Bulletin of the Technical Committee on

Data Engineering

June 2015 Vol. 38 No. 2

IEEE Computer Society

Letters

Special Issue on Location-based Social Media Analysis

Conference and Journal Notices

The TC on Data Engineering

Membership in the TC on Data Engineering is open to all current members of the IEEE Computer Society who are interested in database systems. The TCDE web page is http://tab.computer.org/tcde/index.html.

The Data Engineering Bulletin

The Bulletin of the Technical Committee on Data Engineering is published quarterly and is distributed to all TC members. Its scope includes the design, implementation, modelling, theory and application of database systems and their technology.

Letters, conference information, and news should be sent to the Editor-in-Chief. Papers for each issue are solicited by and should be sent to the Associate Editor responsible for the issue.

Opinions expressed in contributions are those of the authors and do not necessarily reflect the positions of the TC on Data Engineering, the IEEE Computer Society, or the authors' organizations.

The Data Engineering Bulletin web site is at http://tab.computer.org/tcde/bull_about.html.

Letter from the Editor-in-Chief

Computer Society News

The Technical Committe on Data Engineering (TCDE) is part of the IEEE Computer Society, the largest society within the IEEE. As such it is governed by the technical committee framework established by the Computer Society. The Computer Society itself is governed by a Board of Governors (BOG) that decides policy for the society within the constraints permitted by the IEEE. I recently became a BOG member and have some (limited) first hand knowledge of what is happening in the Computer Society and how it may impact technical committees, e.g. TCDE, and the conferences that they sponsor.

The BOG met in the first week of June. At that meeting it adopted a new policy (which I played a role in originating) that directly impacts the TCDE and our flagship ICDE conference. The new policy divides conference surplus into equal thirds, going to conference (for the next conference instance), technical committee, and Computer Society. And, perhaps most importantly, it gives the TCDE the ability to accrue a balance that continues indefinitely. This will permit the TCDE to plan things on a longer time horizon than is currently possible. My hope is that it will stimulate more activity at the TC level of the Computer Society. This would be a very big plus for the entire organization.

The Current Issue

One of the spooky things about the world we live in is how much information is widely available about us and what might be done with it- either with or without our knowledge. One of the great things about this world is all the great services that can be provide to us based on data that is widely available about us. Interesting dilemma. Do we resist this world or seize it with gusto?

We, as citizens in this world need to understand both the personal data that is available and the services that we might usefully exploit based on that data. It is easy to understand the utility of location information– which we give up willingly (at least in circumscribed conditions) to enable GPS systems to provide us with directions. More difficult to get a handle on is social network style data. But varying kinds of machine learning operating over ”big data” make it possible to personalize experience, such as providing restaurant recommendations. Of course, this same data can be used to “foist” annoying ads your way. But no one ever promised that life would be either easy or simple.

The current issue explores this interaction of location data and social data. This is incredibly timely, both from a personal “they know that about me and how do I stop this” or “what a useful service” viewpoint and as a technical research area. As a research area, this is at the intersection of data platforms and data management on the one corner, and machine learning at another, and social media at a third. This is a great place to be as a researcher as there are several directions that are emphatically worth exploring. Xiaofang Zhou has assembled a great collection of papers in precisely this intersection of technical areas. The issue is by no means encyclopedic, but it is a very useful sampling of good work going on right now. Hence it is a great place to start when (1) trying to discover what companies might do with your data, and (2) for exploring research opportunities in this space. I want to thank Xiaofang for a great issue of the Bulletin on a topic of great current interest.

David Lomet
Microsoft Corporation

Letter from the Special Issue Editor

Among all types of big data available to us today, two types of data are in abundance and closely related with each other: social media data and location data. Both types of data are rich in information but are highly challenging to process. The best value derivable from these two types of data can be achieved once they are considered together. This special issue on the topic of location-based social media analysis reports the latest advances from leading researchers in this area for managing and analyzing these two type of data in a holistic approach.

The past decade has witnessed the phenomenal success of online social networks (OSN) with billions of users across various platforms, on which anyone is able to create and share any kind of information (news, articles, images, videos) to her connections, leading to a huge amount of social media data. The current pervasiveness of GPS-enabled mobile devices and the fact that all the giant social networks have also gone mobile have empowered people to add a location dimension to existing online social networks in a variety of ways. For example, users can upload geo-tagged photos/videos to Flickr, Instagram or Vimeo to share their great moment with friends, comment on an event in Twitter with geo-tagged tweets, share what they think about a restaurant on Foursquare, or log bicycle trails for sport analysis and experience sharing on Bikely. These kinds of location-embedded and location-driven social structures are known as location-based social networks (LBSN), while the geo-tagged social media is often referred to as location-based social media. Compared to traditional online social networks where peoples relationships in the virtual world may not necessarily correspond to those in the real world, the location dimension bridges the gap between online social networks (aka virtual world) and their real lives (aka real world). Moreover, as location is one of the most important components of user context, incorporating locational information while analyzing online social networks enables a deeper understanding of user preferences and behavior in the physical world. The enormous volume, fine granularity and heterogeneous formats of location-based social media have brought us unprecedented opportunities to, for the first time, study and understand humans social behavior with the scale and depth that could not possibly be achieved in the past. This special issue consists of seven articles from leading researchers geared towards the recent development and new frontiers of models, algorithms, applications and systems for location-based social media analysis.

The special issue starts with three survey-styled articles that review and summarize the challenges and state-of-the-art technologies in dealing with location-based social media data. The article Discovering Location Information in Social Media introduces some recent analytical techniques that leverage geographical information in social media to make recommendations and predictions. It then moves on to discussing how machine learning methods can be applied to infer the location of a social media post so that the prior analysis can be carried onto the entirety of social media data, rather than just those explicitly tagged with geographic information. In the second article Inferring Real-World Relationships from Spatiotemporal Data, the authors survey the related techniques pursuing the inference of the real-world social connections and social strength from spatio-temporal data that are generated from location-based social networks. The last article in this set Go Beyond Raw Trajectory Data: Quality and Semantics first points out the limitations of traditional techniques for processing raw trajectory data in coping with the spatio-temporal data in the context of LBSNs due to the lack of quality control mechanism and semantic information. The authors then review their recent work on enhancing the quality of trajectory data and utilizing the semantic information that are readily available in location-based social media to improve the interpretability of trajectory search results.

The next set of two papers mainly focuses on location-based social network mining by applying different machine learning models and techniques. In the article Mining Location-based Social Networks: A Predictive Perspective, the authors adopt supervised learning models to predict the future locations for users with regular mobility patterns and irregular mobility patterns respectively. Afterwards they discuss how to characterize the novelty-seeking propensity of LBSN users, which is used to prioritize the corresponding prediction models and rank the locations for recommendation. The paper Clustering in Geo-Social Networks, on the other hand, applies an unsupervised learning method (i.e., clustering) to find groups of places in an LBSN that share similar geo-

social attributes and structures, which can benefit applications like marketing campaign, urban planning, travel recommendation and so on.

The last two articles concern new techniques of analyzing microblogs by taking the space-time attribute into consideration. In the article entitled Space-Time Aware Behavioral-Topic Modeling for Microblog Posts, the authors model the topic of a microblog post where associated information in the form of timestamps, geo-locations and user interactions (i.e., reply, re-tweet) is available. The article Taqreer: A System for Spatio-temporal Analysis on Microblogs introduces their recent development for Taqreer, which is a scalable and efficient system for auto-generation of spatio-temporal analysis reports on large number of microblogs. Database technologies including indexing structures, flushing strategies, query optimization and recovery management have been employed and integrated into the query processing engine in order to deal with microblogs with high arrival rate and volume.

Irrespective of the nature of the papers, collectively they provide a good view of the state-of-the-art thoughts and research in the area of location-based social media analytics. I hope you enjoy reading these articles!

Xiaofang Zhou
The University of Queensland
Australia

Discovering Location Information in Social Media

Fred Morstatter, Huiji Gao, Huan Liu
Arizona State University
{Fred.Morstatter, Huiji.Gao, Huan.Liu}@asu.edu

Abstract

Social media is immensely popular, with billions of users across various platform. The study of social media has allowed for deeper inquiries into questions posed by computer scientists, social scientists, and others. Social media posts tagged with location have provided means for researchers to perform even deeper analysis into their data. While location information allows for rich insight into social media data, very few posts are explicitly tagged with geographic information. In this work, we begin by introducing some state-of-the-art analysis techniques that can be performed using the location of a social media post. Next, we introduce some systems that help first responders provide relief with the help of the location of social media posts. Finally, we discuss how machine learning techniques can be applied to infer the location of a social media post, bringing this analysis to any message posted on social media.

1 Introduction

Social media sites provide ways for their users to conveniently share their lives in real-time, from their current mood to the music they are listening to, and even information pertaining to their physical activities. Among the myriad ways to share information, the ability for users to share their location has come to the forefront of many sites. Sites such as Twitter and Facebook allow users to tag their posts with their current location, either with the venue or "place", or the exact GPS coordinates. Sites such as Foursquare have built their entire platform around users sharing their geographic information.

Increasingly, researchers and practitioners have found ways to make use of this new source of information for novel applications, such as recommending new venues to users, and predicting the number of people who will check in at a certain location. It has also been used to increase the effectiveness of existing problems such as helping deliver the right information to first responders in humanitarian assistance and disaster recovery.

In this work we discuss state-of-the-art challenges in leveraging geographic information in social media research. We begin by discussing how researchers use this information to make recommendations and to predict the next location a user will visit. Next, we discuss systems that have been created to help make sense of users mobility patterns in online social networks and to use these patterns to understand the greater picture of an event of disaster as it unfolds on social media. Finally, we introduce techniques that can predict a user's location in absence of explicit information on their post. These techniques have the potential to bring this analysis to the entirety of social media posts, and not just those explicitly tagged with geographic information.

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

2 Using Location Data for Novel Applications

2.1 Personalized Point of Interest Recommendation on Location-Based Social Networks

The rapid growth of cities has developed an increasing number of points of interest (POIs), e.g., restaurants, theaters, stores, and hotels, providing us with more choices of life experience than before. People are willing to explore the city and neighborhood in their daily life and decide "where to go" according to their personal interest and the various choices of POIs. At the same time, making a satisfying decision efficiently among the large number of POI choices becomes a touch problem for a user. To facilitate a user's exploration and decision making, POI recommendation has been introduced by location-based services such as Yelp and Foursquare. However, such recommendation models are commonly based on majority users' preference on POIs, which ignore a user's personal preference. Comparing to visiting places that best fit a user's interest, visiting places against a user's taste may give him very terrible experience, especially in a situation when the user travels to a new place. Therefore, personalized POI recommendation is proposed to help users filter out uninteresting venues according to their own taste and save their time in decision making.

2.1.1 Background

Before the Web 2.0 era, analyzing user's mobility for personalized POI recommendation is based on cellphone-based GPS data. Due to the lack of mapping information between geographical coordinates and real-world POIs on GPS data, a POI is usually determined by the stay points (geographical points at which a user spent sufficient long time) extracted from hundreds of users' GPS trajectory (a sequence of time-stamped latitude/longtitude pairs collected repeatedly at intervals of a short period) logs [25, 26]. With the rapid development of location-based social networking services, e.g., Foursquare, Yelp, and Facebook Places[1], users are able to check in at real-world locations with specific POI information and share such check-ins with their friends through mobile devices, resulting in more abundant information to improve personalized location recommendation.

This abundance of information has led to a new class of social network, called a "location-based social network". Location-based social networks not only refer to the social connections among users, but also consist of the "location-based" context including geographical check-in POIs, check-in time stamps, and check-in related content (e.g., tips, comments, POI descriptions, etc.), as shown in Figure 1. Compared with other online social networks that consist of user activities interacting with the virtual world, LBSNs reflect a user's geographical action in the real world, residing where the online world and real world intersect, therefore bridging the gap between the real world and the virtual world, providing both opportunities and challenges for researchers to investigate users' check-in behavior for personalized POI recommendation in spatial ("where"), temporal ("when"), social ("who") and content ("what") aspects.

In this work, we use POI, venue, and location as interchangeable terms. Let $\mathbf{u} = \{u_1, u_2, ..., u_m\}$ be the set of users and $\mathbf{l} = \{l_1, l_2, ..., l_n\}$ be the set of POIs where m and n are the numbers of users and POIs, respectively. The problem of personalized POI recommendation on LBSNs is defined as:

Given a user $u \in \boldsymbol{u}$, a set of POIs (locations) $\boldsymbol{l}_u \in \boldsymbol{l}$ that u has checked-in, recommend him some POIs for his future visits based on the LBSN context (e.g., social connections, content information of check-ins, time stamps of check-ins) related to him.

In the last decade, recommender systems have been widely studied among various categories, e.g., movie recommendation on Netflix, dating recommendation on Zoosk, item recommendation on Amazon. However, it is not sufficient to directly apply these technologies as personalized POI recommendation on LBSNs presents unique challenges due to the heterogeneous information layout and the specificity of human mobility. Designing efficient POI recommendation approaches on LBSNs inevitably needs to consider the following properties.

[1] http://www.facebook.com/about/location

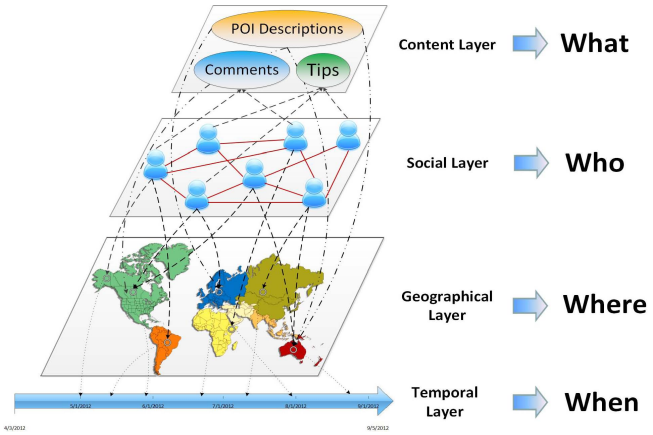


Figure 1: The information layout of location-based social networks. The geographical layer contains the historical check-ins of users, while the social layer contains social friendship information, and the content layer consists of user feedbacks or tips about different places. All these three layers share one timeline, indicating the temporal information of the user "check-in" behavior.

Geographical Property of Social Connections

Geographical property and social connections are coherent and affect each other in human behavior. For example, a user is more likely to be friends with other users who are geographically close to him, e.g, co-workers, colleagues. Likewise, a user may check-in at a location due to the influence from his friends, such as following friends' suggestions to visit a restaurant. Such coherence results in a new property, commonly referred to as socio-spatial properties [21]. Thus, considering the social information together with the geographical property enables us to capture the user preferences more precisely in personalized POI recommendation on LBSNs.

Temporal Patterns of Geographical Check-ins

Human geographical movement exhibits strong temporal patterns [3, 15, 24] and is highly relevant to the location property. For example, a user regularly goes to a restaurant for lunch around 12:00 pm, watches movie on Friday night, and shops during weekends. This is generally referred to as temporal cyclic patterns. Such temporal patterns are not widely observed in other recommender systems. For instance, it is not common to observe a user regularly watching a specific movie (e.g., Batman, Avatar) or purchasing a specific item (e.g., camera, cellphone) at specific hour of the day, or day of the week. (Although birthdays or holidays like Thanksgiving may affect human behavior a bit, they are not commonly considered).

Semantic Indications of Check-in Content

Content information on LBSNs could be related to a user's check-ins, providing a unique opportunity for location recommendation from a conceptual perspective. For example, By observing a user's comment on a Mexican restaurant discussing its spicy food, we observe if the user is interested in spicy food or not. This is an example of *user interests*. By observing a location's description as "vegetarian restaurant", we may infer that the restaurant serves "vegetarian food" and users who check-in at this location might be interested in the vegetarian diet. This is an example of *location properties*. These two types of information are representatives of user-generated content and location-associated content on LBSNs. The former refers to comments that left by users towards specific locations when they check-in; the latter can be descriptive tags associated with specific locations.

The above three properties indicate the three unique relationships between geographical information and so-

cial, temporal, and content information, commonly referred to as *geo-social correlations*, *geo-temporal patterns*, and *geo-content indications*. For more information, please refer to Gao and Liu 2015 [9].

2.2 Geolocated Information for Crisis Response Applications

In this section we discuss how geolocated social media data can be used to help with disaster relief. We introduce two classes of systems: crisis maps that help first responders match need with resources, and tools which help first responders get a deeper understanding of the situation.

2.2.1 Crisis Mapping

Crisis mapping consists of tools that help first responders to coordinate resources in times of disaster. Usually, the requests for assistance are obtained through SMS, as well as through social media. Twitter is often used in such applications [11].

Ushahidi [8] is one of the first crisis mapping systems. It has helped to coordinate relief in Kenya [16], Afghanistan, and Haiti. The system features a request engine that allows for those affected to seek out the resources they need for their specific situation. Volunteers and disaster relief organizations can then use this map to allocate aid and to see where their services can be of the most use.

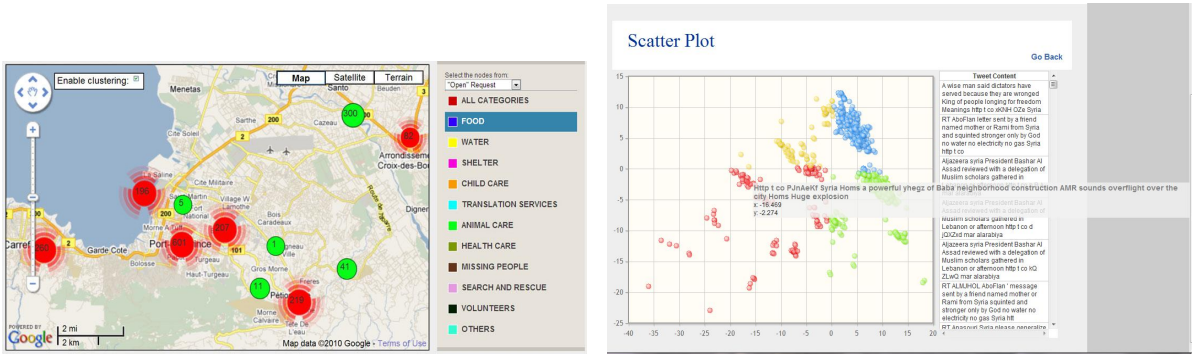
TweetTracker [10] enables a first responder to collect Twitter data pertaining to a crisis by specifying parameters about the crisis. These parameters can come in one of three forms: 1) keywords which describe words that pertain to the crisis, 2) geographical bounding boxes which specify the region or regions affected by the crisis, and 3) user names which can be users that tweet about the crisis. TweetTracker collects tweets that match any of these parameters and shows them to the first responder.

ASU Coordination Tracker (ACT) [7], is designed to collect crowdsourced requests, keep first responders aware of the current situation, and help them coordinate for disaster relief. The main goal of ACT is to analyze crowdsourced requests and promote inter-agency coordination to prevent duplication of effort during crisis. It comprises of five functional modules: request collection, request analysis and visualization, response, coordination, and situational awareness. Figure 2(a) shows an overview of classified requests and the quantity of requests on ACT's crisis map. ACT collects two types of requests: requests from crowds (crowdsourcing) and requests from groups (groupsourcing). Crowdsourcing refers to requests submitted by people (e.g. victims, volunteers) who are not from certified organizations. The groupsourcing [1, 6] requests originate from responding organizations such as United Nation, Red Cross, etc. Specially, crowdsourcing data are collected in forms of web, SMS and tweets collected through TweetTracker. The data analysis takes advantage of both data mining technology and expert knowledge to iteratively capture the essential content of raw requests and classify them into several categories (food, shelter, missing persons, etc.). Figure 2(b) shows a clustering visualization for expert labeling and decision making based on active learning techniques.

2.2.2 Visualizing Crisis Data

Another way that geolocated social media data has helped first responders is by giving them a picture of what is unfolding on the ground. While looking at the raw data is very difficult, several systems have emerged in the past few years to make sense of the massive volume of data that is generated during a crisis.

TweetXplorer [17] is a system that is designed to help first responders get situational awareness. An overview of the system is shown in Figure 3(a). Queries are issued to the system using the query pane on the bottom right. The network in the top right shows the most retweeted users in the dataset. The top left shows a map, which shows the locations that gave the most geotagged tweets about the user's query. This map can also be used to help the user's explore the data. The map can be combined with the network to show the geo-tagged retweets of a particular user, as shown in Figure 3(b). This can help the analysts understand which locations care the most



(a) ACT Request Window

(b) Request Classification and Visualization

Figure 2: ASU Coordination Tracker overview. This shows two important views of the system. On the left we see a request window which shows requests made by specific organizations classified in terms of quantity. On the right we see the active learning module which helps the human expert to classify requests.

about a particular tweet. The map can also be "brushed", causing it to show a tag cloud of the most important words in the selected region.

Another visual analytic system designed to help first responders understand events on the ground is SensePlace2 [13]. The system helps analysts find important tweets by allowing them to query in two ways: through keywords, and through a spatial filtering interface. In this way the users can find both the content they are interested in and where it originates. More details about the system can be found on the project web page[2].

3 Inferring Location Information in Social Media

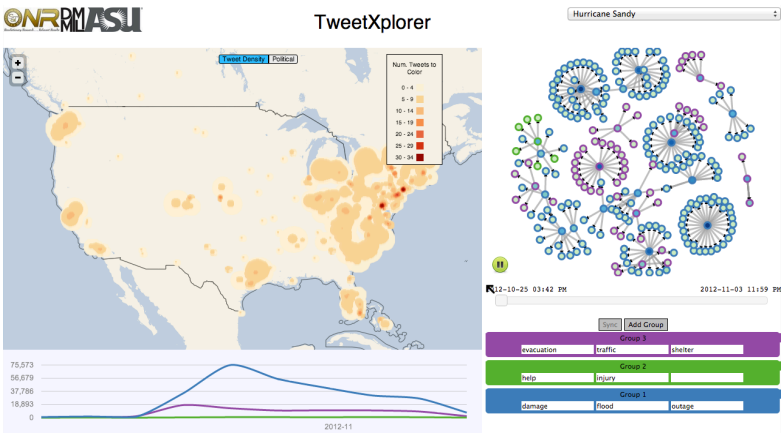
While location-based user analysis has taken off in recent years, the number of users providing their location has not kept pace. Only about 1% of all of the tweets posted on Twitter are geotagged [19]. This is largely due to Twitter's "opt-in" policy for providing user location. The disparity prevents existing techniques being used to assist the vast majority of users of a service who do not use their geographical information.

To bridge the gap analysis and users who lack location information, researchers have focused on uncovering the locations of users who do not share their location on social media. This location can be uncovered from three perspectives: the user's *profile location*, where he lives; the *tweet's location*, where the message was published; and the *event location*, where the message is talking about. In this section we discuss attempts that researchers have made to address these problems.

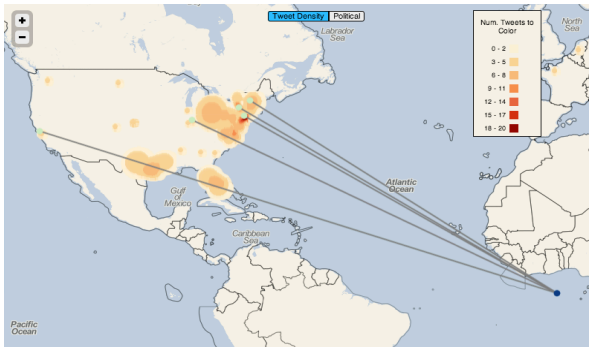
3.1 Inferring a User's Location

A "user's location" refers to the location where the user lives, or the location that he would give in his profile. The problem of user location prediction of Twitter users was first investigated in [4], where the authors used the language of the user's tweets to estimate his home location. The authors manually asses the statistical distribution of words in the tweets to find words that contained a "strong geo-scope". One example of such a word is "Red Sox", which occurs much more in the Boston, Massachusetts area than anywhere else in the USA. This work was extended in [2], where the authors propose an automated approach to finding words with a strong geo-scope. Linguists have also made an attempt at this problem looking for statistical variation in different geographic regions. In [5], the authors propose geographical topic models to model the language across the entire continental United States. [23] and [20] use a grid-based approach to define regions using the data.

[2] http://www.geovista.psu.edu/SensePlace2/



(a) TweetXplorer



(b) Social Network Overlaid on Map

(c) Geo-Aware Tag Cloud

Figure 3: Overview of TweetXplorer. The top pane shows an overview of the system. The bottom left shows how the network can be combined with the map to analyze who is retweeting the user. The bottom right shows a tag cloud that shows the most important words from a particular geographic area.

Researchers have approached this problem not just from one perspective (e.g. text, or network) but instead take a holistic approach trying to incorporate as much signal as possible into their models. In [22], the authors combine signals such as the tweeter's "location field" in his profile, any URLs in his biography (top-level domains such as .uk may provide a country indicator), and the time zone the user has set. This heuristic-based approach is furthered by [14], who adds more information including patterns of users posting time, and points of interest. This work also studies how different types of spatial aggregation and ensemble approaches can lead to better classification results.

In the context of disaster response, user location has been used to help first responders to find "eyewitness accounts": accounts that are both geographically near the disaster and discussing the topics of those affected. Kumar et. al 2013 [12] proposed a method for finding eyewitness accounts by measuring all of the users who tweet about a crisis along two dimensions: the location of their geotagged tweets, and their affinity for a set of topics. An example of these dimensions is shown in Figure 4. The users who score above-average on both dimensions, putting them in the upper-right quadrant, are considered "Q1" users. The authors find that the properties of Q1 users reflect the properties of eyewitness users: they tweet first on pressing topics and often relay relevant information that is not found in the other dimensions.

One important requirement of all of the above work is the number of tweets required in order for the approaches to make accurate predictions. The geo-scope approaches described in [2,4] requires at least 700 tweets for an accurate prediction, while the linguistic [5, 20, 23] and heuristic-based [14, 22] methods require approxi-

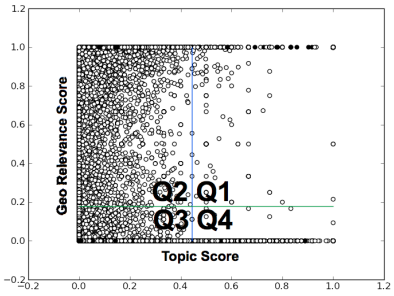
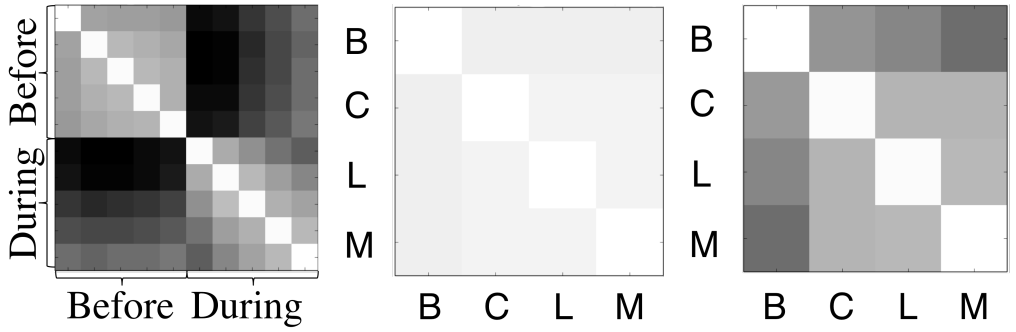


Figure 4: Quadrants used to find eyewitness users. Here, the authors focus on users with an above-average Geo Relevance Score (vertical axis), and an above-average Topic Score (horizontal axis).



(a) Temporal language differences. (b) Geographic language differences: tranquil time. (c) Geographic language differences: crisis time.

Figure 5: Temporal and geographic differences of language (calculated using Jensen-Shannon divergence); darker shades represent greater difference. To illustrate geographic differences, we compare Boston (B) with three other major U.S. cities: Chicago (C), Los Angeles (L), and Miami (M).

mately 200 tweets.

3.2 Inferring a Tweet's Location

While current approaches to user location prediction have shown promising results, one limitation is that they need a substantial history of a user's tweets in order to make accurate predictions. This much data is often unavailable for the vast majority of users on Twitter. Moreover, even for users who have posted this much information, it can be very difficult to collect this history under duress. Here we discuss alternatives to this problem, that allow users to geotag a single tweet. Often this is necessary in times of crisis when it is not feasible to collect a user's entire history to estimate his location.

Disaster response agencies often look to Twitter to understand what is unfolding on the ground in real time. To get a sense of the area most effected by the disaster, these agencies look at geo-tagged tweets. Since geo-tagged tweets only account for 1% of all activity on Twitter, these first responders are left looking for other methods to find a tweet's location. However, with the requirement of hundreds of past tweets for a particular user, existing methods to finding a user's location become infeasible during crisis situations.

In the absence of explicit geographic information, it is unlikely that a single tweet contains enough information to locate its exact position. Instead, to accommodate the lack of information, in [18] we change the problem to reflect what first responders are actually looking for during times of crisis: whether or not a tweet actually originates from within the crisis region. By simplifying the problem from predicting two continuous values to

predicting one boolean value, we make the problem more tractable with such sparse data.

To differentiate the users within a crisis region from those outside by the text of their tweet, we must first verify that the text that is generated from within a crisis region is actually different from the text outside of it. We perform this analysis along two dimensions: within the area of the crisis before and during the crisis, and during the time of the crisis across different locations. The results of this analysis are shown in Figure 5. Figure 5(a) shows the temporal difference by hour over the course of April 15, 2013, the day of the Boston Marathon Bombing. We see that the hours leading up to the bombing are much more similar than the hours after the bombing. Furthermore, in the location comparisons, we see that the cities are similar before the disaster (Figure 5(b)), and exhibit different behavior after the beginning of the crisis (Figure 5(c)). Thus, a linguistic difference exists between the linguistic patterns during the crisis within the crisis location.

Now that we have established that a difference exists between the locations during the crisis, we can continue to build a machine learning model that can capture these differences and aid first responders in finding tweets coming from within the crisis region. To do this, we hypothesize some linguistic features within the tweet that may be useful in identifying whether it originates from within the crisis region: *Word Unigrams and Bigrams*, *Part-of-Speech Tags*, *Shallow Parsing*, and *Crisis-Sensitive Features*. Crisis-sensitive features are some features identified by inspecting the text produced in the tweet. These consist of some part-of-speech patterns that are commonly observed in crises.

