

# A Study on User Location Inference in Social Media

Graduate School of Systems and Information Engineering

University of Tsukuba

March 2014

Yuto Yamaguchi



## Acknowledgements

First of all, I would like to thank my supervisor Professor Hiroyuki Kitagawa. This thesis has been done under the direction of him, and would not have been possible without his helpful advices and encouragement. Great faculty members, Associate Professor Toshiyuki Amagasa, Lecturer Hideyuki Kawashima, Assistant Professor Yasuhiro Hayase, and Assistant Professor Chiemi Watanabe also helped me. I would especially like to thank Associate Professor Toshiyuki Amagasa for his advices and support.

I have spent a great summer at IBM Research Tokyo as an intern. I would like to thank the members of the group I belonged to: Miki Enoki, Yohei Ikawa, Takashi Imamichi, Akiko Murakami, Hidemasa Muta, Toyotaro Suzumura, Michiaki Tatsubori, and Hideo Watanabe. During the internship I got the idea of this work thanks to the helpful advices from them.

I would like to thank my thesis committee members, Professor Hiroyuki Kitagawa, Professor Jiro Tanaka, Professor Mikio Yamamoto, Associate Professor Toshiyuki Amagasa, and Associate Professor Jun Sakuma. Their advices and comments are helpful to improve the thesis. I have been a research fellow of the Japan Society for the Promotion of Science (JSPS) since April 2012; This research has been supported in part by JSPS.

I would also like to thank members of Kitagawa Data Engineering Laboratory. I appreciate Hiroko Odagiri and Yumiko Hisamatsu for their tremendous support. I am grateful to my seniors, Yutaka Kabutoya, Takahiro Komamizu, Hiroaki Shiokawa, and Tsubasa Takahashi for their support and good will.

My college life would not be the same without members of Picnic Tennis Team. Playing tennis, playing TV games, eating dinners, everything became a cherished memory. Thank you so much, tennis gang.

I would like to thank my parents, Ryosuke and Emiko, and my brother, Takuto, for always being supportive. Without your comprehensive support, I would not be here. Thank you very much all this time. Finally, I would like to thank my wife, Shoko, for standing beside me. Your love and encouragement have always helped me.

## Abstract

The exponential growth of online social media allows individuals to transmit information anytime and anywhere. Vast quantities of information are sent in real time from around the globe using social media, enabling the real world to be monitored and modeled through the web. Governments and public institutions are now able to analyze economies, politics, and public health by leveraging this information. In addition, companies benefit from such information to get reviews of their products and to market them.

Residential information of social media users plays a crucial role because governments and companies need to know where information is transmitted from for real-time monitoring or for marketing products to individuals in a specific location. However, most social media users are unwilling to disclose their residences publicly due to several reasons, including privacy concerns. Hence, there are growing needs to accurately infer and utilize residence information. This thesis tackles problems with inferring social media users' residential information to not only enhance the potential of the abovementioned applications, but also to examine how privacy of personal information is divulged to further consider privacy related issues.

In particular, this thesis investigates three variants of the inference algorithm that utilize different types of clues. First, we examine the use of *graph landmarks* in social graphs, which are user accounts receiving local attention (e.g., a local administrative bureau and a local weather report). Second, we leverage real-world *local events* (e.g., earthquake or tornado) for location inference. Third, we take advantage of the content streams, or *social streams*, to allow online location inference. These three algorithms all achieve a higher accuracy of inference than existing ones. This improved accuracy and online algorithm should broaden the field of location-related social media applications.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	The Importance of Location Information . . . . .	3
1.2	Contributions . . . . .	4
1.2.1	Graph Landmarks for Location Inference . . . . .	5
1.2.2	Real-World Local Events for Location Inference . . . . .	5
1.2.3	Online Location Inference Algorithm . . . . .	6
1.3	Overview of the Thesis . . . . .	7
<b>2</b>	<b>Background and Survey</b>	<b>8</b>
2.1	Monitoring the Real World using Social Media . . . . .	8
2.1.1	Economics . . . . .	9
2.1.2	Politics . . . . .	11
2.1.3	Public Health . . . . .	13
2.1.4	Disasters . . . . .	14
2.1.5	Event Detection . . . . .	15
2.2	User Location Inference in Social Media . . . . .	17
2.2.1	Problem Definition . . . . .	18
2.2.2	Content-Based Approach . . . . .	19
2.2.3	Graph-Based Approach . . . . .	21
2.2.4	Hybrid Approach . . . . .	24

2.2.5	Discussion . . . . .	24
<b>3</b>	<b>User Location Inference based on Graph Landmarks</b>	<b>30</b>
3.1	Introduction . . . . .	30
3.2	Problem Statement . . . . .	34
3.3	Graph Landmarks . . . . .	34
3.3.1	Definition . . . . .	34
3.3.2	Preliminary Experiment . . . . .	36
3.3.3	Examples . . . . .	38
3.4	Proposed Method . . . . .	39
3.4.1	Model Formulation . . . . .	40
3.4.2	Parameter Estimation . . . . .	42
3.4.3	Thresholds to Adjust the Trade-offs . . . . .	43
3.5	Experiments . . . . .	45
3.5.1	Experimental Setups . . . . .	45
3.5.2	Comparison with Existing Methods. . . . .	47
3.5.3	Comparison of Target Areas . . . . .	49
3.5.4	Comparison of Variations of the Proposed Method. . . . .	51
3.5.5	Effect of the Confidence Threshold . . . . .	52
3.5.6	Effect of the Degree Threshold . . . . .	54
3.6	Conclusion . . . . .	55
<b>4</b>	<b>User Location Inference based on Local Events</b>	<b>58</b>
4.1	Introduction . . . . .	58
4.2	Problem Statement . . . . .	61
4.3	Proposed Method . . . . .	62
4.3.1	Local Event Detection . . . . .	63
4.3.2	User Location Inference . . . . .	65

4.4	Experiments . . . . .	68
4.4.1	Prototype System . . . . .	69
4.4.2	Dataset . . . . .	69
4.4.3	Validity of Detected Events . . . . .	70
4.4.4	Accuracy of User Location Inference . . . . .	72
4.4.5	Coverage of User Location Inference . . . . .	74
4.4.6	Effect of Parameters . . . . .	76
4.5	Conclusion . . . . .	78
<b>5</b>	<b>Online User Location Inference based on Social Streams</b>	<b>80</b>
5.1	Introduction . . . . .	81
5.2	Problem Statement . . . . .	83
5.3	Proposed Method . . . . .	83
5.3.1	Overview . . . . .	84
5.3.2	Geo Clustering . . . . .	85
5.3.3	Statically-Local Words Extraction . . . . .	88
5.3.4	Temporally-Local Words Extraction . . . . .	92
5.3.5	Location Inference Model . . . . .	93
5.3.6	Sequential Inference . . . . .	97
5.3.7	OLIM Algorithm . . . . .	97
5.4	Experiments . . . . .	98
5.4.1	Experimental Setups . . . . .	99
5.4.2	Comparison with Existing Methods . . . . .	101
5.4.3	Error Reduction over Time . . . . .	103
5.4.4	Effects of Parameters . . . . .	105
5.5	Conclusion . . . . .	109
<b>6</b>	<b>Conclusion and Future Work</b>	<b>110</b>

6.1 Summary of Contributions . . . . . 111

6.2 Future Work . . . . . 112

**Bibliography** **114**

**Reference Papers** **129**

**Other Papers** **129**



# List of Figures

3.1	CDF plot of the distance distribution between home locations of mutually connected users in Twitter. 60% or more of connected users are at least 100km apart. . . . .	31
3.2	Upper right user is regarded as a graph landmark because this user has followers in a small region, while upper left user is not regarded as a graph landmark because followers of this user are dispersed. Lower right user is regarded as a graph landmark because this user has a lot of followers, while lower left user is not regarded as a graph landmark because this user does not have a lot of followers. . . . .	35
3.3	CDF of users over their dispersions for the top 5% users with high degrees. This figure indicates that there are $y\%$ of users that have dispersions lower than $x$ . Although most users have a large dispersion, there exist users with relatively small dispersion, which are regarded as graph landmarks. . . . .	37
3.4	Distributions of both all users and graph landmarks. Red and blue dots represent home locations of all users in our dataset and dominance locations of graph landmarks, respectively. Both populations have similar distributions. Graph landmarks can cover a large segment of the user population; that is, most users can find a graph landmark near their home location. . . . .	38

3.5	Accumulative precision of our method and existing methods at various error distances using US dataset. Our method successfully locates about a half of the users within a 1km error distance, and outperforms the other methods, including the state-of-the-art method. . . . .	49
3.6	Accumulative precision of our method and existing methods at various error distances using JP dataset. Although the results of our method and Backstrom seem to be in the same level, our method outperformed other methods.	50
3.7	Accumulative precision for variations of our method at various error distances. Although considering the geographical dispersion improves the precision, considering the mixture weight does not. Medoid and Centroid do not work well because they simply use the dominant location as the points rather than as the distributions. . . . .	53
3.8	Effect of the confidence threshold. x-axis denotes the value of the threshold $p_0$ . As the value of $p_0$ increases, the precision increases but the coverage decreases. The F-measure achieves the best score around $p_0 = 0.003$ , and the precision is about 0.88 at $p_0 = 0.02$ . . . . .	54
3.9	Effect of the degree threshold. x-axis denotes the value of the threshold $c_0$ . As $c_0$ increases, the ratio of utilized graph landmarks rapidly decreases, preserving the high coverage value. The decrease in the average number of neighbors denotes the reduction of the computational cost of our method. The precision remains high or even increases when we utilize a small number of graph landmarks. . . . .	56

