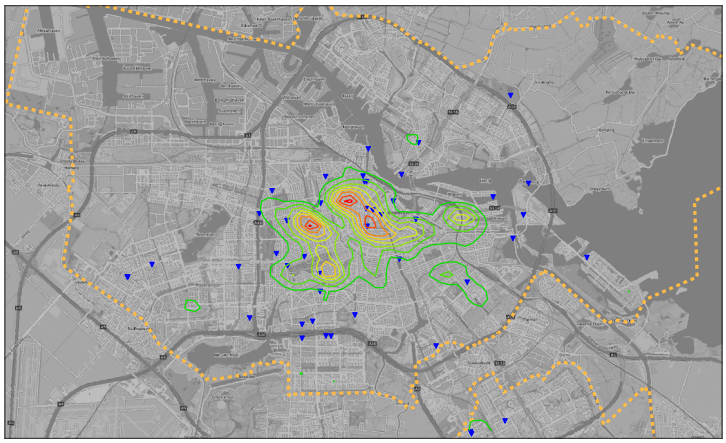


Figure 6.11: The spatial distribution of tweets and the reports in Amsterdam during Jul 11–17, 2012

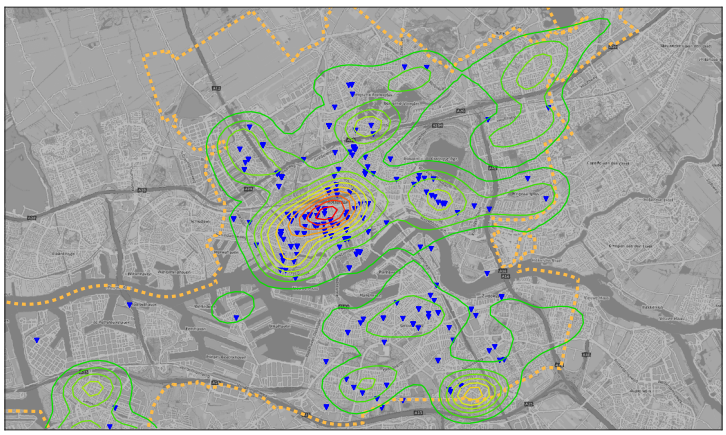


Figure 6.12: The spatial distribution of tweets and the reports in Rotterdam during Oct 11–17, 2013

In both figures, contours are used for depicting the density of reports while the triangle symbols mark the origins of relevant tweets. The two cities have different patterns. In Amsterdam, the tweets are more widely spread than the reports and very few tweets showed up in the most severely affected areas according to the reports. The correlation between tweet locations and report locations is 0.4464. On the other hand, in Rotterdam, the tweets and the reports collocate more closely, *i.e.*, their correlation is 0.6731. In both cities, the spatial correlation between the tweets and the population (0.2649 for Amsterdam and 0.3084 for Rotterdam) are lower than that of the reports and the population (0.4064 for Amsterdam and 0.3613 for Rotterdam). This suggests that for the selected incidents, the tweets we collected are more closely related to reports than to the population density.

Some of the tweets (71%) related to water problem are from the accounts held by the municipality which are dedicated to broadcasting emergencies. These tweets (referred to as *auto tweets*) are also from official channels rather than social channels. So we further investigate the tweets from normal user accounts (refferred to as *non-auto tweets*) by manually labelling the collected tweets. As shown in Figure 6.13, though a few auto tweets are removed from the dense area of reports, the general distribution is close to the reports. In numbers, the correlations between non-auto tweets and the reports (0.6129) and the population (0.3179) do not vary much from that of the general relevant tweets.

The overall findings support the idea that signals from social media may be suited for predictions related to water management issues. As the weather information has relatively course-grained location information (one weather station per city in our case), social media may serve as a complementary source of information to water managers in the municipality.

6.5 Conclusion

Water management is an essential problem for modern cities, *e.g.*, Amsterdam and Rotterdam in the Netherlands. In this chapter we present a case study in both cities analysing the registered records about water damages as well as the messages from Twitter regarding strong precipitation. We find tweets regarding specific incidents of strong precipitation which shows the possibility of using social media as a source of information regarding infrastructure damage

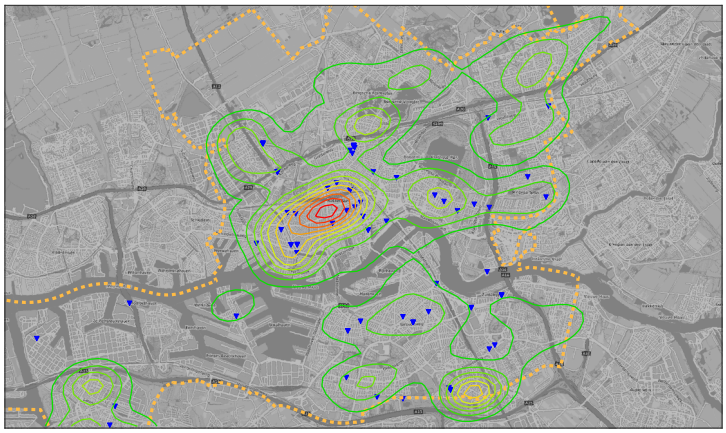


Figure 6.13: The spatial distribution of non-auto tweets and the reports in Rotterdam during Oct 11–17, 2013

management in public space (RQ4a).

The tweets response to the precipitation relatively sooner than the official damage reports which suggests that social media may have advantages in real-time analysis (RQ4b). Additionally, we found that the two information sources of water issues are more closely correlated than either of them with the population density. That is the posts of water issues from social media channel are not severely biased towards population density in space. This shows another evidence that it is possible to use social media to capture water issues in cities. However, we need more effective methods to pick up the signal regarding water problem from social media, as the sparsity of the collected data may lead to bias.

Based on the analysis shown in this chapter, we may devise automated monitoring algorithms for water damage or any other types of defects to the public infrastructures of cities. The gathered information from such automated monitoring systems can be used to improve the management of public space in cities. It may also support emergency management in extreme weather incidents, *i.e.*, prioritizing tasks that may have vast effects and/or predict the propagation of certain problems identified in historical data.

Chapter 7

Social Media Workbench

For scientists studying social media, it is important to get themselves familiarized with the data they have. However, such a process is generally effort consuming and tends to involve ad hoc data management tasks such as creating temporary files for aggregated or filtered data and data visualization. These tasks are effort consuming if researchers use the standard data scientists' toolkit. In this chapter, we present a tool[1] for interactively exploring dataset, especially datasets from social media, with essential ability of customization, extension, *etc.*

7.1 Introduction

For more than a decade, social media services have experienced massive growth rates and are now in regular use by a majority of the developed countries' population. Looking for new ways to understand human behaviour, various disciplines in the science community have turned their attention to analysing the resulting data. For example, Twitter, as one of the most popular online

This work was published as "Interactive Summarization of Social Media" by W. Li, C. Eickhoff and A. de Vries in Proceeding of Fifth Information Interaction in Context Symposium - IIiX'14

[1] https://github.com/spacelis/portraitist2

social media platforms boasts more than 200 million active users producing half a billion tweets (short messages, photos, web links, *etc.*) per day in 2013 [149]. The service inspired a wealth of interesting studies, such as detecting earthquake-related events [131], investigating how people influence each other on social media [12] and predicting users' mobility [87]. Many of these quantitative, data-driven studies are enabled by the high degree of diversity, coverage and scale at which information is available on Twitter. Qualitative approaches, however, may regard these exact properties as obstacles. They often involve the exploration of a set of carefully chosen, focused samples. Under this setting, how should one select the right test subjects among millions of users? How should the available data be adequately partitioned? And, finally, how should the research insights be abstracted and communicated to third parties? The numerous individuals and research groups around the globe that rely on samples from, *e.g.*, Twitter and YouTube, as a scientific resource have come up with their very own, individual solutions to these questions. In this chapter, we briefly describe an extensible open-source workbench to facilitate exploration of data samples in an interactive manner. The workbench provides researchers, data architects and users with an easy handle on understanding data collected from Twitter through private crawls; encouraging reuse of code within the research community.

7.2 Related Tools

In general, three types of tools each of which cover a part of the functions we try to achieve in our Social Media Workbench. The most closely related tools are those visualizing a set of social media data to reveal certain properties of social media which can be used for helping users discover the change of trends or emergent events. Another type of tools are the general purpose charting Web Apps which serve as the online equivalence of spreadsheet chart makers. The third type includes various charting libraries available online which simplify the process of chart making for software developers.

7.2.1 Social Media Visualization

The great popularity of online social media resources among researchers calls for a diverse arsenal of tools for data organization and visualization. Marcus et al. [103] proposed a system called *TwitInfo* for analysing and visualizing the shifting sentiment of Twitter streams. It shows the distribution of tweets

over time, their sentiment categories, a map for geotagged tweets and a list of relevant tweets given a query. *Tweet Sentiment Visualization App*[2] visualizes tweets alongside a sentiment chart for classified tweets, maps of geographic spread and word clouds. *Sentiment140*[3] also shows sentiment classification of tweets, using a number of pie and bar charts. *Socialmention*[4] offers sentiment visualization including social influence indicators. The commercial service *UberVU*[5] provides a wide range of analysis tools for brand management and marketing efforts. Besides those sentiment-centric visualization tools, many other systems exist for social media visualization. *Livehoods*[6] is an interactive map of user mobility based on check-ins from Foursquare in eight selected metropolitan areas. *Vizify*[7] is a personal profiling assistant that can automatically generate infographics showing aspects such as favourite topics, living locations, connectivity, etc. *TwiqsNL*[8] is a Hadoop-based search engine for Dutch tweets collected from Twitter's Streaming API, which depicts tweets matching query filters like keywords or hash-tags in distributions over time, locations and user categories, as well as the textual content in a word cloud. *Crisees* [104] is proposed for monitoring realtime social events on Twitter, especially for crisis discovery. It is composed of a list view of tweets related to a queried event, a map showing the whereabouts of those tweets and a list of related media. *Eddi* [24] is a Twitter client optimized towards better tweet reading experience. It groups tweets into topics and showing them along with a trending timeline and a word cloud.

Each of these tools is tailored for their specific purposes and they are neither open source nor easy to customize for other applications.

7.2.2 Online Chart Generating Tools

There is a wide range of online chart generation solutions that accept arbitrary data streams in standardized formats. Some are Web-based versions of spread-

[2] http://www.csc.ncsu.edu/faculty/healey/tweet_viz/tweet_app
[3] http://www.sentiment140.com
[4] http://www.socialmention.com
[5] https://www.ubervu.com
[6] http://livehoods.net
[7] http://vizify.com
[8] http://twiqs.nl

sheet applications, e.g., *iCharts*[9], *Chart Maker*[10]. Some more sophisticated ones offer greater control over design and layout and facilitate generation of posters and infographics. Popular examples include *Infogra.me*[11], *Piktochart*[12] and *VANNGAGE*[13]. These tools only provide limited means of importing data, and often lack the ability for automatic data aggregation. Users have to manually input data points in each chart and their interactivity is limited.

Many Eyes[14] and *StatPlanet*[15] are two online applications that are similar to our social media workbench. They have both been designed for interactive data visualization. Many Eyes is hosted by IBM as a Java applet-based Web application. It supports 17 types of charts and endorses users for sharing their dataset and inferred charts. The main drawback of the application is that charts cannot be connected for interactive operation, so interactions on one chart do not propagate to other display elements. StatPlanet supports such propagation of interaction on a set of charts, but the Flash based interface is limited to geographical visualization and not customizable. Additionally, both of them require users to manually input data and do not support retrieving data from third party websites.

7.2.3 JavaScript Libraries

Better flexibility can be achieved by programming a tool with existing JavaScript charting libraries, such as *D3.js*[16]. However, existing libraries vary in the required programming effort and the power of data presentation (*e.g.*, colour themes, interactivity). We measure programming effort in four dimensions, namely, the number of types of built-in charts, the related technology, the support of Domain Specific Language based (DSL-based) templates and the interactivity. The built-in charts can save a lot of time and effort if they can fit into the use case. The more types of built-in charts a library supports, the more likely the charts one needs has been already implemented. The related technologies indicate what knowledge is required if one wants to program a

[9] http://www.icharts.net
[10] http://almaer.com/chartmaker
[11] http://infogra.me
[12] http://piktochart.com
[13] http://vanngage.co
[14] http://www-958.ibm.com/software/data/cognos/manyeyes
[15] http://www.statsilk.com
[16] http://d3js.org

tool based on the library. The support of DSL-based templates can reduce the programming effort by allowing one to use simple language to compose a tool rather-than learning a general purposed programming language. The power of a library is measured by the interactivity supported by the charts it makes. It renders how much more a library can provide for data exploration besides plotting a static chart, *e.g.*, zooming into a part of the chart. An ideal library should have as many types of built-in charts as possible, rely on a small set of technologies, support DSL-based templates and provide as much interactivity as possible.

Table 7.1 lists popular charting libraries for Web development and their strengths and limitations in the 4 dimensions. As shown, *DC.js*[17] is a good candidate for researchers to build an exploration tool with. But still it requires one to learn JavaScript to use it in a data exploration tool.

Thus, in this chapter, we present our workbench which is built on top of *DC.js* and we design and integrate a DSL-based template system to simplify tool building process. Our workbench requires only knowledge about the simple DSL we designed for the chart layouts on a user interface.

7.3 System Details

The rest of this Chapter details the Social Media Workbench, a tool developed in response to the shortcomings of the three types of approaches discussed above. Our proposed workbench is designed and implemented towards a system that can help users compose comprehensive interfaces containing interactively connected charts of different types. To this end, it provides a simple domain specific language (DSL) for composing charts so that users can focus more on what types of charts are suitable for their tasks rather than how to program them. Easy modular expansion, embedding and modification are also considered in the first place and a small set of APIs are designed for integrating more charts into the system. For example, we integrate a word cloud drawing library *wordcloud2.js*[18] and an interactive map library *GMaps.js*[19] for plotting over Google Maps. All these charting modules are encapsulated as *directives*

[17] `https://github.com/dc-js/dc.js`
[18] `http://timdream.org/wordcloud2.js`
[19] `http://hpneo.github.io/gmaps`

Table 7.1: Comparison of existing libraries

Name	Built-in Charts	DSL	Programming effort	Interactivity
D3.js	–	No	JS + SVG + CSS	No
DC.js	++	No	JS	Cross Charts, Drilling Down
Chart.js	++	No	JS	Zooming
amCharts	+++	No	JS	Zooming
Google Chart Tools	++	No	JS	Data Visibility
HiCharts	+++	No	JS	Data Visibility
YUI Charts	+++	No	JS	Data Visibility
plot.ly	+++	No	JS or Python or Matlab	Zooming, Data Visibility
Angular-Chartjs	++	Yes	TAG + JS	No

enabled by *AngularJS*[20], which can then be used as a DSL for composing a workbench for exploring social media dataset. We would like to stress that the proposed system is not "yet another library for online chart generation". It is designed for facilitating researchers to efficiently build a prototype interface for aggregating and analysing, for example, Twitter data.

7.3.1 Features

The key features of the proposed Social Media Workbench are:

- Easy composition of new chart layout with built-in directive tags.
- It can be used as a standalone application as well as embedded in another interface.
- It is possible to load data from local storage or 3rd party APIs on the fly.
- Built-in pie, line, bar charts, word clouds and maps.
- Drill down operations for data via charts.
- APIs for extending and embedding the workbench into another Web application.
- Source code available under MIT License.

Figure 7.2 shows an example interface of the Social Media Workbench, summarizing a user's activity in terms of a number of pie charts, bar charts for displaying his/her check-in time lines as well as categorical and geo-spatial spread on an interactive map. All charts in the interface are dynamic; interacting with any of the charts (e.g., clicking on a slice in a pie chart) triggers immediate updates of all other charts (e.g., focusing on the selected share of data). The chart layout can be easily customized through template editing. As an example (see Listing 7.1), a tweet word cloud could be inserted by simply adding a single tag.