To test the effectiveness of our linguistic features, we build basic classifiers to test our features. The model then outputs its prediction of whether the tweet is inside region or outside region. We compare all possible combinations of individual feature classes and find that a combination of Unigram + Bigram + Crisis Sensitive features perform best for both crises.

We see that in both crises all of the top performing feature combinations still contain both the Bigram and Unigram feature classes. These classifiers massively outperform traditional approaches in the geolocation problem. This shows that inferring the tweet's location is possible, and that by modifying the problem to focus on the binary question of "within location" and "outside location" we are able achieve superior performance on this problem.

4 Conclusion

Social media is immensely popular, allowing users to share their lives in new ways. By allowing users to share their location, social media sites have enabled their users with richer means to express themselves. Location information is an important part of social media analysis, allowing researchers to obtain new insights into the behavior of users online and practitioners to develop new applications. In this work, we have shown how location can be leveraged to find users' interests and to predict what location a user will visit next. Furthermore, location can be leveraged to help those affected by disaster, both by helping them to find the right information and by making sure their requests for help are sourced to the correct agencies.

One of the main difficulties with studying location in social media is the lack of explicit information. This comes, in part, from the low number of users who share their information on social media sites. We have presented work which seeks to address this problem. We have also presented an algorithm that helps find users in the region affected by a crisis. By focusing only on whether the user is inside or outside the region, we are able to achieve higher performance than traditional approaches.

The problem of discovering location information in social media is a challenging one with a long way to go. Future work consists of finding the "event location", the location that the user is talking about. This can differ from both the user's tweet location and his user location. Another area for future work is location privacy. In discovering location, we may uncover the location of users who do not want to be discovered, such as users participating in protests. While existing approaches illuminate the potential for privacy concerns, future work will be to address them in a way that does not bring users into harms way. Additionally, data reliability is a

problem within the context of location discovery. Users providing fake and incorrect values for their location add noise to the data. Future work seeks to identify these fake and incorrect locations and remove them to increase the performance of the location discovery task.

Acknowledgments

This work is sponsored in part by Office of Naval Research (ONR) grant N000141410095 and by the Department of Defense under the Minerva Initiative through the ONR grant N000141310835.

References

[1] Geoffrey Barbier, Reza Zafarani, Huiji Gao, Gabriel Fung, and Huan Liu. Maximizing benefits from crowdsourced data. *Computational and Mathematical Organization Theory*, 18(3):257–279, 2012.

[2] Hau-wen Chang, Dongwon Lee, Mohammed Eltaher, and Jeongkyu Lee. @Phillies Tweeting from Philly? Predicting Twitter User Locations with Spatial Word Usage. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, ASONAM '12, pages 111–118, Washington, DC, USA, 2012. IEEE Computer Society.

[3] Z. Cheng, J. Caverlee, K. Lee, and D.Z. Sui. Exploring millions of footprints in location sharing services. In *Proceedings of the Fifth International Conference on Weblogs and Social Media*, 2011.

[4] Zhiyuan Cheng, James Caverlee, and Kyumin 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*, pages 759–768. ACM, 2010.

[5] Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. A Latent Variable Model for Geographic Lexical Variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287. Association for Computational Linguistics, 2010.

[6] H. Gao, G. Barbier, and R. Goolsby. Harnessing the crowdsourcing power of social media for disaster relief. *Intelligent Systems, IEEE*, 26(3):10–14, 2011.

[7] H. Gao, X. Wang, G. Barbier, and H. Liu. Promoting coordination for disaster relief–from crowdsourcing to coordination. *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 197–204, 2011.

[8] Huiji Gao, Geoffrey Barbier, and Rebecca Goolsby. Harnessing the crowdsourcing power of social media for disaster relief, 2011.

[9] Huiji Gao and Huan Liu. *Mining Human Mobility in Location-Based Social Networks*. Data Mining and Knowledge Discovery. Morgan & Claypool, 2015.

[10] Shamanth Kumar, Geoffrey Barbier, Mohammad Ali Abbasi, and Huan Liu. TweetTracker: An Analysis Tool for Humanitarian and Disaster Relief. *ICWSM*, 2011.

[11] Shamanth Kumar, Fred Morstatter, and Huan Liu. *Twitter data analytics*. Springer, 2014.

[12] Shamanth Kumar, Fred Morstatter, Reza Zafarani, and Huan Liu. Whom Should I Follow?: Identifying Relevant Users During Crises. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, HT '13, pages 139–147, New York, NY, USA, 2013. ACM.

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

[14] Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. Home Location Identification of Twitter Users. *Transactions on Intelligent Systems and Technology*, 5(3), 2014.

[15] E. Malmi, T.M.T. Do, and D. Gatica-Perez. Checking in or checked in: Comparing large-scale manual and automatic location disclosure patterns. *The 11th International Conference on Mobile and Ubiquitous Multimedia (MUM 2012)*, 2012.

[16] Patrick Meier and Kate Brodock. Crisis mapping kenyas election violence: Comparing mainstream news, citizen journalism and ushahidi. *iRevolution Blog, October*, 23, 2008.

[17] Fred Morstatter, Shamanth Kumar, Huan Liu, and Ross Maciejewski. Understanding Twitter Data with TweetXplorer. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 1482–1485, New York, NY, USA, 2013. ACM.

[18] Fred Morstatter, Nichola Lubold, Heather Pon-Barry, J"urgen Pfeffer, and Huan Liu. Finding eyewitness tweets during crises. *ACL Workshop on Language Technologies and Computational Social Science*, 2014.

[19] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M. Carley. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. *Proceedings of ICWSM*, 2013.

[20] Stephen Roller, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, and Jason Baldridge. Supervised Text-Based Geolocation using Language Models on an Adaptive Grid. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1500–1510. Association for Computational Linguistics, 2012.

[21] S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo. Socio-spatial properties of online location-based social networks. *Proc. 5th International AAAI Conference on Weblogs and Social Media*, 11, 2011.

[22] Axel Schulz, Aristotelis Hadjakos, Heiko Paulheim, Johannes Nachtwey, and Max Mühlhäuser. A multi-indicator approach for geolocalization of tweets. In *Proceedings of ICWSM*, 2013.

[23] Benjamin Wing and Jason Baldridge. Simple Supervised Document Geolocation with Geodesic Grids. In *ACL*, volume 11, pages 955–964, 2011.

[24] M. Ye, K. Janowicz, C. Mülligann, and W.C. Lee. What you are is when you are: the temporal dimension of feature types in location-based social networks. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 102–111. ACM, 2011.

[25] Vincent W Zheng, Yu Zheng, Xing Xie, and Qiang Yang. Collaborative location and activity recommendations with gps history data. In *Proceedings of the 19th international conference on World wide web*, pages 1029–1038. ACM, 2010.

[26] Y. Zheng, L. Zhang, X. Xie, and W.Y. Ma. Mining interesting locations and travel sequences from gps trajectories. In *WWW*, pages 791–800. ACM, 2009.

Inferring Real-World Relationships from Spatiotemporal Data

Cyrus Shahabi Huy Pham
Integrated Media Systems Center
University of Southern California
Los Angeles, CA 90089-0272

Abstract

The pervasiveness of GPS-enabled mobile devices and the popularity of location-based services have generated, for the first time, massive data that represents the movements of people in the real world at a high resolution, aka spatiotemporal data. Such collections of spatiotemporal data constitute a rich source of information for studying various social behaviors, and in particular, give a boost to the study of inferring the real-world social connections from spatiotemporal data. This article surveys the prominent techniques proposed for deriving social connections and social strength from spatiotemporal data and discusses their formulations.

1 Introduction

Social networks have been studied by social scientists since the pre-Internet era, and their relevance particularly increased in the last decade. We identify three periods in the study of social networks corresponding to the growth in the availability of data over time.

The very first period in social networks started back in 1970s [12] when social scientists realized that it was critical to understand the underlying network that portrays people's social connections and influence relationships. Such information is significant in the analysis of the propagation of information, innovations, practice, rumors and contagious infections, and also in commerce including target advertising and recommendations. However, in the pre-Internet era, the problem of identifying "*who is friend of whom*" was challenging, and studies on social networks in this earlier stage had to confine themselves to extremely small datasets [11], which mostly came from some social surveys at very limited scales.

The second period started along with the Internet revolution in the '90s through the development of web, when our lives have continually expanded to occupy virtual worlds [7]. Towards the end of the last decade, the research on social networks witnessed an explosion. To a large extent, this has been fueled by the spectacular growth of social media and online social networks, such as LinkedIn, Facebook and Twitter, which started in 2003, 2004 and 2006, respectively [11]. These giant networks have produced and continue to produce enormous datasets about hundreds of millions of online connected users in the form of social graphs. Therefore, the "*who is friend of whom*" question, which was a big challenge during the first period, suddenly became a cakewalk. The readily available social graphs collected from online social networks motivated social scientists to move far

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

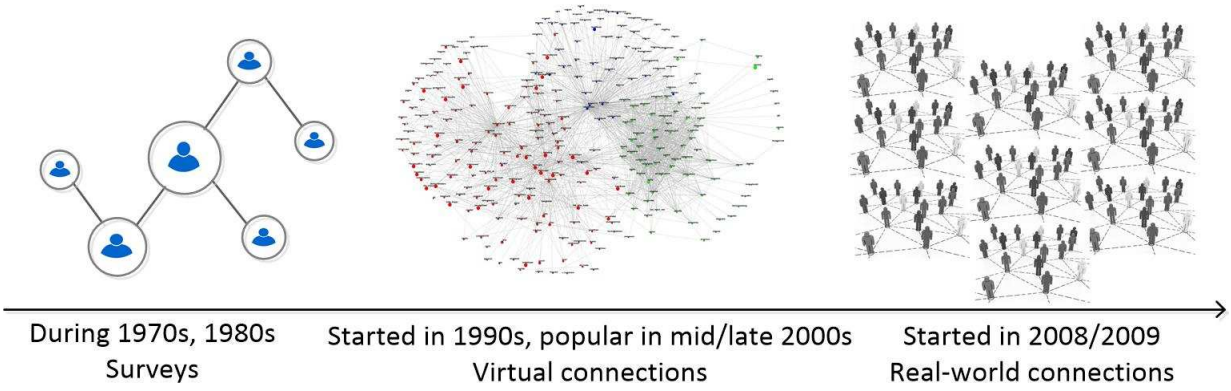


Figure 1: Three periods in social network studies.

beyond the basic question of "*who is friend of whom*" to much more interesting and sophisticated topics. As a result, a large number of studies has been devoted to new questions/solutions related to social networks, including measuring friendships quantitatively [13], identifying most influential people in a network [9], maximizing and speeding up the propagation of information and innovations in a social graph [10], and analyzing the structures and properties of a social network (e.g., density, clusters, stability, etc.) [15] [14]. However, all these achievements may still be considered inadequate in the eyes of the social scientists due to the gap that exists between online social networks (aka the *virtual world*) and the real lives (aka the *real world*). The large volume of studies during this period focused on the virtual world and utilized data collected from online networks. However, the people's relationships in the virtual world may not necessarily correspond to those in the real world.

Subsequently, we are now witnessing the third period as the phase of bridging the gap between the virtual world and the real world. Indeed, the pervasiveness of GPS-enabled mobile devices, and the fact that all the giant social networks have also gone mobile, has introduced massive data that represents the movements of people in the real world at high resolution, specifically by indicating *who* has been *where* and *when* (aka *spatiotemporal data*). Spatiotemporal data can be collected effortlessly from online services, such as geo-tagged contents (tweets from Twitter, pictures from Instagram, Facebook and Flickr, check-ins from Foursquare), or from mobile apps' data (WhatsApp, Glancee), etc. Such collections of spatiotemporal data constitute a rich source of information for studying and inferring various social behaviors, including social connections. For example, for social connections, the intuition is that if two people have been to the same places at the same time (aka *co-occurrences*), there is a good chance that they are socially related. Since these social connections are inferred from people's real world locations, they constitute social connections that occur in the real world, as opposed to those that may take place only in the virtual world.

The goal of this article is to survey the techniques pursuing the inference of the real-world social connections from spatiotemporal data during, what we called earlier, the third period of social networks.

2 Motivation

The ubiquity of mobile devices and the popularity of location-based services have generated, for the first time, rich datasets of people's location information at a very high fidelity. Just a few years ago, it was practically impossible to find any data that could describe people's locations at high resolution and large scale. However, this is no longer the case nowadays since smart phones and Location-Based Services have produced a tremendous corpus of rich spatiotemporal data. For example, Twitter and Foursquare reportedly received millions of spatiotemporal records per day as geo-tagged tweets or check-ins [16]. This newly available location data is useful for investigating various social behaviors, and thus has motivated social scientists to study and to extend the conventional concept of social behaviors to capture people's activities in the real world, particularly by inferring

the implicit networks of social connections based on the actual physical locations of people.

Furthermore, applications for such physically inferred networks of social connections are plenty. First, they include all the applications of online social networks such as marketing applications (e.g., target advertising, recommendation engines such as friendship suggestions), social studies (e.g., identifying influential people) and cultural studies (e.g., to examine the spreading patterns of new ideas, practices and rumors). In addition, the physically inferred social connections also have their own unique applications due to the geo-spatial properties. For example, the inferred social connections can be used to identify the new (or unknown) members of a criminal gang or a terrorist cell or it can be used in epidemiology to study the spread of diseases through human contacts.

3 Challenges

Inferring the implicit social connections is challenging for several reasons.

First, it is not clear what attribute of spatiotemporal data should be measured to infer social connections? If the frequency of co-occurrences (number of times that two people are seen together) is used as the indication of a social connection, one may arrive at a wrong conclusion about their social relationship. To illustrate, suppose two students study at the same library around the same time every day, which results in high frequencies of co-occurrences, but they may not even know each other. This erroneous conclusion can be attributed to coincidences - the fact that the library is a popular location and many students may co-occur frequently there by accident, and thus, the observation that two people only co-occur at the library is not a strong indication of a social connection. On the other hand, a few co-occurrences between two people in a small private place are perhaps a better indication of a friendship. Or alternatively, several co-occurrences at different popular places (e.g., coffeehouses, restaurants) may also be a better indication of friendships.

Second, it is of great interest to quantify social connections, and thus, the goal of inferring social connections from spatiotemporal data is not just to answer the true/false question, whether two people are friends with each other or not? It is more informative to infer a quantitative value that characterizes how strong a social connection is (aka ***social strength***). Lastly, spatiotemporal data is often extremely large, in the order of gigabytes of text, which could render the inference algorithms inefficient, taking too much time and/or resources to perform.

4 Solutions

In this section, we survey the methods proposed for inferring social connections from spatiotemporal data.

4.1 The report-based study

Eagle et al. were among the pioneers to look into the correlation between the location behaviors of users and their social connections. Specifically, in an early study [1], they conducted an analysis on two different sets of data of the same group of users, who were students at a university campus. One dataset collected from mobile phone, called "behavioral", which contained various features of user data, including the spatial proximity of users at work, their proximity on a specific night of the week, the phone communications between the users and the number of unique locations they were seen together [3] [1]. On the other hand, the other dataset was reported by users themselves, called "self-report", in which each user indicated who were his/her actual friends. Subsequently, a regression analysis was conducted over the behavioral dataset to find out possible friendships, which in turn were compared with the self-reported friendships. Their results showed that the social relationships extracted from the behavioral dataset were indeed related to the self-reported relationships. In addition, communications were the most significant predictor of friendships, followed by the number of common locations and spatial proximity.

4.2 Probability model

Crandall et al. [4] created a probability model to infer the probability of a friendship between two people given their co-occurrences in time and space. Specifically, they divided the surface of the earth into N grid-like cells, whose side lengths span s degrees of latitude and longitude. Two users are said to co-occur if they were present within the same cell within t days from each other. The number of unique locations (cells) of the co-occurrences between two people is the only factor used to determine the probability of their friendship. Multiple co-occurrences between two people within the same cell are not considered. Hence, the question becomes: What is the probability that two people have a social connection, given that they have co-occurrences in k distinct locations at a temporal range of t?

To formulate the friendship probability, assume that there are M people, each has one social tie, meaning one friend, and the social graph consists of $M/2$ disjoint edges. Each day, each pair of friends chooses to visit a place (i) *together* with probability β, and (ii) *separately* with probability $1-\beta$, with random choices of location. Let F denote the event that they are friends, and let C_k denote the event that they visit k unique locations together on k consecutive days. Consequently, the conditional probability $P(F|C_k)$ indicates the probability of two users being friends given that they co-occurred in k different locations on k consecutive days, which can be expressed by the Bayes' law and has the final formula as follows:

$$P(F|C_k) = \frac{P(F)P(C_k|F)}{P(C_k)} \tag{1}$$

$$= \frac{1}{M} e^{k \log \beta (N-1)+1} \tag{2}$$

The final formula in Equation 2 is obtained after computing the component factors in Equation 1: $P(F)$, $P(C_k|F)$ and $P(C_k)$, the details of which can be found in [4].

The advantage of this model is that it has a final, concise and simple expression for the friendship probability. The model only considers the number of unique locations where two users co-occurred, therefore it reduces the complexity of the algorithm significantly. The authors showed that even very few co-occurrences could lead to a sharp increase in the probability of a friendship, and this finding shows potential implications for the privacy of users on social media sites, which tells how much of the user data can be released until their privacy becomes exposed.

Despite achieving some promising results, the model still has several limitations. The first limitation is the simplifying assumption about the structure of the social network: each user can have only one friend, which is usually not the case in reality. Second, the model does not consider the frequency of co-occurrences at each location; all the co-occurrences at one location count only once. Finally, the issue related to coincidences was not addressed, that is whether the co-occurrences between two people are an indication of a social connection, or are simply coincidences between two people in time and space?

4.3 Trajectory-based model

Li et al. proposed the HGSM model that measures the similarity between two users based on the similarity between their trajectories [2]. The primary idea of this model is that the more similar the location histories of two users are, the more similar their common interests and preferences are, and thus the more likely that they are related socially. The HGSM model (Hierarchical-Graph-Based Similarity Measurement) first represents each user's location history as a trajectory (both sequentially and hierarchically), and the similarity between the trajectories of two users indicates their social similarity.

4.3.1 Trajectory

The *sequential* aspect of the trajectory allows the representation of discrete points in space (the locations visited by a user) as a continuous sequence. On the other hand, the *hierarchical* aspect allows for finer levels of geo-spatial granularity in a trajectory. For example, when a a *stay point* by a user contains multiple businesses located near each other, that stay point can be further turned into a subsequence of points with a smaller scale and a different (finer) level of granularity, and all such levels are said to form a hierarchy of granularity of location history of a user. A stay point is represented as a *cluster* of points at a smaller scale. Therefore, in HGSM, a user's trajectory consists of a set of graphs $HG = \{G\}$ built on different geo-spatial scales of the hierarchy, where each graph $G_i(V, E) \in HG$ is a set of vertexes $V = \{C\}$ (a set of clusters containing the user's stay points) and edges E.

The trajectories (aka sequences) of two users are presented as follows:

$$seq1 =< a_1(k_1) \overset{\Delta t_1}{\rightarrow} a_2(k_2) \overset{\Delta t_2}{\rightarrow} ...a_m(k_m) >$$

$$seq2 =< b_1(k'_1) \overset{\Delta t'_1}{\rightarrow} b_2(k'_2) \overset{\Delta t'_2}{\rightarrow} ...b_m(k'_m) >$$

where $a_i \in V$ is the cluster ID, k_i is the set of time the user successively visited cluster a_i, and Δt_i is the transition time the user traveled from cluster a_i to cluster a_{i+1}.

These two sequences are considered similar if and only if they satisfy two following conditions:

- For $\forall\ 1 \leq i \leq m$, $a_i = b_i$, meaning nodes a_i and b_i must share the same cluster ID.
- For $\forall\ 1 \leq i \leq m$, $|\Delta t_i - \Delta t'_i| \leq t_{th}$, where t_{th} is a pre-defined time threshold.

Under these two conditions, a common similar sequence ($sseq$) for the two users is defined as follows:

$$sseq =< b_1(min(k_1, k'_1)) \rightarrow b_2(min(k_2, k'_2)) \rightarrow ...b_m(min(k_m, k'_m)) >$$

m is therefore called the length of the common similar sequence of the two users.

4.3.2 Similarity measurement

The similarity of the two sequences $seq1$ and $seq2$ is determined by the measurement or the score of their common similar sequence ($sseq$), which depends on its length m:

$$s_{(m)} = \alpha_{(m)} \sum_{i=1}^{m} min(k_i, k'_i) \tag{3}$$

where $\alpha_{(m)}$ is an m-dependent coefficient, for which the optimal value is determined experimentally to be $\alpha_{(m)} = 2^{m-1}$.

At a single layer of the hierarchy, two users may have multiple, say n, common similar sequences. Their similarity at a single layer is determined by the following equation:

$$S_l = \frac{1}{N_1 \times N_2} \sum_{i=1}^{n} s_i \tag{4}$$

where N_1 and N_2 denote the numbers of stay points of the two users at the layer.

The overall similarity of the two users' trajectories across the set H of multiple layers of the hierarchy is:

$$S = \sum_{l=1}^{H} \beta_l S_l \tag{5}$$

β_l is a layer-dependent coefficient, determined experimentally to be $\beta_l = 2^{l-1}$.

The advantages of this model include the exhaustive representation of users' location histories as trajectories at very fine levels of geo-spatial granularity, and the similarity measurement of two trajectories as a quantity or strength of a friendship. The model clearly shows a high level of correlation between the human movements in the real world and their social relationships. One of the disadvantages of the model is the high complexity of constructing the users' trajectories. Another disadvantage is that the issue related to coincidences was not addressed, that is when two users happen to be in the same location by accident, and possibly on multiple occasions, such as in a crowded shopping mall. Such coincidences can contribute to the similarity between two trajectories of two unrelated users and may cause a misunderstanding of a social tie existing between them.

4.4 Feature-based model

Cranshaw et al. introduced various features extracted from spatiotemporal data that have connections with, or are indications of friendships, and thus inferred friendships based on such features [3]. Similar to the approach by Crandall et al. [4] presented in Section 4.2, in order to find the co-occurrences between people, the authors first divided the space into a grid-like cells, each is of approximately 30m each side. Two people are said to co-occur if they are present in the same cell within a time interval of 10 minutes. Various features of the co-occurrences between two people were introduced, which we summarize below.

4.4.1 Diversity of a location

The primary goal of studying the diversity of a location is to evaluate the impact of a co-occurrence between two people (aka. co-location) on the fact that whether they are friends or not. Specifically, the authors aimed to find out, whether a co-occurrence between two people happened by chance (aka a coincidence) or it happened as the result of a social connection between them. For example, the fact that two people shop at the same popular mall or dine at the same popular restaurant during the same time may happen by chance, and thus they are strangers to each other. On the other hand, co-occurrences between two people at a small place, where there are only a few people, are likely a good indication of a friendship. Thus, the popularity of the location of co-occurrences matters to the prediction of friendships. Three measures are introduced to measure the diversity of a location.

Frequency is the raw number of visits by people to the location. Obviously, the higher the frequency, the more popular the location is. *User-count* is the number of unique people who have visited the location.

Location Entropy measures the diversity of a location by taking into account both the number of of unique visitors to the location, and the relative proportions of their visits. Specifically, let l be a location, let $V_{l,u} = \{< u, l, t >: \forall t\}$ be the set of visits (aka check-ins or spatiotemporal records) in location l by user u, let $V_l = \{< u, l, t >: \forall t, \forall u\}$ be the set of all visits in location l by all users. The probability that a randomly picked check-in from V_l belongs to user u is $P_{u,l} = |V_{l,u}|/|V_l|$. If we define this event as a random variable, then its uncertainty is given by the Shannon entropy as follows:

$$H_l = -\sum_{u, P_{u,l} \neq 0} P_{u,l} \log P_{u,l} \tag{6}$$

This is called Location Entropy. A high value of the location entropy indicates a popular place with many visitors and is not specific to anyone. On the other hand, a low value of the location entropy implies a less popular place with few visitors, e.g., domestic houses, which are often non-crowded.

4.4.2 Features of co-occurrences

Various features of co-occurrences are introduced in this work. *Intensity* and *duration* features measure how actively (frequently) two users co-occurred, and for how long. *Location diversity* introduced in Section 4.4.1 is

characterized by three different measures: frequency, user-count and location entropy. These provide the basis for understanding the impact of co-occurrences on the friendship information. *Specificity* measures how specific a location is to the user pair who co-occurred in the location. This feature is inspired by *tf-idf* [3]; specifically, the specificity of a location to the user pair u_1 and u_2 is defined as the number of co-occurrences between them in the location divided by the total number of visits by all users in the location. Some other features are related to the structural properties, such as (a) the number of people who have co-occurred with both users, (b) that number divided by the number of people who co-occurred with either user, and (c) the total number of unique locations visited by both users together divided by the total number of unique locations visited by either of the users. In addition, the regularity of each user's routine was also measured, the details of which can be found in [3].

4.4.3 The inference of friendship information

As a final step, the above-mentioned features, together with the explicit friendships (the ground truth) are used to train different classifiers, including Random Forest, AdaBoost and Support-Vector machine. The experimental results in the study [3] showed that the Random Forest and AdaBoost classifiers outperformed the Support-Vector machine classifiers.

The advantages of this model include the consideration of the popularity of the locations of co-occurrences, its impact on friendship information, and the consideration of various features of co-occurrences in inferring friendships. There are two main disadvantages of this model. First, the model only infers the binary information of friendships, meaning whether two users are friends or not, but not the strength of a friendship as compared to the two models discussed in Sections 4.2 and 4.3. Second, the use of many features may lead to the difficulty of balancing their relative importance during the training of the classifiers.

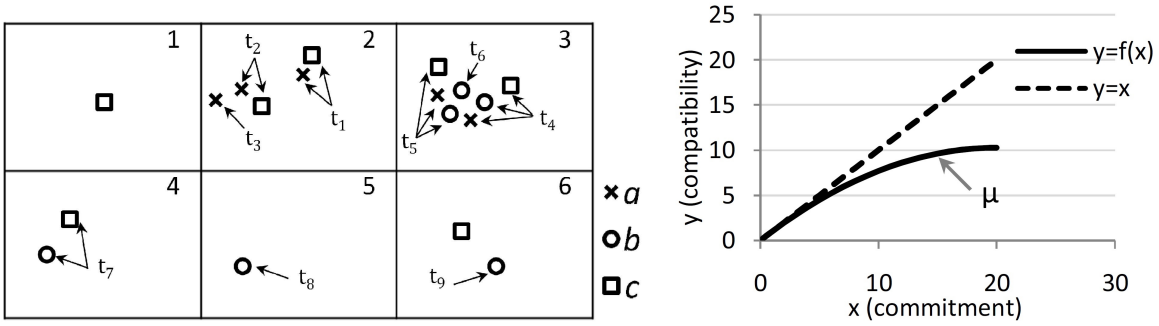
4.5 GEOSO model

In this model, Pham et al. took an entirely different approach to infer social connections from spatiotemporal data by trying to estimate the strength of people's relationships (aka ***social strength***) based on the geometric similarity of their visit patterns (i.e., who has been where and when) [5]. The authors introduced two properties: *commitment* and *compatibility*, which must be considered by any distance measure in order to correctly infer social strength from people's location behaviors.

Commitment is a phenomenon when two people repeatedly co-occurred at the same place on multiple occasions; the level of commitment is the number of times they co-occurred at the place. On the other hand, **compatibility** is a phenomenon when two people co-occurred at multiple different places; we say that two people are compatible to each other because they share a variety of common interests, which, in this case, are the places they co-visited. The main question is which, commitment or compatibility, is a better indication of a friendship? Intuitively, two close friends tend of hang out together in many different locations, and thus should co-occur in various places. On the other hand, if two people co-occurred frequently, but at only one place, they may or may not be friends because their co-occurrences may be coincidences. Therefore, the intuition is that compatibility should have more impact on social strength than commitment. We will see how GEOSO (standing for geo-social) model addresses this issue.

4.5.1 Data representation

Visit vector is a data structure that records the movement history of a user, specifically by indicating what places a person visited in the past, and at what time. To achieve this, the authors also divided the space into grid-like cells (see Section 4.2), where each cell has a unique ID. The grid is considered as a matrix, which is flattened into a vector by traveling from left to right and from top to bottom (row-first order). Correspondingly, for a given



(a) (b)

Figure 2: (a) Visit history of three users a, b and c. An arrow indicates the time that a user visited the cell. (b) Commitment vs. compatibility.

user, each dimension of the visit vector represents one cell of the grid, and the value of the dimension is a list of time showing when the user visited the cell.

For users a and b in Figure 2(a), their visit vectors are following:

$$V_a = (0, < t_1, t_2, t_3 >, < t_4, t_5 >, 0, 0, 0)$$

$$V_b = (0, 0, < t_4, t_5, t_6 >, t_7, t_8, t_9)$$

Co-occurrence vector: Two users are said to co-occur (or to have a co-occurrence) if they were present in the same cell within a time interval τ (a threshold that can be taken as 30 minutes). Correspondingly, a *co-occurrence vector* is a data structure that indicates how many times two users co-occurred, and where they co-occurred. For example, the co-occurrence vector between users a and c is $C_{ac} = (0, 2, 2, 0, 0, 0)$. The formal co-occurrence vector for any two users i and j has the following format:

$$C_{ij} = (c_{i1,j1}, c_{i2,j2}, ..., c_{ik,jk}, ..., c_{iN,jN}) \tag{7}$$

where $c_{ik,jk}$, called ***local frequency***, denotes the number of co-occurrences between users i and j at cell of ID k.

Next, imagine there are two users $\hat{i}$ and $\hat{j}$, who co-occurred more frequently than any other user pairs, both in the number of times they co-occurred in each cell, and in the number of unique cells, in which they co-occurred. Undoubtedly, this user pair would represent the strongest possible social connection among all the user pairs, assuming that the social strength is derived from co-occurrences only. We call the co-occurrence vector of $\hat{i}$ and $\hat{j}$ the **Optimal Vector** (or the Master Vector), which is defined as follows:

$$M = (m, m, ..., m), \qquad m = max\{c_{ik,jk}\}, \quad \forall i, j, k \tag{8}$$

m is the maximum local frequency among all the user pairs in all locations. Note that M is a conceptual co-occurrence vector; there may or may not exist a user pair with the co-occurrence vector M. The useful information we obtain from M is that its length indicates the maximum possible commitment, while its direction corresponds to the maximum possible compatibility.