3.10	Effect of the degree threshold. x-axis denotes the value of the threshold $c_0$ . As $c_0$ increases, the ratio of utilized graph landmarks rapidly decreases, preserving the high coverage value. The decrease in the average number of neighbors denotes the reduction of the computational cost of our method. The precision remains high or even increases when we utilize a small number of graph landmarks. . . . .	57
4.1	(a) The large part of messages are posted from metropolises such as Tokyo and Kyoto. (b) Unlike Figure 4.1(a), most tweets are posted from the Hiroshima prefecture, indicating that users who mention the event tend to have geographical proximity with the event. . . . .	60
4.2	Procedure of the content clustering. Each text is represented as the term vector, and is clustered in Euclidean space. Posts in a sparse region are disposed of as noises. . . . .	64
4.3	Procedure of the spatial filtering. Each post in a cluster is plotted in the Euclidean space. If the dispersion does not exceed the <i>Maxdispersion</i> , the cluster is regarded as an event. . . . .	64
4.4	Overview of the prototype system. Two crawling components collect data that is used for event detection and user location inference are conducted. .	70
4.5	Precision of detected events. The precision of <i>Emergency</i> is low because few local events happened during this experiment in this dataset. As the number of known user locations increases, the precision increases slightly. . . . .	72
4.6	The number of detected events. The number of detected events considerably differs from each other. More events are detected as the number of known user locations increases. . . . .	73
4.7	Accuracy of user location inference. Precision within 160km error distance of our method indicates improvements of 34% and 122% compared to the methods of UDI and Cheng, respectively. . . . .	75

4.8	Efficiency and accuracy as <i>MaxDispersion</i> varies. Results indicate a trade-off between coverage and accuracy. . . . .	77
4.9	Coverage and accuracy with varied <i>WindowSize</i> . Increased <i>WindowSize</i> causes an increase in accuracy but a decrease in coverage. . . . .	78
4.10	Coverage and accuracy with <i>Eps</i> . Although accuracy is not affected by <i>Eps</i> , the coverage decreases when <i>Eps</i> has an extremely value. . . . .	79
5.1	Diagram of OLIM. . . . .	85
5.2	The result of geo clustering. Each circle shows the center of the corresponding region. The size of the circle is proportional to the mixture weight of the region. Large regions are located at metropolises (e.g., New York, Chicago, and Tokyo) both in the United States and in Japan. . . . .	87
5.3	Comparison of different methods. A high value indicates a good result. The best result is from the proposed OLIM method. . . . .	103
5.4	Impact of TL-words on location inference. TL-words can improve location inference. . . . .	104
5.5	Computational time spent for five methods. . . . .	105
5.6	Error reduction over time. Online location inference can reduce the error distance as new tweets arrive. . . . .	106
5.7	Effects of parameters. . . . .	108

# Chapter 1

## Introduction

Historically, the transmission of information has been unidirectional from mass media to people. Ordinary people have not had the ability to convey information to the public. However, the recent rise of online social media has caused a paradigm shift [1]. The exponential growth of social media has enabled individuals to transmit information anytime and anywhere. Unlike traditional mass media where off-the-shelf information is broadcast to all people, individuals can exchange information in social media based on personal interests or preferences. This freedom to share has caused the web to accumulate massive amounts of data about what people think, like, and do.

Although the term of social media can be defined in many different ways, the principal requirement is that social media users can interact with each other by exchanging their information [1]. According to this requirement, social media has existed for hundreds of years. A good example is a coffee shop where people gather, discuss their interests, and even publish their collective knowledge. However, the audience size in traditional social media is relatively small. Today there are many online services where users exchange various types of content, and these services attract massive numbers of people. For example, Myspace [2] and Facebook [3] users interact online. Flickr [4] users exchange photos. Youtube [5] and Ustream [6] users broadcast videos. Delicious [7] users share their bookmarks. Twitter [8]

and Chinese Weibo [9] users exchange short messages, and Foursquare [10] users share their current locations. Today, the term social media usually implies one of these types of online services.

The emergence of online social media has had a great impact.

- **Voice of the people.** Social media contents reflect the voice of the people, describing what people think, like, and do. For example, leveraging these contents, governments can understand citizens' needs and concerns or companies can obtain reviews of their products from actual users. In addition, social media plays a crucial role in launching and maintaining political activities; the largest example is the Arab Spring. Various contents exchanged on social media (e.g., photos, videos, and texts) can be used to model information consumption on the web and to improve search engines [11, 12], recommendation systems [13], etc.
- **Large scale social network.** Sociologists have long studied social networks, which represent friendships and interactions among people. Because past studies have relied solely on methods such as questionnaires and interviews, the analysis of social networks has been limited to a small scale. However, the appearance of online social media has impacted social science. Most online social media have large-scale social networks, or *social graphs*, where nodes, which represent users, are connected if corresponding users have friendships or some kind of interaction. Social graphs realize larger-scale analysis of social networks. For example, the *six degrees of separation* or *small world phenomena* [14, 15] was experimentally demonstrated in several large-scale networks [16, 17, 18]. Various works [16, 18, 19] have investigated the *power-law degree distribution* [20] and *homophily* or *assortative mixing* [21]. In addition to these famous studies, large social networks support diverse and comprehensive social studies.
- **Real-world sensing.** Vast quantities of real-time contents are posted in social media from around the globe. In fact, approximately 1.11 billion Facebook users post content

on a daily basis<sup>1</sup>, and approximately 200 million Twitter users send over 400 million messages per day<sup>2</sup>. Hence, monitoring social media data helps understand real-world phenomena, such as economics [22], politics [23], public health [24], and disasters [25].

## 1.1 The Importance of Location Information

In addition to these aforementioned impacts, location-related information in social media receives much attention. Location information in social media includes:

1. **Home location** (place where a user lives)
2. **Current location** (place where a user is currently)
3. **Posting location** (place where the content is transmitted)
4. **Mentioned location** (place mentioned in the content)

The first three locations are users' physical locations, whereas the last one is a location that users are interested in (e.g., resorts and restaurants). Although mentioned locations can be utilized for several purposes [26], users' physical locations are more important to sense the real world. Integrating social media contents with location information provides insight on what and where events are happening. For example, real-world events (e.g., earthquakes) can be detected using Twitter messages with location information where the messages were sent [27]. In addition, global ailments can be analyzed [24]. The availability of large-scale social media data with the location information reduces the laborious task of conducting traditional surveys to acquire such data.

However, because users' physical locations are not explicitly available in most cases, it is difficult to perform detailed investigations. For example, influenza analysis remains at the state level because city- or zip code level analysis is difficult with the location information currently available from Twitter [24]. Moreover, detecting small-scale events

---

<sup>1</sup>[http://en.wikipedia.org/wiki/Facebook\\_statistics](http://en.wikipedia.org/wiki/Facebook_statistics)

<sup>2</sup><http://en.wikipedia.org/wiki/Twitter>

is also difficult due to the scarcity of contents with location information. For example, more than 99.5% of Twitter messages are not associated with their location information (GPS tags) [28] and approximately 94% of Facebook users do not disclose their residential information [29]. Therefore, this thesis addresses the problem of inferring users' *home locations*. It is considered that users' current locations and posting locations can also be inferred by their home locations assuming that social media users mainly post contents near their home locations.

The impact of the results of this thesis is wide-ranging. Enlarging the amount of location information and ensuring location information is accurate not only expand the capabilities of real-world sensing, but also realize several applications. As for the real-world sensing, most research, including influenza analysis and event detection, should be conducted on finer scales. By classifying social media users into administrative districts, governments can collect public opinions by state or city, which may improve government services. In addition, companies are interested in marketing with regard to location because analyzing where products sell well can facilitate product distribution.

On the other hand, home location inference has some issues, including privacy concerns. A user who wants to keep his/her residence private should be allowed to do so. In fact, this research may help understand how private information is exposed, reducing the risk of unintentional disclosure.

## 1.2 Contributions

A lot of research has investigated issues with user location inference in social media (see Section 2.2). Existing location inference methods can be classified as either graph-based approaches or content-based approaches. Graph-based approaches analyze social networks (e.g., [29]), while content-based approaches analyze user-generated contents in social media (e.g., [28]). This thesis tackles research challenges in both approaches. The following subsections describe the contributions of this thesis by comparing to existing research.



### 1.2.1 Graph Landmarks for Location Inference

Most existing graph-based approaches are based on the *closeness assumption*, which assumes that connected users (i.e., friends) tend to live close each other. For example, a user with many friends who live in Tsukuba may also live in Tsukuba. However, not all social graphs follow this assumption. According to McGee et al. [30], a significant fraction of Twitter users are geographically distant even if they are connected in the social graph. Our preliminary analysis also supports this assertion (see Section 3.1).

Therefore, Chapter 3 introduces another assumption called the *concentration assumption*. The concentration assumption states that there are users that most of their friends are in a small region, which are called *graph landmarks*. Graph landmarks provide another strategy to infer users’ home locations. For example, if a graph landmark in Tsukuba is identified, then a user who is a friend of the graph landmark also lives in Tsukuba. Based on this novel assumption, we propose a method to infer home locations, called the *landmark mixture model*. This method can be applied to any social media that has a social graph where users are connected.

A large-scale experiment using a Twitter dataset demonstrates that our new concept and method improve the accuracy of inferring locations compared to existing methods. Moreover, our method is applicable to social graphs that do not satisfy the closeness assumption, expanding the usability of graph-based approaches.

### 1.2.2 Real-World Local Events for Location Inference

Content-based approaches assume that users tend to post contents related to their residences (e.g., toponyms and dialect). For example, because Twitter users living in Houston, USA tend to post contents that include the word “rockets” [28], “rockets” is regarded as a clue to infer the home locations of these users. Words that are strongly associated with a certain location are called *local words*. Although these location-related contents are informative with respect to home locations, contents contain noise, degrading the accuracy of location

inference. For example, a user who lives in Tokyo may post about a trip to Osaka, creating noise when inferring his/her home location.