Listing 7.1: A example snippet of Social Media Workbench template

```
...
```

[20] http://angularjs.org

```
  <tagcloud name="MyTags" id="tagcloud" data-dimension="text"></
      tagcloud>
...
```

7.3.2 Architecture

The workbench is built as a Web application to make optimal use of state-of-the-art technology and allows for flexible platform-independent deployment either locally or on any popular *Platform as a Service* (PaaS). As shown in Figure 7.1, the workbench consists of a front-end and a back-end.

Computation of statistics and chart plotting are accomplished on the JavaScript front end. From the data stream provided by data APIs, the front end will extract properties of each item in the stream and send them to corresponding charts. When the user selects, for example, a particular category of places, a filter will be applied to the stream of items where only the matched items will account for the update and all other charts will be synchronized to the new selection criteria.

The back end is a dedicated server for hosting data APIs, providing easy handles on data from Google Datastore or other data APIs online (e.g., Twitter). For example, users can feed a list of tweet IDs or user IDs or search queries for the workbench to automatically obtain all the corresponding tweets.

The front end and back end are decoupled by simple REST APIs, through which JSON formatted data is communicated between them. The resulting workbench can either run as a standalone application (with data served by built-in back end running on Python) or be embedded in another application where the front end pulls JSON formatted stream of data from that application. The detailed usage can be found in Appendix A.

Interface Customization

The front end, for the purpose of analyzing existing data from an API access point, can be used for tailoring the interface to one's own needs. The interface is merely a HTML page with embedded directives (DSL). Each directive represents a chart that will be displayed when the page is loaded in a modern browser (*e.g.*, Google Chrome[21]). The datasource directive is a mandatory

[21] https://www.google.com/intl/en/chrome/browser

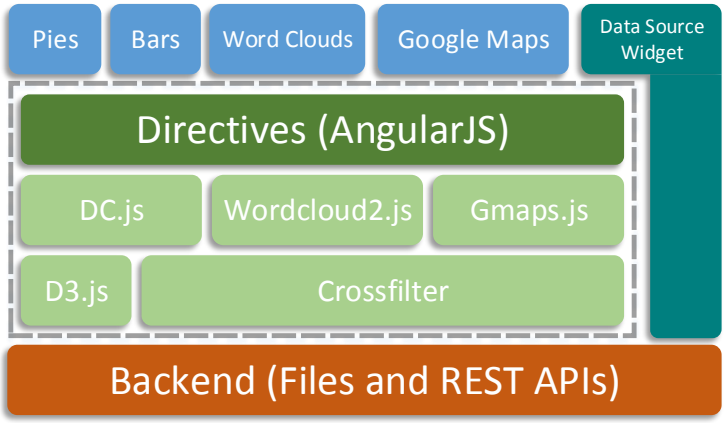


Figure 7.1: The Architecture of Social Media Workbench

component for the interface, as it stores the data coming from the back-end and serves as a data provider to all other charts on the interface. For example, one can customize the interface with a bar chart, a pie chart and a datasource. When a user loads the interface, the datasource directive will ask the back-end for data which is configured when composing the interface. When the data is ready, the data will be pulled by all the chart components for updating their chart.

All the charting and datasource components on the interface as well as other underlying libraries are managed by *RequireJS*[22]. *RequireJS* is a library for managing module loading and dependency in a JavaScript application. Any extension can be easily integrated to the system by adding a dependency declaration in the configuration file of *RequireJS*.

Backend

The back end is decoupled from the front end. The provided back end is a default minimal server suited for simple data crawling from Twitter API and handling the storage of the tweets retrieved. Users can replace it with any Web framework. For example, users can load the interface from their local file system and configure the datasource to pull a local CSV or JSON file as

[22] `http://requirejs.org`

alternative data stream. For more complex cases, users can setup their own data API server with each endpoint (a URL) serving a subset of the data in their database. In general, the front end only requires a URL that points to a REST service that provides data in CSV or JSON format.

7.4 Example Systems

In this section we demonstrate our workbench in two different projects to show its flexibility and ease of customization.

7.4.1 Geo-Spatial Summaries

The workbench was originally used for assessing users' expertise towards *Points of Interest* (POI) discussed in Chapter 5 [86]. The assessment task is a complex process in which assessors are required to look deeply into users' check-in profiles, each of which can be composed of up to thousands of individual geotagged tweets. It would not be effective nor efficient to let assessors go through all available tweets to find out whether the user knows about a given place or a type of places. Instead, we created an interface based on the proposed workbench to show as many dimensions of a profile as possible, without exceeding or cluttering the confines of a single page. The aggregated diagram lets assessors focus on the actual check-ins rather than the content of tweets.

As shown in Figure 7.2, we include in the interface 4 pie charts for showing the distribution of check-ins across different types of locations and a chart of geotagged tweets distributed over time. Additionally, we provide a map to show the spatial distribution of check-ins. Our assessors can use this interface to comfortably inspect different dimensions of a user's check-in profile. For example, they can click on any part of the pie charts to select a type of place. Other charts including the map will immediately show only the check-ins at the selected type of places. After exploring a user's profile, assessors are required to evaluate the user's expertise for each given place or type of places, based on the visualized summarization and assign the scores accordingly at the top of the interface.

We offer a refined and comprehensive interface for the assessors to avoid biased answers caused by lack of information or by obtrusive interfaces. For this project, we include charts and maps in the workbench to display the

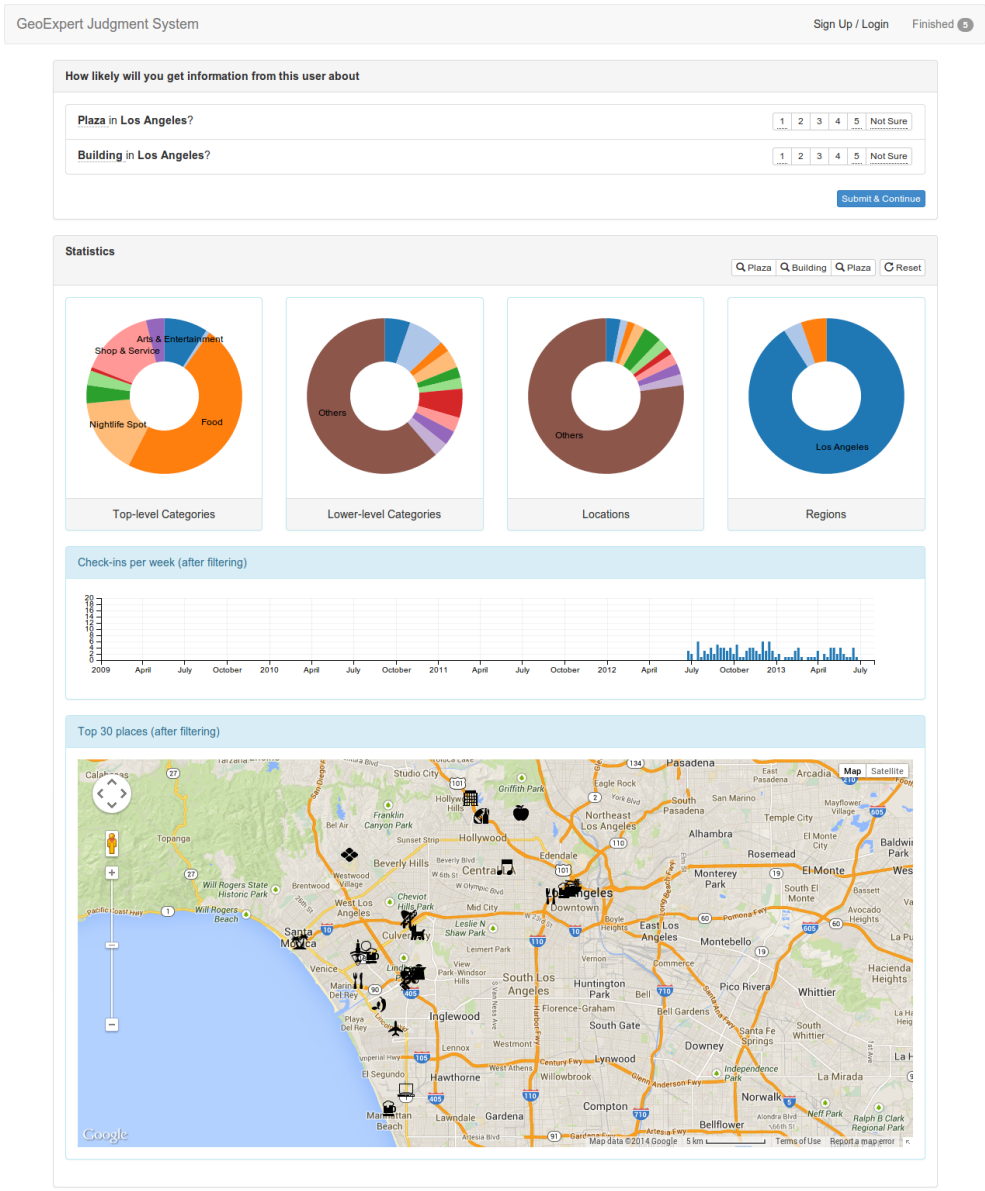


Figure 7.2: The geo-expertise annotation interface

users' profiles in as many dimensions as possible. Moreover, the interactivity of our proposed workbench helps assessors to look into different dimensions and form their own ways of exploring profiles for the tasks. For example, when assessors are evaluating a users' expertise towards a place, they may also want to know users' check-ins at other similar types of places. The selection can be narrowed down to include only the desired type of places with a single click. The category distribution of check-ins and all other charts will be immediately updated to the new selection. In the original study, we encouraged assessors to explore new ways of using the workbench, for example, looking into the check-in distributions only in the past month by selecting check-ins on the bar chart, or investigating check-ins in different cities by clicking on the pie chart for check-in distribution over cities. By giving assessors freedom to explore the data, we hope to minimize the possibility of leading them towards false conclusions.

7.4.2 Complaints Discovery in Social Channel

We used the same workbench for the project discussed in Chapter 6, which leverages social media channels to collect complaints about damages caused by heavy rain in Amsterdam and Rotterdam in the Netherlands. Based on these complaints the municipalities may discover vulnerable locations in urban areas and then try to improve them.

As the first step in the pilot study, we needed a tool for exploring social media for potential textual features characterizing rain damage related tweets. To realize this objective, we composed an interface based on the previously described workbench for summarizing the tweets relevant to a given search query. With this tool, we identified keywords for retrieving tweets regarding heavy rain damages so that we can monitor such keywords on the Twitter Streaming API.

In this application of the Social Media Workbench, we include a bar chart, a word cloud and a structured data table. The bar chart shows the time distribution of relevant tweets, the word cloud contains frequent words in the selected tweets and the data table will show tweets sorted by their relevance towards the filter queries. A thin wrapper over a 3rd party full-text search library (Whoosh[23]) enabled the selection of relevant tweets through a REST

[23] `https://bitbucket.org/mchaput/whoosh/`

API. Alternatively, we could even have used a complete self-contained IR system running in the browser as recently prototyped by Lin [92].

Figure 7.3 shows an example query of *wateroverlast*, the Dutch word for *flood*. As can be seen, the volume of relevant tweets peaks after the 11th of October 2013, the day on which an exceptionally heavy storm hit the city of Rotterdam. We can either zoom into a particular day to check the word distribution of tweets by selecting a specific range on the time line, or select another word as a new query by clicking on that word in the word cloud. With this tool, we refined the keyword set to monitor and collect more tweets about upcoming complaints.

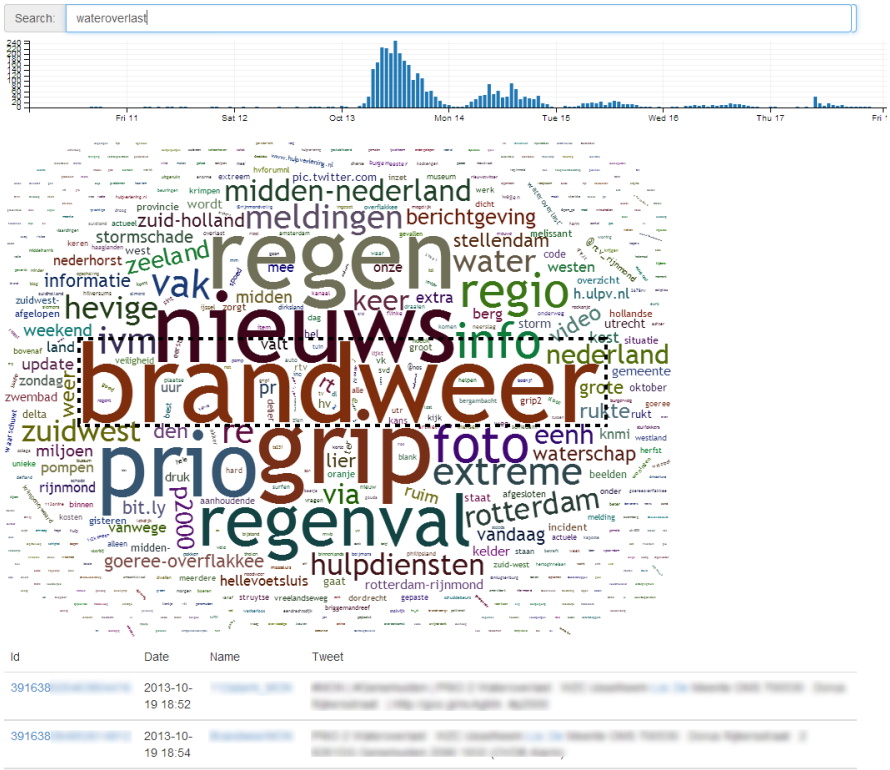


Figure 7.3: Water Complaints profiling tool

7.5 Conclusion

We described the first version of the *Social Media Workbench*, an open-source system that is designed to help researchers explore large social media data sets originating from platforms such as Twitter. The current version of the tool along with the two described use cases focuses mainly on meta information such as geotags that are associated with tweets. The tool is an ongoing engineering effort, and we plan to integrate more content-related aggregation and summarization functions, such as latent semantic analysis or unsupervised clustering functionality based on user defined fields. In this way, we hope to reduce the need for code replication throughout the research community by giving a flexible, easy-to-adapt environment for qualitative research on potentially large-scale datasets.

Part IV

Conclusion

Chapter 8

Conclusion

Social media is a new channel for people to acquire and share information online. The diversity of user contributed information can help both researchers and product developers to better understand users' behaviour. In this dissertation, we have focused on the geographical information presented in geotags (POI-tags) on social media and explored different aspects regarding such information to demonstrate how it can be used for inferring user activities.

8.1 The Answers

At the beginning of this dissertation we identified four research questions, and we explored the solutions in the previous chapters.

8.1.1 Content-based Location Prediction

Our first question regards whether it is possible to predict users' locations based on the content of messages they send. From the experiment, we find it is possible to distinguish locations from each other by the tweets originated from them and predict users' locations based on the content in their tweets. Due to the sparsity of the data collected from Twitter, we proposed to retrieve external evidence for the textual features, *i.e.*, using the text from the Web pages about each location to enrich the models we build for them. The performance of the enriched models shows statistically significant improvement (around 10%

in P@3) over the simple ones when there is only a few tweets. The marginal improvement is due to the divergent vocabulary of the text from web sources. It may relieve the problem if we only include high-quality resources on the Web.

Besides the textual features, we also explored temporal features for the locations, inspired by the fact that places have their own preferred visiting time, *e.g.*, parks during days while bars during nights. As shown in the experiments, the temporal feature can further improve the precision of the prediction but did not improve the performance much. This may relate to the same level of sparsity if we try to build models for less popular locations. It might be useful to smooth the time model from similar locations.