4.5.2 GEOSO distance measure

The social distance d_{ij} between users i and j is defined by the Pure Euclidean Distance (PED) between the co-occurrence vector c_{ij} and the optimal vector M. The shorter the PED distance, the closer c_{ij} and M are (in

both direction and length), and thus the stronger the social connection. Social strength s_{ij} is therefore defined as the inverse of the distance metric d_{ij}.

$$d_{ij} = \sqrt{\sum_{k}(c_{ik,jk} - m)^2}, \quad s_{ij} = \frac{1}{(d_{ij}+1)} \tag{9}$$

In the denominator of the formula of s_{ij}, 1 is added to d_{ij} to make sure that s_{ij} does not become infinity when $d_{ij} = 0$. This also normalizes the value of social strength s_{ij} to the range $[0, 1]$.

4.5.3 Commitment versus Compatibility

The remaining task is to analyze the relative importance of commitment and compatibility using the social strength define by the GEOSO model. Assume that users i and j have only x co-occurrences in one cell (say cell 1), user p and q have only y co-occurrences, all of which took place in different cells; without loss of generality, we can assume that y co-occurrences took place in the first y cells. The co-occurrence vectors are: $c_{ij} = (x, 0, ..., 0)$, $c_{pq} = (1, 1, ..., 1, 0, ...0)$. Clearly, users i and j have pure commitment, while users p and q have pure compatibility. We are interested in knowing how much of x would be equivalent to y in the sense that they both create the same social strength? To achieve this, we equalize the social strengths $s_{ij} = s_{pq}$ in order to find the equivalence relationship between x (commitment) and y (compatibility). From equation $s_{ij} = s_{pq}$, it is not difficult to find that $y = (2mx - x^2)/(2m - 1)$. Figure 2(b) shows this relationship. It is clear that compatibility is more important than commitment as it has more impact on social strength. For example, $x = 20$ would be equivalent to $y = 10$ for both to produce the same value of social strength. This observation is consistent with our intuition as multiple co-occurrences in a single location might just be an indicator of coincidences, such as students study in the same library while they are not friends, and therefore should be limited in contributing to social strength. On the other hand, co-occurrences in multiple locations are seldom coincidences and therefore should have more impact on social strength.

GEOSO model is particularly interesting for introducing the two properties of co-occurrences: commitment and compatibility, and for evaluating their relative importance or impact on social strength. The geometric social distance is intuitive and creates a quantitative value for social strength instead of just indicating the binary information of friendships. The disadvantage is that all locations are considered equally important, meaning a co-occurrence in a private office can have the same impact on social strength as a co-occurrence in a crowded cafe or mall. This same problem also occurs in the models in Sections 4.2 and 4.3.

4.6 EBM model

By proposing the EBM (Entropy-Based Model) model to infer social strength from spatiotemporal data [6], the goal of the authors is to address all the issues that were unsolved or partially addressed in the former studies. These issues include (a) quantifying social connections, (b) discounting the impact of coincidences, (c) evaluating the impact of each co-occurrence depending on its location, (c) addressing the problem of data-sparseness, and (d) improving the efficiency.

The EBM model explores two independent ways: ***diversity*** and ***weighted frequency***, through which co-occurrences contribute to social strength. Specifically, diversity measures how diverse the co-occurrences between two people are in terms of locations, while weighted frequency measures the impact of each co-occurrence individually depending on the popularity of the location of the co-occurrence.

4.6.1 Diversity of co-occurrences

Consider the co-occurrence vectors for 3 different pairs of users:

$$C_{12} = (10, 1, 0, 0, 9\ )$$
$$C_{23} = (\ \ 2, 3, 2, 2, 3\ )$$
$$C_{13} = (10, 0, 0, 0, 10)$$

User 1 and User 2 have 20 co-occurrences, and User 2 and User 3 have only 12. However, in the latter case the co-occurrences are spread over 5 different locations, while in the former case the co-occurrences happened in just 3 different locations. Similarly, User 1 and User 3 co-occurred only in 2 different locations. Hence, C_{23} is *more diverse* than C_{12}, and C_{12} is *more diverse* than C_{13}.

Intuitively, people, who are socially connected, tend to visit *various* places together [4] [3] [1]. This intuition is captured as *how diverse* their co-occurrences are. Below is the definition of diversity of co-occurrences [6]:

Definition 1: Diversity is a measure that quantifies how many effective locations the co-occurrences between two people represent, given the mean proportional abundance of the actual locations.

The goal is to formulate the diversity of co-occurrences by using either Shannon entropy or Renyi entropy. First, let's define some notations.

Let $r_{i,j}^{l,t} =< i, j, l, t >$ be a co-occurrence of User i and User j in location l and at time t. Let $R_{ij}^{l} = \bigcup_t r_{i,j}^{l,t}$ be the set of co-occurrences of User i and User j, which happened in location l. R_{ij} is the set of all co-occurrences of User i and User j in all locations: $R_{ij} = \bigcup_l R_{i,j}^{l} = \bigcup_{l,t} r_{i,j}^{l,t}$

The probability that a randomly picked co-occurrence from the set R_{ij} happened in location l is $P_{ij}^{l} = |R_{ij}^{l}|/|R_{ij}|$. If we randomly pick a co-occurrence from the set R_{ij} and define its location as a random variable, then the uncertainty associated with this random variable is defined by the Shannon entropy for User i and User j as follows (the upper index S denotes *Shannon*):

$$H_{ij}^{S} = -\sum_{l} P_{ij}^{l} \log P_{ij}^{l} = -\sum_{l, c_{ij,l} \neq 0} \frac{c_{ij,l}}{f_{ij}} \log \frac{c_{ij,l}}{f_{ij}} \tag{10}$$

where $f_{ij} = \sum_l c_{ij,l}$ is the total number of co-occurrences of User i and User j, termed *frequency*, and $P_{ij}^{l} = \frac{c_{ij,l}}{f_{ij}}$ is expressed using the notation of the co-occurrence vector of User i and User j. Note the difference between *frequency* f_{ij} and *local frequency* $c_{ij,l}$; the *frequency* of two users is the sum of all their *local frequencies* across all locations.

Similarly, the uncertainty can also be expressed using Renyi entropy - a more generalized type of entropy with a flexibility to control the contribution of each component P_{ij}^{l}.

$$H_{ij}^{R} = \left(-\log \sum_{l} \left(P_{ij}^{l} \right)^{q} \right) / (q-1) \tag{11}$$

$$= \left(-\log \sum_{l} \left(\frac{c_{ij,l}}{f_{ij}} \right)^{q} \right) / (q-1) \tag{12}$$

where $q \geq 0$ is the order of diversity.

Generally, entropy is often regarded to as the *index* of diversity, but not diversity itself [8]. Diversity D is computed as the exponential function of entropy H. Specifically, $D = \exp(H)$. The expressions for diversity using each of the entropies above is:

$$D_{ij}^{S} = \exp\left(-\sum_{l, c_{ij,l} \neq 0} \frac{c_{ij,l}}{f_{ij}} \log \frac{c_{ij,l}}{f_{ij}} \right) \tag{13}$$

$$D_{ij}^{R} = \left(\sum_{l, c_{ij,l} \neq 0} \left(\frac{c_{ij,l}}{f_{ij}} \right)^{q} \right)^{1/(1-q)} \tag{14}$$

The upper index S denotes *Shannon*, R denotes *Renyi*.

Both Shannon entropy and Renyi entropy show how diverse a co-occurrence vector is in terms of locations. It is the *unpredictability* of the location of a co-occurrence. In other words, it is the amount of location information in the co-occurrences of two users. Therefore, their advantage is that its capture of diversity is consistent with the intuitions of friendships. First, the more locations, the higher the entropy. This is intuitive as the more places two users visited together, the stronger their connection. Second, the more uniform the distribution of the co-occurrences across locations (more equal proportion of co-occurrences in each location), the higher entropy. This is also intuitive for social strength, because close friends tend to hang out at various places together, thus their co-occurrences should be spread out over many locations, which results in more uniform co-occurrence vectors.

However, the disadvantage of Shannon entropy is that it may give higher importance to large components (aka outliers) of the co-occurrence vector because each component is weighted by its proportional abundance. For example, in co-occurrence vector $C_{12} = (10, 1, 0, 0, 9)$, 10 co-occurrences in the first cell is an outlier, which contributes more to the value of Shannon entropy as compared to the single co-occurrence in the second cell. This is not always a desired behavior that we want, because a high number of co-occurrences in a single crowded location may indicate coincidences, and their contribution to the social strength should, in fact, be limited rather than amplified.

On the other hand, Renyi entropy can effectively address the problem of coincidences. The elegance of using the Renyi entropy comes from the parameter q, called the ***order of diversity***, which indicates its ***sensitivity*** to the local frequency $c_{ij,l}$. Specifically:

- When $q > 1$ the Renyi entropy H_{ij}^{R} considers the *high* values of $c_{ij,l}$ more favorably. In other words, the higher the local frequency $c_{ij,l}$, the more impact the outliers have on Renyi entropy.
- When $q < 1$, instead, the Renyi entropy gives more weight to the *low* local frequencies $c_{ij,l}$.
- When $q = 0$, the Renyi entropy is completely ***insensitive*** to $c_{ij,l}$ and gives the pure number of co-occurrence locations - a.k.a. *richness*.
- When $q = 1$: As we know by now, the Renyi entropy favors local frequencies $c_{ij,l}$ in opposite directions when $q < 1$ versus when $q > 1$, therefore $q = 1$ is the *cross-over* point where Renyi entropy stops all of its biases and weighs the local frequencies $c_{ij,l}$ by their *own* relative proportions, which is what Shannon entropy does. Thus, at $q = 1$, Renyi entropy becomes Shannon entropy. Indeed, even though Equations (11) and (12) are *undefined* at $q = 1$, their limits exist when $q \to 1$ and become the Shannon entropy.

Advantages: The advantage of Renyi Entropy is its flexibility to limit or increase a particular behavior in co-occurrences. Particularly, it can reduce the impact of coincidences by setting parameter q to low values. An optimal value of q can be obtained experimentally if a ground truth is available. The readers are referred to [6] for how to obtain the optimal order of diversity experimentally.

4.6.2 Weighted frequency

While diversity measures the *breadth* of co-occurrences across locations, weighted frequency, on the other hand, measures the *depth* of co-occurrences and weighs each co-occurrence individually depending on the popularity

of the location. Weighed frequency utilizes Location Entropy, which was discussed in Section 4.4.1. The formula of weighted frequency is given as follows:

$$F_{ij} = \sum_{l} c_{ij,l} \times \exp(-H_l) \tag{15}$$

Weighted frequency tells us how important the co-occurrences at non-crowded places are to social connections. Crowed locations have high Location entropy H_l, thus low $\exp(-H_l)$, and consequently the impact of $c_{ij,l}$ on F_{ij} is decreased. On the other hand, for non-crowded locations, $\exp(-H_l)$ is high and this increases the impact of $c_{ij,l}$. The authors also provided more details about weighted frequency, including its comparison to *tf-idf*, and how weighted frequency addresses the problem of data sparseness [6].

4.6.3 Social strength

Finally, diversity and weighted frequency are combined to create social strength. Let s_{ij} be the ultimate social strength that captures both diversity and weighted frequency. A linear regression is conducted:

$$s_{ij} = \alpha.D_{ij} + \beta.F_{ij} + \gamma \tag{16}$$

where D_{ij} and F_{ij} are defined in Equations (14) and (15), respectively. Parameters α, β and γ can be either learned from dataset, or provided by users, or provided as application-dependent parameters. As a good practice, s_{ij} is generally normalized to $[0, 1]$. The information about how to obtain the parameters of the regression can be found in [6].

The advantages of the EBM model include the capture of the intuition of social connections in co-occurrences, specifically by measuring the diversity of co-occurrences in terms of locations and the weighted frequency. While the diversity offers a flexible mechanism of eliminating the impact of coincidences through Renyi entropy, weighted frequency takes into account the impact of each individual location of co-occurrences by analyzing the popularity of each location through Location Entropy. In general, all the main concerns of the former models we pointed out in the previous sections have been effectively addressed by the EBM model.

5 Conclusion

In this article, we surveyed the solutions proposed for inferring the real-world social connections from spatiotemporal data. Toward this end, we presented various models in details; for each model, we discussed the key ideas/intuitions of how social connections are linked to the location history of users. We also explained the main formulations of social connections and social strength for each model, together with its advantages and disadvantages.

This line of research opens a number of opportunities for future work. For example, the inferred real-world social connections and their strengths can be used to further study other aspects of social networks, such as social influence and information propagation among people in the real world. It is also possible to investigate the type of each social connection, whether two people are in a casual friendship, colleagues or in a family relationship, based on the semantics of the locations, in which they co-occurred. The real-world social connections can also be applied in other fields of study, such as in epidemiology to study the spread of disease through human contacts, or in criminology to investigate the nature, causes, patterns and consequences of a criminal behavior.

6 Acknowledgements

This research has been funded in part by NSF grants IIS-1115153, IIS-1320149, and CNS-1461963, the USC Integrated Media Systems Center (IMSC), and unrestricted cash gifts from Google, Northrop Grumman, Microsoft, and Oracle. Any opinions, findings, and conclusions or recommendations expressed in this material are

those of the author(s) and do not necessarily reflect the views of any of the sponsors such as the National Science Foundation.

References

[1] Eagle, Nathan, Alex Sandy Pentland, and David Lazer. "Inferring friendship network structure by using mobile phone data." Proc. of the National Academy of Sciences 106, no. 36 (2009): 15274-15278.

[2] Li, Quannan, Yu Zheng, Xing Xie, Yukun Chen, Wenyu Liu, and Wei-Ying Ma. "Mining user similarity based on location history." Proc. of the 16th ACM SIGSPATIAL, p. 34. ACM, 2008.

[3] Cranshaw, Justin, Eran Toch, Jason Hong, Aniket Kittur, and Norman Sadeh. "Bridging the gap between physical location and online social networks." Proc. of the 12th ACM Ubicomp, pp. 119-128. ACM, 2010.

[4] Crandall, David J., Lars Backstrom, Dan Cosley, Siddharth Suri, Daniel Huttenlocher, and Jon Kleinberg. "Inferring social ties from geographic coincidences." Proc. of NAS 107, 22436-22441, 2010.

[5] Pham, Huy, Ling Hu, and Cyrus Shahabi. "Towards integrating real-world spatiotemporal data with social networks." Proc. of the 19th ACM SIGSPATIAL, pp. 453-457. ACM, 2011.

[6] Pham, Huy, Cyrus Shahabi, and Yan Liu. "Ebm: an entropy-based model to infer social strength from spatiotemporal data." Proc. of the 2013 ACM SIGMOD, pp. 265-276. ACM, 2013.

[7] http://mashable.com/2012/01/09/real-world-digital-world/

[8] Jost, Lou. "Entropy and diversity." Oikos 113, no. 2 (2006): 363-375.

[9] Kempe, David, Jon Kleinberg, and Eva Tardos. "Maximizing the spread of influence through a social network." Proc. of the ninth ACM SIGKDD, pp. 137-146. ACM, 2003.

[10] Leskovec, Jure, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne VanBriesen, and Natalie Glance. "Cost-effective outbreak detection in networks." Proc. of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 420-429. ACM, 2007.

[11] Chen, Wei, Laks VS Lakshmanan, and Carlos Castillo. "Information and influence propagation in social networks." Synthesis Lectures on Data Management 5, no. 4 (2013): 1-177.

[12] Rogers, Everett M., and F. Floyd Shoemaker. "Communication of Innovations; A Cross-Cultural Approach." (1971).

[13] LibenNowell, David, and Jon Kleinberg. "The linkprediction problem for social networks." Journal of the American society for information science and technology 58, no. 7 (2007): 1019-1031.

[14] Ugander, Johan, Brian Karrer, Lars Backstrom, and Cameron Marlow. "The anatomy of the facebook social graph." arXiv preprint arXiv:1111.4503 (2011).

[15] Wu, Peng, and Dan Tretter. "Close and closer: social cluster and closeness from photo collections." Proc. of the 17th ACM international conference on Multimedia, pp. 709-712. ACM, 2009

[16] Noulas, Anastasios, Salvatore Scellato, Cecilia Mascolo, and Massimiliano Pontil. "An Empirical Study of Geographic User Activity Patterns in Foursquare." ICwSM 11 (2011): 70-573.

[17] Roussopoulos, Nick, Stephen Kelley, and Frdric Vincent. "Nearest neighbor queries." In ACM sigmod record, vol. 24, no. 2, pp. 71-79. ACM, 1995.

[18] Sharifzadeh, Mehdi, and Cyrus Shahabi. "The spatial skyline queries." Proc. of the 32nd international conference on Very large data bases, pp. 751-762. VLDB Endowment, 2006.

Go Beyond Raw Trajectory Data: Quality and Semantics

Kai Zheng Han Su
The University of Queensland, Brisbane, Australia
{uqkzheng, h.su1}@uq.edu.au

Abstract

Past decades have witnessed extensive studies from both academia and industries over trajectory data, which are generated from a diverse range of applications. Existing literature mainly focuses on raw trajectories with spatio-temporal features such as location, time, speed, direction and so on. Recently, the pervasive use of smart mobile devices like smart phones, watches and bands have brought about more generation of trajectory by personal users (instead of companies or organizations) and from online space (instead of physical space), where individuals can decide when and where to log on and share their locations with others. The more discentralized and contextualized trajectory sources have brought some unique challenges for database management with respect to the quality and semantics of trajectories data. With more applications and services relying on trajectory data analysis, it is necessary for us to think about how these new issues will affect the traditional way that trajectories are digested and processed. In this paper we will elaborate on these challenges and introduce our recent progress in the respective directions. The message we try to deliver is that raw trajectories themselves no longer satisfy the requirement of today's mainstream applications. To really release the power of trajectory-based applications, we should go beyond the raw trajectory data by enhancing their quality and semantics, which calls for novel computing architectures, paradigms and algorithms with sufficient capabilities to manage and analyse the enhanced trajectory data.

1 Introduction

The increasing availability of location-acquisition technologies including telemetry attached on wildlife, GPS set on cars, WLAN networks, and mobile phones carried by people have enabled tracking of almost any kind of moving objects, which results in huge volumes of spatio-temporal data in the form of trajectories [36]. Trajectory data consists of rich information about when and where a particular moving object is and offers unprecedented opportunity for discovering its mobility patterns. This inspires tremendous amount of research in trajectory data from a variety of aspects in the past decade, ranging from designing effective indexing structures [24] [8] [22] [9] [13] and efficient query processing algorithms [24] [29] [11] [14], to data mining and knowledge discovery [19] [16] [15] [21]. Despite their significant contributions in this area, traditional research on trajectory data has primarily focused on its raw format, i.e., a sequence of spatio-temporal points collected directly from the location-acquisition devices. While there was nothing wrong with this research philosophy especially back in the days when the source and scale of trajectory data are quite limited, recent advances in sensor technologies and

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

location-based social networks (LBSN) have posed new challenges to this community, which can be summarised in the following two aspects.

- **Challenge 1: Data Quality.** Although a trajectory can be theoretically modelled as a continuous function mapping from time to space, in a database it is actually a discrete sequence of spatio-temporal locations sampled from the movement of an object. In other words, when a raw trajectory is reported to the server and stored, it is just a sample of the original travel history. Therefore different sampling rates can result in completely different raw trajectories even for the same travel history. Since the sources of trajectory data are so diversified nowadays, the sampling rates vary significantly from one application to another. As a few examples, a geologist equipped with specialized GPS-devices can report her locations with very high frequency (e.g., every second) while a casual mobile phone user may only provide one location record every couple of hours or even days (via, for example, a check-in service in LBSN). Such variations can also be imposed by external factors (such as availability of on-device battery and wireless signal) and may change at the users discretion. In this big data era, it is not uncommon that we need to integrate trajectories across multiple sources and analyse them altogether. Nevertheless, our previous study [26] has shown that the great variance in sampling rates can render existing trajectory distance functions (e.g., DTW [17], LCSS [29] or EDR [10], ERP [11]) ineffective, which will in turn affect the algorithms, systems and applications relying on those distance functions. From database perspective, this essentially is a data quality problem that can be present in many analytical tasks involving multi-sourced and heterogeneous data. Systematic approaches are desired in order to gain deeper understanding of its root cause and eventually develop a comprehensive solution.

- **Challenge 2: Data Semantics.** Recent years have witnessed the flourish of location-based social networks (LBSN) that enables people to add a location dimension to existing online social networks in a variety of ways. For example, users can upload geo-tagged photos/videos to Flickr [2], Instagram [4] and/or Vimeo [6] to share their great moment with friends, comment on an event in Twitter [5] with geo-tagged tweets, check-in at a restaurant on Foursquare [3], or log bicycle trails for sport analysis and experience sharing on Bikely [1]. The location dimension serves as glue in LBSN that bridges the gap between physical and digital world. In other words, by aggregating all the geo-tagged contents posted by a user in her cyber-space (i.e., LBSN), we can actually know not only where and when she has been, as in the traditional trajectory database, but also what she was doing by extracting the information from the multimedia contents attached to the locations (e.g., text, images, videos). Moreover, we can even transform raw trajectories collected from GPS modules to semantic trajectories by applying semantic annotation techniques [7, 30]. With such a large volume of trajectory data enriched with semantic and activity information, we are confronted with challenges in terms of managing, analysing and understanding it. Due to the complex and combinatorial nature of this data, techniques across multiple areas including database, multimedia, data mining and natural language processing should be considered.

In this paper we will categorize and introduce our recent progress that has been made with respect to the above challenges. Generally speaking, we have found that the knowledge derived from raw trajectories is quite limited in most cases and even misleading sometimes. We believe the mainstream location-based services should base themselves on a higher level of abstraction for trajectory data, the one that has been dedicatedly processed to acquire better quality and more semantics. Figure 1 demonstrates our opinion about the relative positions of raw trajectories and enhanced trajectories in modern trajectory data management systems and applications. In the remainder of this paper, we will focus on explaining the research philosophy of our work and their relationships, while referring interested readers to the original papers for technique details.

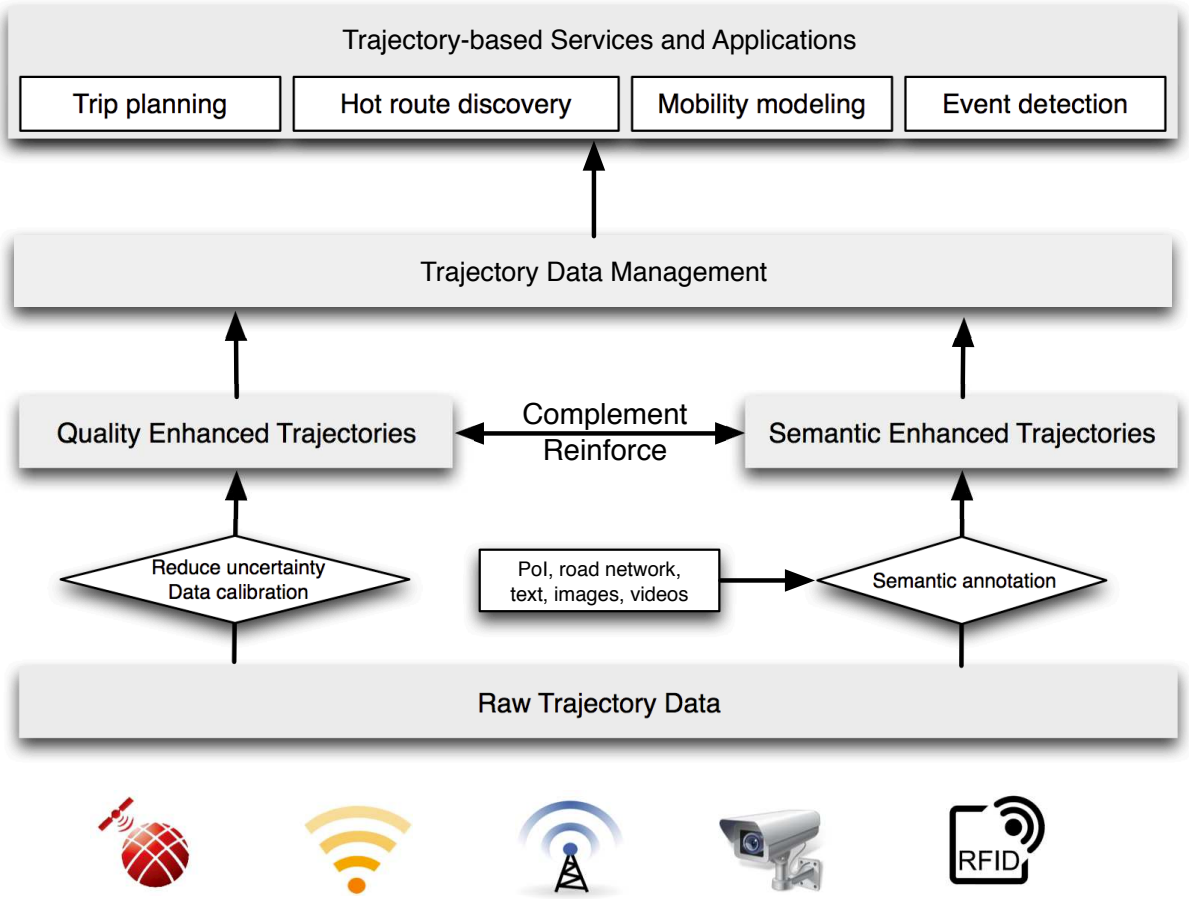


Figure 1: This figure demonstrates the relationship between raw trajectories, quality enhanced trajectories and semantic enhanced trajectories. It also illustrates their relative positions in trajectory data management systems and trajectory-based applications.

2 Quality-Aware Trajectory Management

Nowadays trajectory data can be generated from highly diversified services and applications, resulting in data with different qualities. Generally trajectory data quality issues can arise from two levels: point level and trajectory level. The first one is caused by the inaccuracies of location-acquisition devices and systems, i.e., the reported location is deviated from its actual location. Although this issue seems inherent and inevitable, we normally do not regard it as a major problem due to rapid advances of tracking technologies (e.g., GPS with sub-meter precision). Our focus is then on the second level, which is caused by the sampling rates of trajectories. As mentioned before, a trajectory in database is just a sample of its original travel history. Because nothing is known about the objects' whereabouts in-between two consecutive sampled locations, a trajectory is of low quality or high uncertainty if its sampling rate is low. To deal with an object's location in-between those samples, a typical technique is to apply interpolation [20] by which means the sampled positions become the end points of line segments, and the trajectories are transformed into polylines in 3D (*x-y-time*) space. However, as pointed out by [12, 18, 23], interpolation cannot reflect the exact movement pattern of an object. In theory, a moving object can be located anywhere within a given (bounded) region, as long as it does not violate physical constraints (e.g., maximum allowed velocity). Some efforts have been made to consider this issue when processing trajectory data by proposing probabilistic queries [28, 31] that, instead of reporting the result only,

provides the confidence of the result being true as well. However the quality of the trajectories cannot benefit or be improved from those approaches. In this section, we will introduce our methodologies to tackle this problem – enhancing the trajectory quality, which we believe to be more fundamental and efficient solutions compared to the expensive probabilistic queries.

2.1 Enhancing Trajectory Quality by Reducing Uncertainty

In [32] we aimed at reducing the uncertainty of a trajectory with low sampling rate, which is the main cause of trajectory quality issues. More specifically, given a low-sampled trajectory, our goal is to estimate its original and complete route/path on the underlying road network. At the first glance this seems a mission impossible if we simply act on each low-sampled trajectory separately since no better estimation can be done than linear interpolation for consecutive samples. However we have made two important observations based upon our analysis on real data. First, *travel patterns between certain locations are often highly skewed.* This is due to the fact that, when people travel, they often plan the route based on the experience of their own or others, rather than choosing a path randomly. The skewness of travel pattern distribution makes it feasible to distinguish the possible routes based on their popularity. The second observation, which is more interesting, is that *similar trajectories can often complement each other to make themselves more complete.* This implies that if we consider these trajectories collectively, they may reinforce each other to form a more complete route. These two observations show that the original route of a low-sampled trajectory can be estimated to some extent if a set of historical trajectories within the same spatial domain is available. Now the question is how to leverage this historical data. Intuitively, given a low-sampled trajectory, one can simply search for the historical trajectories that pass by all the sampled locations of the given trajectory and then find the most popular routes. Nevertheless, since the given trajectory can have arbitrary locations, we usually cannot find any historical trajectory that matches the whole part of the query very well. Even if we can, the amount may not be large enough to serve as reliable statistics. Therefore we propose a more practical solution consisting of three steps. Firstly, we divide the whole query into a sequence of sub-queries and search for the *reference trajectories* that can give hints on how each sub-query travels. Then we infer the *local routes* for each sub-query by considering the reference trajectories in a collective manner. At last, we connect consecutive local routes to form the *global routes* and return the ones with the highest scores to the users. As a summary, the essence of the route inference approach in this paper is to extract the travel pattern from history, and infer the possible paths of the query by suggesting a few popular routes. Compared to the original number of possible routes, the uncertainty of the given trajectory is reduced significantly in this way. Please refer to [32] for the detailed algorithms.

2.2 Enhancing Trajectory Quality by Data Calibration

Data quality issues do not just lie in low-sampled trajectories. In [26] we observed that trajectories with inconsistent sampling rate (no matter low or high) are almost incomparable and make the most classical trajectory distance functions less effective. To address this problem, in [26] we take a different philosophy that, instead of manipulating or adjusting the original trajectory data, uses a fixed and data independent set of spatial objects (called reference system) to re-write all the sampled locations of the original trajectories. This process is called trajectory calibration, the aim of which is to reduce the inconsistency in the sampling rates amongst all trajectories and improve the effectiveness of similarity-based trajectory analysis. Nonetheless it is a non-trivial task to perform trajectory calibration. First, building a good reference set is the stepping stone for the entire system. Since our goal is to rewrite the trajectory data using the reference set, we expect a good reference set to be stable, independent of data sources, and have a strong association with the trajectory data. The first and second properties are essential for producing trajectories in a unified form, while the third property ensures that the calibration process will not introduce a large deviation from the original routes. Trajectory calibration may encounter three circumstances when rewriting a trajectory with the reference set: 1) a trajectory point may need

to be shifted and aligned onto the reference; 2) some trajectory points may need to be removed or merged (when the sampling rate is higher than necessary); 3) some new trajectory points may need to be inserted (when the sampling rate is too low), all in the context of the chosen reference system. Further, the criteria to judge the goodness of the calibration results need to be established, for the system to enforce efficiently and effectively and for the users to understand to what extent the calibration can improve the data analysis results. The calibration framework we proposed comprises two components: a reference system and a calibration method. For the first component, we present several reference systems by defining different types of anchor points (space-based, data-based, POI-based and feature-based), which are fixed small regions in the underlying space. A series of strategies are designed for the calibration component, including the methods to insert anchor points to trajectories in order to make them more complete without scarifying geometric resemblance to the original routs. Please refer to [26] for more technical details.