Hence, Chapter 4 focuses on temporal features, namely, real-world local events. Social media users tend to post about real-world local events that occur around them (e.g., earthquakes and typhoons) [27]. For example, when a tornado hit Tsukuba, Japan<sup>3</sup>, many social media users who lived in Tsukuba posted contents about the tornado. Using contents related to such local events, we infer that a user who posts about a tornado when a tornado hits Tsukuba may live in Tsukuba. Therefore, we propose a location inference method based on real-world local events, which is applicable to the social media where event-related contents are posted in real-time (e.g., microblogs).

Our proposed method is the first location inference method to utilize temporal features. The proposed inference method is more accurate than existing methods.

### 1.2.3 Online Location Inference Algorithm

In social media, a lot of contents are posted continuously, and the amount of content grows rapidly. In this situation, content-based approaches should be able to continuously infer home locations of users based on streams of social media contents (i.e., *social streams*). However, existing content-based approaches cannot perform online inference. Consequently, existing methods must repeat the whole inference process based on all the data, which leads to high computational and storage costs.

Due to this inconvenience, Chapter 5 proposes an online location inference algorithm based on social streams. Our novel algorithm can perform online inferences when new contents become available, avoiding the high costs of existing methods. Specifically, the proposed algorithm utilizes two types of local words: *statically-local words* (*SL-words*) and *temporally-local words* (*TL-words*). Statically-local words are steadily associated with a certain location, and are adopted by several existing content-based approaches [28] (e.g.,

---

<sup>3</sup><http://www.bbc.co.uk/news/world-17974487>

toponyms and dialect), whereas temporally-local words, which are introduced in this thesis, are temporally associated with a certain location (e.g., earthquake, typhoon, and thunder). Each time a statically-local or temporally-local word is observed, our algorithm infers and updates home locations of users who post observed local words. Our method is designed for the social media where user-generated contents are posted in real-time (e.g., microblogs).

Our online algorithm reduces the computational cost and the storage cost of online inference. At the same time, our algorithm achieves high accuracy and high coverage by using two types of local words.

### 1.3 Overview of the Thesis

This thesis is organized as follows. Chapter 2 surveys existing studies on real-world monitoring using social media and user location inference in social media. Specifically, the problem of user location inference is precisely defined, and the important points including merits and demerits of location inference are discussed. Chapters 3–5 describe the three proposed methods. In Chapter 3, a novel graph-based method, which is called the *Landmark Mixture Model (LMM)* is described and experimentally compared to other existing major graph-based methods. One experiment compares the accuracy of location inference in two different areas (i.e., the United States and Japan). In Chapter 4, a content-based method based on real-world events called *Event-based Location Inference Method (ELIM)* is proposed and compared to existing methods using a large-scale Twitter dataset composed of event-related tweets in Japan. In Chapter 5, an online algorithm of user location inference called the *Online Location Inference Method (OLIM)* is proposed and compared to other existing major content-based methods. Finally, Chapter 6 concludes this thesis with the future work.

## Chapter 2

# Background and Survey

As mentioned above, various kind of analyses have become possible by the emergence of social media. In this chapter, we start with the survey of the problem of monitoring the real world based on social media (Section 2.1), which has attracted much attention in the past several years. Researches of monitoring the real world can be considered as applications of user location information, because there are a number of researches dealing with geographical analysis. Section 2.2 defines the problem of user location inference in social media, and carefully reviews and discusses existing researches related to this problem.

### 2.1 Monitoring the Real World using Social Media

Social media users tend to transmit the information related to incidents or concerns around them. That kind of information is considered to reflect the real world and is therefore valuable to investigate. This section surveys existing researches that perform the analysis on the real world based on the information transmitted in social media. Especially, we focus on the representative categories of researches below:

- Analysis on economics such as predicting stock prices. (Section 2.1.1 )
- Analysis on politics including predicting the election outcomes. ( Section 2.1.2 )

- Analysis on public health such as influenza epidemics. ( Section 2.1.3 )
- Analysis on effects of disasters such as earthquakes and typhoons. ( Section 2.1.4 )
- Event detection method based on the contents posted in social media. ( Section 2.1.5 )

Although analyses of economics and politics are not related to the location information, there are a lot of geographical analysis on the public health, disasters, and event detection.

Note that most of researches on this topic utilize Twitter because a vast amount of information is posted in the form of text in real time in Twitter, which is an important feature for monitoring the real world. In addition to Twitter, there are several researches taking advantage of Flickr, a photo sharing service, and Weibo, Chinese microblogging service.

### 2.1.1 Economics

Researches that analyze economics based on the web have been conducted since around 2005. Gruhl et al. [31] predicted sales amount of books analyzing weblogs, and Mishne and Glance [32] predicted box-office sales of movies by sentiment analysis [33] on weblogs. In addition, Zhang and Skiena [34] also predicted box-office sales of movies integrating news articles and the IMDB database<sup>1</sup>, and Joshi et al. [35] addressed the same problem by linear regression analysis on various review texts. Following these studies, Twitter has started to be utilized to analyze economics such as predicting stock prices.

Asur et al. [22] proposed a method to predict box-office ratings of movies analyzing Twitter. They reported that the prediction accuracy of a relatively simple model on Twitter outperformed the market based prediction result. Specifically, in the experiments of predicting the stock prices of movies in Hollywood Stock Exchange, the accuracy in the level of 0.97 of adjusted R2 was achieved. Moreover, they showed how sentiment information derived from Twitter improved the prediction accuracy.

---

<sup>1</sup><http://www.imdb.com/>

O'Connor et al. [36] studied the correlation between Gallup CCI (consumer confidence index) and the analysis result on Twitter. According to their results, although the correlation coefficient varied with the dataset, it achieved about 80% in the best case. They argued that this light-weight analysis on Twitter can potentially replace the traditional massive-scale investigations that require much time and effort. This result indicates that it is possible to do more extensive researches with more sophisticated NLP tools.

Bollen et al. [37] addressed the problem of predicting Dow Jones Industrial Average (DJIA) based on the public mood extracted from Twitter. They made use of two sentiment analysis tools: OpinionFinder<sup>2</sup> and GPOMS (Google-Profile of Mood States) to extract public mood from texts. The former can classify texts into positive or negative ones, and the latter can measure mood of texts in terms of 6 dimensions (Calm, Alert, Sure, Vital, Kind, Happy). Their results showed that their method could predict the daily up and down changes in the closing prices of the DJIA with an accuracy of 87.6%.

Zhang et al. [38] predicted Dow Jones, NASDAQ, and S&P 500 stock prices using Twitter messages. They first derived collective *hope*, *fear*, and *worry* from tweets per day, and studied the correlation between these indices and stock market indicators. It was shown that the percentage of emotional tweets is negatively correlated with Dow Jones, NASDAQ, and S&P 500, significantly.

Ruiz et al. [39] studied the problem of correlating Twitter users' activity with stock-market events: changes in the price and trading volume of stocks. They extracted two features. Features in the first category measure the Twitter users' posting activities such as the number of posts and the number of retweets. Features in the second category are properties of an *interaction graph* they proposed, which include the number of components, statistics on the degree distribution, and other graph-based properties. They reported that the most predictive feature is the number of components in the interaction graph. Moreover, the stock trading strategy based on their features outperformed other baseline strategies.

Mao et al. [40] integrated various kinds of data (Twitter, news headlines, and volumes of

---

<sup>2</sup><http://mpqa.cs.pitt.edu/opinionfinder/>

Google search queries) and extracted sentiments from them to predict market indices such as the Dow Jones Industrial Average, trading volumes, and market volatility (VIX), as well as gold prices. They compared the prediction results based on the extracted sentiments with the ones by the traditional offline survey. For example, their results showed that the frequency of occurrences of financial terms in Twitter in the previous 1-2 days are found to be statistically significant predictors of daily market log return.

### 2.1.2 Politics

It is still fresh in our minds that Obama’s campaign analyzed and utilized social media data in the presidential election in 2008 and 2012<sup>3</sup>. In this way, analyzing politics based on social media data has been widely studied.

Tumasjan et al. [23] analyzed tweets that mentioned political parties or politicians during the German federal election. The results showed that the number of mentions to the parties reflected the election outcomes. The results also demonstrated that joint mentions of two parties are in line with real world political ties and coalitions. Additionally, they found the relationship between sentiments of tweets that mention political issues and politicians’ political positions.

Conover et al. [41] examined two networks of political communication on Twitter, comprising of more than 250,000 tweets from the six weeks leading up to the 2010 U.S. congressional midterm elections. Two networks they constructed are the retweet network and the mention network, which are partitioned by graph clustering. According to the result of graph partitioning, the retweet network was clearly partitioned into left- and right-leaning users, while the mention network was not. In the retweet network, it was found that users propagated contents they like in their community. Besides, they also classified Twitter users with regard to their political affiliations [42]. Concretely, they classified manually annotated 1,000 users by SVM (Support Vector Machine [43]). Comparing hashtags and tweets as the

---

<sup>3</sup><http://swampland.time.com/2012/11/07/inside-the-secret-world-of-quants-and-data-crunchers-who-helped-obama-win/2/>

classification features, it was found that the accuracy of utilizing tweets is higher than that of hashtags.

Pennacchiotti et al. [44] dealt with the problem of constructing user profiles including political affiliations. They considered the user profiling problem as a classification problem then leveraged known user profiles, activities of tweeting and following, and social graphs. Their results demonstrated that accurate classification based on user centric features was feasible while the same accuracy level based on graph centric features was not achievable.

Chen et al. [45] predicted election results. They focused on the point that social media users of different groups have significant differences in their tweet contents and tweet behaviors, and compared the prediction accuracy of election results of utilizing each user group. Their study revealed what type of users have a good predictive power, which leads to the effective user sampling for election prediction.