8.1.2 Trail-based Location Prediction

As inspired by the mobile pattern shared by city residents, we looked into the problem of whether it is possible to predict users' future visits based on his/her visiting history at the level of location categories. Since this is the first study on location category prediction, we borrowed ideas from other domain and devised several models based on intuition and observations. We also developed a Collaborative Filtering method based on users' temporal spatial matrices together with a smoothing technique for the problem.

As shown in the experiments, the proposed method successfully improves the performance of predicting users' future locations based on their visiting histories. Specifically, the collaborative filtering based method (CF-K) achieves around 1–5% improvement in MRR scores in predicting users' visits in terms of the top-categories of locations compared to that of Markov Chain models (MC) and 1–7% improvements in predicting visits in terms of lower-level categories of locations. The advantages of the collaborative filtering based methods are that they include longer histories of users' visits and are able to handle the introduced noise. In both experiment settings, the two methods (CF-K and MC) have similar performance on the data from Los Angeles. We suspect it is due to that a large proportion of check-ins in the data are from the international airport in Los Angeles.

8.1.3 Geo-expertise Retrieval

An interesting aspect about the geotags embedded in social media is that it implies users' physical presence at different locations which can be converted into users' knowledge about the locations. Such knowledge is hard to document or digitalize, which necessitates expertise retrieval systems so that users can obtain the knowledge from the experts. Thus we studied how an effective geo-expertise retrieval system can be built. As a pilot study we run surveys to collect people's opinion about the problem and the results confirmed our conjecture about the needs of such systems. As concluded in the survey, three features are considered as the most important models for determining whether a user knows about a location or a type of locations. We build models according to these three criteria based on probabilistic inference. Then an evaluation is carried out for these criteria and also for the methods derived from others' study. According to the annotated data acquired respectively via crowdsourcing and from university students and staff, we found that the proposed methods make better predictions of geo-expertise than the random baselines. Specifically, we found the best methods increase the performance by 106% (WTD+A) in P@1 scores on the data from the CrowdFlower platform while by 274% on the data from the university student/staff. However, the difference between the proposed criteria is much smaller. This may be a result of the open-ended annotation, as assessors might not agree on a single criterion. To further investigate the difference, data from a real running system may provide a better source of evidence.

8.1.4 Monitoring Extreme Weather on Social Media

Besides studies on benefiting individual users online, we also conducted research in helping organizations better understand urban lives. A collaborated project was carried out for studying the reports of water damage caused by storms in social media. As shown in Chapter 6, there were not many Twitter messages about water damage collected during our study. However, the trend and spatial distribution have slight correlation with the official register of water damage reports. The lack of tweets may lead to the low correlation, which may be caused by the different focus of social media.

To improve the quality of the data from the social media channel, it may be useful to raise the awareness of using the channel to report water damage. Another way to collect more evidence from the social media channel is to

improve the methods in discovering water damage reports on Twitter and estimating the locations regarding the reports.

8.1.5 An Interactive Social Media Workbench

In respect to the recurring need for data visualization tools during the course of our study, we developed a Web based interface for visualizing the data collected from the social media. It enabled us to show the data to annotators in the way that they can interactively explore the data. The tool is open source licensed and can be obtained from `https://github.com/spacelis/portraitist2`.

8.1.6 The Privacy

The abundant information that can be derived from geo-spatial data shows promising evidence of helping people in various ways, e.g., recommending places to visit, context-aware searching, and suggesting new social ties. However, such abundant information is also considered as a very important part of users' privacy. As stated in Section 4.1, locations, especially in terms of POIs, usually reflect what kinds of activities users may carry out at different places. The visiting patterns may also reveal other aspects of users' everyday life, *e.g.*, purchasing habits, working and living areas, and level of income. These types of information can be very sensitive and can be exploited for crimes such as identity fraud. The more users engage on those platforms the more personal information they give out and the more likely they may get into trouble when such information is abused.

In the study presented in this dissertation, though the data set collected for the experiments is a small fraction of the entire data ocean accumulated on Twitter, it has already shown the possibility of location inference based on users' check-ins. In Chapter 3, we demonstrated that location information can be inferred from the content of users' tweets. In Chapter 4, we demonstrated that the missing part of users' trails can be recovered based on their visiting histories. The positive results, unfortunately, show the possibility that users can be tracked by their service providers and criminals may exploit the services and look for targets (especially, IR technology is available for such data). This dissertation can be seen as a reminder to both users and social network service providers that the accumulated online personal information is more than just the data itself.

With the dilemma between openness and privacy, researchers are urged to investigate how to prevent the exploitation of users' privacy and how such prevention would affect user experience with various services.

8.2 Future Challenges and Applications

Along the line of the study presented in this dissertation, there are many possible future routes to follow. In the work of Chapter 3 and Chapter 4, users are purely treated as sets of check-ins and other aspects are not considered in the prediction. However, intuitively, other personal information may provide prior knowledge or evidence for estimating users' (future) whereabouts. For example, users put some text on their Twitter page to describe themselves, which may contain the users' hobby, profession, gender, *etc.* The information from this source can be very useful in inferring users' location, but the source can be noisy since there is no constraint on what users can put in the text box. It is worthwhile and challenging to investigate how such information can be integrated to the existing location prediction models.

Though we focused on non-textual information in users' check-in tweets, it would be interesting to investigate how to pick up signals in tweets that may imply a future visit to a location. For example, it would be a strong implication of a future visit when a user tweets about the tickets of his favourite band he just got. Though it appears to be easy for human to learn the connection between the information in the tweet and the future visit, it is hard to generalize such heuristic in the form of algorithms.

The context-aware prediction is also an interesting direction for the problem of location prediction. The situation surrounding a user can provide information about the user's whereabouts and can also be a useful indicator, for example, to deduce the user's mood and feeling which may motivate them to go to certain places. However, the context can vary from application to application as it is usually inferred from other sources of information, such as moving speed, weather condition. It would be challenging to identify the context from the signal collected and integrate the context into the predicting algorithm.

Along the line of research on geo-expertise, an application based on Twitter or Foursquare can be developed so that the needs of geo-expertise and the effectiveness of the system can be studied. As mentioned by several participants in the experiment, trust and common preferences play a very important role in

geo-expertise seeking. Thus evidence from non-geographical features may also help the system achieve better retrieval results by introducing candidates to the users who trust them or share similar interests. It would be challenging to integrate these factors into the probabilistic models we designed in Chapter 5. Besides, some experts may know a lot about the queried locations, but they may be either reluctant or too busy to help the knowledge seekers. It would be interesting to integrate this factor into the model of expertise retrieval since the ultimate goal is to find the right person to answer the question rather than a person knowing a lot but would not share his/her knowledge. Though there are already quite some credit systems in Q&A like platforms, but these credit systems presumably focus on the quality of the answers. It is not clear how effective they are in the setting of geo-expertise retrieval and this needs further study.

Besides the directions along with the work presented in this dissertation, geographical information in social media is also an important information source for other applications. Companies, organizations and government can save a lot of effort if they can make use of the mobility information extracted from social media. For example, users' trails reflected in their geotagged messages on social media can be used to find potential sites for new shops, advertising boards or public facilities. If there is a site that the shoppers (identified via their tweet streams) of a brand of supermarkets visit a lot, it might be a good place for displaying advertisement or even opening a new branch. However, it might require careful inspection on different types of visitors, for example, a user that has many check-ins at a supermarket can just be a staff member in that supermarket.

Bibliography

[1] F. Abel, C. Hauff, G.-J. Houben, R. Stronkman and K. Tao. Twitcident: Fighting Fire with Information from SocialWeb Streams. In *Proceedings of the 21st international conference companion on World Wide Web - WWW '12 Companion*, 305–308, 2012.

[2] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, **17**, 734, 2005.

[3] N. Agarwal, H. Liu, L. Tang and P. S. Yu. Identifying the influential bloggers in a community. In *Proceedings of the international conference on Web search and web data mining - WSDM '08*, 207, 2008.

[4] D. Ahlers. Applying Geographic Information Retrieval. *Datenbank-Spektrum*, **14**, 39, 2014.

[5] E. Allen. Update on the Twitter Archive at the Library of Congress, 2013. `http://blogs.loc.gov/loc/2013/01/update-on-the-twitter-archive-at-the-library-of-congress/`. Last Accessed: 2013-01-21.

[6] O. Alonso and M. Lease. Crowdsourcing for information retrieval. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information - SIGIR '11*, 1299, 2011.

[7] O. Alonso and S. Mizzaro. Using crowdsourcing for TREC relevance assessment. *Information Processing & Management*, **48**, 1053, 2012.

[8] E. Amitay, N. Har'El, R. Sivan and A. Soffer. Web-a-Where: Geotagging-Web Content. In *Proceedings of the 27th annual international conference on Research and development in information retrieval - SIGIR '04*, 273, 2004.

[9] A. Arampatzis, M. van Kreveld, I. Reinbacher, C. B. Jones, S. Vaid, P. Clough, H. Joho and M. Sanderson. Web-based delineation of imprecise regions. *Computers, Environment and Urban Systems*, **30**, 436, 2006.

[10] D. Ashbrook and T. Starner. Learning significant locations and predicting user movement with GPS. In *Proceedings of the Sixth International Symposium on Wearable Computers*, 101–108, 2002.

[11] L. Backstrom, E. Sun and C. Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on World wide web - WWW '10*, 61–70, 2010.

[12] E. Bakshy, J. M. Hofman, W. A. Mason and D. J. Watts. Everyone's an Influencer: Quantifying Influence on Twitter Eytan. In *Proceedings of the fourth ACM international conference on Web search and data mining - WSDM '11*, 65–74, 2011.

[13] K. Balog, L. Azzopardi and M. de Rijke. Formal models for expert finding in enterprise corpora. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '06*, 43–50, 2006.

[14] K. Balog, L. Azzopardi and M. de Rijke. A language modeling framework for expert finding. *Information Processing & Management*, **45**, 1, 2009.

[15] K. Balog and M. de Rijke. Non-local evidence for expert finding. In *Proceeding of the 17th ACM conference on Information and knowledge mining - CIKM '08*, 489, 2008.

[16] K. Balog, Y. Fang, M. de Rijke, P. Serdyukov and L. Si. Expertise Retrieval. *Foundations and Trends in Information Retrieval*, **6**, 127, 2012.

[17] S. S. Banerjee and R. R. Dholakia. Mobile Advertising: Does Location Based Advertising Work? *International Journal of Mobile Marketing*, 2008.

[18] J. Bao, Y. Zheng and M. F. Mokbel. Location-based and preference-aware recommendation using sparse geo-social networking data. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems - SIGSPATIAL '12*, 199, 2012.

[19] R. Bar-Haim, E. Dinur, R. Feldman, M. Fresko and G. Goldstein. Identifying and following expert investors in stock microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing - EMNLP '11*, 1310–1319, 2011.

[20] S. B. Barnes. A privacy paradox: Social networking in the United States. *First Monday*, **11**, 2006.

[21] D. F. Bauer. Constructing Confidence Sets Using Rank Statistics. *Journal of the American Statistical Association*, **67**, 687, 1972.

[22] P. N. Bennett, F. Radlinski, R. W. White and E. Yilmaz. Inferring and using location metadata to personalize web search. In *SIGIR '11*, 135, 2011.

[23] B. Berjani and T. Strufe. A recommendation system for spots in location-based online social networks. In *Proceedings of the 4th Workshop on Social Network Systems - SNS '11*, 1–6, 2011.

[24] M. S. Bernstein, B. Suh, L. Hong, J. Chen, S. Kairam and E. H. Chi. Eddi: Interactive Topic-based Browsing of Social Status Streams. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology - UIST '10*, 303–312, 2010.

[25] A. Bhattacharya and S. K. Das. LeZi-Update: An Information-Theoretic Approach to Track Mobile Users in PCS Networks. In *Proceedings of the 5th annual ACM/IEEE international conference on Mobile computing and networking - MobiCom '99*, 1–12, 1999.

[26] D. M. Boyd and N. B. Ellison. Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication*, **13**, 210, 2007.

[27] O. Buyukokkten, J. Cho, H. Garcia-Molina, L. Gravano and N. Shivakumar. Exploiting Geographical Location Information of Web Pages. In *Proceedings of the ACM SIGMOD Workshop on the Web and Databases - WebDB'99*, 91–96, 1999.

[28] C. S. Campbell, P. P. Maglio, A. Cozzi and B. Dom. Expertise identification using email communications. In *Proceedings of the twelfth international conference on Information and knowledge management - CIKM '03*, 528–531, 2003.

[29] N. Cardoso and M. J. Silva. A GIR architecture with semantic-flavored query reformulation. In *Proceedings of the 6th Workshop on Geographic Information Retrieval - GIR '10*, 1, 2010.

[30] A.-J. Cheng, Y.-Y. Chen, Y.-T. Huang, W. H. Hsu and H.-Y. M. Liao. Personalized travel recommendation by mining people attributes from community-contributed photos. In *Proceedings of the 19th ACM international conference on Multimedia - MM '11*, 83, 2011.

[31] Z. Cheng, J. Caverlee and K. Lee. You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users. In *Proceedings of the 19th ACM international conference on Information and knowledge management - CIKM '10*, 759–768, 2010.

[32] Z. Cheng, J. Caverlee, K. Lee and D. Z. Sui. Exploring Millions of Footprints in Location Sharing Services. In *International AAAI Conference on Weblogs and Social Media, ICWSM '11*, volume 2010, 81–88, 2011.

[33] Z. Cheng, J. Caverlee, H. Barthwal and V. Bachani. Who is the barbecue king of texas? In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval - SIGIR '14*, 335–344, 2014.

[34] E. Cho, S. A. Myers and J. Leskovec. Friendship and Mobility: User Movement in Location-Based Social Networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '11*, 1082, 2011.

[35] M. Clements, P. Serdyukov, A. P. de Vries and M. J. Reinders. Using flickr geotags to predict user travel behaviour. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '10*, 851, 2010.

[36] M. Clements, P. Serdyukov, A. P. de Vries and M. J. T. Reinders. Personalised Travel Recommendation based on Location Co-occurrence. *arXiv*, 2011.

[37] D. J. Crandall, L. Backstrom, D. Huttenlocher and J. Kleinberg. Mapping the world's photos. In *Proceedings of the 18th international conference on World wide web - WWW '09*, 761, 2009.

[38] J. Cranshaw, R. Schwartz, J. Hong and N. Sadeh. The livehoods project: Utilizing social media to understand the dynamics of a city. In *Internation Conference on Weblogs and Social Media*, 2012.

[39] N. Craswell, D. Hawking, A.-M. Vercoustre and P. Wilkins. P@NOPTIC Expert: Searching for Experts Not Just For Documents. In *Ausweb Poster Proceedings*, 2001.

[40] W. B. Croft. Information Retrieval Based on Statistical Language Models. *Foundations of Intelligent Systems*, **1932**, 1, 2010.

[41] M. Daoud and J. X. Huang. Modeling geographic, temporal, and proximity contexts for improving geotemporal search. *Journal of the American Society for Information Science and Technology*, **64**, 190, 2013.

[42] S. Dhar and U. Varshney. Challenges and business models for mobile location-based services and advertising. *Communications of the ACM*, **54**, 121, 2011.

[43] J. Ding, L. Gravano and N. Shivakumar. Computing Geographical Scopes of Web Resources. In *Proceedings of the 26th International Conference on Very Large Data Bases*, 545–556, 2000.

[44] M. Duggan and A. Smith. Social Media Update 2013. Technical report, Pew Research Center, 2014.

[45] C. Eickhoff and A. P. de Vries. Increasing cheat robustness of crowdsourcing tasks. *Information Retrieval*, **16**, 121, 2012.