3 Semantic Enhanced Trajectory Management

A trajectory in its raw format is just a sequence of spatio-temporal locations (e.g., a GPS point is a triplet (longitude, latitude, timestamp)). Although a lot of research have been done towards mining interesting patterns from a collection of trajectories purely based on their spatio-temporal features [15, 16, 19, 21, 34, 35], the results from those mining algorithms are often hard to explain and interpret for humans.This is because raw trajectory data can only reveal *when* and *where* a person was but cannot tell *what* she was doing (i.e., activity) and *how* she went there (i.e., moving behaviour) without leveraging extra information at semantic level. There have been some preliminary studies that enrich GPS locations with semantic entities such as POIs, roads, regions, resulting in semantic trajectories or annotated trajectories [7, 25]. These work focused on how to determine the correct semantic label for each trajectory point when multiple entities are in its vicinity, i.e., the generation of semantic trajectory data. In this section, we will introduce our recent research in managing and processing trajectories that have been enhanced with semantic information.

3.1 Querying Semantic Trajectories

Even though semantic trajectories contain much more information than raw trajectories, their value cannot be derived and utilised until there is an appropriate way to store, manage and process this data efficiently to a large scale. In this light, we developed a database storage framework to support efficient indexing and query processing over activity trajectory [33], which is a specific kind of semantic trajectory with textual information (e.g., keywords, tags, short phrases) describing user's activity at each location. More precisely, we propose a novel similarity query for activity trajectories by incorporating both geometric distance and activity match into the similarity measure, with the goal of returning more meaningful results to the users. However, answering this new query turns to be a more challenging problem since just making use of either location or activity information for pruning the search space will result in bad query performance. Our approach to this problem starts with a novel grid index called *GAT*, which includes a hierarchy of cells for each activity, an inverted list of trajectories containing each activity within each cell, and a summarized sketch of activities for each trajectory. GAT keeps the advantage of hierarchical spatial index while avoiding the flaws of large "dead zones" when indexing trajectories by minimum bounding boxes. In addition, the index not only uses the local information on trajectory segments within the cells but also preserves some global information for the entire trajectory in the activity sketch, so that its pruning power can be boosted. On top of the index, we develop a best-first search strategy with tighter distance lower bound for all "unseen" trajectories in the database and an efficient algorithm to compute the distance between candidates and the query. The interested readers can find more details about the indexing structure and search algorithms in our paper [33].

3.2 Summarising Trajectories with Short Text

The common way to generate semantic trajectories is to mechanically replace the coordinate of each location with a semantic entity, which often yields excessive information for people to digest and interpret. Therefore another direction we have been working on recently is to find a more compact, expressive and interpretable way to represent semantic trajectories. Inspired by text summarisation in information retrieval, we in [27] proposed to use short text to summarise and represent a trajectory by leveraging a diversified source of auxiliary information (e.g., PoI, road network). We found the textual representation can be superior than raw and semantic trajectories in two aspects. First, as the output is a summarization rather than mechanical transformation of raw trajectories (like semantic trajectories), data volume will be reduced significantly. Second, despite of smaller data size, the information conveyed in the text are strategically focused on the most 'interesting' parts of the trajectories, thus making more sense for humans. A partition-and-summarization framework was proposed in our work. The partition phase tries to find an optimal partition according to the user's granularity requirement, which can minimize the variation of predefined features for the trajectory segments within the same partition. The purpose of this optimization is to use more compact representation to summarize each partition. In the summarization phase, we define a novel measure for the unusualness of each feature by employing the common patterns amongst other trajectories, and generate textual description for the most unusual features with a predefined template. Please refer to [27] for more details about this framework.

4 Concluding Remarks

In this paper we have discussed some new challenges in trajectory data management that were brought about by the emergence of location-based services and explosion of smartphone users. Particular attentions are paid on two aspects – quality and semantics, which are believed as vital dimensions to uncover the true value of trajectory data for government, businesses and personal users. We introduce some of our recent studies in addressing these issues from different angles and highlight the connection between our research and the conventional ones. We hope these discussions can trigger more research interest and efforts in developing modern computing platforms and data management systems for trajectories – one of the most ubiquitous and accessible data today.

5 Acknowledgements

This work has been partially supported by ARC Discovery Early Career Researcher Award (DE140100215).

References

[1] Bikely. `http://www.bikely.com/`.

[2] Flickr. `https://www.flickr.com/`.

[3] Foursquare. `https://foursquare.com/`.

[4] Instagram. `https://instagram.com/`.

[5] Twitter. `https://twitter.com/`.

[6] Vimeo. `https://vimeo.com/`.

[7] L.O. Alvares, V. Bogorny, B. Kuijpers, J.A.F. de Macedo, B. Moelans, and A. Vaisman. A model for enriching trajectories with semantic geographical information. In *GIS*, pages 1–8, 2007.

[8] Y. Cai and R. Ng. Indexing spatio-temporal trajectories with chebyshev polynomials. In *SIGMOD*, pages 599–610, 2004.

[9] V.P. Chakka, A.C. Everspaugh, and J.M. Patel. Indexing large trajectory data sets with seti. In *CIDR*, 2003.

[10] L. Chen and R. Ng. On the marriage of lp-norms and edit distance. In *VLDB*, pages 792–803, 2004.

[11] L. Chen, M.T. Özsu, and V. Oria. Robust and fast similarity search for moving object trajectories. In *SIGMOD*, pages 491–502, 2005.

[12] R. Cheng, D.V. Kalashnikov, and S. Prabhakar. Evaluating probabilistic queries over imprecise data. In *SIGMOD*, pages 551–562, 2003.

[13] P. Cudre-Mauroux, E. Wu, and S. Madden. Trajstore: An adaptive storage system for very large trajectory data sets. In *ICDE*, pages 109–120, 2010.

[14] E. Frentzos, K. Gratsias, N. Pelekis, and Y. Theodoridis. Nearest neighbor search on moving object trajectories. *SSTD*, pages 328–345, 2005.

[15] H. Jeung, H.T. Shen, and X. Zhou. Convoy queries in spatio-temporal databases. In *ICDE*, pages 1457–1459, 2008.

[16] H. Jeung, M.L. Yiu, X. Zhou, C.S. Jensen, and H.T. Shen. Discovery of convoys in trajectory databases. *Proceedings of the VLDB Endowment*, 1(1):1068–1080, 2008.

[17] J.B. Kruskal. An overview of sequence comparison: Time warps, string edits, and macromolecules. *SIAM review*, pages 201–237, 1983.

[18] B. Kuijpers and W. Othman. Trajectory databases: data models, uncertainty and complete query languages. *Journal of Computer and System Sciences*, 2009.

[19] J.G. Lee, J. Han, and K.Y. Whang. Trajectory clustering: a partition-and-group framework. In *SIGMOD*, page 604, 2007.

[20] José Antonio Cotelo Lema, Luca Forlizzi, Ralf Hartmut Güting, Enrico Nardelli, and Markus Schneider. Algorithms for moving objects databases. *The Computer Journal*, 46(6):680–712, 2003.

[21] Z. Li, B. Ding, J. Han, and R. Kays. Swarm: Mining relaxed temporal moving object clusters. *Proceedings of the VLDB Endowment*, 3(1-2):723–734, 2010.

[22] J. Ni and C.V. Ravishankar. Indexing spatio-temporal trajectories with efficient polynomial approximations. *TKDE*, 19(5):663–678, 2007.

[23] D. Pfoser and C.S. Jensen. Capturing the uncertainty of moving-object representations. In *SSD*, pages 111–131, 1999.

[24] D. Pfoser, C.S. Jensen, and Y. Theodoridis. Novel approaches to the indexing of moving object trajectories. In *VLDB*, pages 395–406, 2000.

[25] S. Spaccapietra, C. Parent, M.L. Damiani, J.A. De Macedo, F. Porto, and C. Vangenot. A conceptual view on trajectories. *Data & knowledge engineering*, 65(1):126–146, 2008.

[26] Han Su, Kai Zheng, Haozhou Wang, Jiamin Huang, and Xiaofang Zhou. Calibrating trajectory data for similarity-based analysis. In *SIGMOD*, pages 833–844. ACM, 2013.

[27] Han Su, Kai Zheng, Kai Zeng, Jiamin Huang, Nicholas Jing Yuan, and Xiaofang Zhou. Making sense of trajectory data: A partition-and-summarization approach. In *ICDE*. IEEE, 2015.

[28] G. Trajcevski, R. Tamassia, H. Ding, P. Scheuermann, and I.F. Cruz. Continuous probabilistic nearest-neighbor queries for uncertain trajectories. In *EDBT*, pages 874–885, 2009.

[29] M. Vlachos, D. Gunopoulos, and G. Kollios. Discovering similar multidimensional trajectories. In *ICDE*, page 0673, 2002.

[30] Mao Ye, Dong Shou, Wang-Chien Lee, Peifeng Yin, and Krzysztof Janowicz. On the semantic annotation of places in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 520–528. ACM, 2011.

[31] K. Zheng, G. Trajcevski, X. Zhou, and P. Scheuermann. Probabilistic range queries for uncertain trajectories on road networks. In *EDBT*, pages 283–294, 2011.

[32] K. Zheng, Y. Zheng, X. Xie, and X. Zhou. Reducing uncertainty of low-sampling-rate trajectories. In *ICDE*, 2012.

[33] Kai Zheng, Shuo Shang, Nicholas Jing Yuan, and Yi Yang. Towards efficient search for activity trajectories. In *ICDE*, 2013.

[34] Kai Zheng, Yu Zheng, Nicholas Jing Yuan, and Shuo Shang. On discovery of gathering patterns from trajectories. ICDE, 2013.

[35] Kai Zheng, Yu Zheng, Nicholas Jing Yuan, Shuo Shang, and Xiaofang Zhou. Online discovery of gathering patterns over trajectories. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1974–1988, 2014.

[36] Y. Zheng and X. Zhou. *Computing with spatial trajectories*. Springer, 2011.

Mining Location-Based Social Networks: A Predictive Perspective

Defu Lian[†§], Xing Xie[§], Fuzheng Zhang[§], Nicholas J. Yuan[§], Tao Zhou[†], Yong Rui[§]
[†]Big Data Research Center, University of Electronic Science and Technology of China
[§]Microsoft Research, Beijing, China
dove@uestc.edu.cn, {xingx,nichy,v-fuz,yongrui}@microsoft.com, zhutou@ustc.edu

Abstract

With the development of location-based social networks, an increasing amount of individual mobility data accumulate over time. The more mobility data are collected, the better we can understand the mobility patterns of users. At the same time, we know a great deal about online social relationships between users, providing new opportunities for mobility prediction. This paper introduces a novelty-seeking driven predictive framework for mining location-based social networks that embraces not only a bunch of Markov-based predictors but also a series of location recommendation algorithms. The core of this predictive framework is the cooperation mechanism between these two distinct models, determining the propensity of seeking novel and interesting locations.

1 Introduction

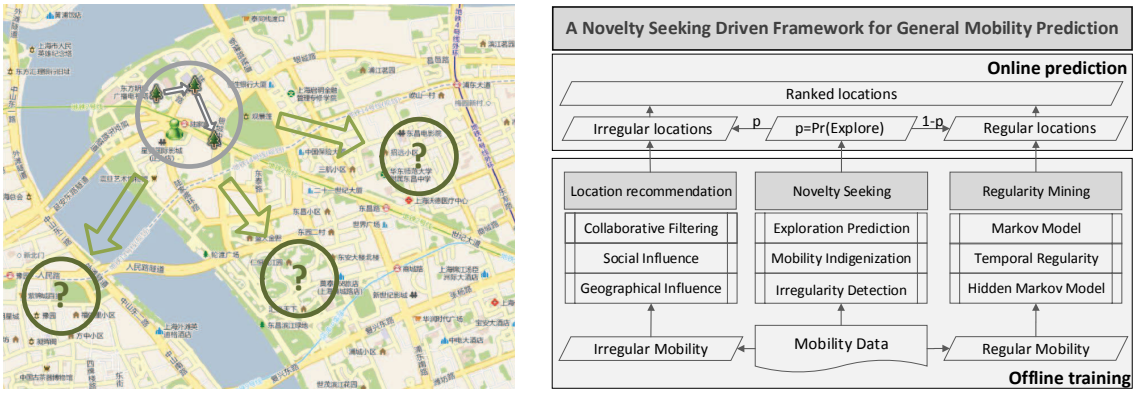
With the proliferation of smart phones and the advance in positioning technologies, location information can be acquired more easily than ever before. This development has led to the flourishing of a new kind of social network service, known as location-based social networks (LBSNs), such as Foursquare, Gowalla, and so on. In these LBSNs, people can not only track and share individual location-related information, but also learn collaborative social knowledge. Thus, a large amount of mobility data, such as check-ins (announcing a user's current location), have been collected, along with online social relationships between users. The more these data are collected, the better we can understand individual and crowd mobility patterns, and the more accurately we can predict future locations.

Mobility prediction plays important roles in urban planning [12], traffic forecasting [13], advertising, and recommendations [36], and has thus attracted lots of attention in the past decade. A typical scenario is shown in Fig 1(a). Past mobility data, such as GPS trajectories, sequences of Wifi access points, and cell tower traces, are either of coarse positioning granularity but passively recorded or only collected actively by a small number of volunteers. Thus, the collected data may be large scale, but redundant, so that the research for mobility prediction has mainly focused on frequent pattern mining. With the development of location-based social networks, mobility prediction is becoming a hot research topic once again. This is, on one hand, because mobility data are actively collected from a large number of users connected by online social networks; on the other hand, the introduction of social relationships provides new opportunities for mobility understanding and prediction since it has been observed that mobility behaviors, particularly long-distance travel, are more influenced by

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

social network ties [6]. At the same time, the locations are of extremely fine granularity (e.g., a room in an office) so that mobility patterns are much less redundant. Since users may not have an impetus to record their regular behaviors, some movement behaviors are missed. Due to these characteristics, mobility prediction on location-based social networks faces several challenges. First, mobility data are extremely sparse, so that only a small number of frequent patterns and only a portion of user preferences are implied. Second, more irregular behaviors are presented in the mobility data from LBSNs, increasing the difficulty of prediction and urgently requiring irregularity mobility prediction. Third, the collected check-ins tend to be noisy since check-ins don't necessarily imply a physical visit, so that mobility behaviors do not reveal an individual's full preferences.

To address these challenges, we start by analyzing the mobility data from location-based social networks in two ways to understand the distinct characteristics of mobility patterns. 1) *Spatial analysis*, is conducted on this mobility data to understand individual spatial distribution and the distance distribution between consecutively visited locations, given that regularly and irregularly visited locations coexist in the mobility data. 2) *Temporal analysis*, is achieved by delving into this mobility data, to determine the significance and strength of temporal regularity and Markov dependence.



(a) (b)

Figure 1: (a) A typical scenario for next check-in location prediction; (b) A novelty-seeking driven framework for general mobility prediction

Following the analysis of mobility data, we introduce a novelty-seeking driven predictive framework for mobility prediction, which consists of three components, as shown in Fig 1(b). 1) *Regularity mining for regular mobility prediction*, which includes a temporal-based regularity model and Markov-based predictors [16]. To address the sparsity challenge, we exploit kernel smoothing for regularity estimation and interpolation techniques for integrating different orders of Markov model. And we further analyze the limit of predictability by calculating the Kolmogorov entropy of trajectories, where the power of all Markov models from zero-order to infinity-order are taken into account [18]. 2) *Recommendation techniques for irregular mobility prediction*. Obviously, it is difficult for Markov-based models to predict the irregular mobility behaviors, such as visiting novel but appealing restaurants, but such behaviors are still subject to geographical restriction and are preference driven. Additionally, irregular mobility behaviors are probably affected by social influence since they may be more likely to involve distant travel. Thus, we introduce into the predictive framework the second component: a series of location recommendation algorithms that capture these three factors. In these algorithms, to alleviate the data sparsity when presenting individual preference, we resort to the histories of similar users and friends for collaboration and use geographical constraint to discover the highest possible negatively preferred locations. To reduce the effect of the noise when presenting user preference, we treat the data as an indication of positive and negative preference with vastly varying confidence. 3) *Mining propensity of novelty seeking*. In order to jointly predict both regular and irregular locations that a user will visit next, we introduce the core component, addressing the cooperation mechanism between these two distinct models by determining the propensity of seeking a novel and attractive location. When people have strong propensity for novelty seeking, more emphasis can be

placed on irregular mobility prediction, but when people are more likely to behave regularly, regularity-based models are assigned larger importance.

2 Related Work

Mobility prediction has been widely studied in two independent fields. One field is statistical physics, by assuming human movement can be equivalent to particles and thus leveraging their well-studied motion model for mobility prediction. For example, statistical physicians analyzed mobile phone data, bank notes, GPS trajectories to understand users' individual mobility patterns at an aggregated level by studying the distribution of displacement and waiting time [2, 11, 25]. They then stimulated or predicted human movement based on the derived motion model, such as continuous time random walk and truncated levy flight. This aggregated scaling law can be analytically predicted by the mixed nature of human travel under the principle of maximum entropy, given the constraint on total traveling cost [31]. The other field is mobile communication and data mining in computer science, by directly modeling the mobility patterns based on the data. For example, in [1,9,23,28], the authors presented Markov models and a frequented pattern tree to capture sequential mobility patterns for mobility prediction. In [6, 8], time-aware regularity was modeled for mobility prediction. Furthermore, concomitant social relationships have brought new opportunities for mobility prediction and thus several novel prediction algorithms that incorporate social networks have been proposed [3, 6, 9, 24, 26]. All of this work has observed a small but significant effect of social relationships on mobility prediction. Although social influence is considered as a kind of collective wisdom, it neglects collaborative social knowledge, e.g., from users with similar mobility patterns. In contrast to this existing work, the proposed framework not only tries to fully capture collaborative social knowledge based on recommendation techniques, but also makes better use of the individual power of the regularity-based model and recommendation based on mining propensity of novelty seeking. Therefore, this framework prevents regularity (individual preference) from always playing a dominant role.

Although there are few research that suggest exploiting this knowledge for prediction, the learning of this collaborative social knowledge has been widely studied in *location recommendation*. For example, in [5, 10, 17, 20, 34], social influence, geographical restriction, and personalized user preference have been used for location recommendation. Since these authors all have observed the significant effect of geographical constraint, they have proposed different models, such as k-means clustering and kernel density estimation, for geographical modeling. In addition, the text content of locations, such as reviews and tips, has been used for further improvement [21, 32] of recommendation. In contrast to existing methods, the proposed framework not only takes into account the implicit feedback characteristics of mobility data but also presents a fully unified matrix factorization for jointly modeling user preference, geographical constraint, and social influence. Through this unified model, we have added more pseudo-negative (disliked) locations into the framework, thus alleviating the sparsity challenge.

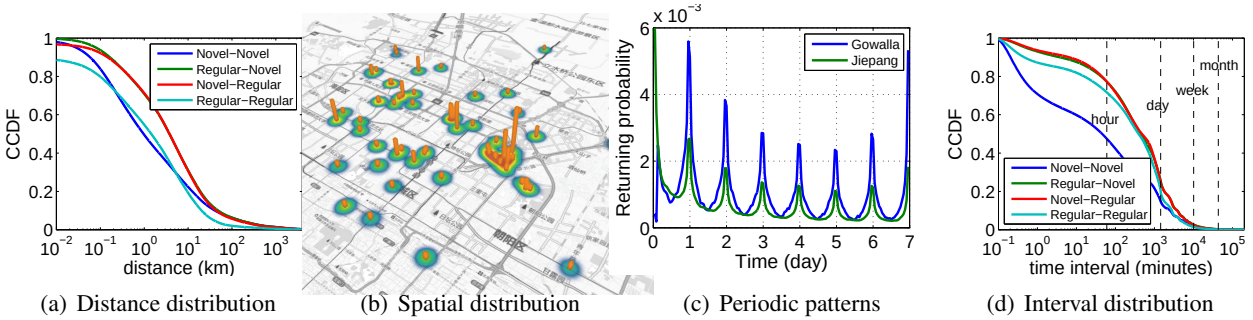
Similar ideas to *mining propensity of novelty seeking* have been proposed in [22, 27], where the probability of novelty seeking is empirically assumed to either be invariant or proportional to the number of distinct visited locations. If novelty seeking is considered to be a deviation from routine, it is related to the work in [29], where deviation from routine is detected by likelihood testing. In contrast, we have summarized our research from three perspectives, two of which are based on supervised learning, which can easily incorporate other features, and the third one is based on unsupervised learning but differentiates several levels of novelty seeking. Additionally, one method of them has a practical explanation, being directly related to the indigenization process of people.

3 Mobility Understanding on Location-based Social Networks

We first understand some basic mobility patterns on location-based social networks from the spatial and temporal perspectives.

3.1 Spatial Analysis

From the spatial perspective, first, we are interested in the distance distribution between consecutive mobility records given regular and irregular (novel) mobility behaviors coexisted, and show the distribution in Fig 2(a). Based on this, we find that 1) most check-ins (over 80%) are within 10 kilometers from the immediately preceding locations; 2) when we already know that users have checked in at regular locations, the next regular location is significantly nearer to them than next novel location; 3) users are more willing to explore continuously. This means that when a user has visited a new attraction, she may also try a nearby restaurant. These three characteristics indicate that spatial analysis can be useful for both regular and irregular location prediction and confirm the need to separate novel locations from regular ones. Second, we are interested in individual spatial density distribution. Thus, we randomly pick one user with more than 100 mobility records and plot her spatial distribution in Fig 2(b). This figure demonstrates that users usually have several major activity areas, such as home and working place, and implies that kernel density estimation is more appropriate for inferring the geographical preference.



(a) Distance distribution (b) Spatial distribution (c) Periodic patterns (d) Interval distribution

Figure 2: Illustration of spatial and temporal analysis

3.2 Temporal Analysis

From the temporal perspective, first, we are interested in periodicity, measured as returning probability [11], which is defined as the probability that a user will revisit a location t hours after her first visit. Its distribution is shown in Fig. 2(c), which indicates that the returning probability is characterized by peaks of each day, capturing a strong tendency to daily revisit regular locations. It thus confirms the existence of temporal regularity, which is thus necessarily introduced into the prediction model. Second, we study the distribution of the time interval between consecutive mobility records, and show the distribution in Fig. 2(d). This shows 1) when a user has visited a regular location, she is less inclined for exploration soon after; 2) users will be more likely to visit novel neighboring locations consecutively within a short interval (e.g., hour). Last, the existence of Markov dependence has been found in the mobility data by comparing the entropy of trajectories with randomly shuffled trajectories under the Markov assumption [30]. We do not elaborate on this here.

4 Mobility Prediction on Location-based Social Networks

Given regular and irregular mobility behaviors coexisting in mobility data, we propose a novelty-seeking driven predictive framework to jointly make use of regularity-based models for predicting regular behaviors and recommendation based algorithms for modeling irregular behaviors. The choice between them is based on people's propensity for novelty seeking, as shown in Fig. 1(b). To be more specific, when people have strong propensity for novelty seeking, recommendation-based algorithms can be relied on more, while when people are more likely to behave regularly, regularity-based models are assigned larger importance.

4.1 Regularity Mining for Regular Mobility Prediction

Regularity-based mining consists of Markov-based predictors for modeling the sequential dependence, temporal regularity for capturing periodical patterns, and a unified Hidden Markov Model for integrating these two models.

4.1.1 Markov-based Predictors

Learning the Markov model mainly depends on the estimation of location transition (due to the small amount of personal data, only first-order Markov models are taken into account). However, maximum likelihood estimation easily suffers from over-fitting due to the insufficiency of training data. Particularly, in most mobility datasets from LBSNs, the number of parameters in the Markov estimator is around $40\times$ 40 since there are 40 POIs for each user on average, while there are only about 60 training instances (mobility records) on average. Although Laplace smoothing techniques can have some effect, they don't differentiate the events of the same observed frequency. Thus, we leverage the widely-used Kneser-Ney smoothing techniques [4], that is

$$P(l|k) = \frac{\max\{n(k,l) - \delta, 0\}}{\sum_{l'} n(k,l')} + \frac{\delta \sum_{l'} \mathbf{1}_{\{n(k,l')>0\}}}{\sum_{l'} n(k,l')} \frac{\sum_{p} \mathbf{1}_{\{n(p,l)>0\}}}{\sum_{l'} \sum_{p} \mathbf{1}_{\{n(p,l')>0\}}} \tag{17}$$

where $\mathbf{1}_{\{\cdot\}}$ is an indication function and $0 \leq \delta \leq 1$ is a discounting parameter that can be set using the empirical formula $\delta = \frac{n_1}{n_1+2n_2}$ (n_1 and n_2 are the number of one-time transitions and two-time transitions across locations, respectively). Intuitively, this equation discounts the observed times of a transition and turns them over to the possibility that some locations cannot be transitioned from location k. Additionally, such an estimation ensures that zero-order distribution (the marginal of the first-order probability distribution) matches the marginals of the training data. Specifically,

$$\sum_{k} P(l|k) P_{ML}(k) = P_{ML}(l) \tag{18}$$

Thus $P_{ML}(l)$ is the stationary distribution of Markov process determined by the stochastic transition matrix $P(l|k)$.

4.1.2 Limit of Predictability

We only consider first-order Markov model above, but it is possible to use higher-order or even infinity-order Markov models. The benefit of using higher-order models can be studied by analyzing the limit of predictability [18]. Such analysis can be achieved by estimating the amount of information in terms of Kolmogorov entropy in mobility trajectories. Since it is difficult to estimate Kolmogorov entropy directly, Lempel-Ziv estimators in data compression [14] are often used for approximation, as they can converge to the real entropy of a time series when the length of trajectories is sufficiently large. One estimator of a trajectory of n points is defined as follows:

$$S \approx \frac{\ln n}{\frac{1}{n}\sum_{i=1}^{n} \Lambda_i^i} \tag{19}$$

where Λ_i^i is the length of the shortest substring starting at position i without appearing from position 1 to $i-1$. Without a sufficiently long mobility trace, the entropy will be overly estimated since some frequent patterns have not been observed yet. After estimating the entropy, we then resort to Fano's inequality [7] to transform it into the limit of predictability since this inequality connects the error probability of prediction with the sequential entropy. The overly estimated entropy will incur the lower predictability due to the concavity and monotonic decrease of the Fano function. The key problem of Fano's inequality should guarantee that the maximal prediction probability should be much higher than the random probability. The larger the difference between them is, the closer the upper bound is to the actual predictability. In other words, the more regular

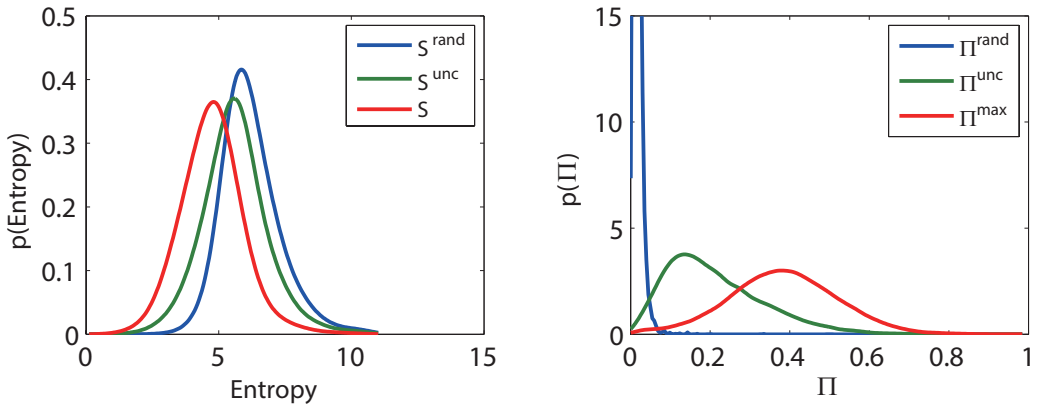


Figure 3: Left: The distribution of Kolmogorov entropy S, sequential uncorrelated entropy S^{unc} without and random entropy S^{rand} across user population. Right: The distribution of predictability of three entropies

the mobility behaviors are, the smaller the error between the upper bound and actual predictability is. Fig.3 shows examples of the distribution of estimated entropy and predictability on the Gowalla dataset [6], which only indicate 38% potential predictability.

4.1.3 Temporal Regularity

In temporal regularity, the conditional probability $P(l|d,h)$ must be estimated accurately, where d is the day of week and h is the hour of day. Assuming the conditional independence d and h given location l, this conditional probability can be transformed as

$$P(l|d,h) = \frac{P(d|l)P(h|l)P(l)}{\sum_l P(d|l)P(h|l)P(l)}. \tag{20}$$

The probability to estimate becomes $P(h|l)$ and $P(d|l)$. However, without sufficient training data, the MLE tends to be overfit. Also, the difference in the probability between neighbor hours of the day and between neighbor days of the week can not be guaranteed to be small. For example, assume a user has been to a Chinese restaurant at 6 p.m. only once. If this user returns to this restaurant in the near future, the distribution of the revisit time should be centered around 6 p.m. rather than at 6 p.m. exactly. Thus we exploit Gaussian kernel smoothing function for smoothing the MLE to the parameters.

$$\tilde{P}(h|l) = \frac{\sum_{g=0}^{23} K(\frac{d(h,g)}{\sigma_{g,l}}) P_{ML}(g|l)}{\sum_{h'=0}^{23}\sum_{g=0}^{23} K(\frac{d(h',g)}{\sigma_{g,l}}) P_{ML}(g|l)}, \quad \tilde{P}(d|l) = \frac{\sum_{e=0}^{6} K(\frac{d(d,e)}{\sigma_{e,l}}) P_{ML}(e|l)}{\sum_{d'=0}^{6}\sum_{e=0}^{6} K(\frac{d(d',e)}{\sigma_{e,l}}) P_{ML}(e|l)} \tag{21}$$

where $d(h,g) = \min(|h-g|, 24-|h-g|)$ is the distance between the h^{th} and g^{th} hour of day and $d(d,e) = \min(|d-e|, 7-|d-e|)$ is the distance between the d^{th} and e^{th} day of week. The reason for defining distance in this way is that there is a cyclic property among them (the probability of 0 a.m. is close to 1 a.m. and 23 p.m. and the probability of Sunday is also close to Saturday and Monday). $K(x)$ is a truncated standard Gaussian distribution over $x \in [0, +\infty)$.