Weber et al. [46] developed a system that computes political polarization of Twitter hashtags. They first identified *seed users* with known political leanings (e.g., Barak Obama), and then propagated the leanings to users who retweeted the seed users. After political leanings are propagated to users, the system assigns leaning scores by calculating the fraction of leanings of retweeting users.

However, these studies were criticized for overstating by several papers. Metaxas et al. [47] tested predictive powers of existing studies for election results. As a result, it was reported that predicting methods of existing studies did not achieve statistically significant results. Besides, they proposed a set of standards that studies aiming to predict election results using social media should follow. Gayo-Avello [48] surveyed studies dealing with the problem of predicting elections, and also concluded that results of existing studies were greatly exaggerated.



### 2.1.3 Public Health

A milestone paper that tracked influenza epidemics using Google query logs published in Nature [49]. Starting from this paper, a lot of studies have been conducted to analyze public health using Twitter. Researches in early days [50, 51, 52] mainly performed time-series analysis to track the period of influenza epidemics. Although these works did not conduct geographical analysis, they argued that it was important to do so.

After the time-series analyses, public health has been geographically analyzed by associating tweets with location information (e.g., Geo tags and home locations). Takahashi et al. [53] developed a system that maps degrees of spreads of hay fever. In their system, user location profiles were converted into coordinates using the gazetteer, and all tweets were considered to be posted from users' home locations. Poul and Dredze [24] examined various ailments (e.g., flu, obesity, and hay fever) state-by-state. They also used user home locations as locations of tweets, and they argued that it was difficult to do more detailed analyses with the current amount and granularity of location information. Aramaki et al. [54] proposed a method to classify tweets posted by users who caught the flu by SVM. Moreover, they developed a service named as *Influ-kun* [55] which maps degree of epidemics of influenza in a certain area. Lampos et al. [56] also performed a geographical analysis of influenza epidemics in large cities in U.K.

These analyses of public health were made possible by the fact that social media users tend to post about their lives. However, according to Poul and Dredze [24], there are several limitations of such analyses. For example, although Twitter data made it possible to analyze public health at a mass level, it is still difficult to conduct such analyses at a personal level. Besides, it is also hard to study seniors with respect to ailments based on Twitter because most of Twitter users are young people. Therefore, it is considered to be important to understand these limitations and use these online analyses in conjunction with the traditional surveys.

#### 2.1.4 Disasters

In social media, especially in Microblog services such as Twitter, users tend to post the information about disasters when they occur. Hence, situations after disasters can be potentially captured from social media data.

Longueville et al. [57] geographically studied the forest fire that occurred in the southern part of France in 2009 using Twitter data. They experimentally demonstrated that users can be categorized into three types, namely citizens, media, and aggregators with respect to their patterns of information transmission. For example, aggregators do not transmit information by themselves, but they assemble already transmitted information and retransmit it. In addition, they argued that it was difficult to distinguish primary and secondary information because they are mixed in Twitter.

Vieweg et al. [25] examined tweets during the Oklahoma Grass fire of April 2009 and the Red River Floods that occurred in March and April 2009. Their results indicated that it was possible to identify tweets enabling situation awareness from disaster-related tweets. Moreover, they found that disaster-related tweets tend to contain words that help us locate these tweets. Crooks et al. [58] geographically analyzed the earthquake that happened in U.S. in August 2011, and concluded that affected area could be identified based on Twitter data.

Sakaki et al. [59] studied amount of tweets from several parts of Japan after the Great East Japan Earthquake, and found that there were planned blackouts in the Kanto area and that the amount of tweets from the affected area drastically decreased.

The above mentioned studies aimed to identify the affected area or to increase situation awareness. In contrast to these studies, several researchers have investigated social media users' behaviors of information transmission and consumption under the disasters. Starbird et al. [60] showed that how information about the Red River Flood in 2009 was produced and consumed in Twitter. According to their result, Twitter has developed a self-organized mechanism in which users themselves evaluate the value and credibility of information.

Mendoza et al. [61] examined how rumors spread during a few hours after the Chili earthquake in 2010 in Twitter. As a result, rumors and news are different from each other with regard to the form of spread. Specifically, rumors tend to be mentioned by a lot of users with doubt while right information (e.g., news article) does not, which leads to the identification of rumors and preventing the spread of them. Castillo et al. [62] also studied information credibility on Twitter by carefully analyzing information propagation, and concluded that rumors could be detected with about 80% of accuracy based on the propagating patterns.

Qu et al. [63] utilized Chinese microblogging service Weibo to analyze contents about 2010 Yushu Earthquake. Their study includes content analysis of microblog messages, trend analysis of different topics, and an analysis of the information spreading process, and concluded that people used microblog for four purposes, namely situation update, opinion expression, emotional support, and calling for actions. In addition, they also found that messages of situation update tend to be readily posted right after disasters and spread faster than other kind of information. However, it was reported that a filtering mechanism is needed to filter the situation updates if they are overwhelmingly posted.

Miyabe et al. [64] and Toriumi et al. [65] studied how Twitter was used after the Great East Japan Earthquake. According to the result of Miyabe et al., it was found that users in the affected areas tend to communicate each other, while users in the other areas tend to retweet the former users [64]. Toriumi et al. [65] reported that users normally use Twitter as a communication tool, but they use it as an information sharing tool in the emergence.

### **2.1.5 Event Detection**

Social media users tend to post about events around their lives when the events happen. Leveraging these posts, many researches aiming to detect real-world events have been published.

In the early days of this kind of researches, photos uploaded in the photo sharing service

Flickr were utilized to detect events. Rattenbury et al. [66] proposed a method to detect spatio-temporal bursts based on geotagged photos in Flickr. Their method computes the spatio-temporal distributions of tags attached to photos (e.g., New York, World Cup, and dogs) and identifies tags with significant geographical and temporal skews as events. Chen et al. [67] also detected events using photos in Flickr. They adopted wavelet transformation to reduce tags that are not related to events. Periodic and aperiodic events are identified and classified. Experimental result showed that their proposed method more accurately detected periodic events from Flickr photos.

With the exponential growth of Twitter, there appeared a lot of researches dealing with the problem of event detection based on a huge amount of data in Twitter [68, 69, 70, 71, 72, 73, 74] .

Lee et al. [69] devised a method for event detection by monitoring moving behaviors of Twitter users. Their method assumes that local events happen at the area where a lot of Twitter users move into. They conducted experiments using geotagged tweets and showed that the method could detect local events such as local festivals.

Ritter et al. [71] also addressed the problem of event detection leveraging Twitter data. Their proposed method, which is based on the unsupervised topic model, could detect various kinds of events (open domain events) and categorize them. It was experimentally shown that the accuracy of the proposed method was higher than that of the supervised baseline.

Packer et al. [72] tackled the problem of identifying what topics tweets mention, leading to improving the accuracy and the recall of the event detection. They employed the query expansion technique and the semantic web. In their experiments, tweets that were written about a music festival were collected and associated with the bands playing at the festival.

Event detection specialized in some particular domains (e.g., traffic events and earthquakes) have been widely investigated. Daly et al. [75], Ribeiro et al. [76], and Schulz et al. [77] dealt with traffic events. Daly et al. [75] identified the causes of the traffic accidents

in certain areas by using geocoded tweets. Ribeiro et al. [76]’s method detects traffic events, geocodes them into the coordinates of longitude and latitude, and maps them in real time. It was shown by their experiments that it is difficult to detect finer-grained events without mentions to POIs (point-of-interest) such as shopping malls and bars. Although the above mentioned studies dealt with the problem of detecting large-scale events, Schulz et al. [77] addressed the problem of detecting small-scale incidents such as traffic accidents. Their method achieved 82.2% accuracy and 82% recall.

As for the earthquakes, there are a lot of researches including Okazaki et al. [78], Sakaki et al. [27], and Robinson et al. [79]. Okazaki et al. [78] developed a method to classify tweets related to particular events by SVM. Features of the classification method are contents and contexts of tweets, and the number of words contained in tweets. Sakaki et al. [27] applied Okazaki et al.’s method to classify tweets, and utilized them to detect real-world events. Their extended method could track the trajectories of moving events such as typhoons by *particle filters*. They also developed a system that sends alerts of detected earthquakes to registered users, which could send alerts of large earthquakes before alerts by JMA (Japan Meteorological Agency). Robinson et al. [79] also devised a system to detect earthquakes in Australia and New Zealand, which is based on their prior ESA system [80]. Although the system showed a certain level of accuracy, it was reported that its sensitivity gave false positives.

## 2.2 User Location Inference in Social Media

Geographical analyses of web resources has been conducted for over ten years. In 1999, Buyukkokten et al. [81] published a paper to estimate geographical scopes of web documents, which is the first research of such kind of analysis of web resources to the best of our knowledge. A geographical scope of a web document means the location that the document mentions. For example, the geographical scope of a document is Tsukuba if its main topic is an Italian restaurant in Tsukuba, while the geographical scope of a document about

the history of the earth is the entire world. Geographical scopes have been considered to be important by many researchers to improve web search results using the geographical search intent, and hence there are a lot of researches dealing with the problem of estimating locations which web documents are written about [82, 83, 84, 85, 86].

Inspired by geographical analyses of web documents, there have been a large number of researches geographically analyzing various types of web resources. They include identifying locations which search queries intend [87], locations where search queries are issued from [88], locations of photos uploaded to Flickr [89, 90], locations which blog articles describe [91], locations of editors of Wikipedia [92], and so on. In particular, inferring home locations of social media users has been an active research field (e.g., [28, 29]).