[46] T. Emrich, H.-P. Kriegel, N. Mamoulis, M. Renz and A. Züfle. Indexing uncertain spatio-temporal data. In *Proceedings of the 21st ACM international conference on Information and knowledge management - CIKM '12*, 395, 2012.

[47] M. Eskevich, R. Aly, D. N. Racca, R. Ordelman, S. Chen and G. J. F. Jones. The Search and Hyperlinking Task at MediaEval 2014. In *CEUR Workshop Proceedings: Working Notes Proceedings of the MediaEval 2014 Workshop*, volume 1263, 2014.

[48] H. Fang and C. Zhai. Probabilistic Models for Expert Finding. In G. Amati, C. Carpineto and G. Romano (eds.), *Advances in Information Retrieval (ECIR '07)*, *Lecture Notes in Computer Science*, volume 4425, 418–430, 2007.

[49] Y. Fang, L. Si and A. P. Mathur. Discriminative models of integrating document evidence and document-candidate associations for expert search. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '10*, 683, 2010.

[50] G. Ference, M. Ye and W.-C. Lee. Location recommendation for out-of-town users in location-based social networks. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management - CIKM '13*, 721–726, 2013.

[51] C. Fink, C. Piatko, J. Mayfield, D. Chou, T. Finin and J. Martineau. The Geolocation of Web Logs from Textual Clues. In *2009 International Conference on Computational Science and Engineering*, volume 4, 1088–1092, 2009.

[52] D. Flatow, M. Naaman, K. E. Xie, Y. Volkovich and Y. Kanza. On the Accuracy of Hyper-local Geotagging of Social Media Content. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15*, 127–136, 2015.

[53] J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, **76**, 378, 1971.

[54] Foursquare. Personalized search categories, a faster app, and a map made just for you, 2012. `http://blog.foursquare.com/2012/09/26/personalized-search-categories-a-faster-app-and-a-map-made-just-for-you-and-iphone-5-ready/`. Last Accessed: 2013-02-08.

[55] S. Gaitan, L. Calderoni, P. Palmieri, M.-c. ten Veldhuis, D. Maio and M. B. van Riemsdijk. From Sensing to Action: Quick and Reliable Access to Information in Cities Vulnerable to Heavy Rain. *IEEE Sensors Journal*, 2014.

[56] H. Gao, J. Tang and H. Liu. gSCorr: Modeling Geo-Social Correlations for New Check-ins on Location-Based Social Networks. In *Proceedings of the 21st ACM international conference on Information and knowledge management - CIKM '12*, 1582, 2012.

[57] F. Giannotti, M. Nanni, D. Pedreschi, F. Pinelli, C. Renso, S. Rinzivillo and R. Trasarti. Unveiling the complexity of human mobility by querying and mining massive trajectory data. *The VLDB Journal*, **20**, 695, 2011.

[58] E. Gilbert and K. Karahalios. Predicting tie strength with social media. In *Proceedings of the 27th international conference on Human factors in computing systems - CHI 09*, 211, 2009.

[59] M. C. González, C. A. Hidalgo and A.-L. Barabási. Understanding individual human mobility patterns. *Nature*, **453**, 779, 2008.

[60] P. A. Grabowicz, J. J. Ramasco, B. Gonçalves and V. M. Eguíluz. Entangling mobility and interactions in social media. *PloS one*, **9**, e92196, 2014.

[61] C. Hauff, B. Thomee and M. Trevisiol. Working Notes for the Placing Task at MediaEval 2013. In *CEUR Workshop Proceedings: Working Notes Proceedings of the MediaEval 2013 Workshop*, volume 1043, 2013.

[62] J. Hays and A. A. Efros. Im2gps: estimating geographic information from a single image. In *Proceedings of 2008 IEEE Conference on Computer Vision and Pattern Recognition*, 1–8, 2008.

[63] R. Heatherly, M. Kantarcioglu and B. Thuraisingham. Preventing Private Information Inference Attacks on Social Networks. *IEEE Transactions on Knowledge and Data Engineering*, **25**, 1849, 2013.

[64] B. Hecht, L. Hong, B. Suh and E. H. Chi. Tweets from Justin Bieber's Heart : The Dynamics of the "Location" Field in User Profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '11*, 237–246, 2011.

[65] E. Herder and P. Siehndel. Daily and Weekly Patterns in Human Mobility. In *CEUR Workshop Proceedings*, 2012.

[66] D. Horowitz and S. D. Kamvar. The anatomy of a large-scale social search engine. In *Proceedings of the 19th international conference on World wide web - WWW '10*, 431, 2010.

[67] L. Hu, A. Sun and Y. Liu. Your neighbors affect your ratings. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval - SIGIR '14*, 345–354, 2014.

[68] A. Java, X. Song, T. Finin and B. Tseng. WhyWe Twitter: Understanding Microblogging Usage and Communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis - WebKDD/SNA-KDD '07*, 56–65, 2007.

[69] A. Jeffries. The man behind Flickr on making the service 'awesome again', 2013. `http://www.theverge.com/2013/3/20/4121574/flickr-chief-markus-spiering-talks-photos-and-marissa-mayer`. Last Accessed: 2014-10-29.

[70] M. Johanson. How Burglars Use Facebook To Target Vacationing Homeowners, 2013. `http://www.ibtimes.com/how-burglars-use-facebook-target-vacationing-homeowners-1341325`. Last Accessed: 2014-11-19.

[71] A. M. Kaplan and M. Haenlein. Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, **53**, 59, 2010.

[72] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, **46**, 604, 1999.

[73] J. Kulshrestha, F. Kooti, A. Nikravesh, P. Gummadi and K. P. Gummadi. Geographic Dissection of the Twitter Network. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2012.

[74] T. Kurashima, T. Iwata, G. Irie and K. Fujimura. Travel route recommendation using geotags in photo sharing sites. In *Proceedings of the 19th ACM international conference on Information and knowledge management - CIKM '10*, 579, 2010.

[75] T. Kurashima, T. Iwata, T. Hoshide, N. Takaya and K. Fujimura. Geo Topic Model: Joint Modeling of User's Activity Area and Interests for

Location Recommendation. In *Proceedings of the sixth ACM international conference on Web search and data mining - WSDM '13*, 375, 2013.

[76] H. Kwak, C. Lee, H. Park and S. Moon. What is Twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web - WWW '10*, 591, 2010.

[77] M. Larson, M. Soleymani and P. Serdyukov. Automatic tagging and geotagging in video collections and communities. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval - ICMR '11*, 1–8, 2011.

[78] J. Lehmann and C. Castillo. Finding news curators in twitter. In *Proceedings of the 22nd international conference on World Wide Web companion*, 863–870, 2013.

[79] J. L. Leidner, G. Sinclair and B. Webber. Grounding spatial named entities for information extraction and question answering. In *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references*, volume 1, 31–38, 2003.

[80] K. W.-T. Leung, D. L. Lee and W.-C. Lee. CLR: A Collaborative Location Recommendation Framework based on Co-Clustering. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information - SIGIR '11*, 305, 2011.

[81] A. Leuski and V. Lavrenko. Tracking dragon-hunters with language models. In *Proceedings of the 15th ACM international conference on Information and knowledge management - CIKM '06*, 698, 2006.

[82] C. Li and A. Sun. Fine-grained location extraction from tweets with temporal awareness. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval - SIGIR '14*, 43–52, 2014.

[83] H. Li, R. K. Srihari, C. Niu and W. Li. Location normalization for information extraction. In *Proceedings of the 19th international conference on Computational linguistics* -, volume 1, 1–7, 2002.

[84] H. Li, R. K. Srihari, C. Niu and W. Li. InfoXtract location normalization. In *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references* -, volume 1, 39–44, 2003.

[85] N. Li and G. Chen. Analysis of a Location-Based Social Network. In *2009 International Conference on Computational Science and Engineering*, volume 4, 263–270, 2009.

[86] W. Li, C. Eickhoff and A. de Vries. Geo-spatial Domain Expertise in Microblogs. In M. de Rijke, T. Kenter, A. de Vries, C. Zhai, F. de Jong, K. Radinsky and K. Hofmann (eds.), *Advances in Information Retrieval - ECIR '14*, *Lecture Notes in Computer Science*, volume 8416, 487–492. Springer International Publishing, 2014.

[87] W. Li, C. Eickhoff and A. P. de Vries. Want a coffee? Predicting Users' Trails. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '12*, 1171–1172, 2012.

[88] W. Li, P. Serdyukov, A. P. de Vries, C. Eickhoff and M. Larson. The where in the tweet. In *Proceedings of the 20th ACM international conference on Information and knowledge management - CIKM '11*, 2473–2476, 2011.

[89] X. Li, M. Larson and A. Hanjalic. Geo-visual ranking for location prediction of social images. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval - ICMR '13*, 81, 2013.

[90] X. Li and T.-a. N. Pham. Rank-GeoFM : A Ranking based Geographical Factorization Method for Point of Interest Recommendation. In *Proceedings of the 38th international ACM SIGIR conference on Research & development in information retrieval - SIGIR '15*, 2015.

[91] Q. V. Liao, C. Wagner, P. Pirolli and W.-T. Fu. Understanding experts' and novices' expertise judgment of twitter users. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*, 2461, 2012.

[92] J. Lin. Building a Self-Contained Search Engine in the Browser. In *Proceedings of the 2015 International Conference on Theory of Information Retrieval - ICTIR '15*, 309–312, 2015.

[93] G. Linden, B. Smith and J. York. Amazon.com recommendations: item-to-item collaborative filtering. *IEEE Internet Computing*, **7**, 76, 2003.

[94] X. Liu, W. B. Croft and M. Koll. Finding experts in community-based question-answering services. In *Proceedings of the 14th ACM international conference on Information and knowledge management - CIKM '05*, 315–316, 2005.

[95] X. Liu, Y. Liu, K. Aberer and C. Miao. Personalized point-of-interest recommendation by mining users' preference transition. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management - CIKM '13*, 733–738, 2013.

[96] G. R. Loftus. Evaluating forgetting curves. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **11**, 397, 1985.

[97] X. Lu, C. Wang, J.-M. Yang, Y. Pang and L. Zhang. Photo2Trip - Generating Travel Routes from Geo-Tagged Photos for Trip Planning. In *Proceedings of the international conference on Multimedia - MM '10*, 143, 2010.

[98] X. Lu, E. Wetter, N. Bharti, A. J. Tatem and L. Bengtsson. Approaching the limit of predictability in human mobility. *Scientific reports*, **3**, 2923, 2013.

[99] M. Lv, L. Chen and G. Chen. Discovering personally semantic places from GPS trajectories. In *Proceedings of the 21st ACM international conference on Information and knowledge management - CIKM '12*, 1552, 2012.

[100] Y. Lv, D. Lymberopoulos and Q. Wu. An exploration of ranking heuristics in mobile local search. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '12*, 295, 2012.

[101] J. Mahmud and J. Nichols. Where Is This Tweet From? Inferring Home Locations of Twitter Users. In *AAAI Conference on Weblogs and Social Network - ICWSM 2011*, 511–514, 2012.

[102] J. Mahmud, J. Nichols and C. Drews. Home Location Identification of Twitter Users. *ACM Transactions on Intelligent Systems and Technology*, **5**, 1, 2014.

[103] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden and R. C. Miller. TwitInfo: Aggregating and Visualizing Microblogs for Event Exploration. In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*, 227–230, 2011.

[104] D. Maxwell, S. Raue, L. Azzopardi, C. Johnson and S. Oates. Crisees: Real-Time Monitoring of Social Media Streams to Support Crisis Management. In *Advances in Information Retrieval - ECIR '12*, 573–575, 2012.

[105] K. S. McCurley. Geospatial mapping and navigation of the web. In *Proceedings of the tenth international conference on World Wide Web - WWW '01*, 221–229, 2001.

[106] MediaTV. Zware regen lijdt tot beheersbare overlast, 2013. `http://www.mediatv.nl/nieuws/10359/Zware-regen-lijdt-tot-beheersbare-overlast.html`. Last Accessed: 2014-09-28.

[107] E. Meeuwissen, P. Reinold and C. Liem. Inferring and predicting context of mobile users. *Bell Labs Technical Journal*, **12**, 79, 2007.

[108] Q. Mei, C. Liu, H. Su and C. Zhai. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *Proceedings of the 15th international conference on World Wide Web - WWW '06*, 533, 2006.

[109] R. Minch. Privacy issues in location-aware mobile devices. In *37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the*, 10 pp., 2004.

[110] A. Monreale, F. Pinelli, R. Trasarti and F. Giannotti. WhereNext: a Location Predictor on Trajectory Pattern Mining. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09*, 637, 2009.

[111] M. Naaman, A. Paepcke and H. Garcia-Molina. From Where to What: Metadata Sharing for Digital Photographs with Geographic Coordinates. In R. Meersman, Z. Tari and D. Schmidt (eds.), *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE SE - 14*, *Lecture Notes in Computer Science*, volume 2888, 196–217. Springer Berlin Heidelberg, 2003.

[112] NOS. Rain leads to flooding A'dam, 2012. http://nos.nl/artikel/394938-regen-leidt-tot-wateroverlast-adam.html. Last Accessed: 2014-09-28.

[113] A. Noulas, S. Scellato, C. Mascolo and M. Pontil. An empirical study of geographic user activity patterns in Foursquare. In *International AAAI Conference on Weblogs and Social Media, ICWSM'11*, 570–573, 2011.

[114] Nu.nl. Veel overlast door aanhoudende regen, 2013. http://www.nu.nl/binnenland/3600587/veel-overlast-aanhoudende-regen.html. Last Accessed: 2014-09-28.

[115] I. Ounis, C. Macdonald, J. Lin and I. Soboroff. Overview of the TREC-2011 Microblog Track. In *TREC '11*, 2011.

[116] L. Overbey, C. Paribello and T. Jackson. Identifying Influential Twitter Users in the 2011 Egyptian Revolution. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, volume 7812, 377–385, 2013.

[117] N. O'Hare and V. Murdock. Modeling locations with social media. *Information Retrieval*, **16**, 30, 2012.

[118] A. Pal and S. Counts. Identifying topical authorities in microblogs. In *Proceedings of the fourth ACM international conference on Web search and data mining - WSDM '11*, 45–54, 2011.

[119] C. Parent, N. Pelekis, Y. Theodoridis, Z. Yan, S. Spaccapietra, C. Renso, G. Andrienko, N. Andrienko, V. Bogorny, M. L. Damiani, A. Gkoulalas-Divanis and J. Macedo. Semantic trajectories modeling and analysis. *ACM Computing Surveys*, **45**, 1, 2013.

[120] A. Popescu and G. Grefenstette. Mining social media to create personalized recommendations for tourist visits. In *Proceedings of the 2nd International Conference on Computing for Geospatial Research & Applications - COM.Geo '11*, 1–6, 2011.

[121] R. Purves, P. Clough and H. Joho. Identifying imprecise regions for geographic information retrieval using the web. In *Proceedings of the 13th GIS Research UK Annual Conference*, 2005.

[122] A. Rae and P. Kelm. Working Notes for the Placing Task at MediaEval 2012. In *CEUR Workshop Proceedings: Working Notes Proceedings of the MediaEval 2012 Workshop*, volume 927, 2012.

[123] A. Rae, V. Murdock, P. Serdyukov and P. Kelm. Working notes for the Placing Task at MediaEval 2011. In *CEUR Workshop Proceedings: Working Notes Proceedings of the MediaEval 2011 Workshop*, volume 807, 2011.

[124] K. Ren, S. Zhang and H. Lin. Where Are You Settling Down: Geolocating Twitter Users Based on Tweets and Social Networks. In Y. Hou, J.-Y. Nie, L. Sun, B. Wang and P. Zhang (eds.), *Lecture Notes in Computer Science*, *Lecture Notes in Computer Science*, volume 7675, 150–161, 2012.

[125] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom and J. Riedl. GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work - CSCW '94*, 175–186, 1994.