4.1.4 Hidden Markov Model

Temporal regularity and Markov model can be integrated in a unified Hidden Markov Model, where locations are considered as hidden states and the temporal information is considered as the observations of Hidden Markov Model. The supervised learning to estimate the parameters corresponds to the above estimation process, except

the initial probability of the hidden state is not estimated. Actually, we can simply use MLE for the initial state probability. Note that we don't take social relationship into account since social network ties are more likely to influence long-distance travel according to [6] while long-distance travel may more probably involve irregular mobility behaviors we will introduce next.

4.2 Location Recommendation for Irregular Mobility Prediction

Obviously, regularity-based models will fail to predict irregular mobility behaviors, but such behaviors are still subject to geographical restriction, and are driven by both user preference and social influence. Below, we introduce how to leverage these factors for irregular behavior prediction.

4.2.1 User Preference Learning

Learning user preference mainly involves collaborative filtering techniques, which take the user-location matrix as input and mine the commonality between users. Each element of the matrix can either be visit frequency or a binary value indicating whether the visit has occurred or not. Below, we introduce two approaches for collaborative filtering that mines user commonality from different perspectives.

User-based collaborative filtering [16], directly measures user's commonality in terms of similarity on behavior data. According to our analysis, considering the element of matrix as a binary value to define the similarity is empirically optimal for recommendation. In this case, a user u is represented as $\mathbf{r}_u \in \{0,1\}^N$, where there are N locations in total and her similarity with another user v is defined as follows,

$$s_{u,v} = \frac{\mathbf{r}_u^T \mathbf{r}_v}{\|\mathbf{r}_u\| \|\mathbf{r}_v\|}. \tag{22}$$

The scoring function of user u to location i is in proportion to $\mathbf{s}_u^T \mathbf{r}_i$.

Matrix factorization is a dimension reduction technique such that the dot product between users, between items, and between user and item in the reduced latent space can measure the commonality. However, since mobility data only include the locations where users have been and are likely to prefer, while unattractive locations and undiscovered but potentially appealing ones are mixed in unvisited locations, mobility data are actually a kind of implicit feedback. In this case, we need to use a special class of matrix factorization algorithms, which treat all unvisited locations as pseudo-negative and assign them a significantly lower confidence. User preference is thus learned by solving the following optimization problem,

$$\min_{P,Q} \sum_{u,i} w_{u,i}(r_{u,i} - p_u^T q_i)^2 + \lambda(\|P\|_F^2 + \|Q\|_F^2)$$

where $p_u \in \mathbb{R}^K$ and $q_i \in \mathbb{R}^K$ represent the preferences of user u and POI i. The weight $w_{u,i}$ is empirically set as $\alpha(c_{u,i}) + 1$ if $c_{u,i} > 0$; and 1 otherwise, where $\alpha(c_{u,i})$ is monotonic increasing w.r.t visit frequency $c_{u,i}$, indicating the visit frequency reflect confidence that the users are fond of them.

4.2.2 Geographical Constraint

Kernel density estimation. The geographical information of location requires physical interactions with users to foster the universality of Tobler's First Law of Geography: "Everything is related to everything else, but near things are more related than distant things." The key for capturing this phenomenon is geographical modeling. We use two-dimensional kernel density estimation, which infers the probability a user will show up around location l_j, i.e.,

$$P(l_j) = \frac{1}{|L_i|h} \sum_{l_k \in L_i} K(\frac{d_{j,k}}{h}),$$

where $K(\cdot)$ is a kernel function. The setting of bandwidth h in the kernel function is determined by the requirement that the influence of candidate locations on the border of the influence circle is close to zero. If the probability on the border is at most ϵ times smaller than the maximum possible check-in probability, it is subject to $K(\frac{d}{h}) < \epsilon K(0)$.

Learning-based geographical inference is proposed for the sake of seamlessly integrating geographical modeling with matrix factorization based user preference learning. This is achieved by splitting the whole world into grids of approximately the same size and pre-computing the received influence of each grid from all the locations, and then converting kernel density estimation to the following optimization problem,

$$\min_{x_u} \sum_i w_{u,i}(x_u^T y_i - r_{u,i})^2 + \lambda\Omega(x_u), \text{subject to } x_u \geq 0$$

In this objective function, y_i is an influence vector of a location i, and each element corresponds to a grid's influence received from this location; and x_u is an activity area vector of user u, in which every element represents the possibility that this user will appear in a certain grid. Thus, the dot product between them can be considered to be the possibility that user u will show up around location i. $\Omega(x_u)$ is a regularized term for avoiding over-fitting.

4.2.3 Social Influence

Social-based filtering [16] is similar to user-based collaborative filtering, except it captures user commonality based on social network information. The simplest commonality/similarity between two users is defined as 1 if they are friends and 0 otherwise. In this case, a user's preference score for a location can be expressed as the number of her friends who have visited. To more accurately distinguish the importance of friends based on their closeness, we exploit another strategy, which is in proportion to the number of common friends, i.e.,

$$s_{i,l} = \frac{|F_i \cap F_l|}{|F_i \cup F_l|},$$

where F_i and F_l represent the friend sets of user u_i and u_l, respectively.

Graph Laplacian regularization [15] is more often exploited for capturing social influence for the sake of seamless integration with matrix factorization based preference learning, although social-based filtering tends to be more intuitive. Given all users' symmetric similarities S based on social network ties, such as the ratio of common friends [19], this regularizer can be defined as follows:

$$\Omega_S(P) = \frac{1}{2}\sum_{i,l} s_{i,l}\|p_i - p_l\|^2 = tr(P^T L P)$$

where $D_{i,i} = \sum_l s_{i,l}$ and $L = D - S$ is a Laplacian matrix.

4.2.4 Hybrid Recommendation

Given the factors affecting the prediction of irregular behaviors, there are many methods for empirical integration. Since geographical modeling is converted into an optimization problem, it can be seamlessly incorporated into user preference learning in terms of matrix factorization, as shown in Fig. 4. In this model, the influence areas of a POI are considered as an extra part of the POI's latent factors and the activity areas of a user are considered as an extra part of the user's latent factors. Since they are aligned in position, the dot product between them indicates two-dimensional kernel density estimation. At this moment, because unvisited locations around visited ones share similar geographical influence, user preference for them needs to offset the geographical influence. Thus, such an integration allows us to find more potential disliked locations and plays an important role in

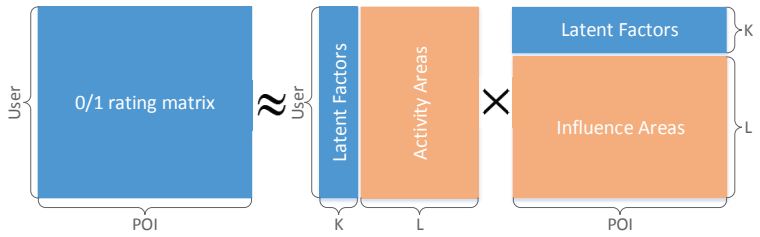


Figure 4: The augmented model for matrix factorization, where the dimension of the latent space is K and the number of grids is L.

alleviating the data sparsity. Furthermore, combining this with graph Laplacian regularization for incorporating social relationships, the overall objective function becomes as follow:

$$\min_{P,Q,X} \|W \odot (R - PQ^T - XY^T)\|_F^2 + \gamma(\|P\|_F^2 + \|Q\|_F^2) + \eta\Omega_S(P) + \lambda\|X\|_1, \text{subject to } X \geq 0$$

where X is a matrix stacking a user's activity area by columns and Y is a matrix stacking the items' influence areas vector by columns. ℓ_1 norm of matrix X, $\|X\|_1$, constrains that users usually stay around several important locations, such as home and working places.

4.3 Mining Propensity of Novelty Seeking

Mining individual propensity of novelty seeking is conducted from three perspectives: exploration prediction, mobility indigenization, and irregularity detection. Exploration prediction is spatially and temporally dependent while mobility indigenization is only with respect to cities. However, irregularity detection is independent to both spatial and temporal contexts.

4.3.1 Exploration Prediction

Exploration prediction determines whether people will seek novel (irregular) locations next. Given mobility data, whether a visit to a location is regular or not can be determined by searching the mobility history of the user. If the visit location has already been visited earlier, the visit is considered as regular; otherwise, it is irregular. Exploration prediction is thus boiled down to a binary classification problem, which can output a classification result (regular or not) or exploration tendency (e.g., a probability of classifying the next location as irregular). In the classifiers, we consider the following three types of features.

Historical features not only summarize the personality traits of novelty seeking, i.e, how often they check in, but also reflect a user's current status of neophilia, including whether a user is currently doing exploration and how many opportunities a user has left to seek novel locations. The more locations near her activity area are visited, the smaller the number of opportunities are left, and the smaller the propensity of seeking novel locations is becoming.

Temporal features are introduced to consider the effect of this temporal information since users usually have distinct degrees of novelty seeking at different times. As we have discovered, 1) users may prefer to do exploration during weekends; 2) the time interval between consecutive records also affects novelty seeking.

Spatial features are also taken into account for exploration prediction because users also exhibit different propensity of novelty seeking at locations with distinct degrees of familiarity. For example, if a user has arrived in an unfamiliar location (e.g., city), her propensity for novelty seeking will increase.

4.3.2 Mobility Indigenization

When considering irregular mobility behaviors as mainly occurring out of town, we can use a more interesting index, i.e., indigenization coefficients, for integration [33]. This index quantifies what extent an individual

behaves like a native. Therefore, this index is opposite to the propensity of novelty seeking. The smaller the index of a user in a city is, the more likely she is non-native to the city, so that irregular-based models should be given higher emphasis.

We have proposed two coefficients for this indigenization index. The first one is an individual behavioral index, $I_i(u)$, which counts the ratio of repeated mobility records in a city, inspired by the fact that a native is more likely to visit some locations many times than a non-native. That is, for a user, N_T indicates the total number of her mobility records and N_D the number of different locations visited by her. The index is then defined as

$$I_i = 1 - \frac{N_D}{N_T}. \tag{23}$$

The second one is a collaborative behavioral index I_c, measured as the average normalized popularity of a user's visit locations, which is inspired by the fact that a native is less likely to visit popular locations than a non-native. Given that $R(l_k)$ is the normalized rank of location l_k (dividing the rank by the total number of locations in a city), this index is formally defined as

$$I_c = \frac{1}{N_T} \sum_{k=1}^{N_T} R(l_k). \tag{24}$$

These two indigenization coefficients can be used to define an integrated coefficient

$$I = \frac{1}{1 + \exp(-w_i I_i - w_c I_c)}, \tag{25}$$

where the parameters w_i and w_c can be learned from the logistic regression that best classifies natives and non-natives. In other words, these two coefficients are taken as features for classifying people as native and non-natives. After learning these two parameters, we obtain a probabilistic value for the indigenization level and thus obtain a probability (i.e.,$1 - I$) for novelty seeking.

4.3.3 Irregularity Detection

Irregularity detection [35] further distinguishes several levels of propensity of novelty seeking, and detects the level of novelty seeking by measuring the popularity of the visit locations and the transition frequency to visiting location with respect to individual mobility history before the visit time. When both the popularity and transition frequency are smaller at the same time, the level of novelty seeking tends to be higher. After determining the level of novelty seeking for each visit in the mobility data, we can measure the novelty seeking trait for each user. For example, we can use the average level of novelty seeking. In other words, such an algorithm will give each user the same but distinct propensity of novelty seeking at any time and any location. In order to leverage it in the general mobility prediction framework, we can normalize it by dividing the maximum level of novelty seeking to get a pseudo probability value. A larger value indicates a higher possibility of novelty seeking.

4.4 A Novelty-Seeking Driven Framework for General Mobility Prediction

Provided the probabilistic output of the regularity mining algorithm $P_r(l)$ (r indicates regular) and recommendation algorithm $P_n(l)$ (n indicates novel), we exploit novelty seeking to combine them based on the probability of exploration $Pr(Explore)$ as follows:

$$P(l) = Pr(Explore)P_n(l) + (1 - Pr(Explore))P_r(l), \tag{26}$$

If $Pr(Explore) \in \{0, 1\}$, i.e., novelty seeking just classifies the next location as novel or not, we can switch between location recommendation and the regularity-based model. Due to the discrete value of $Pr(Explore)$, we denote this case as "hard" integration. If $Pr(Explore) \in [0, 1]$, i.e., representing the propensity of novelty

seeking, we can interpolate the regularity-based model with location recommendation. In other words, both novel and regular locations are ranked together in this case for the final location prediction. Due to the continuous value of $Pr(Explore)$, we denote this case as "soft" integration.

5 Conclusions

In this paper, we have introduced a novelty-seeking driven framework for incorporating regularity-based prediction algorithms and recommendation algorithms for predicting irregular mobility behaviors. In regularity-based prediction, we exploit Hidden Markov model for modeling location transition and temporal dependence. For recommendation algorithms, we propose a unified recommendation framework to integrate social influence, geographical restriction, and user preference based on the implicit feedback characteristics of mobility data. And the central idea of this predictive framework is the mechanism of cooperation between these two distinct models, by exploiting exploration prediction, indigenization coefficient and irregularity detection to characterize the propensity of seeking a novel and appealing location.

References

[1] D. Ashbrook and T. Starner. Learning significant locations and predicting user movement with gps. In *Proceedings of the 6th IEEE International Symposium on Wearable Computers(ISWC'02)*, pages 101–108. IEEE, 2002.

[2] D. Brockmann, L. Hufnagel, and T. Geisel. The scaling laws of human travel. *Nature*, 439(7075):462–465, 2006.

[3] J. Chang and E. Sun. Location3: How users share and respond to location-based data on social. In *Proceedings of ICWSM'11*, 2011.

[4] S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. In *Proceedings of ACL'96*, pages 310–318. ACL, 1996.

[5] C. Cheng, H. Yang, I. King, and M. Lyu. Fused matrix factorization with geographical and social influence in location-based social networks. In *Proceedings of AAAI'12*, 2012.

[6] E. Cho, S. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of KDD'11*, pages 1082–1090, 2011.

[7] R. Fano. *Transmission of information: a statistical theory of communications*. M.I.T. Press, 1961.

[8] H. Gao, J. Tang, X. Hu, and H. Liu. Modeling temporal effects of human mobile behavior on location-based social networks. In *Proceedings of CIKM'13*, 2013.

[9] H. Gao, J. Tang, and H. Liu. Exploring social-historical ties on location-based social networks. In *Proceedings of ICWSM'12*, 2012.

[10] H. Gao, J. Tang, and H. Liu. gscorr: modeling geo-social correlations for new check-ins on location-based social networks. In *Proceedings of CIKM'12*, pages 1582–1586. ACM, 2012.

[11] M. Gonzalez, C. Hidalgo, and A. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[12] M. Horner and M. O'Kelly. Embedding economies of scale concepts for hub network design. *Journal of Transport Geography*, 9(4):255–265, 2001.

[13] R. Kitamura, C. Chen, R. Pendyala, and R. Narayanan. Micro-simulation of daily activity-travel patterns for travel demand forecasting. *Transportation*, 27(1):25–51, 2000.

[14] I. Kontoyiannis, P. Algoet, Y. Suhov, and A. Wyner. Nonparametric entropy estimation for stationary processes and random fields, with applications to english text. *IEEE Transactions on Information Theory*, 44(3):1319–1327, 1998.

[15] D. Lian and X. Xie. Mining check-in history for personalized location naming. *ACM Trans. Intell. Syst. Technol.*, 5(2):32:1–32:25, Apr. 2014.

[16] D. Lian, X. Xie, V. W. Zheng, N. J. Yuan, F. Zhang, and E. Chen. Cepr: A collaborative exploration and periodically returning model for location prediction. *ACM Trans. Intell. Syst. Technol.*, 6(1):8:1–8:27, Apr. 2015.

[17] D. Lian, C. Zhao, X. Xie, G. Sun, E. Chen, and Y. Rui. Geomf: joint geographical modeling and matrix factorization for point-of-interest recommendation. In *Proceedings of KDD'14*, pages 831–840. ACM, 2014.

[18] D. Lian, Y. Zhu, X. Xie, and E. Chen. Analyzing location predictability on location-based social networks. In *Proceedings of PAKDD'14*, 2014.

[19] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031, 2007.

[20] B. Liu, Y. Fu, Z. Yao, and H. Xiong. Learning geographical preferences for point-of-interest recommendation. In *Proceedings of KDD'13*, pages 1043–1051. ACM, 2013.

[21] B. Liu and H. Xiong. Point-of-interest recommendation in location based social networks with topic and location awareness. In *Proceedings of SDM'13*, pages 396–404. SIAM, 2013.

[22] J. McInerney, S. Stein, A. Rogers, and N. R. Jennings. Breaking the habit: Measuring and predicting departures from routine in individual human mobility. *Pervasive and Mobile Computing*, 2013.

[23] A. Monreale, F. Pinelli, R. Trasarti, and F. Giannotti. Wherenext: a location predictor on trajectory pattern mining. In *Proceedings of KDD'09*, pages 637–646. ACM, 2009.

[24] A. Noulas, S. Scellato, N. Lathia, and C. Mascolo. Mining user mobility features for next place prediction in location-based services. In *Proceedings of ICDM'12*, pages 1038–1043. IEEE, 2012.

[25] I. Rhee, M. Shin, S. Hong, K. Lee, S. Kim, and S. Chong. On the levy-walk nature of human mobility. *IEEE/ACM Trans. Netw. (TON)*, 19(3):630–643, 2011.

[26] A. Sadilek, H. Kautz, and J. Bigham. Finding your friends and following them to where you are. In *Proceedings of WSDM'12*, pages 723–732. ACM, 2012.

[27] C. Song, T. Koren, P. Wang, and A. Barabási. Modelling the scaling properties of human mobility. *Nature Physics*, 6(10):818–823, 2010.

[28] L. Song, D. Kotz, R. Jain, and X. He. Evaluating location predictors with extensive wi-fi mobility data. In *Proceedings of INFOCOM'04*, volume 2, pages 1414–1424. IEEE, 2004.

[29] M. Szell, R. Sinatra, G. Petri, S. Thurner, and V. Latora. Understanding mobility in a social petri dish. *Scientific reports*, 2, 2012.

[30] C. Wang and B. A. Huberman. How random are online social interactions? *Scientific reports*, 2, 2012.

[31] X.-Y. Yan, X.-P. Han, B.-H. Wang, and T. Zhou. Diversity of individual mobility patterns and emergence of aggregated scaling laws. *Scientific reports*, 3, 2013.

[32] D. Yang, D. Zhang, Z. Yu, and Z. Wang. A sentiment-enhanced personalized location recommendation system. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media(HT'13)*, pages 119–128. ACM, 2013.

[33] Z. Yang, N. J. Yuan, X. Xie, D. Lian, Y. Rui, and T. Zhou. Indigenization of urban mobility. *arXiv preprint arXiv:1405.7769*, 2014.

[34] M. Ye, P. Yin, W.-C. Lee, and D.-L. Lee. Exploiting geographical influence for collaborative point-of-interest recommendation. In *Proceedings of SIGIR'11*, pages 325–334. ACM, 2011.

[35] F. Zhang, N. J. Yuan, D. Lian, and X. Xie. Mining novelty-seeking trait across heterogeneous domains. In *Proceedings of the 23rd international conference on World wide web*, pages 373–384. International World Wide Web Conferences Steering Committee, 2014.

[36] V. Zheng, B. Cao, Y. Zheng, X. Xie, and Q. Yang. Collaborative filtering meets mobile recommendation: A user-centered approach. In *Proceedings of AAAI'10*. AAAl Press, 2010.

Clustering in Geo-Social Networks

Dingming Wu, Nikos Mamoulis, and Jieming Shi
Department of Computer Science, The University of Hong Kong
Pokfulam Road, Hong Kong
{dmwu,nikos,jmshi}@cs.hku.hk

Abstract

The rapid growth of Geo-Social Networks (GeoSNs) provides a new and rich form of data. Users of GeoSNs can capture their geographic locations and share them with other users via an operation named checkin. Thus, GeoSNs can track the connections (and the time of these connections) of geographic data to their users. In addition, the users are organized in a social network, which can be extended to a heterogeneous network if the connections to places via checkins are also considered. The goal of this paper is to analyze the opportunities in clustering this rich form of data. We first present a model for clustering geographic locations, based on GeoSN data. Then, we discuss how this model can be extended to consider temporal information from checkins. Finally, we study how the accuracy of community detection approaches can be improved by taking into account the checkins of users in a GeoSN.

1 Introduction

Clustering is a common task of data mining, which divides a set of objects into groups such that objects in the same group (called a cluster) are similar to each other while objects in different clusters are dissimilar. Clustering finds applications in machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics. Specific applications include grouping homologous sequences into gene families in bioinformatics, partitioning the general population of consumers into groups in market research, recognizing communities within large groups of people in social networks, dividing a digital image into distinct regions for border detection or object recognition. Clustering can be achieved by various algorithms that may differ significantly in how they define clusters. Popular definitions of clusters are groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. The distance function, the density threshold or the number of expected clusters to use depend on the data to be clustered and the intended use of the clustering results.

The enormous growth of Geo-Social Networks (GeoSNs) not only brings more interesting data to clustering, but also poses challenges. In GeoSNs, such as Gowalla[1], Foursquare[2], and Facebook Places[3], users are allowed to capture their geographic locations and share them by an operation named *checkin*. A checkin is a triplet

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

[1]http://gowalla.com
[2]https://foursquare.com
[3]https://www.facebook.com/about/location

$\langle u, p, time\rangle$ modeling the fact that user u visited place with point location $p = \langle x, y\rangle$ at a certain $time$. Thus, on one hand, GeoSNs provide geographic places (e.g. points of interest) an opportunity to be (temporally) connected with social networks. On the other hand, the users of social networks are associated with their checkin point locations. The purpose of this paper is to investigate how clustering can be applied on this rich form of data.

Different from the traditional clustering of geographic locations, where only the spatial dimension is considered, clustering places in a GeoSN involves geo-social or geo-social-temporal dimensions. The geo-social place clusters discovered in a GeoSN find important application in the generalization and characterization of places. For example, discovering regions populated with similar places with respect to the people who live in them or visit them is a common task in geographic data analysis. Taking another example in urban planning, land managers are interested in identifying regions which have consistent demographic statistics, e.g., areas where elderly people prefer to visit, or, in general, people who belong to certain communities and have special transportation or living needs. The place clusters found in GeoSNs may benefit marketing as well. The fact that two (or more) commercial places belong to the same cluster indicates that there is a high likelihood that a user who likes one place would also be interested to visit the other(s). Therefore, campaigns may be initiated to users who visited other places in the same cluster, or a set of places could do collaborative promotion (e.g., a discount for users who visit multiple places in the cluster). By considering also the temporal information in the data (i.e., when did users checkin at the various places), the discovered clusters can be further refined and can become valuable for urban activity analysis, local authorities, service providers, decision makes, etc. For example, a certain set of places (e.g., shopping spots) may be characterized as a cluster for only restricted time periods or intervals (e.g., during Saturday morning hours). In addition, the user-groups that are relevant to a cluster could be relative to certain time periods. For example, shopping places in downtown are visited during the evening by people who have to work and could not shop at daytime, while supermarkets and small shops in the suburbs are usually visited by housewives in the daytime. Such geo-social-temporal clusters can be useful to marketing or advertising companies, which may benefit from understanding the (time sensitive) shopping habits of various social groups.

GeoSN data can also be used for clustering social network users. Different to classic social networks, which do not have checkin information, GeoSNs allow users to be clustered not only based on their social links but also based on their checkin behavior. Using both the social relationships and the checked in places by users can help discovering user clusters (called *local communities*) such that users in the same cluster not only have close social relationships, but also have similar mobility behavior in terms of their checkin places. The discovered local communities may provide useful information to local advertisers and social travel recommendation services such as facebook.com/36hrs.in and gogobot.com.

In this paper we investigate the possibilities of clustering geographic locations (i.e., places) and users based on the rich information tracked by GeoSNs. We first present the Density-based Clustering Places in Geo-Social Networks (DCPGS) model in Section 2 that detects geo-social place clusters in GeoSNs, considering both the *spatial* and the *social distances* between places. The DCPGS model (originally, proposed in [12]) extends traditional density-based clustering for spatial locations to consider the social relationships of users who visit them in a GeoSN. Next, we discuss possible definitions and future research directions for the geo-social-temporal place clustering and the local community detection problems in GeoSNs, in Sections 3 and 4, respectively. Finally, Section 5 concludes the paper.

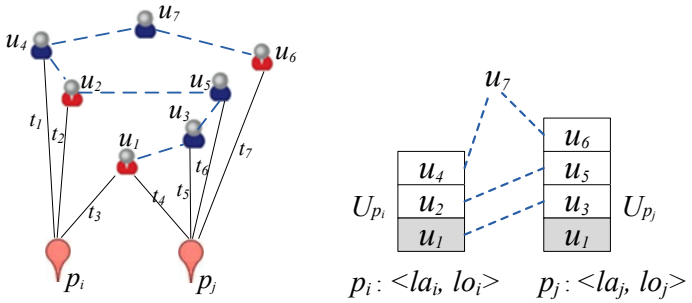
2 Geo-Social Place Clustering

Among various clustering techniques, density-based clustering [4] is an effective approach for spatial data with low dimensionality [13]. It discovers arbitrary shaped clusters and is robust to outliers. The DCPGS model extends the density-based clustering framework by introducing a new distance function that takes both the spatial proximity and the social relationship between places into account. Section 2.1 formulates the DCPGS problem

and defines the social distance measure between places that we use. DCPGS algorithms based on R-tree and grid partitioning are proposed in Section 2.2. We report part of our findings in Section 2.3.

2.1 Model and Definitions

The input of the DCPGS model includes a social network G and the set of checkins CK of a set of users U to a set of places P. The social network is an undirected graph $G = (U, E)$, where U is the set of all users and each edge $(u_i, u_j) \in E$ indicates that users $u_i, u_j \in U$ are friends. Each place $p_k \in P$ is identified by a unique GPS coordinate. Set $CK = \{\langle u_i, p_k, t_r\rangle | u_i \in U, p_k \in P\}$ includes all checkins generated by users in U. For a place p_k, its *visiting user set* is defined by $U_{p_k} = \{u_i | \langle u_i, p_k, *\rangle \in CK\}$, where $*$ means any time.



(a) A toy example (b) Abstraction

Figure 1: Example and storage structure of GeoSNs

2.1.1 DCPGS Model

DCPGS extends the model of DBSCAN [4]; for each place $p_i \in P$, DCPGS finds its *geo-social ϵ-neighborhood* $N_\epsilon(p_i)$, which includes all places p_j such that $D_{gs}(p_i, p_j) \leq \epsilon$, $D_S(p_i, p_j) \leq \tau$, and $E(p_i, p_j) \leq maxD$. For two places p_i, p_j, $E(p_i, p_j)$ is the *Euclidean distance*, $D_S(p_i, p_j)$ is the *social distance*, and $D_{gs}(p_i, p_j) = f(D_S(p_i, p_j), E(p_i, p_j))$ is the *geo-social* distance, defined as a function of $E(p_i, p_j)$ and $D_S(p_i, p_j)$. Parameter ϵ is geo-social distance threshold, while τ and $maxD$ are two *sanity* constraints for the social and the spatial distances between places, respectively. If the geo-social ϵ-neighborhood of a place p_i contains at least $MinPts$ places, then p_i is a *core* place; in this case, p_i and all places in its geo-social ϵ-neighborhood should belong to a cluster $r(p_i)$. If another core place p_j belongs to cluster $r(p_i)$, then $r(p_i) = r(p_j)$, i.e., the clusters defined by p_i and p_j are merged. After identifying all core places and merging the corresponding clusters, DCPGS ends up with a set of (disjoint) clusters and a set of outliers (i.e., places that do not belong to the geo-social ϵ-neighborhood of any core place).

Parameters. ϵ and $MinPts$ are the main parameters of DCPGS. $MinPts$ (i.e., the minimum number of places in the neighborhood of a core place) is set as in the original DBSCAN model (see [4]); a typical value is 5. ϵ takes a value between 0 and 1, because, as we explain later on, we define $D_{gs}(p_i, p_j)$ to take values in this range. Since the geo-social distance $D_{gs}(p_i, p_j)$ is a function of a spatial and a social distance, τ and $maxD$ constrain these individual distances to avoid the following two cases that negatively affect the quality of geo-social clusters.

- The geo-social distance between two places p_i and p_j could be less than ϵ if they are extremely close to each other in space, but have no social connection at all. This may lead to putting places close to each other spatially, but having no social relationship, into the same cluster.

- The geo-social distance between two places p_i and p_j could be less than ϵ if they have very small social distance, but they are extremely far from each other spatially. This may lead to putting places with close social distances, but large spatial distances, into the same cluster.

Constraints τ and $maxD$ are defined for quality control and can be set by experts or according to the analyst's experience. We experimentally study how clustering quality is affected by the two constraints and ϵ in Section 2.3.

Distance Functions. The social distance $D_S(p_i, p_j)$ takes in the visiting user sets U_{p_i} and U_{p_j} of places p_i and p_j, respectively, and returns a value between 0 and 1. In Section 2.1.2, we present our definition for $D_S(p_i, p_j)$. Before defining the geo-social distance $D_{gs}(p_i, p_j)$, we *normalize* the Euclidean distance $E(p_i, p_j)$ to a *spatial* distance $D_P(p_i, p_j) = \frac{E(p_i,p_j)}{maxD}$ that takes values between 0 and 1. Finally, $D_{gs}(p_i, p_j)$ is defined as weighted sum of $D_S(p_i, p_j)$ and $D_P(p_i, p_j)$, i.e.,

$$D_{gs}(p_i, p_j) = \omega \cdot D_P(p_i, p_j) + (1 - \omega) \cdot D_S(p_i, p_j), \tag{27}$$

where $\omega \in [0, 1]$.