This section discusses and surveys the problem of user home location inference in social media in detail. First, the problem is defined in Section 2.2.1. Then existing researches dealing with this problem are surveyed being categorized into three approaches below:

- Content-based approach (Section 2.2.2)
- Graph-based approach (Section 2.2.3)
- Hybrid approach (Section 2.2.4)

Finally, Section 2.2.5 discusses several points of interest about user location inference, including merits and demerits of location inference, and advantages and disadvantages of existing approaches including the time complexity.

### 2.2.1 Problem Definition

This section defines the terminology and the problem of user location inference.

**User accounts and home locations.** Each user account  $u \in U$  has its own home location  $l_u$ . Locations are defined as coordinates of latitude and longitude, which are denoted as  $l_u = (lat, lng)$ . The user set contains two types of components  $U = U^L \cup U^N$  where  $U^L$  is

a set of *labeled users* whose home locations are known, while  $U^N$  is a set of *unlabeled users* whose home locations are unavailable.

**Social graphs.** Social graph  $G = (U, E)$  is an undirected (e.g., Facebook) or a directed graph (e.g., Twitter), where the vertex set  $U$  is the user set which is defined above. In the undirected graph, each edge  $e = \{u_i, u_j\} \in E$  is undirected, while in the directed graph, each edge  $e = (u_i, u_j) \in E$  is directed. In the former, a connected pair of users are called *friends*. In the latter, similar to Twitter 's vocabulary, we adopt the terms of *follower* and *followee*. When a user *follows* another user, then the former user is called a follower of the latter user, while the latter user is called a followee of the former user.

**Posts.** Each *post*  $p \in P$  is defined via three elements; timestamp  $s$ , text  $t$ , and user  $u$ , and is denoted as  $p = (s, t, u)$ , which means that user  $u$  posts text  $t$  at timestamp  $s$ .

Using these notations, the problem of user location inference can be stated in two settings as follows:

**Problem 1 (Content-Based User Location Inference)** *Given a set of users  $U = U^L \cup U^N$ , and a set of posts  $P$ , infer the home location of each unlabeled user  $u \in U^N$  so that the inferred location  $\hat{\mathbf{l}}_u$  is close to the true location  $\mathbf{l}_u$ .*

**Problem 2 (Graph-Based User Location Inference)** *Given a set of users  $U = U^L \cup U^N$ , a social graph  $G = (U, E)$ , infer the home location of each unlabeled user  $u \in U^N$  so that the inferred location  $\hat{\mathbf{l}}_u$  is close to the true location  $\mathbf{l}_u$ .*

Note that hybrid approaches utilize both a set of posts and a social graph.

### 2.2.2 Content-Based Approach

Content-based approach leverages user generated contents. The idea of this approach is that users are likely to post contents about their residential areas more often than other

areas. For example, it is natural to believe that a user who lives in Tsukuba is more likely to post about Mt. Tsukuba, Tsukuba express, and JAXA.

The content-based approach has its roots in Cheng et al. [28], in 2010. Their method proposed a concept of *local words* and utilized them to infer home locations of Twitter users. Local words are ones that mainly appear in posts by users whose home locations are in some specific region or area. For example, it is reported that a word “*rockets*” is a local word because it is frequently posted by users living in Houston, USA [28]. With this local word, Cheng et al. inferred that users who tend to use it in their posts live in Houston.

Around the same time as Cheng et al., Eisenstein et al. [93] proposed a topic model to explain the generative process of contents and the geographical features of tweets. Their model can also infer home locations of Twitter users. It was reported through their experiments that although the inference accuracy of home locations by their method was not so high as Cheng et al.’s method, their topic model could describe the process of generating tweets with regard to locations and contents. Hong et al. [94] also devised a topic model to describe the process that users post tweets with respect to locations and contents. Their experiments showed that Hong et al.’s model outperformed Eisenstein et al.’s model in the problem of user location inference. They said that Hong et al.’s model could suppress overfitting by introducing the general topic independent of locations and users.

Moreover, inference methods extending Cheng et al.’s proposal have been proposed [95, 96, 97]. Kinsella et al. [95] developed a language model utilizing tweets with GPS tags and proposed an inference method based on this model. It was reported by Kinsella et al. that leveraging geotagged tweets leads to a more robust model against user movements. Chandra et al. [96] focused on conversations of Twitter users, and proposed a language model assuming that tweets in the same conversation share the same topic. Their results demonstrated that the inference method taking advantage of user conversations outperformed an existing method which does not utilize conversations. Chang et al. [97] modeled the geographical distribution of words in tweets by GMM (Gaussian Mixture Model) and inferred user home



locations based on that model. Their method does not require manually annotated data unlike Cheng et al.’s method. They experimentally showed that the proposed unsupervised method, which used a relatively small number of local words, achieved as high accuracy as that of Cheng et al.’s supervised method.

There have been a plenty of other methods tackling with this problem. Hecht et al. [98] surveyed the nature of Twitter users’ location profiles<sup>4</sup> in detail. They manually investigated location profiles and found that about 34% of Twitter users did not enter their home locations in location profiles. They also proposed a method for user location inference by the Naive Bayes Classifier [99], which used only local words as classifier features leading to higher accuracy than using all the words.

Ikawa et al. [100] dealt with a slightly different problem of inferring locations that tweets are posted from. Their proposed method tracked individual Twitter users’ movements using Foursquare, which made it possible to infer spatial behavioral patterns of individuals. For example, suppose that a user regularly checks-in a coffee shop and tends to post a word “coffee” there. Their method can infer that the user is at the coffee shop when the user posts about coffee.

Schulz et al. [101] took advantage of various indicators such as words in tweets, URLs in tweets, home pages of users, users’ time zones, and so on. Their method can infer both user home locations and locations that tweets are posted from. It was reported that the accuracy of inferring tweet locations was approximately 92%, and the median error distance was approximately 30km.

### 2.2.3 Graph-Based Approach

Graph-based approaches analyze social graphs to infer home locations assuming that home locations of connected users in social graph, namely friends, followers, or followees, are likely to be close to each other. For example, with the knowledge that most of user  $u$ ’s friends

---

<sup>4</sup>Twitter users can describe their home locations as free-form texts.

live in Tsukuba, graph-based approaches can infer that  $u$  also lives in Tsukuba.

The graph-based approach has its roots in Backstrom et al. [29], in 2010. Their graph-based method, which aims to infer residential locations of Facebook users, maximizes the likelihood of generating edges in social graph assuming that short-distance edges have a higher chance than long-distance ones. The accuracy of the proposed method was shown to be higher than that of baseline approach which is based on IP addresses.

Abrol et al. [102] proposed a graph-based method called TweetHood. TweetHood infers home locations of Twitter users utilizing a subset of followees of the target user. The subset is composed of the target user’s top  $k$  followees that have large numbers of common neighbors with the target user. During the inference process of the target user  $u$ , if the home location of followee  $v$  is unknown, TweetHood recursively infers that location until the predefined depth and utilize it for the inference of  $u$ ’s home location. It was shown that with deeper recursive inference, the accuracy becomes higher but the computational cost becomes larger.

Clodoveu et al. [103] also proposed an inference method for Twitter users, which simply takes the majority vote of followees’ home locations. Their proposed method is simple and they discussed how many followees are required to effectively infer home locations. They concluded that too few followees give less clues for home location inference, while too many followees imply that the target user is likely to be a celebrity or a bot, and inferring the target user’s home location does not make sense because celebrities and bots usually do not have their home locations.

Jurgens [104] also developed a recursive inference method in an analogous fashion to Abrol et al. [102]. Jurgens’s method, which is based on the label propagation method [105], calculates the medoid point<sup>5</sup> of home locations of adjacency users to propagate home locations of labeled users. In their experiments, although the proposed method showed the high accuracy, it was not compared with existing methods.

While all of above mentioned methods treat all locations as the same, a method pro-

---

<sup>5</sup>The closest data point to the centroid of the dataset.

posed by Rout et al. [106] considers the populations of locations. Specifically, their method regards friends that live in a “small location” (i.e., a location with a small number of residents) as the strong clue because it is more likely that users in a small city are real friends than users in a large city. Moreover, they also take into account “tie strength” [107] by considering *triangles*<sup>6</sup> of users in a social graph because users forming a triangle are likely to have strong friendship which leads to the strong clue for location inference. Taking various characteristics including above as the feature set, they inferred home locations SVM classifiers.

McGee et al. [108] proposed a method which also considers the tie strength based on their previous study [30]. They classified neighboring users of the target user in terms of their effectiveness for inferring the home location of the target user. Specifically, they trained a decision tree using several features including the frequency of messages. After the classification, their method infers the home location of the target user using classified users based on the method of Backstrom et al. [29]. Experimental results showed that McGee et al.’s method outperformed Backstrom et al.’s method.

Sadilek et al. [109] addressed the integrated problem of inferring Twitter users’ trajectories and predicting links in the Twitter social graph. Their idea is that friends are more likely to move together than others. Their result cautioned about the risk of disclosure of one’s behavioral records.

Existing graph-based approaches are all based on the idea that home locations of connected users in a social graph tend to be close to each other, which is called *closeness assumption*. In contrast to these existing approaches, this thesis introduces a novel assumption called *concentration assumption* in Chapter 3. The concentration assumption states that there are users such that most of their neighbors are in a small region, which are called *graph landmarks*. Graph landmarks can be regarded as strong clues for location inference.

---

<sup>6</sup>Three vertices that are mutually connected in the graph.

### 2.2.4 Hybrid Approach

Hybrid approaches, which utilize both user generated contents and social graphs, are proposed by Li et al. [110, 111]. They first proposed the state-of-the-art *unified discriminative influence model* (*UDI*) to infer users’ home locations [110]. UDI models user-relationships and venue names extracted from user-generated contents as a heterogeneous graph assuming that each node (i.e., user or venue) has its own influence scope. Nodes with a large influence scope (e.g., Lady Gaga) do not provide good clues for locations inference because numerous users in diverse locations follow them.