[126] M. Richardson and R. W. White. Supporting synchronous social q&a throughout the question lifecycle. In *Proceedings of the 20th international conference on World wide web - WWW '11*, 755, 2011.

[127] O. Roick and S. Heuser. Location Based Social Networks - Definition, Current State of the Art and Research Agenda. *Transactions in GIS*, **17**, 763, 2013.

[128] A. Sadilek, H. Kautz and J. P. Bigham. Finding Your Friends and Following Them to Where You Are Categories and Subject Descriptors. In *Proceedings of the fifth ACM international conference on Web search and data mining - WSDM '12*, 723–732, 2012.

[129] D. Saez-Trumper, D. Quercia and J. Crowcroft. Ads and the city: considering geographic distance goes a long way. In *Proceedings of the sixth ACM conference on Recommender systems - RecSys '12*, 187, 2012.

[130] M. Sahami and T. 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 - WWW '06*, 377, 2006.

[131] T. Sakaki, M. Okazaki and Y. Matsuo. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors Takeshi. In *Proceedings of the 19th international conference on World wide web - WWW '10*, 851–860, 2010.

[132] R. Sarver. Think Globally, Tweet Locally, 2009. `https://blog.twitter.com/2009/think-globally-tweet-locally`. Last Accessed: 2014-10-10.

[133] B. Sarwar, G. Karypis, J. Konstan and J. Reidl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the tenth international conference on World Wide Web - WWW '01*, 285–295, 2001.

[134] E. Schonfeld. Gowalla Versus Foursquare: Why Pretty Doesn't Always Win, 2011. `http://techcrunch.com/2011/12/05/gowalla-versus-foursquare/`. Last Accessed: 2013-02-08.

[135] P. Serdyukov, V. Murdock and R. van Zwol. Placing flickr photos on a map. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '09*, 484, 2009.

[136] Y. Shi, M. Larson and A. Hanjalic. Collaborative Filtering beyond the User-Item Matrix: A Survey of the State of the Art and Future Challenges. *ACM Computing Surveys*, **47**, 1, 2014.

[137] Y. Shi, P. Serdyukov and A. Hanjalic. Personalized Landmark Recommendation Based on Geotags from Photo Sharing Sites. In *Fifth International AAAI*, 622–625, 2011.

[138] T. H. Silva, P. O. V. de Melo, J. M. Almeida, J. Salles and A. A. Loureiro. Visualizing the Invisible Image of Cities. In *2012 IEEE International Conference on Green Computing and Communications*, 382–389, 2012.

[139] E. Smirnova. A model for expert finding in social networks. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information - SIGIR '11*, 1191, 2011.

[140] C. Song, Z. Qu, N. Blumm and A.-L. Barabási. Limits of predictability in human mobility. *Science (New York, N.Y.)*, **327**, 1018, 2010.

[141] D. Soper. Is human mobility tracking a good idea? *Communications of the ACM*, **55**, 35, 2012.

[142] K. Starbird, L. Palen, A. L. Hughes and S. Vieweg. Chatter on The Red: What Hazards Threat Reveals about the Social Life of Microblogged Information. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work - CSCW '10*, 241, 2010.

[143] B. Sun and V. Ng. Identifying Influential Users by Their Postings in Social Networks. In M. Atzmueller, A. Chin, D. Helic and A. Hotho (eds.), *Ubiquitous Social Media Analysis SE - 7*, *Lecture Notes in Computer Science*, volume 8329, 128–151. Springer Berlin Heidelberg, 2013.

[144] K. P. Tang, J. I. Hong and D. P. Siewiorek. Understanding how visual representations of location feeds affect end-user privacy concerns. In *Proceedings of the 13th international conference on Ubiquitous computing - UbiComp '11*, 207–216, 2011.

[145] L.-A. Tang, Y. Zheng, X. Xie, J. Yuan, X. Yu and J. Han. Retrieving k-nearest neighboring trajectories by a set of point locations. 223–241, 2011.

[146] Telegraaf. Veel wateroverlast in regio Amsterdam, 2012. `http://www.telegraaf.nl/binnenland/20791883/__Wateroverlast_Amsterdam__.html`. Last Accessed: 2014-09-28.

[147] R. Tinati, L. Carr, W. Hall and J. Bentwood. Identifying communicator roles in twitter. In *Proceedings of the 21st international conference companion on World Wide Web - WWW '12 Companion*, 1161, 2012.

[148] E. Tjong Kim Sang and A. V. D. Bosch. Dealing with big data: The case of Twitter. *Computational Linguistics in the Netherlands Journal*, **3**, 121, 2013.

[149] Twitter. FORM S-1 REGISTRATION STATEMENT, 2013. `http://www.sec.gov/Archives/edgar/data/1418091/000119312513390321/d564001ds1.htm`. Last Accessed: 2014-11-19.

[150] M. A. Vasconcelos, S. Ricci, J. Almeida, F. Benevenuto and V. Almeida. Tips, dones and todos: uncovering user profiles in foursquare. In *Proceedings of the fifth ACM international conference on Web search and data mining - WSDM '12*, 653–662, 2012.

[151] M. Veloso, S. Phithakkitnukoon and C. Bento. Urban mobility study using taxi traces. In *Proceedings of the 2011 international workshop on Trajectory data mining and analysis - TDMA '11*, 23, 2011.

[152] S. Vieweg, A. L. Hughes, K. Starbird and L. Palen. Microblogging During Two Natural Hazards Events: What Twitter May Contribute to Situational Awareness. In *Proceedings of the 28th international conference on Human factors in computing systems - CHI '10*, 1079, 2010.

[153] C. Wagner, V. Liao, P. Pirolli, L. Nelson and M. Strohmaier. It's Not in Their Tweets: Modeling Topical Expertise of Twitter Users. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, 91–100, 2012.

[154] C. Wang, J. Wang, X. Xie and W.-Y. Ma. Mining geographic knowledge using location aware topic model. In *Proceedings of the 4th ACM workshop on Geographical information retrieval - GIR '07*, 65, 2007.

[155] L. Wang, C. Wang, X. Xie, J. Forman, Y. Lu, W.-Y. Ma and Y. Li. Detecting dominant locations from search queries. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '05*, 424, 2005.

[156] J. Weng, E.-P. Lim, J. Jiang and Q. He. TwitterRank: Finding Topic-sensitive Influential Twitterers. In *Proceedings of the third ACM international conference on Web search and data mining - WSDM '10*, 261, 2010.

[157] S. Whiting, K. Zhou, J. Jose, O. Alonso and T. Leelanupab. CrowdTiles : Presenting Crowd-based Information for Event-driven Information Needs. In *Proceedings of the 21st ACM international conference on Information and knowledge management - CIKM '12*, 288024, 2698–2700, 2012.

[158] X. Xiao, Y. Zheng, Q. Luo and X. Xie. Inferring Social Ties between Users with Human Location History. *Journal of Ambient Intelligence and Humanized Computing*, **1**, 1, 2012.

[159] A. Y. Xue, R. Zhang, Y. Zheng, X. Xie, J. Huang and Z. Xu. Destination prediction by sub-trajectory synthesis and privacy protection against such prediction. In *2013 IEEE 29th International Conference on Data Engineering (ICDE)*, 254–265, 2013.

[160] M. Ye, R. Xiao, W.-C. Lee and X. Xie. On theme location discovery for travelogue services. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information - SIGIR '11*, 465, 2011.

[161] D. Yimam-Seid and A. Kobsa. Expert-Finding Systems for Organizations: Problem and Domain Analysis and the DEMOIR Approach. *Journal of Organizational Computing and Electronic Commerce*, **13**, 1, 2003.

[162] Y. Yu and X. Chen. A Survey of Point-of-Interest Recommendation in Location-Based Social Networks. In *AAAI Workshops - Trajectory-Based Behavior Analytics*, 53–60, 2015.

[163] J. Yuan, Y. Zheng and X. Xie. Discovering regions of different functions in a city using human mobility and POIs. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '12*, 186, 2012.

[164] J. Yuan, Y. Zheng, C. Zhang, W. Xie, X. Xie, G. Sun and Y. Huang. T-Drive: Driving Directions Based on Taxi Trajectories. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '10*, 99, 2010.

[165] J. Yuan, Y. Zheng, X. Xie and G. Sun. T-Drive: Enhancing Driving Directions with Taxi Drivers' Intelligence. *IEEE Transactions on Knowledge and Data Engineering*, **25**, 220, 2013.

[166] Q. Yuan, G. Cong, K. Zhao, Z. Ma and A. Sun. Who, Where, When, and What. *ACM Transactions on Information Systems*, **33**, 1, 2015.

[167] C. Zhai. Statistical Language Models for Information Retrieval A Critical Review. *Foundations and Trends in Information Retrieval*, **2**, 137, 2008.

[168] C. Zhang, L. Shou, K. Chen, G. Chen and Y. Bei. Evaluating geo-social influence in location-based social networks. In *Proceedings of the 21st ACM international conference on Information and knowledge management - CIKM '12*, 1442, 2012.

[169] J. Zhang, M. S. Ackerman and L. Adamic. Expertise Networks in Online Communities: Structure and Algorithms. In *Proceedings of the 16th international conference on World Wide Web - WWW '07*, 221–230, 2007.

[170] K. Zheng, Y. Zheng, X. Xie and X. Zhou. Reducing Uncertainty of Low-Sampling-Rate Trajectories. In *2012 IEEE 28th International Conference on Data Engineering*, 1144–1155, 2012.

[171] V. W. Zheng, Y. Zheng, X. Xie and Q. Yang. Collaborative location and activity recommendations with GPS history data. In *Proceedings of the 19th international conference on World wide web - WWW '10*, 1029, 2010.

[172] V. W. Zheng, Y. Zheng, X. Xie and Q. Yang. Towards mobile intelligence: Learning from GPS history data for collaborative recommendation. *Artificial Intelligence*, **184-185**, 17, 2012.

[173] Y. Zheng and X. Zhou. *Computing with Spatial Trajectories*, 2011.

[174] Y. Zheng, L. Liu, L. Wang and X. Xie. Learning transportation mode from raw gps data for geographic applications on the web. In *Proceeding of the 17th international conference on World Wide Web - WWW '08*, 247, 2008.

[175] Y. Zheng, L. Zhang, X. Xie and W.-Y. Ma. Mining correlation between locations using human location history. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '09*, 472, 2009.

[176] Y. Zheng, L. Zhang, Z. Ma, X. Xie and W.-Y. Ma. Recommending friends and locations based on individual location history. *ACM Transactions on the Web*, **5**, 1, 2011.

[177] R. Zhong, J. Fan, G. Li, K.-L. Tan and L. Zhou. Location-aware instant search. In *Proceedings of the 21st ACM international conference on Information and knowledge management - CIKM '12*, 385, 2012.

[178] K. Zickuhr. Location - Based Services. Technical report, Pew Research Center's Internet & American Life Project, 2013.

[179] W. Zong, D. Wu, A. Sun, E.-P. Lim and D. H.-L. Goh. On assigning place names to geography related web pages. In *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries - JCDL '05*, 354, 2005.

Part V

Appendix

Appendix A

Social Media Workbench Manual

A customizable and extensible work bench for exploring data from social media, a.k.a. Social Media Workbench[1].

A.1 How to use

The minimum example of using the Social Media Workbench can be found in `client/example` directory. The `server` directory contains a server based on the Flask framework[2] which serves data to the client and also acts as a proxy to Twitter's API. Simply running the server will demonstrate the example.

A.1.1 Setup Javascript Libraries for Client

The client relies on several third party Javascript libraries managed by Bower[3]. The configuration of dependent libraries is in `bower.json`. Use the following command to grab a copy of all required libraries:

```
bower install
```

[1] https://github.com/spacelis/portraitist2
[2] http://flask.pocoo.org
[3] http://bower.io

A.1.2 Configuration and Running Server

For running the server, some python libraries listed in `requirement.txt` are needed. It can be simply installed by the following command.

```
pip install -r requirement.txt
```

Running the server by issuing:

```
python twitter_proxy.py
```

Then go to the link `http://localhost:9090/examples/example.html`.

A.1.3 An Example

The code in Listing A.1 shows how to defined the UI that corresponds to Figure A.1 where a set of tweets retrieved from Twitter is rendered in the charts:

- a tag cloud showing the words in the tweets,
- a timeline showing when these tweets were posted,
- a pie chart of users' tweets,
- and a map showing the origin of the tweets.

This example uses Twitter's Search API as a source of data. When user click on the button "Load", a set of tweets in JSON format will be retrieved from Twitter. An example of a JSON tweet is shown in Listing A.1.

Listing A.1: An example of UI layout configuration

```
<!DOCTYPE html>
<html>
  <head>
  </head>
  <body>
    <div ng-controller="ResourceCtrl as ctrl">
      <datasource name="Data Source" id="datasource" data-url="http
          ://localhost:9090/tp/1.1/search/tweets.json?geocode
          =37.781157,-122.398720,1mi"></datasource>
      <tagcloud name="MyTags" id="tagcloud" data-dimension="text"></
          tagcloud>
```

```
      <timeline name="Timeline" id="timeline" data-dimension="
          created_at" data-scale="hour"></timeline>
      <piechart name="User" id="piechart" data-dimension="user.
          screen_name"></piechart>
      <googlemap name="TweetMap" id="map" data-dimension="geo.@1,geo
          .@0" ></googlemap>
    </div>
    <script data-main="/js/portraitist" src="/lib/requirejs/require.
        js"></script>
    <link rel="stylesheet" href="/lib/bootstrap/dist/css/bootstrap.
        css" type="text/css">
    <link rel="stylesheet" href="/lib/dc.js/dc.css" type="text/css">
  </body>
</html>
```

Listing A.2: A JSON example

```
{
   "created_at":"Tue Mar 18 09:14:13 +0000 2014",
   "id":445850650027786240,
   "id_str":"445850650027786240",
   "text":"I'm at Faculty Electrical Engineering, Mathematics and
       Computer Science - @delftuniversity (Delft, Zuid-Holland)
       http:\/\/t.co\/UNP9m7cApn",
   ...
   "user":{
      "id":127747814,
      "name":"SpaceLi",
      "screen_name":"spacelis",
      ...
    }
   },
   "geo":{
      "type":"Point",
      "coordinates":[
         51.99882337,
         4.37355868
      ]
   },
   "coordinates":{
     ...
   },
   "place":{
     ...
   },
   ...
}
```