2.1.2 Social Distance Between Places

The social distance $D_S(p_i, p_j)$ between p_i and p_j naturally depends on the social network relationships between the visiting user sets U_{p_i} and U_{p_j} of places p_i and p_j, respectively. Our definition for $D_S(p_i, p_j)$ is based on the set CU_{ij} of *contributing users* between two places p_i and p_j:

Definition 1: (Contributing Users) Given two places p_i and p_j with visiting user sets U_{p_i} and U_{p_j}, respectively, the set of contributing users CU_{ij} for the place pair (p_i, p_j) is defined as $CU_{ij} = \{u_a \in U_{p_i} | u_a \in U_{p_j} \vee \exists u_b \in U_{p_j}, (u_a, u_b) \in E\} \cup \{u_a \in U_{p_j} | u_a \in U_{p_i} \vee \exists u_b \in U_{p_i}, (u_a, u_b) \in E\}$

Specifically, if a user u_a has visited both p_i and p_j, then u_a is a contributing user. Also if u_a has visited place p_i, u_b has visited p_j, and u_a and u_b are friends, both u_a and u_b are contributing users. Users in CU_{ij} contribute positively (negatively) to the social similarity (distance) between p_i and p_j. Formally:

Definition 2: (Social Distance) Given two places p_i and p_j with visiting users U_{p_i} and U_{p_j}, respectively, the *social distance* between p_i and p_j is defined as

$$D_S(p_i, p_j) = 1 - \frac{|CU_{ij}|}{|U_{p_i} \cup U_{p_j}|} \tag{28}$$

The above definition of $D_S(p_i, p_j)$ takes both the set similarity between sets U_{p_i} and U_{p_j} and the social relationships among users in U_{p_i} and U_{p_j} into account. In addition, the distance measure penalizes pairs of places p_i and p_j which are popular (i.e., U_{p_i} and/or U_{p_j} are large) but their set of contributing users is relatively small (see Equation 28). The reason is that such place pairs are not characteristic to their (loose) social connections.

As an example, consider places p_i and p_j of Figure 1. Figure 1(b) shows U_{p_i} and U_{p_j} for the two places p_i and p_j of the toy example in Figure 1(a). The figure also connects the user pairs in the two sets who are linked by friendship edges in the social network. Note that user u_8 does not belong to either U_{p_i} or U_{p_j}, but connects users u_4 and u_7 in the social graph.

To compute $D_S(p_i, p_j)$, we first set $U_{p_i} = \{u_1, u_2, u_4\}$ and $U_{p_j} = \{u_1, u_3, u_5, u_6\}$. All users in U_{p_i} and U_{p_j} are checked one by one to obtain the contributing users between p_i and p_j. We derive $CU_{ij} = \{u_1, u_2, u_3, u_5\}$, since (i) u_1 have visited both p_i and p_j, (ii) user u_2, who visited p_i, has a friend u_5 who visited p_j, (iii) symmetrically, user u_5, who visited p_j, has a friend u_2 who visited p_i, and (iv) u_3 ($\in U_{p_j}$) has a friend u_1 having been to p_i. According to Definition 2, the social distance $D_S(p_i, p_j)$ between p_i and p_j in Figure 1 is $1 - |CU_{ij}|/(|U_{p_i} \cup U_{p_j}|) = 1 - 4/6 \approx 0.3333$.

2.2 Algorithms

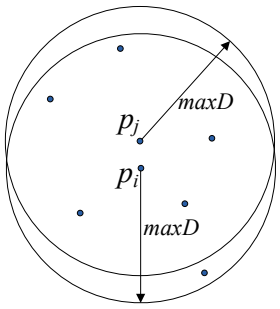
We propose two algorithms for computing geo-social clusters using DCPGS model. Algorithm DCPGS-R (Section 2.2.1) is based on the R-tree index, while algorithm DCPGS-G (Section 2.2.2) uses a grid partitioning.

2.2.1 Algorithm DCPGS-R: R-tree based

Algorithm DCPGS-R is a direct extension of the DBSCAN algorithm; it uses an R-tree to facilitate the search of the geo-social ϵ-neighborhood for a given place. Initially, all places are bulk-loaded into an R-tree. Then, DCPGS-R examines all places and, given a place p_i, it performs a range query centered at p_i with radius $maxD$ to get a set of *candidate places* that may fall in the geo-social ϵ-neighborhood of p_i, i.e., $N_\epsilon(p_i)$. Recall that $maxD$ is the maximum allowed spatial distance between place p_i and places in its geo-social ϵ-neighborhood. Then, DCPGS-R keeps in $N_\epsilon(p_i)$ only the candidates that satisfy the social distance constraint τ and the geo-social distance threshold ϵ. Clusters are identified by merging core places and their geo-social ϵ-neighborhoods.

2.2.2 Algorithm DCPGS-G: Grid based

DCPGS-R conducts a spatial range query for each place to obtain the candidate places for the purpose of discovering geo-social clusters. Even though individual R-tree based range queries are very efficient, discovering geo-social clusters in a GeoSN with millions of places requires millions of such queries (e.g., there are 1,280,969 places in the Gowalla dataset used in our experiments). Given two places p_i and p_j that are spatially close to each other, as Figure 2(a) shows, the results of the two range queries with radius $maxD$ centered at p_i and p_j, respectively, are almost identical. In algorithm DCPGS-R, independently issuing similar range queries on the R-tree searches almost the same space, resulting in redundant traversing paths and computations. To overcome this drawback, we develop a dynamic grid partitioning technique and a new algorithm DCPGS-G.



(a) Nearby spatial range queries (b) Grid partitioning

Figure 2: Nearby spatial range queries and grid partitioning

Grid Partitioning. The area covered by the whole data set is partitioned using a grid of size $maxD/\sqrt{2} \times maxD/\sqrt{2}$. The non-empty grid cells are indexed by a hash table with the grid cell coordinates as search keys.

Neighbor Cells. The neighbor cells of a cell c are the cells that intersect the union of four circles, each centered at a corner of cell c with radius $maxD$. For example, in Figure 2(b), the 20 gray cells (except c) are the neighbor cells of c, denoted as $NC(c)$. We can trivially show that for any place p inside c, the content of p's geo-social ϵ-neighborhood is contained in $NC(c)$ and c itself.

Cluster Discovery. Algorithm DCPGS-G includes three phases. First, it maps all places into grid cells. Second, it computes the geo-social ϵ-neighborhoods of places at the grid cell level. Specifically, for each non-empty and *unprocessed* cell c, its neighbor cells $NC(c)$ are retrieved. This operation filters out the pairs of places (p_i, p_j) with spatial distance greater than $maxD$. A cell is 'unprocessed' if its neighbor cells have not been retrieved

before. Then the pairs of places (p_i, p_j) that satisfy the social distance constraint τ and the geo-social distance threshold ϵ are identified and the geo-social ϵ-neighborhoods $N_\epsilon(p_i)$ and $N_\epsilon(p_j)$ are updated. After all cells have been processed, meaning that the geo-social ϵ-neighborhoods of all places in the GeoSN are acquired, the third phase discovers all geo-social clusters following the framework of algorithm DCPGS-R, except that the $N_\epsilon(p_i)$ of each place p_i has already been computed in the second phase.

Complexity. With the help of grid partitioning, the geo-social ϵ-neighborhood of all places in cell c can be obtained by checking all c's neighbor cells; the whole process can be completed within a single pass of the data. Thus, the complexity of DCPGS-G is $O(n)$, as each of its three phases makes one pass over the data. However, algorithm DCPGS-R computes the geo-social ϵ-neighborhoods of each place one by one. Hence its cost is $O(n \log n)$, given that the expected cost of a single range query on the R-tree is $O(\log n)$.

2.3 Evaluation Results

In [12], we evaluated the DCPGS model and algorithms using two data sets from real geo-social networks[4] from two perspectives: effectiveness and efficiency. To assess effectiveness, we conducted a visualization-based analysis and a social quality evaluation in terms of two measures: social entropy and community score. In general, it has been demonstrated that the social relationships between users who visit places have great impact in place clustering and cannot be overlooked. The social distance measure we propose is more effective compared to competitor measures. To evaluate efficiency, we implemented the R-tree based and the grid-based DCPGS algorithms to apply using alternative distance measures and compared their performance under various parameter settings. The results show that the grid-based implementation is more efficient than the R-tree based implementation and our proposed social distance measure between places is more efficient to compute compare to more complex alternatives. The detailed evaluation results can be found in [12]. In this section, we show part of our visualization-based analysis, which compares the clusters found by DCPGS and competitor methods in the area of Manhattan on the Gowalla dataset (Figures 3(a)–3(f)) and also in the area of Chicago, on the Brightkite dataset (Figures 4(a)–4(b)).

Competitor DBSCAN [15] disregards the social network and finds density-based clusters using only the Euclidean distance between places. Competitor PureSocialDistance is an extreme case of DCPGS where ω is set to 0 in Equation 27. Competitor LinkClustering constructs a place network PN where two places p_i and p_j are connected if $E(p_i, p_j) \leq maxD$ and $D_S(p_i, p_j) \leq \tau$. The edge weight is set to $W_{gs}(p_i, p_j) = 1 - D_{gs}(p_i, p_j)$. Then, an offline community detection algorithm [1, 2] is applied on PN to discover place clusters. Competitor Jaccard replaces the social distance in DCPGS with the Jaccard similarity between the visiting user sets of two places. Finally, competitor SimRank applies the Minimax version of SimRank [7] to measure the similarity between the visiting user sets of two places. Compared to these five competitors, our proposed DCPGS finds geo-social clusters with the following features.

Geo-Social Splitting/Merging Criteria. Clusters found by DBSCAN due to their spatial closeness are split by DCPGS because of their weak social relationships, while clusters split by DBSCAN due to relatively low spatial density are merged by DCPGS due to their strong social ties. For example, Figure 3(a) and 3(b) shows that the layouts of the clusters discovered by DCPGS and DBSCAN are totally different. Specifically, comparing region A in Figures 3(a) with the corresponding region A′ in Figure 3(b), DCPGS and DBSCAN detect different cluster structures. The clusters found by DCPGS cannot be discovered by DBSCAN even if the parameters are tuned, since the densities of the small clusters in the half bottom of region A′ are similar and they are close to each other. Hence, DBSCAN will consider the places in the half bottom of region A′ as either a single cluster or several fragmented clusters (Figure 3(b)), under different parameter settings. Sometimes, DCPGS is able to split spatially dense clusters due to some natural barriers, such as rivers, and walls. It is inconvenient for the users to travel from one side of the barrier to the other side, so that the social ties between the places from the two sides

[4] snap.stanford.edu/data/index.html

of the barrier are weak, resulting in a splitting effect. As an example, in Figures 4(a) and 4(b), a cluster (region C) found by DBSCAN is split into two DCPGS clusters (regions C_1 and C_2) by the river. Although it might be possible for DBSCAN to detect the two DCPGS clusters by reducing the value of eps, such parameter settings will make some existing significant clusters disappear, resulting in too many outliers.

Spatially Loose Clusters. Some geo-social clusters found by DCPGS in region B of Figure 3(a) are considered as outliers by DBSCAN, shown as region B′ in Figure 3(b), since the places in region B′ is spatially too sparse to satisfy the density requirement of DBSCAN, and thus most places inside it are considered as outliers. However, these places (in region B of Figure 3(a)) are grouped into clusters by DCPGS due to the reason that the users who checked in those places have strong social relationships. If reducing the density parameters of DBSCAN, such spatially loose clusters can also be discovered. Nevertheless, many other clusters may be merged, making denser clusters indistinguishable.

Fuzzy Boundary Clusters. The boundaries of some DCPGS geo-social clusters are fuzzy, which makes sense in the real world, since groups of socially connected users may spatially overlap. In contrast, the clusters detected by DBSCAN have clearly strict boundaries. For instance, in Figure 3(a), no strict boundary exists between the four clusters enclosed in region A. Competitor PureSocialDistance also produces clusters with fuzzy boundaries (shown in Figure 3(c)). However, these clusters are spatially indistinguishable and of no interest, i.e., for the applications mentioned in the Introduction.

Competitor LinkClustering produces thousands of small clusters with average size around 3, shown in Figure 3(d), which are typically not well-separated spatially. Because of the sparse geo-social network data, the constructed place network consists of a lot of connected components that are disconnected with each other. For example, the place network built given $\tau = 0.7$, $maxD = 100$, and $\omega = 0.5$ contains 34,496 connected components with 4.3 nodes and 8.2 edges on average.

Competitors Jaccard and SimRank replace our D_S definition (Definition 2) by the Jaccard and the SimRank based measures. Figures 3(e) and 3(f) shows their clustering results. Competitor Jaccard produces small clusters and too many outliers, since large distance values are given for most pairs of places p_i and p_j due to the reason that the set of common users for two places in Jaccard (i.e., $U_{p_i} \cap U_{p_j}$) is expected to be small. On the contrary, competitor SimRank produces clusters of slightly larger sizes compared to DCPGS. We observed that the probability distribution of the SimRank-based measure is skewed towards small values, so that a lot of pairs of places are given low bipartite minimax SimRank social distance.

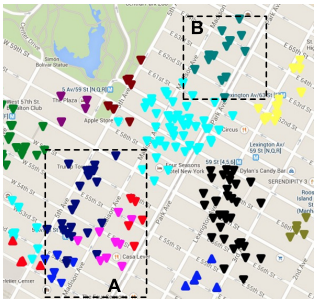
3 Geo-Social-Temporal Place Clustering

A checkin in GeoSNs is a triplet $\langle u, p, time \rangle$ modeling the fact that user u visited place with point location $p = \langle x, y \rangle$ at a certain $time$. The geo-social clusters found by the DCPGS model (presented in the previous section) compute the social distance between places based on the social network relationships between the visiting user sets of the places. However, temporal information is completely disregarded by the DCPGS model. It would be interesting to extend the DCPGS model such that the temporal information is taken into account in clustering and investigate how the temporal information affects the clustering result. In this section, we investigate the discovery of *geo-social-temporal* clusters in GeoSNs, which are spatio-temporal regions visited by groups of socially connected users.

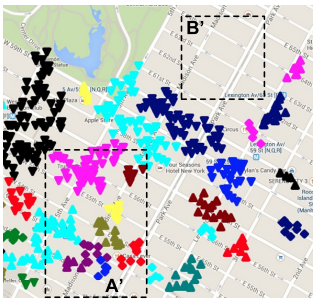
In order to compute such geo-social-temporal clusters, a possible method would be to extend the definition of social distance between places to a socio-temporal distance D_{ST}. Using the social-temporal distance D_{ST}, the DCPGS model can then replace the geo-social distance D_{gs} by a newly defined geo-social-temporal distance as follows:

$$D_{gst}(p_i, p_j) = \omega \cdot D_P(p_i, p_j) + (1 - \omega) \cdot D_{ST}(p_i, p_j). \tag{29}$$

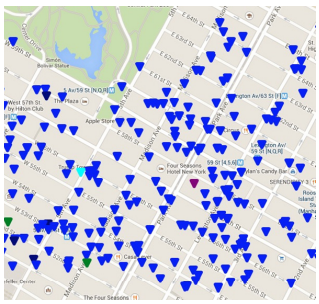
An intuitive definition of the socio-temporal distance D_{ST} would be to consider a pair of places socio-temporally close if they share many visiting users that have checked in the places within a small time period. On the other



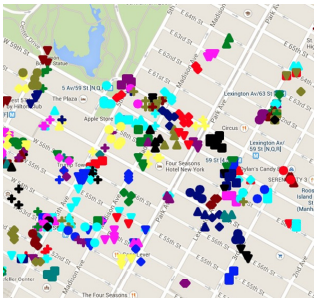
(a) DCPGS: $\epsilon = 0.4$, $\tau = 0.7$, $maxD = 100m$



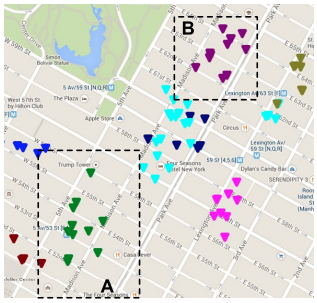
(b) DBSCAN: $eps = 40m$



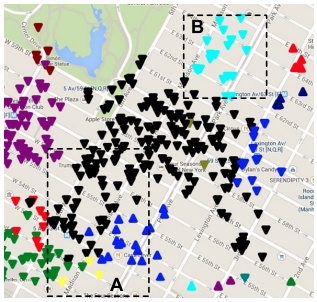
(c) PureSocialDistance: $\epsilon = 0.2$, $\tau = 1$, $maxD = 1000m$



(d) LinkClustering: $\tau = 0.7$, $maxD = 100m$



(e) Jaccard: $\epsilon = 0.4$, $\tau = 0.7$, $maxD = 100m$



(f) SimRank: $\epsilon = 0.3$, $\tau = 0.7$, $maxD = 100m$

Figure 3: Place clusters of Gowalla found in Manhattan

hand, two places are socio-temporally far from each other if they do not have common visitors within a short time interval. The temporal dimension captures the evolution of place visits, and thus reflects the changes of the social distance between places. Based on the above, we suggest that the following three possible definitions of D_{ST} should be investigated.

Temporal Threshold. The socio-temporal distance extends the social distance (Equation 28) by replacing the contributing users CU_{ij} with the temporally contributing users TCU_{ij}, i.e.,

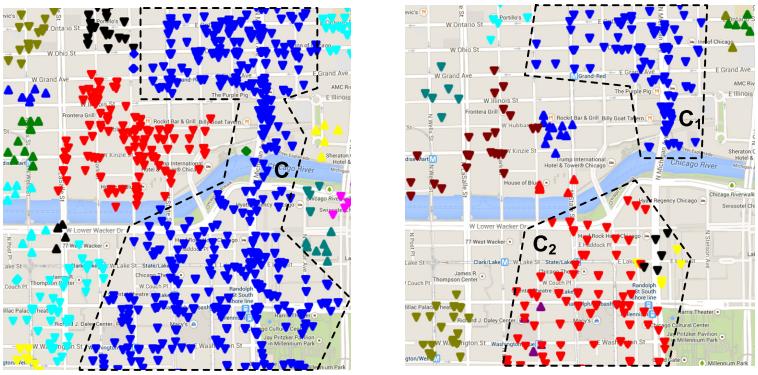
$$D_S(p_i, p_j) = 1 - \frac{|TCU_{ij}|}{|U_{p_i} \cup U_{p_j}|}. \tag{30}$$

The temporally contributing users are socially connected users who checked in p_i and p_j within a time interval θ. Let $T(u_a, p_i)$ be the time when user u_a checked in place p_i; formally:

Definition 3: (Temporally Contributing Users) Given two places p_i and p_j with visiting user sets U_{p_i} and U_{p_j}, respectively, the set TCU_{ij} of temporally contributing users for the place pair (p_i, p_j) is defined as $TCU_{ij} = \{u_a \in U_{p_i} | (u_a \in U_{p_j} \wedge |T(u_a, p_i) - T(u_a, p_j)| \leq \theta) \vee (\exists u_b \in U_{p_j}, (u_a, u_b) \in E \wedge |T(u_a, p_i) - T(u_b, p_j)| \leq \theta\} \cup \{u_a \in U_{p_j} | (u_a \in U_{p_i} \wedge |T(u_a, p_i) - T(u_a, p_j)| \leq \theta) \vee (\exists u_b \in U_{p_i}, (u_a, u_b) \in E \wedge |T(u_b, p_i) - T(u_a, p_j)| \leq \theta)\}$.

This definition of TCU_{ij} favors place pairs to which socially connected users paid visits what were close in time.

Damping Window. This method assigns each contributing user u_a an exponential decay factor $e^{t_c - T(u_a, p_i)}$, where t_c is the current time and $T(u_a, p_i)$ is the time when user u_a checked in place p_i. The contributing users who made checkins recently are weighed high. Instead of counting 1 for each user when computing the social



(a) DBSCAN: $eps = 60m$

(b) DCPGS: $\epsilon = 0.4$ s.t. $\tau = 0.7$, $maxD = 120m$

Figure 4: Clusters of Brightkite found by DBSCAN and DCPGS in Chicago

distance (Equation 28), this method counts the exponential decay factor of each user when computing the social-temporal distance D_{ST}. This definition favors place pairs to which the socially connected users have paid recent visits.

History-frame Clustering. This method performs geo-social clustering for each time period separately. For example, we can generate a different clustering of places for each month, by only using the checkin data recorded in that month. The clustering results would be useful in finding out how the place clusters evolve over time. It is also possible to track which place enters or leaves a cluster at a particular month and which parts of the clusters are time-insensitive.

4 Local Community Detection in GeoSNs

Community detection is an analytics tool for studying the social relationships among users. When detecting communities, there are two possible sources of information one can use: the social network structure and the features and attributes of users. Existing algorithms, however, typically focus on one of these two data modalities: community detection algorithms traditionally consider only on the network structure, while clustering algorithms mostly consider only user (node) attributes. Recently, algorithm CESNA [16] has been proposed to detect overlapping communities in networks with node attributes. CESNA statistically models the interaction between the network structure and the node attributes, which leads to more accurate community detection as well as improved robustness in the presence of noise in the structure. Later, Shakarian et al. [11] used a variant of Newman-Girvan modularity with the Louvain algorithm to address the problem of mining for geographically dispersed communities.

In GeoSNs, it would be interesting to detect local user communities taking both the social network structure and the checkin information into account, so that groups of socially connected users that checkin in the same or geographically close places are discovered. Existing algorithms that use either the network structure or node attributes cannot achieve the goal of local community detection. In addition, although CESNA could be applied by considering the checkin places as attributes of nodes (i.e., users), it may not achieve satisfactory results, because the proximity of places is not taken into account (i.e., only users that check in identical places would be considered as similar). Typically, the probability that two users have a significant overlap in their visiting places is low, therefore it makes sense to consider proximity as a factor of similarity between users in local community detection. Finally, although Shakarian et al. [11] provide a way to leverage spatial information in addition to network connection topology when mining networks for communities, they assume that each node in the social network is associated with only one home location. This approach is not applicable for the case when the users in GeoSNs have multiple check-in locations.

To implement local community detection in GeoSNs, we first need to model the users' mobile behaviors according to their checkin locations. Next, in the social network, we assign weights on the edges based on the similarity between the mobility behaviors of the corresponding users. Then, the resulting edge-weighted graph can be fed to existing community detection algorithms for weighted graphs, such as ISCoDe [6] and the algorithm proposed by Liu et al. [9], to identify the local communities. We suggest the following three ways for modeling the mobility behaviors of users.

Trajectory-based. The checkin locations of each user can be connected according to the time of the checkins to form a trajectory for the user. This trajectory models the mobility behavior of the user. Trajectory similarity can then be used to model the similarity between two users that are connected in a social network. Measures for trajectory similarity include Euclidean distance [10], dynamic time warping [8], edit distance [3], and longest common subsequence [5].

Image-based. The mobile behavior of each user is modeled as a black-and-white image where each pixel corresponds to the coordinates of a checkin location. The light intensity (gray value) of a pixel is determined by the frequency of visits at the corresponding place. The similarity between two images can be computed using the Minkowski metric on their contained pixels or more complicated measures incorporating specific task-dependent features [14].

Frequently Visited Region based. By analyzing the checkin locations of users, we can identify one or multiple frequently visited regions or areas for each of them. The granularity of the frequently visited regions can be determined by superimposing a grid on the map that includes all checkin locations. Then, each user is associated with one or multiple regions (cells) which s/he has frequently visited. For two users, we can use the spatial relationships (e.g., Euclidean distance or overlapping ratio) between their frequently visited regions to determine the similarity between the users.

5 Conclusions

Although geographic data clustering and community detection have been extensively studied for decades and many effective algorithms have been proposed, the rapid growth of the geo-social networks bring to these two problems a new and rich form of data together with new challenges. Clustering places by considering both their spatial proximity and the users who visit them (as well as the ties between these users) results in significantly different clusters compared to just using place locations. The time of the user checkins to places can be used to further refine the clusters. Differences and interesting insights can also be found in the user communities discovered when both the social relationships between users and the proximity between places they check in are considered.

In this paper, we have presented the Density-based Clustering Places in Geo-Social Networks (DCPGS) model [12] that discovers spatially and socially relevant place clusters. Our empirical studies prove the effectiveness of the model. We also discussed how to extend the DCPGS model to consider temporal information in the check-in data. Finally, we introduced the local community detection problem in GeoSNs, where the users forming a cluster are not only socially close but also exhibit similar mobility behavior in terms of their check-in locations.

References

[1] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann. Link communities reveal multiscale complexity in networks. *Nature*, 466:761–764, 2010.

[2] I. R. Brilhante, M. Berlingerio, R. Trasarti, C. Renso, J. A. F. de Macêdo, and M. A. Casanova. Cometogether: Discovering communities of places in mobility data. In *MDM*, 2012.

[3] L. Chen and R. Ng. On the marriage of lp-norms and edit distance. In *VLDB*, pages 792–803. VLDB Endowment, 2004.

[4] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, 1996.

[5] T. Ichiye and M. Karplus. Collective motions in proteins: a covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins*, 11(3):205–217, 1991.

[6] E. Jaho, M. Karaliopoulos, and I. Stavrakakis. Iscode: A framework for interest similarity-based community detection in social networks. In *Computer Communications Workshops (INFOCOM WKSHPS), IEEE Conference on*, pages 912–917, 2011.

[7] G. Jeh and J. Widom. Simrank: A measure of structural-context similarity. Technical Report 2001-41, Stanford InfoLab, 2001.

[8] J. B. Kruskal. An overview of sequence comparison. In *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, pages 1–44. Addison-Wesley, Reading, MA, 1983.

[9] R. Liu, S. Feng, R. Shi, and W. Guo. Weighted graph clustering for community detection of large social networks. *Procedia Computer Science*, 31(0):85 – 94, 2014.

[10] A. C. Sanderson and A. K. C. Wong. Pattern trajectory analysis of nonstationary multivariate data. *IEEE Transactions on Systems, Man, and Cybernetics*, 10(7):384–392, 1980.

[11] P. Shakarian, P. Roos, D. Callahan, and C. Kirk. Mining for geographically disperse communities in social networks by leveraging distance modularity. In *KDD*, pages 1402–1409, 2013.

[12] J. Shi, N. Mamoulis, D. Wu, and D. W. Cheung. Density-based place clustering in geo-social networks. In *SIGMOD*, pages 99–110, 2014.

[13] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison-Wesley, 2005.

[14] A. B. Watson, editor. *Digital Images and Human Vision*. MIT Press, Cambridge, MA, USA, 1993.

[15] D.-N. Yang, C.-Y. Shen, W.-C. Lee, and M.-S. Chen. On socio-spatial group query for location-based social networks. In *KDD*, 2012.

[16] J. Yang, J. J. McAuley, and J. Leskovec. Community detection in networks with node attributes. In *ICDM*, pages 1151–1156, 2013.

Space-Time Aware Behavioral Topic Modeling for Microblog Posts

Qiang Qu [†] Cen Chen [‡] Christian S. Jensen [#] Anders Skovsgaard [♮]
[†] Department of Computer Science, Innopolis University
[‡] School of Information Systems, Singapore Management University
[#] Department of Computer Science, Aalborg University
[♮] TrustSkills, Denmark
[†] qu@innopolis.ru [‡] cenchen.2012@phdis.smu.edu.sg [#] csj@cs.aau.dk [♮] anders@trustskills.com

Abstract

How can we automatically identify the topics of microblog posts? This question has received substantial attention in the research community and has led to the development of different topic models, which are mathematically well-founded statistical models that enable the discovery of topics in document collections. Such models can be used for topic analyses according to the interests of user groups, time, geographical locations, or social behavior patterns. The increasing availability of microblog posts with associated users, textual content, timestamps, geo-locations, and user behaviors, offers an opportunity to study space-time dependent behavioral topics. Such a topic is described by a set of words, the distribution of which varies according to the time, geo-location, and behaviors (that capture how a user interacts with other users by using functionality such as reply or re-tweet) of users. This study jointly models user topic interest and behaviors considering both space and time at a fine granularity. We focus on the modeling of microblog posts like Twitter tweets, where the textual content is short, but where associated information in the form of timestamps, geo-locations, and user interactions is available. The model aims to have applications in location inference, link prediction, online social profiling, etc. We report on experiments with tweets that offer insight into the design properties of the papers proposal.

1 Introduction

Microblogging services that enable the posting and browsing of messages containing, e.g., news or local events, are increasingly being used for social interactions.

For example, Twitter has several hundred million active users from around the world who post half a billion messages each day (https://about.twitter.com/company) and is arguably the most important microblogging service. Twitter messages, called tweets, are timestamped and are limited to 140 characters. Twitter supports reply, retweet, and mention functions for tweets, thus enabling social interactions around tweets. We are also witnessing an increased use of geo-enabled mobile devices, most notably smartphones [12]. They offer not only a timely way of using Twitter, but they also offer the ability to associate user location with tweets, yielding geo-tagged tweets.

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

The resulting tweets offer the following information: 1) who posted the tweet; 2) textual content; 3) the time when the tweet was posted; 4) the geo-location from which the tweet was posted; and 5) an associated social behavior (i.e., post, reply, retweet, or mention).

For example, tweet T1 in Table 1 was posted by "@ohcindyoh"; has text that suggest that the tweet concerns a movie; was posted on April 29, 2013; was posted from location @cinema21[1]; and was posted as an original tweet (using "post"). Put differently, tweets may be viewed as being 5-dimensional.

ID	Date	Author	Textual Content
T1	April 29, 2013	@ohcindyoh	Watching Iron Man 3 (with Geng Depo Bangunan at @cinema21).
T2	April 29, 2013	@imabieberchicka	@brailleman89 What are you doing?
T3	April 30, 2013	@imivycaparas	Gorg sis!! Daniel's concert tomorrow): huehuehue im jelly! Buy smth for me!!! Shirt okay): @jiannex

Table 1: Example tweets.

The availability of large collections of such 5-dimensional microblog posts makes it relevant to study an integrated model of social behavioral patterns that exploits all five dimensions. Existing studies have, however, proposed to model topics of social data based on only some of the 5 dimensions [4, 13–15, 18]. These models can be used in applications such as topic mining [3], followee recommendation [4], and location prediction [13]. However, to the best of our knowledge, this study is the first to consider all of the 5 dimensions of microblog posts. More specifically, we consider behaviors that correspond to the social functions offered by the microblogging service, i.e., post, retweet, reply, mention, for tweets. We consider this user behavior together with space and time because all three describe the context in which posts are generated by users.

To exemplify this context, consider again Table 1. Here, tweet T2 is a reply (tweets staring with @"user") from @imabieberchicka to @brailleman89 that concerns their relationship. Some Twitter users use Twitter as a chatting app for interacting with their friends, so that most tweets are associated with the reply behavior. If we model user behaviors and topics jointly as in one study [4], we will find a "reply-daily" topic that concerns mostly daily issues and appears in user replies. By looking at this topic, we can find users that interact with other users using "reply". The distribution of topics can depend on the kind of user behavior (e.g., post, reply), for which reason the topics are called behavioral.