Li et al. proposed another model, *multiple location profiling model* (*MLP*) [111], which assigns different locations for a single user (e.g., home, work, and former home locations). Li et al.’s two methods use a gazetteer to extract venue names from user-generated contents.

### 2.2.5 Discussion

This section discusses four points of user location inference below:

- Merits and demerits of user location inference.
- Classifying user home locations into administrative districts.
- Advantages and disadvantages of content-based and graph-based approaches.
- Time complexities of inference methods.

**Merits and demerits of user location inference.** User location inference is considered to have various kinds of merits. We discuss those merits from the standpoint of recipients: governments, companies, and individuals.

Governments can get the voice of the people by monitoring the information transmitted in social media. With home location information, it is possible to identify the dissatisfaction and demands from certain areas. Besides, as mentioned above, situations affected by disasters can be traced, which makes it possible to capture demands from disaster victims.

In the case of companies, they can obtain reviews of their products from social media users in certain areas. With such location-aware reviews, they can track where their products are received well or unpopular, which enables them to recommend products for users in a certain location. In addition to that, companies that have distribution networks can control them knowing where and what is in demand.

Individuals using social media benefit by above services from governments and companies. These services not only enhance living standard but also protect them from danger.

As discussed above, location information can be applied to several purposes. City level location inference will be enough for these purposes with respect to a degree of accuracy required by these purposes. Besides, in terms of feasibility, it is difficult to infer home locations at finer-grained levels than the city level (e.g., ZIP code level) using the current social media contents and graphs.

Although these applications are indeed useful, one may think that it is enough to directly ask users for their location information rather than location inference. There are actually *reporting services*, for example, where registered users report weather with their location information, and where individual users make disaster maps in their home town. However, compared to these services, home location inference has several advantages below:

- It is able to associate location information to all the contents not matter what they are intended for.
- In the reporting services, users actively register, and transmitted information is strongly associated with particular intention. Unlike such intended contents, social media contents with inferred location information do not have biases, which can be leveraged for various purposes.
- It is possible to associate inferred location information to the contents posted in the past.

In spite of the above discussed merits of location inference, there are demerits, which include the privacy issues. For the users who want to hide their residences, location inference may lead to the disclosure of privacy. However, we believe that our findings in this thesis can be exploited to comprehend how privacy is exposed in social media. Moreover, the proposed techniques can be utilized to send alerts to users at risk of privacy exposure.

**Classifying user home locations into administrative districts.** It is important to classify social media users into administrative districts such as states, prefectures, and cities, which makes it possible for governments to get the voice of people per districts. For this reason, many researches on identifying web document locations assign administrative districts to the documents by classification techniques [81, 82, 83, 84, 85, 86].

On the other hand, this research and major existing researches on inferring social media users' home locations do not explicitly classify users into districts. They rather aim at reducing error distances between true locations and inferred locations. However, even if we do not explicitly consider administrative districts, we can discuss whether our inference method can classify users into states, prefectures, or cities based on results of error distances. For example, if the average error distance of an inference method is 50km or 10km, the method is considered to be able to classify users at the prefecture level or city level, respectively. Most of existing inference methods aim to accomplish the city level inference because of the above mentioned reasons. Similarly, this thesis also sets the goal of inferring home locations at the city level.

**Advantages and disadvantages of content-based and graph-based approaches.** Most of content-based approaches are designed and targeted for some specific types of social media because they depend on the form of contents (e.g., photos, movies, and texts). This means that, for example, a content-based approach designed for Flickr cannot directly be applied for Twitter. Even in the case of texts, there are varieties of languages, and the target language depends on the target area, which makes content analyses more complicated.

On the other hand, graph-based approaches do not depend on the form of contents and languages, and therefore graph-based approaches can be applied to many types of social media if they have a social graph.

Social graphs are relatively static (i.e., friendships do not frequently change), indicating that the inference result by graph-based approaches also do not change drastically. In contrast, contents are continuously generated in social media because of its real-time feature. This temporal feature can be utilized by content-based approaches, which can be regarded as an advantage of content-based approaches. However, all existing methods do not consider the temporal feature of social media contents. In Chapter 4, we propose a content-based approach that takes advantage of the temporal feature focusing on the real-world events.

In addition, content-based approaches can continuously obtain new clues for location inference because social media contents are posted in real time. Although content-based approaches can exploit these newly obtained contents to update the inference result continuously, there has been no method proposed with such functionality. In Chapter 5, we develop an online inference method exploiting streams of social media contents, or social streams.

**Time complexity.** Most of existing studies did not discuss the time complexity of the inference method. However, it is a crucially important point. Table 2.1 describes time complexities of major existing methods and the proposed methods in this thesis. Note that *LMM*, *ELIM*, and *OLIM* are the methods proposed in Chapters 3, 4, and 5, respectively. Notations used in Table 2.1 are described in Table 2.2.

Regarding computational costs, we discuss several points in this thesis. In Chapter 3, we analyze the trade-off between costs and accuracy, and that between costs and coverage. In Chapter 5, computational costs of online inference and batch inference are experimentally compared.

Table 2.1: Time complexities of location inference methods.

	Preprocess	Inference process
Cheng [28]	Classification of local words by CART [112]	$O( U^N  \cdot ( W_u  +  W_u^L  \cdot  \Lambda ))$
Kinsella [95]	$O( U  \cdot  S_u  \cdot  W_p  \cdot  \Lambda )$	$O( U^N  \cdot  S_u  \cdot  W_p  \cdot  \Lambda )$
Hecht [98]	$O( U  \cdot  W_u  +  \Lambda  \cdot  W  +  U  \cdot  W_u^L )$	$O( U^N  \cdot  W_u^L  \cdot  \Lambda )$
Backstrom [29]	$O( U  \cdot  \Lambda )$	$O( U^N  \cdot k^2)$
Jurgens [104]	No preprocess	$O(r \cdot  U^N  \cdot (k +  M ^2))$
UDI [110]	$O( U  \cdot  W_u  \cdot  \Sigma^L )$	$O(r_{out} \cdot ( U  \cdot k^{in} +  \Sigma^L  \cdot l^{in} + r_{in} \cdot ( U^N  \cdot (k^{in} + k^{out} + l^{out}))))$
LMM (Chapter 3)	$O( U  \cdot ((k^{in})^2 + (k^{out})^2))$	$O( U^N  \cdot (k^{in} + k^{out})^2)$
ELIM (Chapter 4)	$O(n_t \cdot ( U_t  \log  U_t  +  C  \cdot  U_c ^2))$	$O( U^N  \cdot  E_u  \cdot  S_e )$
OLIM (Chapter 5)	$O( S  \cdot  W_p  + K \cdot  \Sigma  \cdot  U_w )$	$O(K \cdot ( S_t  \cdot  W_p  +  \Sigma  \cdot  U_w ))$



Table 2.2: Notations.

$U$	Set of all users in a dataset.
$U^N$	Set of unlabeled users in a dataset.
$U_t$	Set of users that write a post in a time window in ELIM (Chapter 4).
$U_c$	Set of users that write a post belonging to a cluster in ELIM (Chapter 4).
$U_w$	Set of users that write a post containing word $w$ .
$S_u$	Set of posts written by a user.
$\Lambda$	Set of locations.
$k$	Degree in the undirected social graph.
$k^{in}, k^{out}$	In- or out-degree in the directed social graph.
$l^{in}, l^{out}$	The number of posts that mention a location, or the number of posts that mention locations by a user.
$W$	Set of all words in a dataset.
$W_p$	Set of words contained in a post.
$W_u$	Set of words contained in posts written by user $u$ .
$W_u^L$	Set of local words contained in posts written by user $u$ .
$S$	Set of all posts in a dataset.
$S_t$	Set of posts in a time window.
$S_u$	Set of posts written by user $u$ .
$S_e$	Set of posts belonging to event $e$ in ELIM (Chapter 4).
$\Sigma^L$	Set of toponym vocabulary.
$\Sigma$	Set of vocabulary.
$r$	The number of loops in Jurgens.
$r_{out}$	The number of outer loops in UDI.
$r_{in}$	The number of inner loops in UDI.
$C$	Set of clusters in ELIM (Chapter 4).
$E_u$	Set of events mentioned by user $u$ in ELIM (Chapter 4).
$n_t$	The number of time windows in ELIM (Chapter 4).
$K$	The number of divisions (regions) in OLIM (Chapter 5).

## Chapter 3

# User Location Inference based on Graph Landmarks

A large portion of existing studies assume that connected users (i.e., friends) in social graphs are located in close proximity. Although this assumption holds for some fraction of connected pairs, sometimes connected pairs live far from each other. To address this issue, we introduce a novel concept of *graph landmarks*, which are defined as users with a lot of friends who live in a small region. Graph landmarks have desirable features to infer users' home locations such as providing strong clues and allowing the locations of numerous users to be inferred using a small number of graph landmarks. Based on this concept, we propose *Landmark Mixture Model (LMM)* for the problem of user location inference in the graph-based setting. The experimental results using two Twitter datasets in different regions show that our method improves the accuracy of the state-of-the-art method by about 27%.