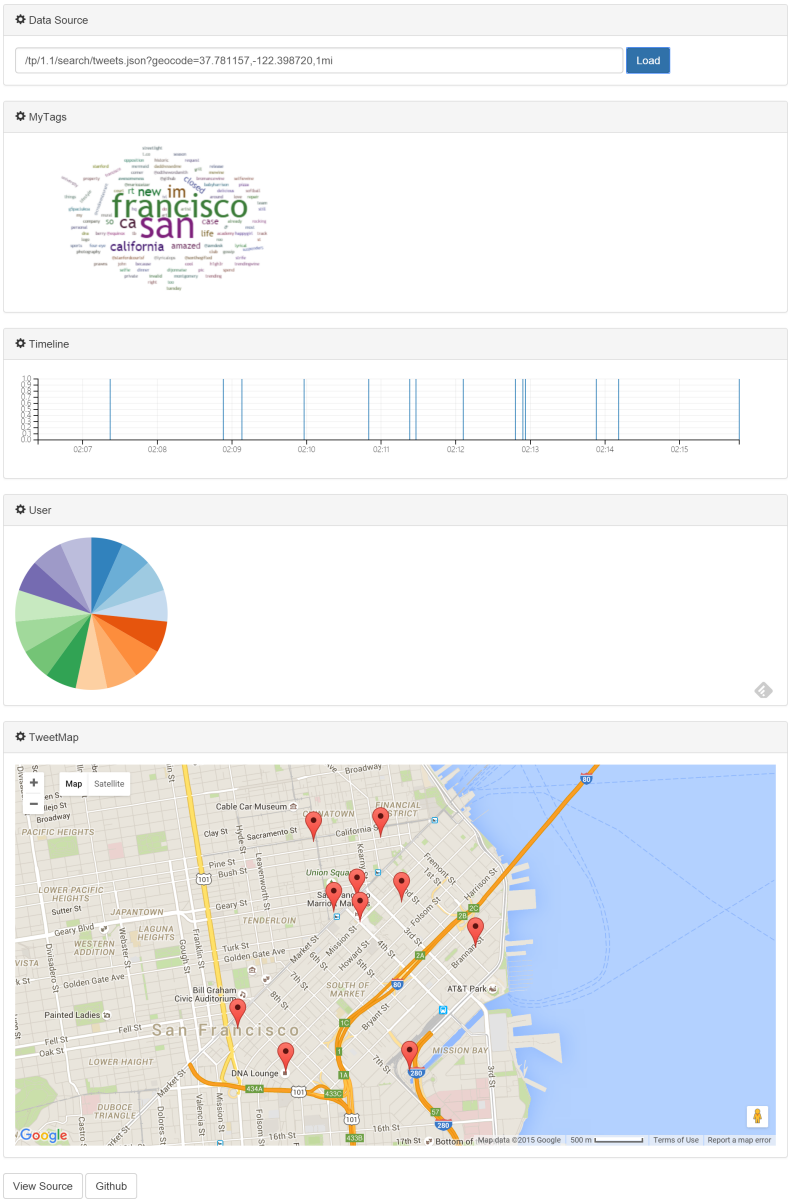


Figure A.1: An example of UI

A.2 Customization

The charting layouts relies on `directive`s facilitated by AngularJS[4]. Thus, it is easy to customize the layout of charts and/or the dimensions each chart depicts. They can be arranged by a HTML page with the following links ensuring import of the RequireJS[5] and the CSS of DC.js[6] and Bootstrap[7].

```
<script data-main="/js/portraitist"
        src="/lib/requirejs/require.js"></script>
<link rel="stylesheet"
      href="/lib/bootstrap/dist/css/bootstrap.css"
      type="text/css">
<link rel="stylesheet"
      href="/lib/dc.js/dc.css" type="text/css">
```

A.2.1 Directives

Each chart is loaded in a directive with some settings that can be passed through both attributes of directives and setting dialogues in the interfaces. Currently, 4 types of charts are available and more are coming. They are pie charts, timelines, tagclouds, maps. They all rely on a data source directive for gathering data from API endpoints.

- `datasource` is a widget for grabbing data from API endpoints, mapping to subset of fields, currently supporting both JSON and CSV formats.
- `tagcloud` can capture the vocabulary of textual fields and depict the words used in the fields.
- `timeline` is a bar chart depicting time series which groups data into different bins according to the given unit of time, e.g., hours, days.
- `piechart` can be used to show the distribution of values in a give field.
- `googlemap` shows a map of geographical data which requires pairs of coordinates.

[4] https://anguarjs.org
[5] http://requirejs.org
[6] https://github.com/dcjs/dcjs
[7] https://getbootstrap.com

All these directives can be arranged by customized HTML pages for interactive exploration or annotation.

A.2.2 Integration of New Charts

The data managing is based on Crossfilter[8] and all the charts should create a `dimension` by passing a function for value accessing and then using that dimension for accessing the filtered data with `dimension.top()` or `dimension.bottom()` A directive can be created by following the built-in ones with a few adaptations. The new chart should also register itself to the global chart render/redraw handler manager via `register_renderer` and `register_redrawer` so that the new chart will be redrawed on filtered data when users interact with other charts.

[8] https://github.com/square/crossfilter

Acknowledgements

The work presented in this book would be much harder without the help and support from all my dear families, friends and colleagues.

I am always grateful to my wife and parents for their effort in making my life easier and better. My special thanks to my wife Jie who is really an amazing partner on my long journey of life. It would be an impossible mission for me to finish my PhD and this book if I did not have them standing by my side.

Studying and working in the MMC group (a.k.a. DMIR) at TU Delft was a great experience. Not only because it is the first stop of my oversea life, but also because I met a lot of wonderful and interesting people.

My supervisor Prof. Arjen de Vries is a great mentor and very thoughtful in research. I learnt a lot during my study with him which I really appreciate. It was also a great pleasure to work with Carsten Eickhoff, who was always helpful and patient to my questions.

Here are some other tips I learnt. Raynor Vliegendhart is the must-go-to guy for all geek stuff and his small volunteer talks are always inspiring. Yue Shi has a lot of interesting thoughts about paper community and writing good research papers. Xinchao Li and Peng Xu are very kind and it is always interesting to hear their discussion about all kinds of stuff.

There are many more names besides the aforementioned and they all become my valuable memory in the jar labelled Delft.

THANK YOU ALL, GUYS!

To my parents:

感谢父母将我带到这个繁华的世界，抚育成人。

Summary

This thesis carried out two major investigations on how the geographical information generated by users in social media can be used to model and predict human behaviours. The first investigation regards the mobility patterns of social media users, which utilizes the relations between geographical information and other sources of information, e.g., textual information and temporal information (Chapter 3) as well as users' trajectory patterns, in predicting the locations of users' future visits (Chapter 4). The experimental results confirm the predictability of human movement and the proposed methods for predicting users' locations are demonstrated to outperform the state-of-the-art via evaluations with the data collected from social media platforms (Twitter and Foursquare). The second investigation regards the geo-knowledge that users may gain from physical visits to locations (Chapter 5). The thesis first presented a discussion on how such knowledge can be modelled from users' check-in posts on social media and then approached this problem from both theoretical and practical points of view. Probabilistic models were built and tested via a configurable system on which all the models were implemented. To evaluate the proposed models, a data set of ground truth was collected by annotation from two different sources (i.e., a crowdsourcing platform and university students/staff). The evaluation show that the proposed models outperform the approaches from related research.

Summanvent

Het proefschrift draagt twee vooraanstaande onderzoeken over hoe, door gebruikers gegenereerde, geografische informatie in sociale media gebruikt kan worden om menselijk gedrag te voorspellen en modelleren. Het eerste onderzoek omvat mobiliteitspatronen van sociale media gebruikers, welke de relatie tussen geografische informatie en andere informatie bronnen benut. Bijvoorbeeld tekstuele- en tijdsgebonden informatie (Hoofdstuk 3), maar ook traject patronen van gebruikers om de locatie van toekomstige bezoeken te voorspellen (Hoofdstuk 4). De experimentele resultaten bevestigen de voorspelbaarheid van menselijke bewegingspatronen. Empirische evaluatie van de verzamelde sociale media data (Twitter en Foursquare) onthult dat de voorgestelde benadering beter presteert dan de gevestigde methodes. Het tweede onderzoek omvat de geografische kennis die gebruikers winnen door het fysiek bezoeken van locaties (Hoofdstuk 5). Het proefschrift bediscussieerd hoe deze kennis van gebruiker 'check-in posts' op sociale media, gemodelleerd kan worden. Vervolgens benaderen wij het probleem van een zowel theoretisch als praktisch standpunt. Onze probabilistische modellen werden getest op een configureerbaar systeem. Om de voorgestelde modellen te evalueren, hebben wij een data set samengesteld via een enquete aan een universiteit en op een crowdsourcing platform. De evaluatie laat zien dat de voorgestelde methodes beter werken dan benadering van gerelateerde onderzoeken.

Resume

Wen Li was born in Xi'an, China on January 2nd, 1985. Upon completing high school education in 2003, he started studying Computer Science at Xi'an Jiaotong University in China. He obtained his bachelor's degree in 2007, and continued to study towards a master's degree at Xi'an Jiaotong University in the research area of data mining. To carry on his interest in how computer technologies can facilitate obtaining knowledge from massive data, Wen started his PhD research at Delft University of Technology in September 2010, with a special focus on geographical information posted by users of social media. In his study, he modeled users on social media platform (e.g. Twitter) by the location they visits and studied how these models can be used for predicting users' future visits and their knowledge about locations. He was also involved in the project SHINE led by Delft University of Technology which endeavour to integrate modern technology and data analytics in urban management such as water damage management. In the project, he analysed the patterns between the signals from social media and existing registration system regarding water problems in urban areas.

SIKS Dissertation Series

2009

2009-01: **Rasa Jurgelenaite** (RUN), *Symmetric Causal Independence Models*

2009-02: **Willem Robert van Hage** (VU), *Evaluating Ontology-Alignment Techniques*

2009-03: **Hans Stol** (UvT), *A Framework for Evidence-based Policy Making Using IT*

2009-04: **Josephine Nabukenya** (RUN), *Improving the Quality of Organisational Policy Making using Collaboration Engineering*

2009-05: **Sietse Overbeek** (RUN), *Bridging Supply and Demand for Knowledge Intensive Tasks - Based on Knowledge, Cognition, and Quality*

2009-06: **Muhammad Subianto** (UU), *Understanding Classification*

2009-07: **Ronald Poppe** (UT), *Discriminative Vision-Based Recovery and Recognition of Human Motion*

2009-08: **Volker Nannen** (VU), *Evolutionary Agent-Based Policy Analysis in Dynamic Environments*

2009-09: **Benjamin Kanagwa** (RUN), *Design, Discovery and Construction of Service-oriented Systems*

2009-10: **Jan Wielemaker** (UVA), *Logic programming for knowledge-intensive interactive applications*

2009-11: **Alexander Boer** (UVA), *Legal Theory, Sources of Law & the Semantic Web*

2009-12: **Peter Massuthe** (TUE, Humboldt-Universitaet zu Berlin), *Operating Guidelines for Services*

2009-13: **Steven de Jong** (UM), *Fairness in Multi-Agent Systems*

2009-14: **Maksym Korotkiy** (VU), *From ontology-enabled services to service-enabled ontologies (making ontologies work in e-science with ONTO-SOA)*

2009-15: **Rinke Hoekstra** (UVA), *Ontology Representation - Design Patterns and Ontologies that Make Sense*

2009-16: siksauthorFritz Reul (UvT), *New Architectures in Computer Chess*

2009-17: **Laurens van der Maaten** (UvT), *Feature Extraction from Visual Data*

2009-18: **Fabian Groffen** (CWI), *Armada, An Evolving Database System*

2009-19: **Valentin Robu** (CWI), *Modeling Preferences, Strategic Reasoning and Collaboration in Agent-Mediated Electronic Markets*

2009-20: **Bob van der Vecht** (UU), *Adjustable Autonomy: Controling Influences on Decision Making*

2009-21: **Stijn Vanderlooy** (UM), *Ranking and Reliable Classification*

2009-22: **Pavel Serdyukov** (UT), *Search For Expertise: Going beyond direct evidence*

2009-23: **Peter Hofgesang** (VU), *Modelling Web Usage in a Changing Environment*

2009-24: **Annerieke Heuvelink** (VUA), *Cognitive Models for Training Simulations*

2009-25: **Alex van Ballegooij** (CWI), *RAM: Array Database Management through Relational Mapping*

2009-26: **Fernando Koch** (UU), *An Agent-Based Model for the Development of Intelligent Mobile Services*

2009-27: **Christian Glahn** (OU), *Contextual Support of social Engagement and Reflection on the Web*

2009-28: **Sander Evers** (UT), *Sensor Data Management with Probabilistic Models*

2009-29: **Stanislav Pokraev** (UT), *Model-Driven Semantic Integration of Service-Oriented Applications*

2009-30: **Marcin Zukowski** (CWI), *Balancing vectorized query execution with bandwidth-optimized storage*

2009-31: **Sofiya Katrenko** (UVA), *A Closer Look at Learning Relations from Text*

2009-32: **Rik Farenhorst and Remco de Boer** (VU), *Architectural Knowledge Management: Supporting Architects and Auditors*

2009-33: **Khiet Truong** (UT), *How Does Real Affect Affect Affect Recognition In Speech?*

2009-34: **Inge van de Weerd** (UU), *Advancing in Software Product Management: An Incremental Method Engineering Approach*

2009-35: **Wouter Koelewijn** (UL), *Privacy en Politiegegevens; Over geautomatiseerde normatieve informatie-uitwisseling*

2009-36: **Marco Kalz** (OUN), *Placement Support for Learners in Learning Networks*

2009-37: **Hendrik Drachsler** (OUN), *Navigation Support for Learners in Informal Learning Networks*

2009-38: **Riina Vuorikari** (OU), *Tags and self-organisation: a metadata ecology for learning resources in a multilingual context*

2009-39: **Christian Stahl** (TUE, Humboldt-Universitaet zu Berlin), *Service Substitution – A Behavioral Approach Based on Petri Nets*

2009-40: **Stephan Raaijmakers** (UvT), *Multinomial Language Learning: Investigations into the Geometry of Language*

2009-41: **Igor Berezhnyy** (UvT), *Digital Analysis of Paintings*

2009-42: siksauthorToine Bogers (UvT), *Recommender Systems for Social Bookmarking*

2009-43: **Virginia Nunes Leal Franqueira** (UT), *Finding Multi-step Attacks in Computer Networks using Heuristic Search and Mobile Ambients*

2009-44: **Roberto Santana Tapia** (UT), *Assessing Business-IT Alignment in Networked Organizations*

2009-45: **Jilles Vreeken** (UU), *Making Pattern Mining Useful*

2009-46: **Loredana Afanasiev** (UvA), *Querying XML: Benchmarks and Recursion*

2010

2010-01: **Matthijs van Leeuwen** (UU), *Patterns that Matter*

2010-02: **Ingo Wassink** (UT), *Work flows in Life Science*

2010-03: **Joost Geurts** (CWI), *A Document Engineering Model and Processing Framework for Multimedia documents*

2010-04: **Olga Kulyk** (UT), *Do You Know What I Know? Situational Awareness of Co-located Teams in Multidisplay Environments*

2010-05: **Claudia Hauff** (UT), *Predicting the Effectiveness of Queries and Retrieval Systems*

2010-06: **Sander Bakkes** (UvT), *Rapid Adaptation of Video Game AI*

2010-07: **Wim Fikkert** (UT), *Gesture interaction at a Distance*

2010-08: **Krzysztof Siewicz** (UL), *Towards an Improved Regulatory Framework of Free Software. Protecting user freedoms in a world of software communities and eGovernments*

2010-09: **Hugo Kielman** (UL), *A Politiele gegevensverwerking en Privacy, Naar een effectieve waarborging*

2010-10: **Rebecca Ong** (UL), *Mobile Communication and Protection of Children*

2010-11: **Adriaan Ter Mors** (TUD), *The world according to MARP: Multi-Agent Route Planning*

2010-12: **Susan van den Braak** (UU), *Sensemaking software for crime analysis*

2010-13: **Gianluigi Folino** (RUN), *High Performance Data Mining using Bio-inspired techniques*

2010-14: **Sander van Splunter** (VU), *Automated Web Service Reconfiguration*

2010-15: **Lianne Bodenstaff** (UT), *Managing Dependency Relations in Inter-Organizational Models*

2010-16: **Sicco Verwer** (TUD), *Efficient Identification of Timed Automata, theory and practice*

2010-17: **Spyros Kotoulas** (VU), *Scalable Discovery of Networked Resources: Algorithms, Infrastructure, Applications*

2010-18: **Charlotte Gerritsen** (VU), *Caught in the Act: Investigating Crime by Agent-Based Simulation*

2010-19: **Henriette Cramer** (UvA), *People's Responses to Autonomous and Adaptive Systems*

2010-20: **Ivo Swartjes** (UT), *Whose Story Is It Anyway? How Improv Informs Agency and Authorship of Emergent Narrative*

2010-21: **Harold van Heerde** (UT), *Privacy-aware data management by means of data degradation*