Next, some users use Twitter as a news channel and often retweet news events. For instance, movie fans often talk about new movies, and music fans may often talk about concerts. In this case, we may find that topics are associated with events. For example, tweet T1 is about watching a movie on April 29, 2013. If we observe many tweets talking about the movie "Iron Man" on the same day then there may be an event related to "Iron Man" on that day. It is thus important to consider time information.

Last but not least, some users may be interested only in events happening close to their locations. It is thus beneficial to consider the geo-location of behavioral topics. For example, tweet T3 concerns a concert in the Philippines. It is then likely to be most appropriate to recommend this event to users in the Philippines. This shows that it is also necessary to consider space information.

In sum, it is important to model users, textual content, behaviors, space, and time jointly for microblog posts. We thus propose a space-time dependent behavioral topic model. However, it is difficult to simply aggregate the dimensions of tweets in a regression model as they are of different types.

The proposed modeling has three notable benefits. First, we can identify user groups at similar locations with similar topics during a time period, but with different social behaviors. For example, Twitter users are likely to check-into geographical locations when posting tweets concerning local events. The identification of different

[1]T1 contains a check-in that is regarded as a geo-location tag. If a tweet has no check-in, we use its lat-long as its geo-location.

user behaviors may help us understand their motivations for using Twitter and how actively they interact with the local events. Second, we can profile users and locations according to social behaviors (e.g., reply and retweet behaviors) and the changes of topics over time. Third, we can predict user locations at a specific time given topics and behaviors.

We compare our model with existing models and propose methods for estimating the parameters of the model. Experimental findings from experiments with tweets show that our model is capable of identifying interesting space-time dependent behavioral topics of users and of predicting user locations. The results also suggest that the proposed model is effective for the applications considered.

The rest of the paper is organized as follows. Following a coverage of related work in Section 2, Section 3 presents the proposed model and means of estimating model parameters. The experimental study is presented in Section 4. Section 5 concludes and discusses future work.

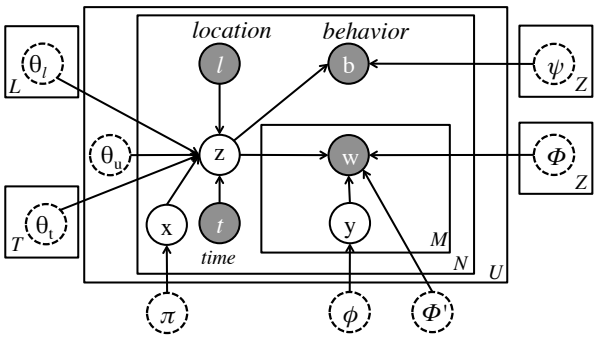
2 Related Work

Recently, geo-tagged and time-stamped social media has drawn much attention [1, 6, 7, 9, 10, 17]. Some studies propose to model topics of microblog posts to understand their social content. Topic models like LDA [2] have been used widely to find hidden "topics" in documents. In these models, each document can be represented in a semantic topic space, which also enables tasks like text classification and document clustering. There is growing interest in adapting topic models to short texts like microblog posts [3, 14, 18].

Twitter-LDA (T-LDA) [3] addresses the shortness of tweets while making two assumptions: 1) one tweet has one hidden topic assignment; and 2) a given tweet may contain both topical words and background words, where the former are words specific to the topic of the tweet and the latter are words that are popular in many tweets. Experiments suggest that T-LDA can capture more meaningful topics than LDA in Twitter data [3], and T-LDA is further extended into Behavior-LDA (B-LDA) [4] to jointly model the topic interests and interactions of a user. B-LDA assumes a universal behavior distribution instead of a personalized behavior distribution for each topic, as the former ensures the behavior information is a property of "topic". In this case, by examining a user's "topic" distribution, one may find personal behavior patterns and topic interests. In other words, a "topic" here is a behavioral topic. To avoid confusion, we refer to a behavioral topic as a *topic* in this study.

One study [13] reviews some of the previous studies that integrate some of the 5 dimensions considered in this paper, and the proposed model (W^4) supports four dimensions (who, when, where, and what). However, W^4 cannot distinguish varying user behaviors. In other words, the model is unable to identify topics from posts by the differences in how the users interact with the content. Moreover, W^4 models time as categorical values consisting of week and weekend days, which is very coarse when aiming to find timely topics. To the best of our knowledge, our study is the first that integrates the 5 dimensions in one model. Further, our model considers location and time at a fine granularity. In experiments, we show results based on the use of fine geographical regions and precise time-stamps of tweets.

Another category of studies relevant to our problem is multi-view clustering [8, 16], where each independent view is able to cluster the data. Generally, the method aims to exploit the multiple views to discover the clusters that agree across the views. For example, the Co-EM algorithm [8] is an expectation maximization algorithm that iteratively performs the expectation step in one view, the result of which is passed to the maximization step in another view. In multi-view clustering, each view corresponds to a representation of the same data with different features, and the goal is to cluster the data by making use of multiple features. Our problem is not a clustering problem, and it has a different goal than unsupervised clustering.



l	a location identifier $1 \leq l \leq L$
b	a behavior in $\mathcal{B} = \{post, retweet, reply, mention\}$
z	a topic identifier $1 \leq z \leq Z$
u	a user identifier $1 \leq u \leq U$
n	a tweet identifier $1 \leq n \leq N$
w	a word identifier $1 \leq w \leq M$
t	a time-stamp identifier $1 \leq t \leq T$
x, y	switches, $x \in [0, 2]$, $y \in [0, 1]$
ψ_z	topic-specific behavior distribution
ϕ_z	topic-specific word distribution
ϕ'	background word distribution
θ_u, θ_l, θ_t	user, location, and time dependent topic distributions
α_u, α_l, α_t	priors of θ_u, θ_l, and θ_t
π, φ	Bernoulli and multinomial distributions

Figure 1: Plate notation for our space-time dependent behavioral topic model for microblog posts. The dashed variables will be collapsed out during Gibbs sampling [11]. Priors over all the multinomial or binomial distributions are omitted for clarity.

3 Model

In this section, we present our space-time dependent behavioral topic model as shown in Figure 1.

3.1 Space-Time Dependent Behavioral Topic Model

In B-LDA [4], it is assumed that all tweets posted by a user concern the user's own interests. However, in many cases, users will not only post tweets according to their own topics of interest, but may also post tweets that concern temporal events and location-dependent topics. As a result, at least two additional dimensions may be built into the model, i.e., *space* and *time*. We propose a probabilistic model that jointly models the space and time of tweets for behavior-topic analysis. In the space-time dependent behavior-topic model, we assume to have three types of topic distributions, i.e., user-dependent, space-dependent, and time-dependent topic distributions. Below we present the full model that considers 5 dimensions: user, text, time, location, and behavior.

We assume to have a data set that contains U users. A user u has N_u ($1 \leq u \leq U$) tweets. We use $M_{u,n}$ ($1 \leq n \leq N_u$) to denote the number of words in nth tweet of uth user, and $w_{u,n,m}$ ($1 \leq m \leq M_{u,n}$) to denote the mth word in nth tweet of uth user, where $1 \leq w_{u,n,l} \leq V$ and V is the vocabulary size. Next, $l_{u,n}$ and $t_{u,n}$ denote the location and time, respectively, of the nth tweet of the uth user. Similar to B-LDA, our model assumes a space $\mathcal{B}$ containing all possible types of behaviors. In the case of Twitter, $\mathcal{B} = \{post, retweet, reply, mention\}$. We use $b_{u,n} \in \mathcal{B}$ to denote the behavior of the nth tweet of the uth user.

We now present our model. First, we assume that there are Z hidden topics, where each topic has a multinomial word distribution ϕ_z and a multinomial behavior distribution ψ_z. We pose Dirichlet priors η and β on ϕ_z and ψ_z, respectively.

$$\forall z(\phi_z \sim \text{Dir}(\beta) \text{ and } \psi_z \sim \text{Dir}(\eta)) \tag{31}$$

Recall that we assume to have three types of topic distributions, i.e., a user-dependent distribution θ_u, a space-dependent distribution θ_l, and a time-dependent distribution θ_t. Similarly, we pose Dirichlet priors $\alpha_u, \alpha_l, \alpha_t$ on these distributions.

$$\forall u(\theta_u \sim \text{Dir}(\alpha_u)), \forall l(\theta_l \sim \text{Dir}(\alpha_l)), \text{ and } \forall t(\theta_t \sim \text{Dir}(\alpha_t)) \tag{32}$$

Each tweet has a single topic that is sampled from one of the three topic distributions θ_u, θ_l, and θ_t. Let $\text{Multi}(\pi) \sim \text{Dir}(\gamma)$. We then use a switch $x_{u,n} \sim \text{Multi}(\pi)$ to choose a topic from the three distributions (values

0, 1, and 2 of $x_{u,n}$ indicate switches of the user, location, and time dependent distributions).

$$z_{u,n} \sim \begin{cases} \text{Multi}(\theta_u) & \text{if } x_{u,n} = 0 \\ \text{Multi}(\theta_l) & \text{if } x_{u,n} = 1 \\ \text{Multi}(\theta_t) & \text{if } x_{u,n} = 2 \end{cases}$$

For a tweet with a topic label $z_{u,n}$ ($1 \leq z_{u,n} \leq Z$), the words in this tweet are generated from two multinomial distributions, namely a background word distribution ϕ' and a topic specific word distribution ϕ. Similarly, they are with Dirichlet priors $\phi' \sim \text{Dir}(\beta')$ and $\phi \sim \text{Dir}(\beta)$. Let $\text{Multi}(\varphi) \sim \text{Dir}(\rho)$. We then use a switch $y_{u,n,m} \sim \text{Multi}(\varphi)$ to choose a word from the two distributions indicated by 0 and 1 values of $y_{u,n,m}$.

$$w_{u,n,m} \sim \begin{cases} \text{Multi}(\phi') & \text{if } y_{u,n,m} = 0 \\ \text{Multi}(\phi_{z_{u,n}}) & \text{if } y_{u,n,m} = 1 \end{cases}$$

3.2 Learning and Parameter Estimation

We use the Collapsed Gibbs sampler [11] to obtain samples of the hidden variable assignments and to estimate the model parameters from these samples. We show the derived Gibbs sampling formulas in the following. Proofs are similar to those given in related work [4].

For tweet n of user u, we jointly sample a switch $x_{u,n}$ and its topic label $z_{u,n}$.

$$\begin{aligned} &p(z_{u,n} = z, x_{u,n} = x \mid \mathbf{Z}_{\neg_{u,n}}, \mathbf{X}, \mathbf{L}, \mathbf{T}, \mathbf{B}) \\ &= \frac{n^x + \gamma}{\sum_{x' \in [0,2]}(n^{x'} + \gamma)} \cdot \frac{n^b_z + \eta}{\sum_{z'}(n^b_z + \eta)} \cdot \left[\frac{n^z_u + \alpha_u}{\sum_{z'}(n^{z'}_u + \alpha_u)}\right]^{x=0} \cdot \left[\frac{n^z_l + \alpha_l}{\sum_{z'}(n^{z'}_l + \alpha_l)}\right]^{x=1} \cdot \left[\frac{n^z_t + \alpha_t}{\sum_{z'}(n^{z'}_t + \alpha_t)}\right]^{x=2}, \end{aligned} \tag{33}$$

where l, t, and b denote the location, time, and behavior information; $n^{z'}_u$ refers to the number of times topic z' co-occurring with user u; and other ns are defined in the same way.

For each word $w_{u,n,m} = w$ in tweet n of user u, we sample its switch $y_{u,n,m}$ as follows.

$$p(y_{u,n,m} = y \mid \mathbf{Y}_{\neg_{u,n}}, \mathbf{X}, \mathbf{Z}, \mathbf{L}, \mathbf{T}) = \frac{n^y + \rho}{\sum_{y' \in [0,1]}(n^{y'} + \rho)} \cdot \left[\frac{n^w_{y=0} + \beta}{\sum_{w'}(n^{w'}_{y=0} + \beta)}\right]^{y=0} \cdot \left[\frac{n^w_z + \beta}{\sum_{z'}(n^w_{z'} + \beta)}\right]^{y=1}, \tag{34}$$

where $n^w_{y=0}$ refers to the number of times word w being labeled as a background word.

With the Collapsed Gibbs sampler, we can make the following estimation of the model parameters:

$$\theta_{u,z} = \frac{n^z_u + \alpha_u}{\sum_{z'} n^{z'}_u + Z\alpha_u} \quad \text{user-topic distribution} \tag{35}$$

$$\theta_{l,z} = \frac{n^z_l + \alpha_l}{\sum_{z'} n^{z'}_l + Z\alpha_l} \quad \text{location-topic distribution} \tag{36}$$

$$\theta_{t,z} = \frac{n^z_t + \alpha_t}{\sum_{z'} n^{z'}_t + Z\alpha_t} \quad \text{time-topic distribution} \tag{37}$$

$$\psi_{z,b} = \frac{n^b_z + \eta}{\sum_{b'} n^{b'}_z + B\eta} \quad \text{topic-behavior distribution} \tag{38}$$

$$\phi_{z,w} = \frac{n^w_{z,y=1} + \beta}{\sum_{w'} n^{w'}_{t,y=1} + V\beta}, \quad \text{topic-word distribution} \tag{39}$$

where n^z_u is the number of times z is sampled for user u and n^b_z is the number of times behavior b co-occurs with topic z.

4 Experimental Study

We proceed to evaluate the proposed model. We first describe the datasets and then present the experimental setup. Finally, we report on findings of a set of experiments.

4.1 Data and Settings

We collected all world-wide geo-tagged tweets from the public Twitter Streaming API from April 29 to July 2, 2013, and we choose 10,000 users at random and use all their tweets. We further select 90% of all the tweets at random for training our model and use the remaining 10% of all tweets for evaluating our model.

Our model is able to find user-specific, space and time dependent behavioral topics, making it useful for several real-world tasks. To evaluate the model, we

1. qualitatively analyze the learned word distributions and topic distributions from the model, and we
2. quantitatively evaluate the model against baseline models for the task of location prediction.

In this study, we focus on location and time relevant topics. Our model inherits its behavior dimension from B-LDA. We thus do not discuss behavioral topics.

We ran 1000 iterations of Monte Carlo EM. For the Gibbs sampling steps, we ran 400 iterations for burn-in, and we sampled every 10 iterations to reduce auto-correlation. We fixed the number of topics at 20. (We varied this number from 10 to 100 with a step size of 10 and found the resulting topics to be most meaningful at around 20 by manual examination). For our models and competing baselines, we use grid search on a development set to select the model parameters.

4.2 Qualitative Analysis

(Topics.) Table 2 presents top topic-specific words for some sample topics. The experimental findings show that Twitter users often talk about themselves, for example, topic "daily life" is a popular topic that mostly concerns the users' daily updates. Similarly, topic "school" looks to be on updates about school. The topic "music" is about songs, country music, pandora, etc. All these topics are readily identified based on their top topical words. They can also serve as interpretable labels for the corresponding tweets or users.

We note that some of the extracted topics are featured with location information. For example, tweets related to topic "movie" are mostly posted from locations close to a cinema. This suggests that some locations have their own topics and relevant words; thus, based on the words used by users, we can draw clues about users' locations. In light of this, we study location prediction in Section 4.3.

(Location and Time Dependent Events.) Unlike related work [13], the proposed model considers temporal information at a fine granularity. This allows us to discover bursty events, i.e., topics with a sudden increase of usage. We define a burstiness score of topic t on day d as $s(t,d) = \frac{c_{t,d} - c_{t,d-1} + 1}{c_{t,d-1} + 1}$, where $c_{t,d}$ denotes the number of tweets with topic t on day d.

Table 3 visualizes top bursty topics sorted by $s(t,d)$ as obtained using our model. We find that all these bursty events are meaningful. The first bursty event is about the release of the movie Iron Man 3. The second concerns a concert. The third one concerns a political event. Note that in the proposed model, each topic has a location distribution; Thus, all the bursty events above have a location dimension. Close examination shows that the first bursty event has tweets are from all over the world and is global. The second one is more localized as its tweets are from the Philippines. The third one happened in the UK when the UKIP leader Nigel Farage was on a campaign visit to Edinburgh. By using our model, we find that the locations associated with the event are indeed from the UK. In all, we find that by considering space and time in the modeling of topics of microblog posts, we can obtain better insights into the behavioral topics of users, locations, and times.

"daily life"	"god"	"cars"	"school"	"music"	"movie"	"food"	"drink"
good	god	car	school	song	movie	food	drink
today	lord	drive	year	music	watch	eat	smoke
tired	world	ride	class	shows	watching	ice	water
early	jesus	house	summer	trend	show	cream	beer
nap	bless	hit	days	listen	funny	pizza	drinking
ready	live	driving	hate	listening	fast	chicken	bottle
day	man	street	test	album	movies	breakfast	smoking
school	beautiful	walking	exam	favorite	game	chocolate	blunt
wake	give	walk	back	world	favorite	hot	cold
shower	woman	bus	start	songs	guy	cheese	juice
long	good	road	ill	love	episode	ate	drunk
feeling	blessed	hate	homework	country	purge	hungry	coffee
woke	love	work	final	pandora	wolf	dinner	smell
night	life	gas	math	topic	teen	cake	cup
awake	pray	truck	english	taylor	family	ill	drank

Table 2: Top topic-specific words from $\phi_{z,w}$ for sample topics. Labels are assigned manually.

Dates	Tweets	Label
April 29, 2013	Iron man 3 wiff @rasekarini (@ Studio 21 - @cinema21 w/ 18 others) Uuurgh Can't wait to watch "IronMan", Seems like it's awesome movie Hype for this new Iron Man movie....I ♡ Marvel Watching Iron Man 3 (at @cinema21) http://t.co/QMVlvpkgIk	Iron Man 3
April 30, 2013	Daniel padilla live paperview yipee < 3 Okaaay so like, naa man daw payperview sa concert ni daniel padilla =)) Daniel padilla invades not just araneta,but also the twitter world.. Rocking my souvenir!! Daniel padilla concert #DanielLiveAtTheBigdome	Daniel Live! Concert
May 16, 2013	Yes. Yes, he is. RT @juliahobsbawm: Nigel Farage is a Black Swan. Nigel Farage has a great taste in suits and hats it must be said! Well that's my vote. Viva Nigel Farage! http://t.co/MC7k84YgIO Johnny don't do Nigel Farage as he would look exactly the same #UKIP	Nigel Farage is heckled

Table 3: Bursty topics found by our model and sample tweets. Labels are assigned manually.

4.3 Location Prediction

We apply our model to a location prediction task. Specifically, given a tweet from a user, the task is to predict the location where the user posted it. The intuition is that many locations have their own topics. For example, if a location is a food court, people tend to tweet more about food in this location. Our method is that we first obtain location-dependent topics learned by the proposed model; then, given a tweet with a set of words and a behavior, we estimate its topics and find the most relevant location, detailed as follows.

For tweet n from user u with words $w_{u,n}$ and behavior $b_{u,n}$, we predict its location by using this formula: $l_{u,n} = \text{argmax}_l\, p(l|w_{u,n}, b_{u,n}, \psi, \phi, \phi', \theta)$. Here, ψ, ϕ, ϕ', and θ are learned using Equations 36-39. We further

compute $p(l|w_{u,n}, b_{u,n}, \psi, \phi, \phi', \theta)$ as follows.

$$\begin{aligned} p(l|w_{u,n}, b_{u,n}, \psi, \phi, \phi', \theta) &\propto p(l)p(w_{u,n}, b_{u,n}|l) \\ &= p(l)\sum_z p(z|l)p(w_{u,n}, b_{u,n}|z) \\ &= p(l)\sum_z \theta_{l,z}p(w_{u,n}|z)p(b_{u,n}|z) \\ &= p(l)\sum_z \theta_{l,z}\psi_{z,b_{u,n}}\prod_{w\in w_{u,n}}\phi_{z,w} \end{aligned} \tag{40}$$

For simplicity, we assume all the words are topic-specific, and we approximate $p(l)$ by using the popularity of location l.

Recall that we use 90% of all tweets for learning and that we have held out 10% of all tweets for testing. For each tweet in the test data, we compute its probability of belonging to a certain location l using the above method. We then sort the locations based on the probabilities. The higher the real location of the tweet is ranked, the better our method is. We consider three baseline methods:

1. Random. By using random guessing, the expected ranking of the real locations will average at around 50%.

2. Majority. The majority baseline always ranks the locations by their popularity. This works well when the held-out tweet locations are from popular locations.

3. Clustering method. This method treats all tweets with the same location as a cluster, and for a new tweet, we compute its similarity to all the clusters and rank all the locations according to the similarity scores. To measure the similarity between a tweet and a cluster of tweets, we use the averaged Jaccard index score.

As for evaluation, we use these metrics: average ranking r_{average} (the lower the better), median ranking r_{median} (the lower the better), and mean reciprocal rank MRR (the higher the better), defined as follows. $r_{\text{average}} = \frac{1}{|D_{test}|}\sum_{d\in D_{test}}\frac{r_d}{|L|}$, $MRR = \frac{1}{|D_{test}|}\sum_{d\in D_{test}}\frac{1}{r_d}$. Metric r_{median} is similar to r_{average}, but uses the median ranking instead of average. Here, r_d refers to the real location's ranking for tweet d, $|L|$ is the total number of locations, and D_{test} is the test set. These criteria have also been used for a similar task, followee recommendation [4, 19].

Table 4 shows the results. The majority method performs worse than the random method. This means that the held-out tweet locations are often from less frequent locations. By Wilcoxon signed-rank test [5] and the results in Table 4, we obtain that the clustering method outperforms both majority and random methods at 0.1% significance level. This implies that many locations indeed have their own location-specific behavioral topics. Our method also outperforms the other methods at 0.1% significance level by Wilcoxon signed-rank test. Using average ranking, our method ranks the real locations in the top 12.6% of all the locations, and with a median ranking at around 2.7% that means that the real location of a given tweet is ranked in the top 2.7% of all the locations. Since the location set size is large in our data set, the findings show that our method can learn good location specific topics and topic specific words.

Metric	Our Model	Random	Majority	Clustering
r_{average}	**0.126**	0.5	0.55	0.132
r_{median}	**0.027**	0.5	0.54	0.103
MRR	**0.090**	0.0001	0.0003	0.015

Table 4: Comparison of the methods used for location prediction.

A given tweet may not necessarily be location-specific but could be user-specific or time-specific. To address this, we propose a simple way to compute a confidence score to measure whether a tweet is location-specific or not. We define the confidence score $s(d)$ as the aggregated probability of the tweet d belonging to a certain location: $s(d) = \sum_l p(l|w_d, b_d, \psi, \phi, \phi', \theta)$. Below we show the results of our model in terms of different confidence scores.

Metric	10%	20%	50%	80%	100%
r_{average}	**0.076**	0.104	0.126	0.130	0.126
r_{median}	0.034	0.036	0.039	0.034	**0.027**
MRR	0.073	0.071	0.070	0.081	**0.090**

Table 5: Findings for location prediction. We use set of tweets with different confidence scores. Thus, n% means that the tweets with the n% highest confidence scores are used.

Using r_{average}, our method is best at tweets with top 10% confidence scores. Our method ranks the real locations in the top 7.6% when using the top 10% most confident tweets, while our method ranks the real locations in the top 12.6% when considering all tweets. Results at 10% are better than at 20% and 50% in terms of both r_{median} and MRR, which indicates that top confident tweets often benefit location prediction. The table also shows that the MRR score at 10% is not as high as at 100%. The reason may be that on the tweets with top 80–100% confidence scores, the variance is smaller than that using tweets with top 10–50% confidence scores. Similar observation can be found for r_{median}, and the results also show that r_{median} seems to be rather insensitive to the percentage.

5 Conclusion and Future Work

In this study, we propose to model space and time dependent behavioral topics of microblog posts that associated with text, timestamps, geographical locations, and user behaviors with users. The experiments on Twitter data demonstrate our model is able to identify useful and insightful user behavioral topics with a fine spatial and temporal granularity.

There may exist a range of applications of our model, including user location inference, link prediction, user or location profiling by the changes of topics over time, burst event detection, and automatic tagging semantic text to geographical locations. Applications such as these deserve exploration in future work. Moreover, it may also be promising to integrate other contextual types, such as popularity of images on Instagram, in our model, or to find a generalized way to integrate social context with textual content in the model.

Acknowledgments

We thank Minghui Qiu and Anna Tigunova for the contribution in parts of the discussion and programming. This research is supported by the Russian Science Foundation under Grant No. 15-11-10032.

References

[1] X. Cao, L. Chen, G. Cong, C. S. Jensen, Q. Qu, A. Skovsgaard, D. Wu, and M. L. Yiu. *Spatial Keyword Querying*. In ER, 2012, pp. 16–29.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. *Latent dirichlet allocation*. J. Mach. Learn. Res., vol. 3, pp. 993–1022, 2003.

[3] X. Zhao, J. Jiang, J. Weng, J. He, E. P. Lim, H. Yan, X. Li. *Comparing Twitter and Traditional Media Using Topic Models*. In ECIR, 2011, pp. 338–349.

[4] M. Qiu, F. Zhu, and J. Jiang. *It Is Not Just What We Say, But How We Say Them: LDA-based Behavior-Topic Model*. In SDM, 2013, pp. 794–802.

[5] I. C. A. Oyeka and G. U. Ebuh. *Modified Wilcoxon Signed-Rank Test*. Open Journal of Statistics, no. 2, pp. 172–176, 2012.

[6] Q. Qu, S. Liu, B. Yang, and C. S. Jensen. *Integrating non-spatial preferences into spatial location queries*. In SSDBM, 2014, Article 8.

[7] C. R. Vicente, D. Freni, C. Bettini, and C. S. Jensen. *Location-Related Privacy in Geo-Social Networks*. IEEE Internet Computing, vol. 15, no. 3, pp. 20–27, 2001.

[8] S. Bickel and T. Scheffer. *Multi-view clustering*. In ICDM, 2004, pp. 19–26.

[9] Y. Zhou, G. Cong, B. Cui, C. S. Jensen, and J. Yao. *Routing Questions to the Right Users in Online Communities*. In ICDE, 2009, pp. 700–711.

[10] Q. Qu, S. Liu, B. Yang, and C. S. Jensen. *Efficient Top-k Spatial Locality Search for Co-located Spatial Web Objects*. In MDM, 2014, pp. 269–278.

[11] J. S. Liu, *The Collapsed Gibbs Sampler in Bayesian Computations with Applications to a Gene Regulation Problem*. Journal of the American Statistical Association, vol. 89, no.427, pp. 958–966.

[12] C. Hage, C. S. Jensen, T. B. Pedersen, L. Speicys, and I. Timko. *Integrated data management for mobile services in the real world*. In VLDB, 2003, pp. 1019–1030.

[13] Q. Yuan, G. Cong, Z. Ma, A. Sun, and N. Magnenat-Thalmann. *Who, where, when and what: discover spatio-temporal topics for twitter users*. In SIGKDD, 2013, pp. 605–613.

[14] Y. Wang, E. Agichtein, and M. Benzi. *TM-LDA: efficient online modeling of latent topic transitions in social media*. In SIGKDD, 2012, pp. 123–131.

[15] M. C. Gonzlclez, C. A. Hidalgo, and A. L. Barabasi. *Understanding individual human mobility patterns*. Nature 453, pp. 479–482, 2008.

[16] S. Sun. *A survey of multi-view machine learning*. Neural Computing and Applications, vol. 23, no. 7–8, pp. 2031–2038, 2013.

[17] Q. Qu, S. Liu, C. S. Jensen, F. Zhu, and C. Faloutsos. *Interestingness-Driven Diffusion Process Summarization in Dynamic Networks*. In ECML/PKDD, 2014, pp. 597–613.

[18] L. Hong, A. Ahmed, S. Gurumurthy, A. J. Smola, and K. Tsioutsiouliklis. *Discovering geographical topics in the twitter stream*. In WWW, 2012, pp. 769–778.

[19] W.-Y. Chen, J.-C. Chu, J. Luan, H. Bai, Y. Wang, and E. Y. Chang, *Collaborative filtering for orkut communities: discovery of user latent behavior,* in WWW, 2009, pp. 681–690.

Taqreer: A System for Spatio-temporal Analysis on Microblogs*

Amr Magdy[2] Mashaal Musleh[1] Kareem Tarek[1] Louai Alarabi[2]
Saif Al-Harthi[1] Hicham G. Elmongui[1,4] Thanaa M. Ghanem[3]
Sohaib Ghani[1] Mohamed F. Mokbel[2]

[1]KACST GIS Technology Innovation Center, Umm Al-Qura University, Makkah, KSA
[2]Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN, USA
[3]Department of Information and Computer Sciences, Metropolitan State University, Saint Paul, MN, USA
[4]Department of Computer and Systems Engineering, Alexandria University, Alexandria, Egypt
{amr,louai,mokbel}@cs.umn.edu, {mmusleh,ktarek,sharthi,elmongui,sghani}@gistic.org,
thanaa.ghanem@metrostate.edu

Abstract

Motivated by the wide popularity of microblog services (e.g., Twitter and Facebook) along with the sheer sizes of microblogs issued every second, this article introduces Taqreer *as a scalable and efficient system for auto-generation of spatio-temporal analysis reports on microblogs.* Taqreer *is composed of two main modules: The* Taghreed *query engine, which is a scalable and efficient query processing engine for spatio-temporal keyword queries on microblogs and a* Report Generation Tool*, which receives the user analytic report request and divides it into a set of queries sent to the* Taghreed *engine, and a set of analysis tasks executed on top of the returned query answers. As of now,* Taqreer *is able to produce three analysis report types, namely comparative reports, categorical reports, and image gallery reports. Other report types will be added in the future.*

1 Introduction

Microblogs, e.g., tweets and Facebook comments, have become incredibly popular in the last few years. Everyday, over a billion of users post about four billions microblogs on Twitter and Facebook [5, 20]. As user-generated data, microblogs are associated with various types of rich contents, including user locations, used language, event updates, news items, opinions, reviews, and/or discussions. With the importance of temporal aspect in microblogs [4, 11] (i.e., more recent microblogs are more important than older ones), and the wide availability of location information of microblogs [3, 12], a high fraction of analysis applications on microblogs rely on *spatio-temporal* analysis. Examples of such analysis include user analysis for geo-targeted advertising [14], event detection [1, 7, 13, 16, 21], news extraction [2, 15, 17], and analysis [6, 18, 19]. Unfortunately, existing systems cannot manage microblogs data efficiently as they are designed for managing either fast or

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

*This work is supported by KACST GIS Technology Innovation Center at Umm Al-Qura University (GISTIC), under project GISTIC-13-06, and was all done when all the authors were only affiliated with GISTIC.