### 3.1 Introduction

Major graph-based inference approaches, such as [29] [111] [110] assume that connected users on social graphs are located close to each other, which is called *closeness assumption*. Connected users in Facebook are friends where every edge is mutual, or in Twitter are either

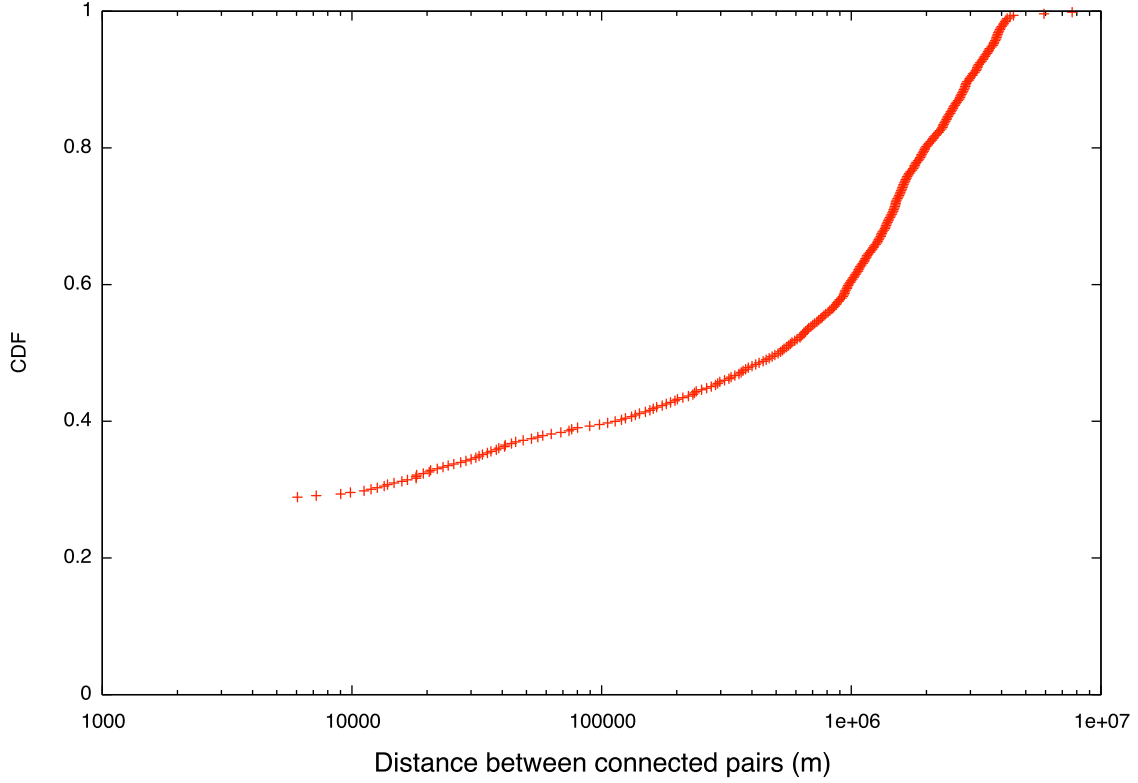


Figure 3.1: CDF plot of the distance distribution between home locations of mutually connected users in Twitter. 60% or more of connected users are at least 100km apart.

friends or followers where each edge has a direction. Although the closeness assumption holds for some fraction of connected pairs, a significant percentage are geographically distant from each other.

Figure 3.1 shows the CDF (Cumulative Distribution Function) plot of the distribution of distances between home locations of mutually connected users in Twitter. This figure uses the US dataset described in Section 3.5.1. 60% or more of connected pairs are at least 100km apart from each other. Hence, this closeness assumption between connected users may not provide us with strong clues for location inference.

Herein to improve the accuracy of the graph-based inference method, we introduce a novel concept of *graph landmarks*. Graph landmarks are users who have the following two characteristics: 1) having a lot of friends and 2) the home locations of friends are close to

each other. For example, if user  $u$  has a lot of friends whose locations are mostly in Boston, we regard user  $u$  as a graph landmark in Boston. After identifying this graph landmark, if another user  $v$  whose home location is unavailable follows this graph landmark  $u$ , we infer that  $v$  also lives in Boston. In this case, the city of Boston is this graph landmark's *dominance location*.

Graph landmarks have two desired features for location inference:

1. **Strong clues:** Due to the friends' geographical proximity, solid inferences can be made utilizing graph landmarks.
2. **Wide coverage:** Due to the large *degree* of graph landmarks in a social graph, only a few graph landmarks are necessary to infer the most part of users in the graph.

Suppose that 80% of a graph landmark's friends live in Boston, and then this graph landmark provides a clue with an 80% confidence level that a new friend also lives in Boston even if the new friend's location is unavailable. Compared to the traditional closeness assumption, we call this novel assumption *concentration assumption*. Based on this concept, we propose *Landmark Mixture Model (LMM)* to infer users' home locations. LMM models both dominance locations of graph landmarks and the home locations of users as continuous probability distributions over a geographical space. Specifically, the distributions of home locations are modeled as mixtures of the distributions of dominance locations of graph landmarks.

In addition to above basic idea, LMM can adjust the two trade-offs as follows. First, LMM allows the trade-off between precision and coverage to be exploited. In this context, precision refers to the ratio of correctly inferred users, while the coverage is the ratio of inferred users versus all users. Because home locations are modeled as probability distributions, decisions can be made based on the distribution shape. If the home location distribution of a user has a clear peak at a certain location, the user's location can be confidently determined at the location. On the other hand, if the distribution lacks a clear peak, we cannot confidently infer the home location of the corresponding user because we do not have enough clues for

the user, which can be avoided by imposing the *confidence threshold* (Section 3.4.3).

Second, LMM can adjust the trade-off between computational cost and coverage. Finding the mode point in the mixture model is inherently costly, and hence the location inference process based on LMM may also be costly because it also needs to find the mode point from the mixture. However, LMM can reduce the cost based on the observation that users' home locations can be inferred by using only a small number of graph landmarks with large degrees, leading to lower computational cost. This can be achieved by imposing the *degree threshold* (Section 3.4.3).

The contributions of this paper can be summarized as follows:

- We introduce a novel assumption called *concentration assumption* that states that there are users called *graph landmarks* with a lot of friends in a small region, which can be utilized for the problem of user location inference.
- We propose *Landmark Mixture Model (LMM)* to infer users' home locations. This model is based on the concentration assumption other than the traditional closeness assumption, and can adjust the trade-offs between precision and coverage, and between computational cost and coverage.
- We experimentally evaluate the performance of the proposed method by comparing the existing methods including the state-of-the-art method, which is based on two Twitter datasets in different regions: the United States and Japan.

Our results show that LMM significantly improves the precision of existing methods including the state-of-the-art, and preserves high coverage. The results also demonstrate that LMM flexibly adjusts the abovementioned trade-offs by imposing two thresholds; raising the precision to about 90% while preserving 60% of the coverage; and the cost is reduced to 10% while preserving 85% of coverage.

The rest of this chapter is organized as follows. Section 3.2 states the problem addressed in this chapter and defines the terminology. The concept of graph landmarks is introduced

in Section 3.3, and then our LMM is proposed in Section 3.4. Section 3.5 describes the experiments conducted to verify the effectiveness of our method compared to the other existing methods including the state-of-the-art one. Finally, Section 3.6 concludes the chapter.

## 3.2 Problem Statement

The problem addressed in this chapter is in the graph-based setting, and undirected social graph is provided, which is stated as:

**Problem 2 (Graph-Based User Location Inference)** *Given a set of users  $U = U^L \cup U^N$ , a social graph  $G = (U, E)$ , infer the home location of each unlabeled user  $u \in U^N$  so that the inferred location  $\hat{\mathbf{l}}_u$  is close to the true location  $\mathbf{l}_u$ .*

Notations used in this chapter are the same as those in Section 2.2.1.

Instead of the widely adopted closeness assumption, we employ the concentration assumption to tackle this problem in this chapter. Section 3.3 introduces the concept of graph landmarks, while Section 3.4 describes our landmark mixture model (LMM) to solve the user location inference problem.

## 3.3 Graph Landmarks

### 3.3.1 Definition

To introduce graph landmarks, two measurements of graph landmarks, *degree* and *dispersion*, need to be defined. The degree is the same concept as the term in the graph theory. The dispersion of user  $u$  means how far  $u$ 's neighbors (i.e., friends or followers) are located from each other. Note that dispersion does not depend on user  $u$ 's own home location. Based on degree and dispersion, we define graph landmarks below.

**Definition 1 (Graph landmarks)** *A graph landmark is a user account  $u$  with a large*

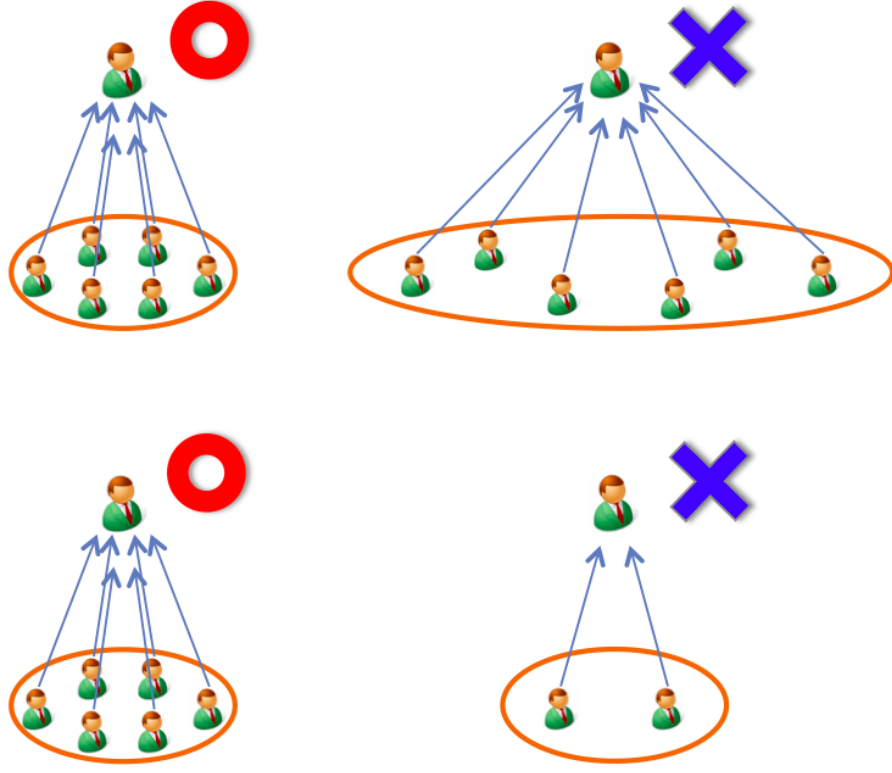


Figure 3.2: Upper right user is regarded as a graph landmark because this user has followers in a small region, while upper left user is not regarded as a graph landmark because followers of this user are dispersed. Lower right user is regarded as a graph landmark because this user has a lot of followers, while lower left user is not regarded as a graph landmark because this user does not have a lot of followers.