2010-22: **Michiel Hildebrand** (CWI), *End-user Support for Access to Heterogeneous Linked Data*

2010-23: **Bas Steunebrink** (UU), *The Logical Structure of Emotions*

2010-24: **Dmytro Tykhonov** *Designing Generic and Efficient Negotiation Strategies*

2010-25: **Zulfiqar Ali Memon** (VU), *Modelling Human-Awareness for Ambient Agents: A Human Mindreading Perspective*

2010-26: **Ying Zhang** (CWI), *XRPC: Efficient Distributed Query Processing on Heterogeneous XQuery Engines*

2010-27: **Marten Voulon** (UL), *Automatisch contracteren*

2010-28: **Arne Koopman** (UU), *Characteristic Relational Patterns*

2010-29: **Stratos Idreos** (CWI), *Database Cracking: Towards Auto-tuning Database Kernels*

2010-30: **Marieke van Erp** (UvT), *Accessing Natural History - Discoveries in data cleaning, structuring, and retrieval*

2010-31: **Victor de Boer** (UVA), *Ontology Enrichment from Heterogeneous Sources on the Web*

2010-32: **Marcel Hiel** (UvT), *An Adaptive Service Oriented Architecture: Automatically solving Interoperability Problems*

2010-33: **Robin Aly** (UT), *Modeling Representation Uncertainty in Concept-Based Multimedia Retrieval*

2010-34: **Teduh Dirgahayu** (UT), *Interaction Design in Service Compositions*

2010-35: **Dolf Trieschnigg** (UT), *Proof of Concept: Concept-based Biomedical Information Retrieval*

2010-36: **Jose Janssen** (OU), *Paving the Way for Lifelong Learning; Facilitating competence development through a learning path specification*

2010-37: **Niels Lohmann** (TUE), *Correctness of services and their composition*

2010-38: **Dirk Fahland** (TUE), *From Scenarios to components*

2010-39: **Ghazanfar Farooq Siddiqui** (VU), *Integrative modeling of emotions in virtual agents*

2010-40: **Mark van Assem** (VU), *Converting and Integrating Vocabularies for the Semantic Web*

2010-41: **Guillaume Chaslot** (UM), *Monte-Carlo Tree Search*

2010-42: **Sybren de Kinderen** (VU), *Needs-driven service bundling in a multi-supplier setting – the computational e3-service approach*

2010-43: **Peter van Kranenburg** (UU), *A Computational Approach to Content-Based Retrieval of Folk Song Melodies*

2010-44: **Pieter Bellekens** (TUE), *An Approach towards Context-sensitive and User-adapted Access to Heterogeneous Data Sources, Illustrated in the Television Domain*

2010-45: **Vasilios Andrikopoulos** (UvT), *A theory and model for the evolution of software services*

2010-46: **Vincent Pijpers** (VU), *e3alignment: Exploring Inter-Organizational Business-ICT Alignment*

2010-47: **Chen Li** (UT), *Mining Process Model Variants: Challenges, Techniques, Examples*

2010-48: **Milan Lovric** (EUR), *Behavioral Finance and Agent-Based Artificial Markets*

2010-49: **Jahn-Takeshi Saito** (UM), *Solving difficult game positions*

2010-50: **Bouke Huurnink** (UVA), *Search in Audiovisual Broadcast Archives*

2010-51: **Alia Khairia Amin** (CWI), *Understanding and supporting information seeking tasks in multiple sources*

2010-52: **Peter-Paul van Maanen** (VU), *Adaptive Support for Human-Computer Teams: Exploring the Use of Cognitive Models of Trust and Attention*

2010-53: **Edgar Meij** (UVA), *Combining Concepts and Language Models for Information Access*

2011

2011-01: **Botond Cseke** (RUN), *Variational Algorithms for Bayesian Inference in Latent Gaussian Models*

2011-02: **Nick Tinnemeier** (UU), *Organizing Agent Organizations. Syntax and Operational Semantics of an Organization-Oriented Programming Language*

2011-03: **Jan Martijn van der Werf** (TUE), *Compositional Design and Verification of Component-Based Information Systems*

2011-04: **Hado van Hasselt** (UU), *Insights in Reinforcement Learning; Formal analysis and empirical evaluation of temporal-difference learning algorithms*

2011-05: **Base van der Raadt** (VU), *Enterprise Architecture Coming of Age - Increasing the Performance of an Emerging Discipline*

2011-06: **Yiwen Wang** (TUE), *Semantically-Enhanced Recommendations in Cultural Heritage*

2011-07: **Yujia Cao** (UT), *Multimodal Information Presentation for High Load Human Computer Interaction*

2011-08: **Nieske Vergunst** (UU), *BDI-based Generation of Robust Task-Oriented Dialogues*

2011-09: **Tim de Jong** (OU), *Contextualised Mobile Media for Learning*

2011-10: **Bart Bogaert** (UvT), *Cloud Content Contention*

2011-11: **Dhaval Vyas** (UT), *Designing for Awareness: An Experience-focused HCI Perspective*

2011-12: **Carmen Bratosin** (TUE), *Grid Architecture for Distributed Process Mining*

2011-13: **Xiaoyu Mao** (UvT), *Airport under Control. Multiagent Scheduling for Airport Ground Handling*

2011-14: **Milan Lovric** (EUR), *Behavioral Finance and Agent-Based Artificial Markets*

2011-15: **Marijn Koolen** (UvA), *The Meaning of Structure: the Value of Link Evidence for Information Retrieval*

2011-16: **Maarten Schadd** (UM), *Selective Search in Games of Different Complexity*

2011-17: **Jiyin He** (UVA), *Exploring Topic Structure: Coherence, Diversity and Relatedness*

2011-18: **Mark Ponsen** (UM), *Strategic Decision-Making in complex games*

2011-19: **Ellen Rusman** (OU), *The Mind ' s Eye on Personal Profiles*

2011-20: **Qing Gu** (VU), *Guiding service-oriented software engineering - A view-based approach*

2011-21: **Linda Terlouw** (TUD), *Modularization and Specification of Service-Oriented Systems*

2011-22: **Junte Zhang** (UVA), *System Evaluation of Archival Description and Access*

2011-23: **Wouter Weerkamp** (UVA), *Finding People and their Utterances in Social Media*

2011-24: **Herwin van Welbergen** (UT), *Behavior Generation for Interpersonal Coordination with Virtual Humans On Specifying, Scheduling and Realizing Multimodal Virtual Human Behavior*

2011-25: **Syed Waqar ul Qounain Jaffry** (VU), *Analysis and Validation of Models for Trust Dynamics*

2011-26: **Matthijs Aart Pontier** (VU), *Virtual Agents for Human Communication - Emotion*

Regulation and Involvement-Distance Trade-Offs in Embodied Conversational Agents and Robots

2011-27: **Aniel Bhulai** (VU), *Dynamic website optimization through autonomous management of design patterns*

2011-28: **Rianne Kaptein** (UVA), *Effective Focused Retrieval by Exploiting Query Context and Document Structure*

2011-29: **Faisal Kamiran** (TUE), *Discrimination-aware Classification*

2011-30: **Egon van den Broek** (UT), *Affective Signal Processing (ASP): Unraveling the mystery of emotions*

2011-31: **Ludo Waltman** (EUR), *Computational and Game-Theoretic Approaches for Modeling Bounded Rationality*

2011-32: **Nees-Jan van Eck** (EUR), *Methodological Advances in Bibliometric Mapping of Science*

2011-33: **Tom van der Weide** (UU), *Arguing to Motivate Decisions*

2011-34: **Paolo Turrini** (UU), *Strategic Reasoning in Interdependence: Logical and Game-theoretical Investigations*

2011-35: **Maaike Harbers** (UU), *Explaining Agent Behavior in Virtual Training*

2011-36: **Erik van der Spek** (UU), *Experiments in serious game design: a cognitive approach*

2011-37: **Adriana Burlutiu** (RUN), *Machine Learning for Pairwise Data, Applications for Preference Learning and Supervised Network Inference*

2011-38: **Nyree Lemmens** (UM), *Bee-inspired Distributed Optimization*

2011-39: **Joost Westra** (UU), *Organizing Adaptation using Agents in Serious Games*

2011-40: **Viktor Clerc** (VU), *Architectural Knowledge Management in Global Software Development*

2011-41: **Luan Ibraimi** (UT), *Cryptographically Enforced Distributed Data Access Control*

2011-42: **Michal Sindlar** (UU), *Explaining Behavior through Mental State Attribution*

2011-43: **Henk van der Schuur** (UU), *Process Improvement through Software Operation Knowledge*

2011-44: **Boris Reuderink** (UT), *Robust Brain-Computer Interfaces*

2011-45: **Herman Stehouwer** (UvT), *Statistical Language Models for Alternative Sequence Selection*

2011-46: **Beibei Hu** (TUD), *Towards Contextualized Information Delivery: A Rule-based Architecture for the Domain of Mobile Police Work*

2011-47: **Azizi Bin Ab Aziz** (VU), *Exploring Computational Models for Intelligent Support of Persons with Depression*

2011-48: **Mark Ter Maat** (UT), *Response Selection and Turn-taking for a Sensitive Artificial Listening Agent*

2011-49: **Andreea Niculescu** (UT), *Conversational interfaces for task-oriented spoken dialogues: design aspects influencing interaction quality*

2012

2012-01: **Terry Kakeeto** (UvT), *Relationship Marketing for SMEs in Uganda*

2012-02: **Muhammad Umair** (VU), *Adaptivity, emotion, and Rationality in Human and Ambient Agent Models*

2012-03: **Adam Vanya** (VU), *Supporting Architecture Evolution by Mining Software Repositories*

2012-04: **Jurriaan Souer** (UU), *Development of Content Management System-based Web Applications*

2012-05: **Marijn Plomp** (UU), *Maturing Interorganisational Information Systems*

2012-06: **Wolfgang Reinhardt** (OU), *Awareness Support for Knowledge Workers in Research Networks*

2012-07: **Rianne van Lambalgen** (VU), *When the Going Gets Tough: Exploring Agent-based Models of Human Performance under Demanding Conditions*

2012-08: **Gerben de Vries** (UVA), *Kernel Methods for Vessel Trajectories*

2012-09: **Ricardo Neisse** (UT), *Trust and Privacy Management Support for Context-Aware Service Platforms*

2012-10: **David Smits** (TUE), *Towards a Generic Distributed Adaptive Hypermedia Environment*

2012-11: **J.C.B. Rantham Prabhakara** (TUE), *Process Mining in the Large: Preprocessing, Discovery, and Diagnostics*

2012-12: **Kees van der Sluijs** (TUE), *Model Driven Design and Data Integration in Semantic Web Information Systems*

2012-13: **Suleman Shahid** (UvT), *Fun and Face: Exploring non-verbal expressions of emotion during playful interactions*

2012-14: **Evgeny Knutov** (TUE), *Generic Adaptation Framework for Unifying Adaptive Web-based Systems*

2012-15: **Natalie van der Wal** (VU), *Social Agents. Agent-Based Modelling of Integrated Internal and Social Dynamics of Cognitive and Affective Processes*

2012-16: **Fiemke Both** (VU), *Helping people by understanding them – Ambient Agents supporting task execution and depression treatment*

2012-17: **Amal Elgammal** (UvT), *Towards a Comprehensive Framework for Business Process Compliance*

2012-18: **Eltjo Poort** (VU), *Improving Solution Architecting Practices*

2012-19: **Helen Schonenberg** (TUE), *What's Next? Operational Support for Business Process Execution*

2012-20: **Ali Bahramisharif** (RUN), *Covert Visual Spatial Attention, a Robust Paradigm for Brain-Computer Interfacing*

2012-21: **Roberto Cornacchia** (TUD), *Querying Sparse Matrices for Information Retrieval*

2012-22: **Thijs Vis** (UvT), *Intelligence, politie en veiligheidsdienst: verenigbare grootheden?*

2012-23: **Christian Muehl** (UT), *Toward Affective Brain-Computer Interfaces: Exploring the Neurophysiology of Affect during Human Media Interaction*

2012-24: **Laurens van der Werff** (UT), *Evaluation of Noisy Transcripts for Spoken Document Retrieval*

2012-25: **Silja Eckartz** (UT), *Managing the Business Case Development in Inter-Organizational IT Projects: A Methodology and its Application*

2012-26: **Emile de Maat** (UVA), *Making Sense of Legal Text*

2012-27: **Hayrettin Gurkok** (UT), *Mind the Sheep! User Experience Evaluation & Brain-Computer Interface Games*

2012-28: **Nancy Pascall** (UvT), *Engendering Technology Empowering Women*

2012-29: **Almer Tigelaar** (UT), *Peer-to-Peer Information Retrieval*

2012-30: **Alina Pommeranz** (TUD), *Designing Human-Centered Systems for Reflective Decision Making*

2012-31: **Emily Bagarukayo** (RUN), *A Learning by Construction Approach for Higher Order Cognitive Skills Improvement, Building Capacity and Infrastructure*

2012-32: **Wietske Visser** (TUD), *Qualitative multi-criteria preference representation and reasoning*

2012-33: **Rory Sie** (OUN), *Coalitions in Cooperation Networks (COCOON)*

2012-34: **Pavol Jancura** (RUN), *Evolutionary analysis in PPI networks and applications*

2012-35: **Evert Haasdijk** (VU), *Never Too Old To Learn – On-line Evolution of Controllers in Swarm- and Modular Robotics*

2012-36: **Denis Ssebugwawo** (RUN), *Analysis and Evaluation of Collaborative Modeling Processes*

2012-37: **Agnes Nakakawa** (RUN), *A Collaboration Process for Enterprise Architecture Creation*

2012-38: **Selmar Smit** (VU), *Parameter Tuning and Scientific Testing in Evolutionary Algorithms*

2012-39: **Hassan Fatemi** (UT), *Risk-aware design of value and coordination networks*

2012-40: **Agus Gunawan** (UvT), *Information Access for SMEs in Indonesia*

2012-41: **Sebastian Kelle** (OU), *Game Design Patterns for Learning*

2012-42: **Dominique Verpoorten** (OU), *Reflection Amplifiers in self-regulated Learning*

2012-43: *Withdrawn*

2012-44: **Anna Tordai** (VU), *On Combining Alignment Techniques*

2012-45: **Benedikt Kratz** (UvT), *A Model and Language for Business-aware Transactions*

2012-46: **Simon Carter** (UVA), *Exploration and Exploitation of Multilingual Data for Statistical Machine Translation*

2012-47: **Manos Tsagkias** (UVA), *Mining Social Media: Tracking Content and Predicting Behavior*

2012-48: **Jorn Bakker** (TUE), *Handling Abrupt Changes in Evolving Time-series Data*

2012-49: **Michael Kaisers** (UM), *Learning against Learning – Evolutionary dynamics of reinforcement learning algorithms in strategic interactions*

2012-50: **Steven van Kervel** (TUD), *Ontologogy driven Enterprise Information Systems Engineering*

2012-51: **Jeroen de Jong** (TUD), *Heuristics in Dynamic Sceduling; a practical framework with a case study in elevator dispatching*

2013

2013-01: **Viorel Milea** (EUR), *News Analytics for Financial Decision Support*

2013-02: **Erietta Liarou** (CWI), *MonetDB/DataCell: Leveraging the Column-store Database Technology for Efficient and Scalable Stream Processing*

2013-03: **Szymon Klarman** (VU), *Reasoning with Contexts in Description Logics*

2013-04: **Chetan Yadati** (TUD), *Coordinating autonomous planning and scheduling*

2013-05: **Dulce Pumareja** (UT), *Groupware Requirements Evolutions Patterns*

2013-06: **Romulo Goncalves** (CWI), *The Data Cyclotron: Juggling Data and Queries for a Data Warehouse Audience*

2013-07: **Giel van Lankveld** (UvT), *Quantifying Individual Player Differences*