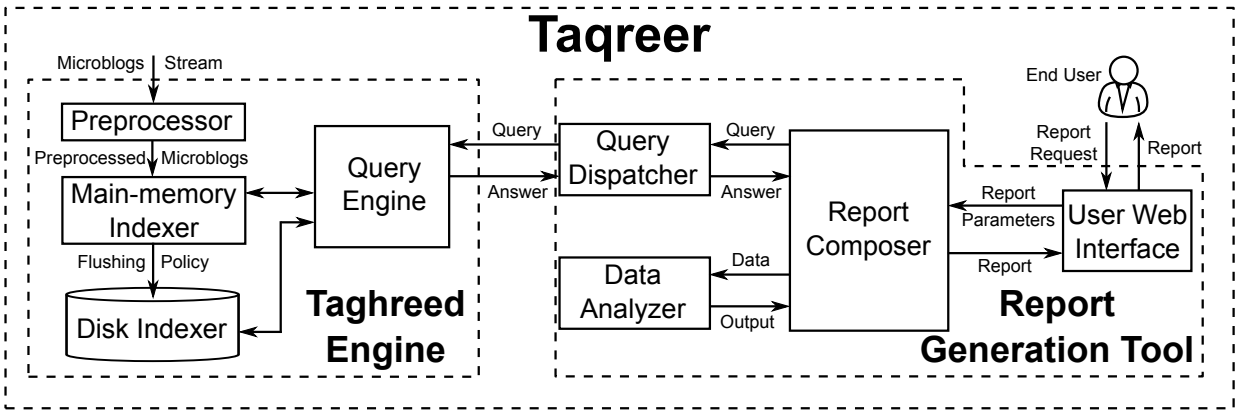


Figure 1: Taqreer Architecture.

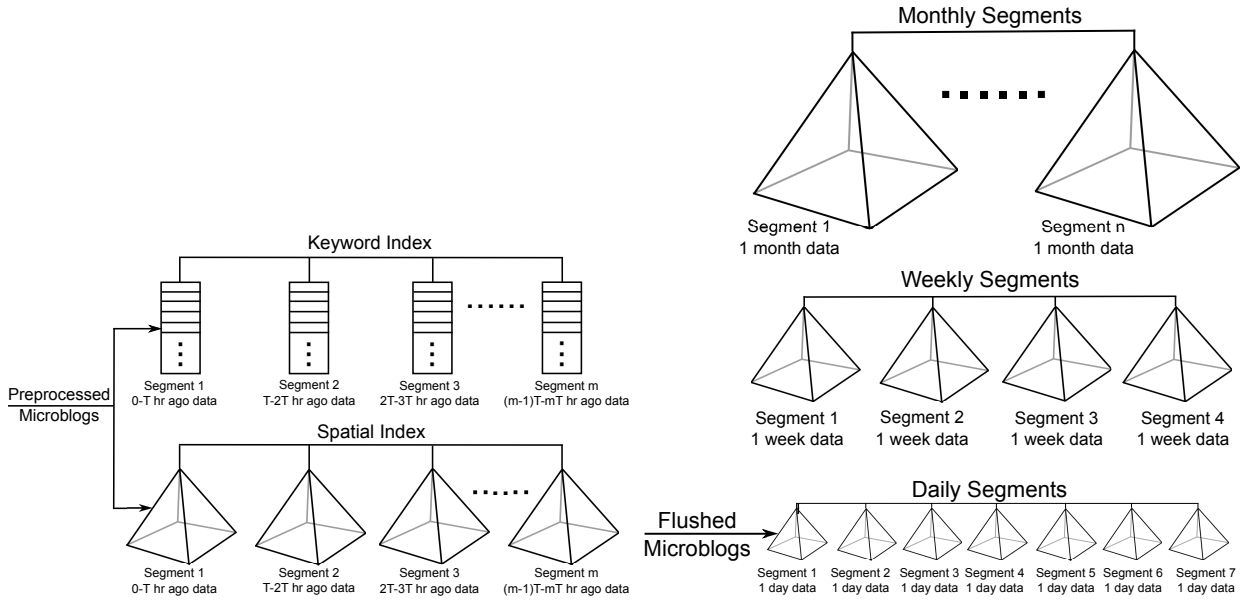
large data. Thus, none of existing systems provide indexing-based data management for data that is fast and large simultaneously, like microblogs.

In this article, we report on our work for the *Taqreer* system. *Taqreer* is a system for generic spatio-temporal analysis and report generation on large numbers of microblogs. *Taqreer* users can generate an analysis report that tracks the appearance of a certain keyword over space and time. *Taqreer* is so scalable that it can generate such report for the whole world over a period of more than a year. Such reporting functionality is important for tracking interest in diseases (e.g., Ebola) or natural disasters (e.g., Nepal earthquake). *Taqreer* users can also generate scalable and interactive comparison reports that compares the trending of various keywords over space and time, which can be used to analyze the status of election candidates, the interest in various soccer teams, or other comparisons based on social media discussions. Other *Taqreer* reports include categorical reports that study the used languages over space and time, which is a powerful report in understanding the diversity of various countries and cities worldwide. In general, *Taqreer* is an extensible system as it provides a rich platform that allows adding various reports easily.

2 System Overview of Taqreer

Figure 1 gives *Taqreer* system architecture, which is composed of two main system components, the *Taghreed* query engine [8, 9] (detailed in Section 3) and a *Report Generation Tool* (detailed in Section 4). *Taqreer* users submit their report generation requests to a *User Web Interface* module. Then, the report parameters are forwarded to a *Report Composer* module, which parses the parameters and divides the report into: (a) a set of spatio-temporal keyword queries that retrieve the necessary data to generate the requested report, and (b) a set of analysis tasks to run on the retrieved answer of the spatio-temporal keyword queries. All spatio-temporal keyword queries are sent to the *Taghreed* query engine through a *Query Dispatcher* module.

Meanwhile, the *Taghreed* query engine continuously receives an incoming stream of microblogs with high arrival rates of up to 5,000 microblogs per second. The incoming stream is processed and digested in a highly scalable and efficient main-memory index structure. Once the memory becomes full, a flushing policy is invoked to move a portion of memory contents to the disk storage in another disk-based highly scalable and efficient index structure. The *Taghreed* query engine answers its incoming spatio-temporal queries from both in-memory and disk-based index structures, based on where the data needed for the query answer reside. Once the query answer is collected, it is sent back to the *Report Generation Tool*, which invokes the *Data Analyzer* module to perform the required data analysis tasks. Finally, the report composer gets the analysis output, composes the report in its final form as an interactive web page and sends it back to the requesting user.



(a) Memory-resident Indexes (b) Disk-resident Spatial Index

Figure 2: Taghreed Indexes Organization.

3 Taghreed Query Engine

Taghreed [8,9] is the query engine behind *Taqreer*, which has two main responsibilities: (1) Digesting incoming stream of microblogs with high arrival rates, and (2) Efficient support for spatio-temporal keyword queries over large set of microblogs. *Taghreed* is composed of five main components, namely, *in-memory index structure*, *disk-based index structure*, *flushing policy*, *query optimizer*, and *recovery manager*, described below.

In-memory index structure. *Taghreed* employs two in-memory index structures; a keyword index and a spatial index. Both indexes are segmented into temporal segments that partition data based on arrival timestamp. Figure 2(a) gives the organization of the memory indexes. Each index is segmented into disjoint segments, where each segment includes the data of the last T hours, where T is a system parameter. Incoming microblogs are digested in the most recent segment. Once the segment spans T hours of data, it is concluded and a new empty segment is introduced to digest the new data. Index segmentation has two main advantages: (a) new microblogs are digested in a smaller index, which is the most recent one, and hence we can support higher digestion rates, and (b) it makes it easier to flush data from memory to disk.

Disk-based index structure. Similar to main-memory index structures, disk-based indexing supports both spatial and keywords attributes, where each index embeds the temporal aspect in its organization. However, the disk-based index structures are a bit different from the main-memory ones. Figure 2(b) gives the organization of *Taghreed* disk-based spatial index structure. The disk-based keyword index has a similar structure. The index is organized in temporally partitioned segments. The temporal segments are replicated in a hierarchy of three levels, namely, daily segments, weekly segments, and monthly segments. The daily segments level stores data of each calendar day in a separate segment. Each weekly segment level consolidates the data in daily segments of one calendar week. Similarly, each monthly segment level consolidates data of weekly segments of a whole calendar month. The main reason behind the three levels replication is to minimize the number of accessed index segments while processing queries for different temporal periods. For example, for an incoming query asking about data of two months, if only daily segments are stored, then the query processor would access sixty index structures to answer the query. On the contrary, the query processor would access only two index structures of

the two months time horizon of the query. This significantly reduces the query processing time for queries on relatively long periods.

Flushing policy. The main goal of *Taghreed* flushing policy is to determine which microblogs should be flushed from main-memory indexes to disk indexes when the memory becomes full. *Taghreed* flushing manager allows the system administrator to employ one of multiple available flushing policies. The default flushing policy is *flush-temporal*, which depends only on timestamps to select its victim microblogs to be flushed. In particular, we expel a portion of the oldest microblogs to empty a room for the newly real-time incoming microblogs. To reduce the flushing overhead, this policy requires the main-memory indexing to partition the data into segments with the same flushing unit. Referring to the main-memory index organization in Figure 2(a), the flushing unit is defined as T hours, i.e., the oldest T hours of data are flushed periodically. T is adjusted by a system administrator based on the available memory resources, the rate of incoming microblogs, and the desired frequency of flushing.

Query optimizer. *Taghreed* query optimizer selects which index segment(s) should be accessed to retrieve the query answer. Specifically, *Taghreed* provides two index structures in both main-memory and disk: a keyword index and a spatial index. In addition, disk-resident data is replicated in three temporal levels, daily, weekly, and monthly index segments. Consequently, the query processor may have different ways to process the same query based on: (1) the order of hitting keyword or spatial indexes, and (2) the number of disk index segments to hit. For example, a query that spans from May 1 to May 9 can be answered from disk indexes by either accessing nine daily index segments, or accessing one weekly and two daily index segments. Each of these is called a query plan. The costs of different query plans are different. The main task of the query optimizer is to generate a plan to execute so that the estimated cost is minimal. To this end, the query optimizer employs two cost estimation models, one for estimating the cost of accessing a keyword index segment and the other for estimating the cost of accessing a spatial index segment. Using these cost models, the query optimizer proposes two selection criteria, one for the main-memory index, where each index has a single level of disjoint segments, and one for the disk index, where each index has three levels of overlapped segments.

Recovery manager. With millions of microblogs managed in main-memory, *Taghreed* system accounts for memory failures that may lead to data loss. *Taghreed* employs a simple, yet effective, triple-redundancy model where the main-memory data is replicated three times over different machines. The core idea of this model is similar to Hadoop redundancy model that replicates the data three times. In particular, when *Taghreed* is launched, all the main-memory modules are initiated on three different machines. Each machine is fed with exactly the same stream of microblogs, thus they form two backup copies of the main-memory system status. Any flushing from memory to disk leads to throwing the data out from the memory of backup machines. On memory failure, the backup machines continue to digest the real-time microblogs and one of them work as a replacement to the system memory contents. Replicating the data three times significantly reduces the probability of having the three machines down simultaneously and lose all the main-memory data.

4 Taqreer Report Generation Tool

Taqreer report generation tool is composed of four modules, namely, *User Web Interface*, *Report Composer*, *Query Dispatcher*, and *Data Analyzer*. The user web interface is the interfacing module between *Taqreer* and its end users. The user input parameters are forwarded to the report composer module to sync the work among other modules. In particular, the composer goes through the following four steps: (1) Based on the report type and parameters, the composer determines the set of queries that retrieve the required data and a set of analysis tasks to be performed on that data, (2) The report composer calls the query dispatcher module to submit spatio-temporal keyword queries to *Taghreed* query engine, (3) The retrieved query answers are forwarded to the data analyzer module to perform the required analysis, (4) The report composer adds all the output to an interactive web page and sends it as the final report to the user.

Report Type	Parameters	Queries	Analysis Tasks
Comparative Reports	• n spatial regions R_i, $1 \leq i \leq n$ • m keywords (topics/entities) W_j, $1 \leq j \leq m$ • Time range [T_s,T_e]	$n \times m$ queries, each takes: • Spatial region R_i • Keyword W_j • Time range [T_s,T_e]	None
Categorical Reports	• Spatial region R, auto divided into n sub-regions of fixed default area • Time range [T_s,T_e] • Categorical attribute A • Optional m keywords W_j, $1 \leq j \leq m$	n queries, each takes: • Spatial region $R_i \subset R$ • Time range [T_s,T_e] • Optional m keywords W_j, $1 \leq j \leq m$	• Count categories of attribute A for each query microblogs • Aggregate counts over less granular spatial levels
Image Gallery Reports	• m keywords W_j, $1 \leq j \leq m$ • Time range [T_s,T_e] • Optional spatial region R	One query that takes: • m keywords W_j, $1 \leq j \leq m$ • Time range [T_s,T_e] • Optional spatial region R	• Extract photos

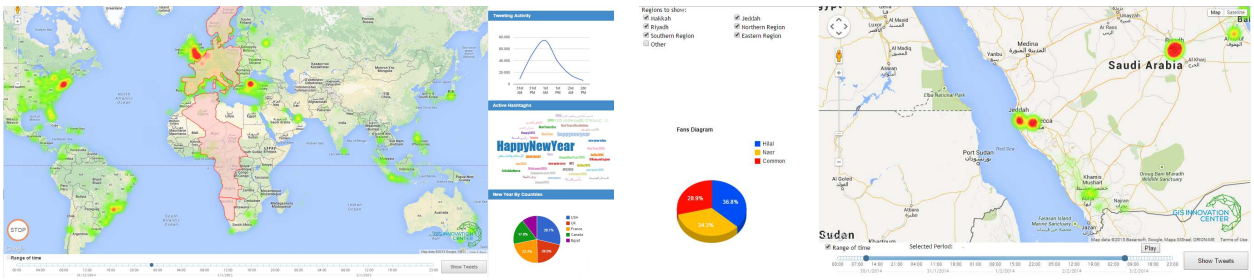
Table 6: Parameters, Queries, and Analysis of Different Report Types

Taqreer supports three types of spatio-temporal reports, namely, *Comparative Reports*, *Categorical Reports*, and *Image Gallery Reports*. Table 6 gives the parameters, queries, and analysis tasks for each of the three reports. Details of these reports are described in the rest of this section.

4.1 Comparative Reports

Comparative reports aim to compare individual microblogs that are related to different topics (or entities) in different spatial regions within a certain time range. Topics (or entities) are defined by a set of keywords/hashtags. An example of these reports is to compare tweets about the two Spanish soccer teams *Real Madrid C.F.* and *FC Barcelona* in different cities in Spain during the week of their soccer game. This can also include analysis related to presidential candidates, political parties, product trademarks, or events. The first row in Table 6 gives the parameters, queries, and analysis tasks of comparative reports. The user inputs n spatial regions of interest R_i, $1 \leq i \leq n$, m entities or topics (identified by keywords W_j, $1 \leq j \leq m$), and an arbitrary time range [T_s,T_e]. A set of $n \times m$ queries are submitted to *Taghreed* query engine to retrieve the report data, each query takes a spatial region R_i, a keyword W_j, and the time range [T_s,T_e]. The retrieved data is displayed in an interactive web page that allows arbitrarily inclusion/exclusion of microblogs of certain spatial regions and navigation along the whole report timeline, either for a single point of time or on a time range. The analysis of these reports include creating a heatmap for the microblogs, optional pie charts that show percentage analysis for the displayed microblogs, optional tag cloud that shows popular topics in the displayed microblogs, and locating and displaying individual microblogs on a geographical map with full text and user information. In the rest of this section, we present using comparative reports for two purposes: (1) analyzing event-specific tweets, and (2) general-purpose social media analysis.

Analyzing event-specific tweets. One of the most important and consistent behavior of Twitter users is posting a plethora of tweets about *events* of different types, e.g., Oscars ceremony, soccer games, and natural disasters. For such kind of event-specific tweets, geotagged tweets grab a high attention as all events, by nature, have a spatial extent. For example, while Boston Marathon explosions were going on in April 2013, users rush to Twitter seeking tweets from the marathon location [2]. An event is generally defined by a temporal horizon, a set of hashtags, and a spatial extent. Events fall in one of two categories: (a) multi-side events, e.g., sports games or elections, and (b) independent events, e.g., Oscars ceremony or New Year Eve. Figure 3 gives two examples of event-specific reports. Figure 3(a) gives a heatmap for New Year Eve tweets over different points of time. The figure shows popular hashtags in a tag cloud and enables going through exact tweets with full text and user information. Figure 3(b) gives tweets of a soccer game in Saudi Arabia, where an extra option is added



(a) New Year Eve 2015 tweets (b) Soccer game tweets

Figure 3: Spatio-temporal Analysis of Event-specific Tweets.

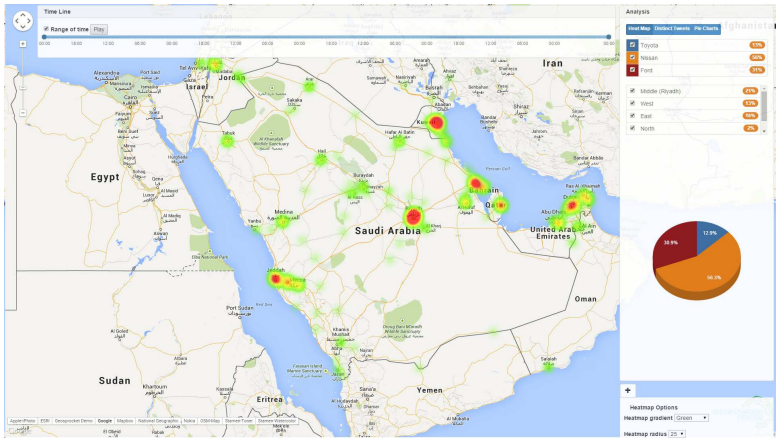


Figure 4: Analyzing tweets mentioning different car brands in Saudi cities.

to classify visualized tweets based on local cities and show percentage of tweets that support each team.

General-purpose social media analysis. With its generic usage, comparative reports can be used as a powerful tool to analyze social media contents. Figure 4 gives an example for a generated report by *Taqreer* that has: (1) January-March 2015 as the temporal horizon, (2) Riyadh, West, East, and North districts as sub-regions within Saudi Arabia, and (3) Toyota, Nissan, and Ford as entities to analyze. The figure shows heatmap as well as individual tweets of each car brand on a geographical map, a time line that allows navigation in different time instances and/or ranges, percentage of tweets that mention each car brand in a pie chart, and spatial filters to include/exclude tweets of each district. Such a generic tool for analyzing social media contents is very helpful in getting insights from the public discussions in different contexts and applications.

4.2 Categorical Reports

The plethora of social media active users enables meaningful analysis tasks that can deduce fruitful conclusions for actual population. One of the underutilized attributes are the *categorical attributes*: the attributes that can take one of multiple discrete values. Prime examples of important categorical attributes in Twitter data include the language attribute that indicates the language used in each tweet and the tweet source attribute, which determines from which operating system, device, or application the tweet is posted. Categorical reports in *Taqreer* perform spatial aggregate analysis over a categorical attribute for microblogs that lie within certain spatial and temporal ranges. As the second row in Table 6 shows, the user inputs a spatial region of interest R, an arbitrary time range $[T_s,T_e]$, a categorical attribute A, and an optional set of keywords. The report composer divides the space

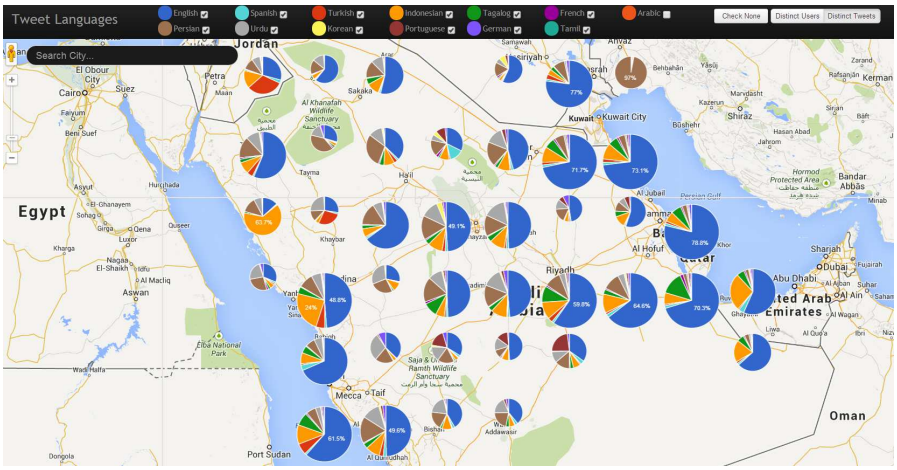


Figure 5: Tweets Languages Spatial Analysis in Arab Gulf Countries.

into n small spatial regions of default fixed size. Then, a set of n queries are submitted to *Taghreed*, each query takes one of the small regions, the time range [T_s,T_e], and the set of keywords. Each query retrieves individual microblogs that lie within the query parameters. The retrieved data is forwarded to the data analyzer module to count microblogs in different categories of attribute A. After counting is performed for all queries, the counts are then aggregated at higher levels of spatial granularity to support zoom in/out analysis in the final report. Finally, the report composer puts all the aggregates on pie charts aligned with latitude/longitude coordinates of a geographical map and embed all of this in a web page. This forms an interactive web page that is sent as the final report to the user.

Figure 5 gives an example of analyzing tweets languages in Arab Gulf countries. The figure gives a pie chart for each sub-region/city. Each pie chart shows the distribution of tweets languages in its region. Zooming in/out gives a finer/coarser granular analysis for language distributions up to the street level. Users can arbitrarily include/exclude languages from the top bar to focus on a subset of all languages. This language analysis, combined with ground truth data enables a full study on language diversity and minorities in local communities [10].

4.3 Image Gallery Reports

Image gallery reports exploit the availability of many photos on the social media to summarize certain topics or entities through creating a photo gallery for their microblogs. An example of such reports is to extract and organize photos that are posted in response to a certain event, e.g., human crisis, terrorist attack, elections, or sports game. Topics and entities are defined by a set of keywords/hashtags. Analyzed microblogs should lie within a certain time range and can optionally be filtered based on a spatial region of interest. As described in the third row in Table 6, users input m keywords W_j, $1 \leq j \leq m$, a time range [T_s,T_e], and an optional spatial region R. A single query with the input parameters is submitted to *Taghreed* to retrieve the report data. The retrieved microblogs are scanned to extract their images. Extracted images are organized and displayed in an interactive web page that allow users to navigate, enlarge, and share portions of the report on social media websites. Figure 6 gives an image gallery for the event of 2015 Chapel Hill Shooting. The shown images are extracted for the hashtag #ChapelHillShooting for 11 days after the accident happened. Thus, *Taghreed* is queried with time range of February 10 to February 20, 2015 and hashtag #ChapelHillShooting. The returned tweets are analyzed to extract their images and organized them as the figure shows.

Figure 6: Tweets Image Gallery for 2015 Chapel Hill Shooting.

5 Conclusion

This article presented *Taqreer*; a scalable and efficient system for auto-generation of spatio-temporal analysis reports on microblogs. *Taqreer* is composed of two main modules, the *Taghreed* query engine, which is responsible for efficiently supporting spatio-temporal keyword queries on microblogs, and a *Report Generation* tool, which is responsible for receiving the user requests, extracting the required queries for the report, sending them to the *Taghreed* query engine, and performing a set of analysis and visualization tasks on top of the returned query results. We have presented three report types as example of what *Taqreer* can generate, namely, comparative reports, categorical reports, and image gallery reports. For each report type, we show the user input parameters, the queries that will be sent to the *Taghreed* query engine, and the set of analysis tasks that will be performed on top of the returned query answers. Other report types can be defined within the *Taqreer* system in a similar way.

References

[1] Hamed Abdelhaq, Christian Sengstock, and Michael Gertz. EvenTweet: Online Localized Event Detection from Twitter. In *Proceedings of the International Conference on Very Large Data Bases, VLDB*, 2013.

[2] After Boston Explosions, People Rush to Twitter for Breaking News. http://www.latimes.com/business/technology/la-fi-tn-after-boston-explosions-people-rush-to-twitter-for-breaking-news-20130415,0,3729783.story, 2013.

[3] Ceren Budak, Theodore Georgiou, Divyakant Agrawal, and Amr El Abbadi. GeoScope: Online Detection of Geo-Correlated Information Trends in Social Networks. In *Proceedings of the International Conference on Very Large Data Bases, VLDB*, 2014.

[4] Michael Busch, Krishna Gade, Brian Larson, Patrick Lok, Samuel Luckenbill, and Jimmy Lin. Earlybird: Real-Time Search at Twitter. In *Proceedings of the IEEE International Conference on Data Engineering, ICDE*, 2012.

[5] Facebook Statistics. http://newsroom.fb.com/company-info/, 2015.

[6] Harvard Tweet Map. worldmap.harvard.edu/tweetmap/, 2013.

[7] Rui Li, Kin Hou Lei, Ravi Khadiwala, and Kevin Chen-Chuan Chang. TEDAS: A Twitter-based Event Detection and Analysis System. In *Proceedings of the IEEE International Conference on Data Engineering, ICDE*, 2012.

[8] Amr Magdy, Louai Alarabi, Saif Al-Harthi, Mashaal Musleh, Thanaa Ghanem, Sohaib Ghani, Saleh Basalamah, and Mohamed Mokbel. Demonstration of Taghreed: A System for Querying, Analyzing, and Visualizing Geotagged Microblogs. In *Proceedings of the IEEE International Conference on Data Engineering, ICDE*, 2015.

[9] Amr Magdy, Louai Alarabi, Saif Al-Harthi, Mashaal Musleh, Thanaa Ghanem, Sohaib Ghani, and Mohamed Mokbel. Taghreed: A System for Querying, Analyzing, and Visualizing Geotagged Microblogs. In *Proceedings of the ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM GIS*, 2014.

[10] Amr Magdy, Thanaa M. Ghanem, Mashaal Musleh, and Mohamed F. Mokbel. Exploiting Geo-tagged Tweets to Understand Localized Language Diversity. In *the International ACM Workshop on Managing and Mining Enriched Geo-spatial Data, GeoRich. In conjunction with SIGMOD*, 2014.

[11] Amr Magdy and Mohamed Mokbel. Towards a Microblogs Data Management System. In *Proceedings of the International Conference on Mobile Data Management, MDM*, 2015.

[12] Amr Magdy, Mohamed F. Mokbel, Sameh Elnikety, Suman Nath, and Yuxiong He. Mercury: A Memory-Constrained Spatio-temporal Real-time Search on Microblogs. In *Proceedings of the IEEE International Conference on Data Engineering, ICDE*, 2014.

[13] Adam Marcus, Michael S. Bernstein, Osama Badar, David R. Karger, Samuel Madden, and Robert C. Miller. Twitinfo: Aggregating and Visualizing Microblogs for Event Exploration. In *Proceedings of the International Conference on Human Factors in Computing Systems, CHI*, 2011.

[14] New Enhanced Geo-targeting for Marketers. https://blog.twitter.com/2012/new-enhanced-geo-targeting-for-marketers.

[15] Owen Phelan, Kevin McCarthy, and Barry Smyth. Using Twitter to Recommend Real-Time Topical News. In *Proceedings of the ACM Conference on Recommender Systems, RecSys*, 2009.

[16] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors. In *Proceedings of the International Conference on World Wide Web, WWW*, 2010.

[17] Jagan Sankaranarayanan, Hanan Samet, Benjamin E. Teitler, Michael D. Lieberman, and Jon Sperling. TwitterStand: News in Tweets. In *Proceedings of the ACM Symposium on Advances in Geographic Information Systems, ACM GIS*, 2009.

[18] Topsy Analytics: Find the insights that matter. www.topsy.com, 2014.

[19] TweetTracker: track, analyze, and understand activity on Twitter. tweettracker.fulton.asu.edu/, 2014.

[20] Twitter Statistics. https://about.twitter.com/company, 2015.

[21] Kazufumi Watanabe, Masanao Ochi, Makoto Okabe, and Rikio Onai. Jasmine: A Real-time Local-event Detection System based on Geolocation Information Propagated to Microblogs. In *Proceedings of the ACM International Conference on Information and Knowledge Management, CIKM*, 2011.

Data Engineering

It's FREE to join!

TCDE

tab.computer.org/tcde/

The Technical Committee on Data Engineering (TCDE) of the IEEE Computer Society is concerned with the role of data in the design, development, management and utilization of information systems.

- Data Management Systems and Modern Hardware/Software Platforms
- Data Models, Data Integration, Semantics and Data Quality
- Spatial, Temporal, Graph, Scientific, Statistical and Multimedia Databases
- Data Mining, Data Warehousing, and OLAP
- Big Data, Streams and Clouds
- Information Management, Distribution, Mobility, and the WWW
- Data Security, Privacy and Trust
- Performance, Experiments, and Analysis of Data Systems

The TCDE sponsors the International Conference on Data Engineering (ICDE). It publishes a quarterly newsletter, the Data Engineering Bulletin. If you are a member of the IEEE Computer Society, you may join the TCDE and receive copies of the Data Engineering Bulletin without cost. There are approximately 1000 members of the TCDE.

Join TCDE via Online or Fax

ONLINE: Follow the instructions on this page:

www.computer.org/portal/web/tandc/joinatc

FAX: Complete your details and fax this form to **+61-7-3365 3248**

Name _____

IEEE Member # _____

Mailing Address _____

_____

Country _____

Email _____

Phone _____

TCDE Mailing List

TCDE will occasionally email announcements, and other opportunities available for members. This mailing list will be used only for this purpose.

Membership Questions?

Xiaofang Zhou
School of Information Technology and Electrical Engineering
The University of Queensland
Brisbane, QLD 4072, Australia
zxf@uq.edu.au

TCDE Chair

Kyu-Young Whang
KAIST
371-1 Koo-Sung Dong, Yoo-Sung Ku
Daejeon 305-701, Korea
kywhang@cs.kaist.ac.kr

IEEE Computer Society
1730 Massachusetts Ave, NW
Washington, D.C. 20036-1903

Non-profit Org.
U.S. Postage
PAID
Silver Spring, MD
Permit 1398