2013-08: **Robbert-Jan Merk** (VU), *Making enemies: cognitive modeling for opponent agents in fighter pilot simulators*

2013-09: **Fabio Gori** (RUN), *Metagenomic Data Analysis: Computational Methods and Applications*

2013-10: **Jeewanie Jayasinghe Arachchige** (UvT), *A Unified Modeling Framework for Service Design*

2013-11: **Evangelos Pournaras** (TUD), *Multi-level Reconfigurable Self-organization in Overlay Services*

2013-12: **Marian Razavian** (VU), *Knowledge-driven Migration to Services*

2013-13: **Mohammad Safiri** (UT), *Service Tailoring: User-centric creation of integrated IT-based homecare services to support independent living of elderly*

2013-14: **Jafar Tanha** (UVA), *Ensemble Approaches to Semi-Supervised Learning Learning*

2013-15: **Daniel Hennes** (UM), *Multiagent Learning - Dynamic Games and Applications*

2013-16: **Eric Kok** (UU), *Exploring the practical benefits of argumentation in multi-agent deliberation*

2013-17: **Koen Kok** (VU), *The PowerMatcher: Smart Coordination for the Smart Electricity Grid*

2013-18: **Jeroen Janssens** (UvT), *Outlier Selection and One-Class Classification*

2013-19: **Renze Steenhuizen** (TUD), *Coordinated Multi-Agent Planning and Scheduling*

2013-20: **Katja Hofmann** (UvA), *Fast and Reliable Online Learning to Rank for Information Retrieval*

2013-21: **Sander Wubben** (UvT), *Text-to-text generation by monolingual machine translation*

2013-22: **Tom Claassen** (RUN), *Causal Discovery and Logic*

2013-23: **Patricio de Alencar Silva** (UvT), *Value Activity Monitoring*

2013-24: **Haitham Bou Ammar** (UM), *Automated Transfer in Reinforcement Learning*

2013-25: **Agnieszka Anna Latoszek-Berendsen** (UM), *Intention-based Decision Support. A new way of representing and implementing clinical guidelines in a Decision Support System*

2013-26: **Alireza Zarghami** (UT), *Architectural Support for Dynamic Homecare Service Provisioning*

2013-27: **Mohammad Huq** (UT), *Inference-based Framework Managing Data Provenance*

2013-28: **Frans van der Sluis** (UT), *When Complexity becomes Interesting: An Inquiry into the Information eXperience*

2013-29: **Iwan de Kok** (UT), *Listening Heads*

2013-30: **Joyce Nakatumba** (TUE), *Resource-Aware Business Process Management: Analysis and Support*

2013-31: **Dinh Khoa Nguyen** (UvT), *Blueprint Model and Language for Engineering Cloud Applications*

2013-32: **Kamakshi Rajagopal** (OUN), *Networking For Learning; The role of Networking in a Lifelong Learner's Professional Development*

2013-33: **Qi Gao** (TUD), *User Modeling and Personalization in the Microblogging Sphere*

2013-34: **Kien Tjin-Kam-Jet** (UT), *Distributed Deep Web Search*

2013-35: **Abdallah El Ali** (UvA), *Minimal Mobile Human Computer Interaction*

2013-36: **Than Lam Hoang** (TUe), *Pattern Mining in Data Streams*

2013-37: **Dirk Börner** (OUN), *Ambient Learning Displays*

2013-38: **Eelco den Heijer** (VU), *Autonomous Evolutionary Art*

2013-39: **Joop de Jong** (TUD), *A Method for Enterprise Ontology based Design of Enterprise Information Systems*

2013-40: **Pim Nijssen** (UM), *Monte-Carlo Tree Search for Multi-Player Games*

2013-41: **Jochem Liem** (UVA), *Supporting the Conceptual Modelling of Dynamic Systems: A Knowledge Engineering Perspective on Qualitative Reasoning*

2013-42: **Léon Planken** (TUD), *Algorithms for Simple Temporal Reasoning*

2013-43: **Marc Bron** (UVA), *Exploration and Contextualization through Interaction and Concepts*

2014

2014-01: **Nicola Barile** (UU), *Studies in Learning Monotone Models from Data*

2014-02: **Fiona Tuliyano** (RUN), *Combining System Dynamics with a Domain Modeling Method*

2014-03: **Sergio Raul Duarte Torres** (UT), *Information Retrieval for Children: Search Behavior and Solutions*

2014-04: **Hanna Jochmann-Mannak** (UT), *Websites for children: search strategies and interface design - Three studies on children's search performance and evaluation*

2014-05: **Jurriaan van Reijsen** (UU), *Knowledge Perspectives on Advancing Dynamic Capability*

2014-06: **Damian Tamburri** (VU), *Supporting Networked Software Development*

2014-07: **Arya Adriansyah** (TUE), *Aligning Observed and Modeled Behavior*

2014-08: **Samur Araujo** (TUD), *Data Integration over Distributed and Heterogeneous Data Endpoints*

2014-09: **Philip Jackson** (UvT), *Toward Human-Level Artificial Intelligence: Representation and Computation of Meaning in Natural Language*

2014-10: **Ivan Salvador Razo Zapata** (VU), *Service Value Networks*

2014-11: **Janneke van der Zwaan** (TUD), *An Empathic Virtual Buddy for Social Support*

2014-12: **Willem van Willigen** (VU), *Look Ma, No Hands: Aspects of Autonomous Vehicle Control*

2014-13: **Arlette van Wissen** (VU), *Agent-Based Support for Behavior Change: Models and Applications in Health and Safety Domains*

2014-14: **Yangyang Shi** (TUD), *Language Models With Meta-information*

2014-15: **Natalya Mogles** (VU), *Agent-Based Analysis and Support of Human Functioning in Complex Socio-Technical Systems: Applications in Safety and Healthcare*

2014-16: **Krystyna Milian** (VU), *Supporting trial recruitment and design by automatically interpreting eligibility criteria*

2014-17: **Kathrin Dentler** (VU), *Computing healthcare quality indicators automatically: Secondary Use of Patient Data and Semantic Interoperability*

2014-18: **Mattijs Ghijsen** (VU), *Methods and Models for the Design and Study of Dynamic Agent Organizations*

2014-19: **Vincius Ramos** (TUE), *Adaptive Hypermedia Courses: Qualitative and Quantitative Evaluation and Tool Support*

2014-20: **Mena Habib** (UT), *Named Entity Extraction and Disambiguation for Informal Text: The Missing Link*

2014-21: **Kassidy Clark** (TUD), *Negotiation and Monitoring in Open Environments*

2014-22: **Marieke Peeters** (UU), *Personalized Educational Games - Developing agent-supported scenario-based training*

2014-23: **Eleftherios Sidirourgos** (UvA/CWI), *Space Efficient Indexes for the Big Data Era*

2014-24: **Davide Ceolin** (VU), *Trusting Semi-structured Web Data*

2014-25: **Martijn Lappenschaar** (RUN), *New network models for the analysis of disease interaction*

2014-26: **Tim Baarslag** (TUD), *What to Bid and When to Stop*

2014-27: **Rui Jorge Almeida** (EUR), *Conditional Density Models Integrating Fuzzy and Probabilistic Representations of Uncertainty*

2014-28: **Anna Chmielowiec** (VU), *Decentralized k-Clique Matching*

2014-29: **Jaap Kabbedijk** (UU), *Variability in Multi-Tenant Enterprise Software*

2014-30: **Peter de Cock** (UvT), *Anticipating Criminal Behaviour*

2014-31: **Leo van Moergestel** (UU), *Agent Technology in Agile Multiparallel Manufacturing and Product Support*

2014-32: **Naser Ayat** (UvA), *On Entity Resolution in Probabilistic Data*

2014-33: **Tesfa Tegegne** (RUN), *Service Discovery in eHealth*

2014-34: **Christina Manteli** (VU), *The Effect of Governance in Global Software Development: Analyzing Transactive Memory Systems*

2014-35: **Joost van Ooijen** (UU), *Cognitive Agents in Virtual Worlds: A Middleware Design Approach*

2014-36: **Joos Buijs** (TUE), *Flexible Evolutionary Algorithms for Mining Structured Process Models*

2014-37: **Maral Dadvar** (UT), *Experts and Machines United Against Cyberbullying*

2014-38: **Danny Plass-Oude Bos** (UT), *Making brain-computer interfaces better: improving usability through post-processing*

2014-39: **Jasmina Maric** (UvT), *Web Communities, Immigration, and Social Capital*

2014-40: **Walter Omona** (RUN), *A Framework for Knowledge Management Using ICT in Higher Education*

2014-41: **Frederic Hogenboom** (EUR), *Automated Detection of Financial Events in News Text*

2014-42: **Carsten Eijckhof** (CWI/TUD), *Contextual Multidimensional Relevance Models*

2014-43: **Kevin Vlaanderen** (UU), *Supporting Process Improvement using Method Increments*

2014-44: **Paulien Meesters** (UvT), *Intelligent Blauw. Met als ondertitel: Intelligence-gestuurde politiezorg in gebiedsgebonden eenheden*

2014-45: **Birgit Schmitz** (OUN), *Mobile Games for Learning: A Pattern-Based Approach*

2014-46: **Ke Tao** (TUD), *Social Web Data Analytics: Relevance, Redundancy, Diversity*

2014-47: **Shangsong Liang** (UVA), *Fusion and Diversification in Information Retrieval*

2015

2015-01: **Niels Netten** (UVA), *Machine Learning for Relevance of Information in Crisis Response*

2015-02: **Faiza Bukhsh** (UvT), *Smart auditing: Innovative Compliance Checking in Customs Controls*

2015-03: **Twan van Laarhoven** (RUN), *Machine learning for network data*

2015-04: **Howard Spoelstra** (OU), *Cooperations in Open Learning environments*

2015-05: **Christoph Bösch** (UT), *Cryptographically Enforced Search Pattern Hiding*

2015-06: **Farideh Heidari** (TUD), *Business Process Quality Computation – Computing Non-Functional Requirements to Improve Business Processes*

2015-07: **Maria-Hendrike Peetz** (UVA), *Time-Aware Online Reputation Analysis*

2015-08: **Jie Jiang** (TUD), *Organizational Compliance: An agent-based model for designing and evaluating organizational interactions*

2015-09: **Randy Klaassen** (UT), *HCI Perspectives on Behavior Change Support Systems*

2015-10: **Henry Hermans** (OUN), *OpenU: design of an integrated system to support lifelong learning*

2015-11: **Yongming Luo** (TUE), *Designing algorithms for big graph datasets: A study of computing bisimulation and joins*

2015-12: **Julie M. Birkholz** (VU), *Modi Operandi of Social Network Dynamics: The Effect of Context on Scientific Collaboration Networks*

2015-13: **Giuseppe Procaccianti** (VU), *Energy-Efficient Software*

2015-14: **Bart van Straalen** (UT), *A cognitive approach to modeling bad news conversations*

2015-15: **Klaas Andries de Graaf** (VU), *Ontology-based Software Architecture Documentation*

2015-16: **Changyun Wei** (UT), *Cognitive Coordination for Cooperative Multi-Robot Teamwork*

2015-17: **André van Cleeff** (UT), *Physical and Digital Security Mechanisms: Properties, Combinations and Trade-offs*

2015-18: **Holger Pirk** (CWI), *Waste Not, Want Not! - Managing Relational Data in Asymmetric Memories*

2015-19: **Bernardo Tabuenca** (OUN), *Ubiquitous Technology for Lifelong Learners*

2015-20: **Loïs Vanhée** (UU), *Using Culture and Values to Support Flexible Coordination*

2015-21: **Sibren Fetter** (OUN), *Using Peer-Support to Expand and Stabilize Online Learn-*

ing

2015-22: **Zhemin Zhu** (UT), *Co-occurrence Rate Networks*

2015-23: **Luit Gazendam** (VU), *Cataloguer Support in Cultural Heritage*

2015-24: **Richard Berendsen** (UVA), *Finding People, Papers, and Posts: Vertical Search Algorithms and Evaluation*

2015-25: **Steven Woudenberg** (UU), *Bayesian Tools for Early Disease Detection*

2015-26: **Alexander Hogenboom** (EUR), *Sentiment Analysis of Text Guided by Semantics and Structure*

2015-27: **Sándor Héman** (CWI), *Updating compressed colomn stores*

2015-28: **Janet Bagorogoza** (TiU), *KNOWLEDGE MANAGEMENT AND HIGH PERFORMANCE; The Uganda Financial Institutions Model for HPO*

2015-29: **Hendrik Baier** (UM), *Monte-Carlo Tree Search Enhancements for One-Player and Two-Player Domains*

2015-30: **Kiavash Bahreini** (OU), *Real-time Multimodal Emotion Recognition in E-Learning*

2015-31: **Yakup Koç** (TUD), *On the robustness of Power Grids*

2015-32: **Jerome Gard** (UL), *Corporate Venture Management in SMEs*

2015-33: **Frederik Schadd** (TUD), *Ontology Mapping with Auxiliary Resources*

2015-34: **Victor de Graaf** (UT), *Gesocial Recommender Systems*

2015-35: **Jungxao Xu** (TUD), *Affective Body Language of Humanoid Robots: Perception and Effects in Human Robot Interaction*

2016

2016-01: **Syed Saiden Abbas** (RUN), *Recognition of Shapes by Humans and Machines*

2016-02: **Michiel Christiaan Meulendijk** (UU), *Optimizing medication reviews through decision support: prescribing a better pill to swallow*

2016-03: **Maya Sappelli** (RUN), *Knowledge Work in Context: User Centered Knowledge Worker Support*

2016-04: **Laurens Rietveld** (VU), *Publishing and Consuming Linked Data*

2016-05: **Evgeny Sherkhonov** (UVA), *Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers*

2016-06: **Michel Wilson** (TUD), *Robust scheduling in an uncertain environment*

2016-07: **Jeroen de Man** (VU), *Measuring and modeling negative emotions for virtual training*

2016-08: **Matje van de Camp** (TiU), *A Link to the Past: Constructing Historical Social Networks from Unstructured Data*

2016-09: **Archana Nottamkandath** (VU), *Trusting Crowdsourced Information on Cultural Artefacts*

2016-10: **George Karafotias** (VUA), *Parameter Control for Evolutionary Algorithms*

2016-11: **Anne Schuth** (UVA), *Search Engines that Learn from Their Users*

2016-12: **Max Knobbout** (UU), *Logics for Modelling and Verifying Normative Multi-Agent Systems*

2016-13: **Nana Baah Gyan** (VU), *The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach*

2016-14: **Ravi Khadka** (UU), *Revisiting Legacy Software System Modernization*

2016-15: **Steffen Michels** (RUN), *Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments*

2016-16: **Guangliang Li** (UVA), *Socially Intelligent Autonomous Agents that Learn from Human Reward*

2016-17: **Berend Weel** (VU), *Towards Embodied Evolution of Robot Organisms*

2016-18: **Albert Meroño Peñuela** (UVA), *Re-*

fining Statistical Data on the Web

2016-19: **Julia Efremova** (TUE), *Mining Social Structures from Genealogical Data*

2016-20: **Daan Odijk** (UVA), *Context & Semantics in News & Web Search*

2016-21: **Alejandro Moreno Célleri** (UT), *From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground*

2016-22: **Grace Lewis** (VU), *Software Architecture Strategies for Cyber-Foraging Systems*

2016-23: **Fei Cai** (UVA), *Query Auto Completion in Information Retrieval*

2016-24: **Brend Wanders** (UT), *Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach*

2016-25: **Julia Kiseleva** (TUE), *Using Contextual Information to Understand Searching and Browsing Behavior*

2016-26: **Dilhan Thilakarathne** (VU), *In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains*

2016-27: **Wen Li** (TUD), *Understanding Geospatial Information on Social Media*

ISBN 978-94-6186-665-3