*degree  $c_u$  and a small dispersion  $d_u$ .*

In this research, 2-dimensional spatial variance over the geographical space is used as dispersion value  $d_u$ . Figure 3.2 illustrate the graph landmark and the non-graph landmark. In Figure 3.2, the upper right user is regarded as a graph landmark because this user has followers in a small region, while the upper left user is not regarded as a graph landmark because followers of this user are dispersed. In addition, the lower right user is regarded as a graph landmark because this user has a lot of followers, while the lower left user is not regarded as a graph landmark because this user does not have a lot of followers. Detailed definition is described in Section 3.4.2.

In the context of the social media where a social graph is directed like Twitter, there are two types of graph landmarks, *in-landmarks* and *out-landmarks*. In-landmarks and out-landmarks are users that satisfy the definition of graph landmarks with regard to their followers and followees, respectively. To deal with these two types of graph landmarks, we also introduce the terms of *in-degree*  $c_u^{in}$ , *out-degree*  $c_u^{out}$ , *in-dispersion*  $d_u^{in}$ , and *out-dispersion*  $d_u^{out}$ . In-degree and in-dispersion are measured using a user’s followers (i.e., vertices with edge directed toward the user). Inversely, out-degree and out-dispersion are measured using a user’s followees. Dominance locations of in-landmarks and out-landmarks are the center points of home locations of their followers and followees, respectively.

### 3.3.2 Preliminary Experiment

To demonstrate the presence of graph landmarks, we investigate the degree and the dispersion of users in our Twitter dataset that is described in Section 3.5.1.

If there exist graph landmarks in the dataset, we can say that the concentration assumption holds. In other words, there are some user groups whose locations are near each other and these users follow the same user (i.e., graph landmark). In this case, we can infer home locations of these users if their home locations are unknown, by propagate home locations of location-known users in the same group.

Figure 3.3 shows the CDF plot of dispersion values of users in the dataset. The dispersion value of user  $u$  is calculated by using  $u$ ’s labeled followers or followees. This figure shows that there are  $y\%$  of users that have dispersions lower than  $x$ . Plotted users in this figure are limited to top 5% of users whose degrees are high in the dataset.

If we define users with dispersions less than 10 as graph landmarks, then 14% of plotted users in Figure 3.3 can be regarded as in-landmarks and 17% as out-landmarks. Although the number of graph landmarks is rather small, graph landmarks do exist.

Figure 3.4 maps the above graph landmarks (i.e., users of the top 5% degree, and less than 10 dispersion value). Red dots represent the home locations of all users in our dataset,



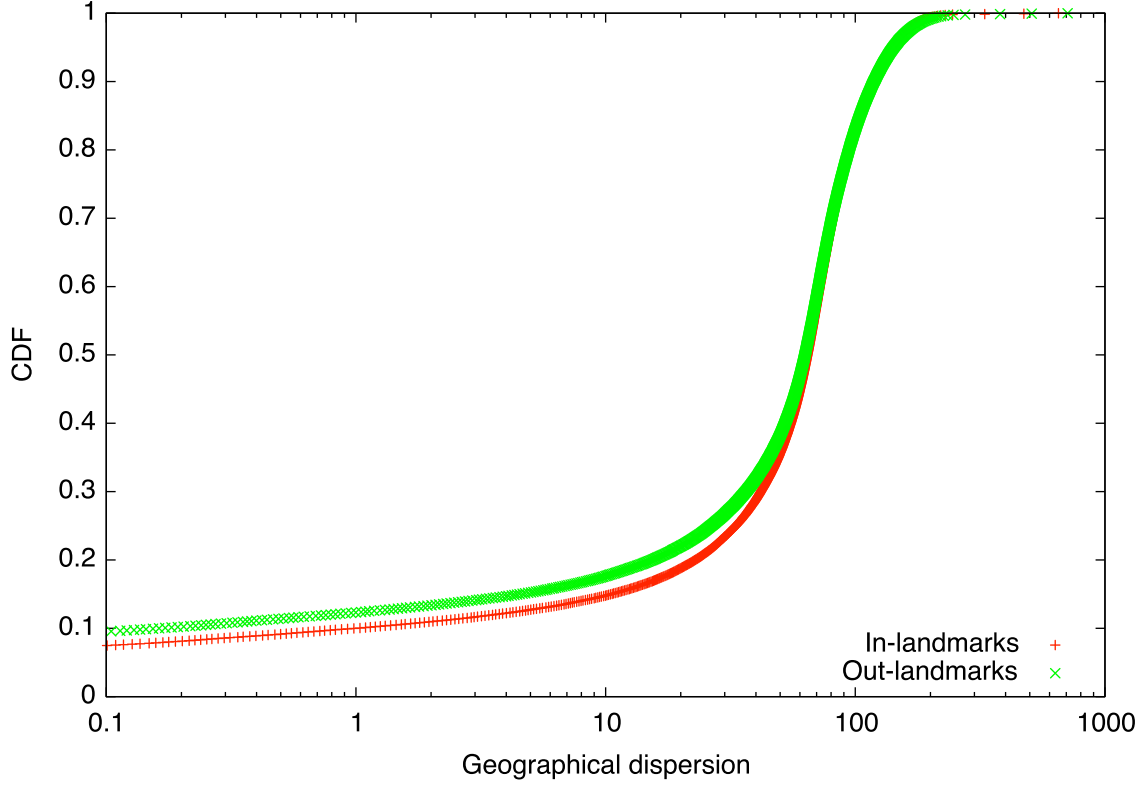


Figure 3.3: CDF of users over their dispersions for the top 5% users with high degrees. This figure indicates that there are  $y\%$  of users that have dispersions lower than  $x$ . Although most users have a large dispersion, there exist users with relatively small dispersion, which are regarded as graph landmarks.

while blue dots represent dominance locations of graph landmarks. First of all, most users, including graph landmarks, are located in the eastern part (e.g., east of the Mississippi River) of the United States, which is consistent with most of other works. Second, the distributions of red and blue dots are similar. Metropolises with a lot of users also have a lot of graph landmarks. Although most graph landmarks lie in the east part, some cities in the west part (e.g., Denver, Phoenix, and Salt Lake City) have a relatively large number of graph landmarks. Thus, graph landmarks can cover large segments of the user population; that is, most users can find graph landmarks near their home locations.

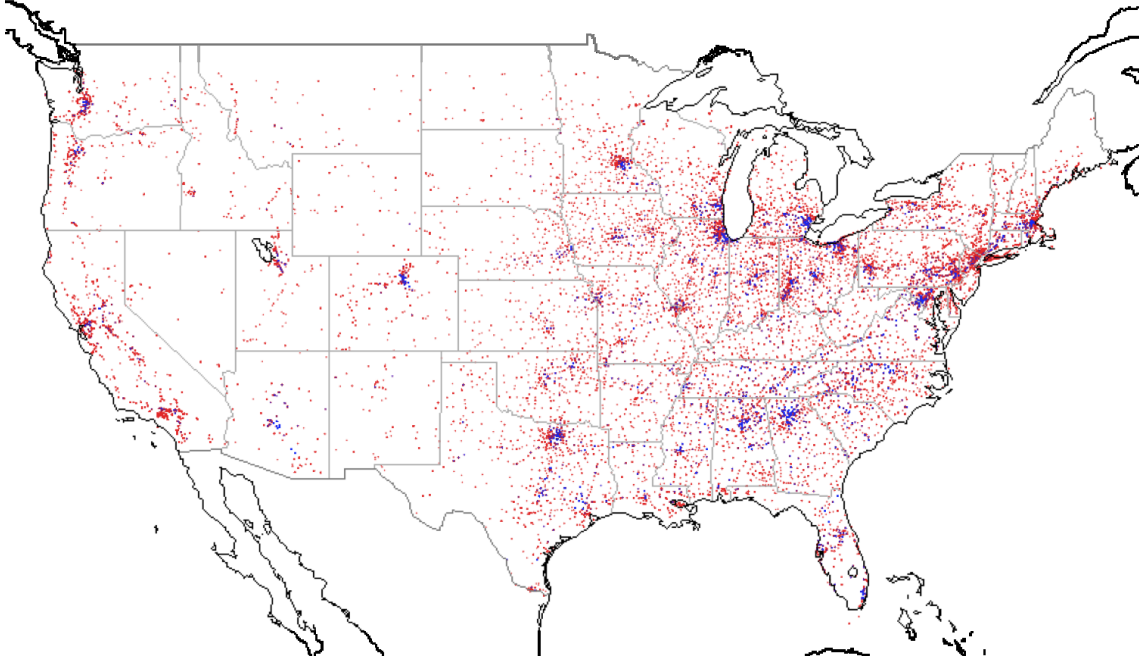


Figure 3.4: Distributions of both all users and graph landmarks. Red and blue dots represent home locations of all users in our dataset and dominance locations of graph landmarks, respectively. Both populations have similar distributions. Graph landmarks can cover a large segment of the user population; that is, most users can find a graph landmark near their home location.

### 3.3.3 Examples

Table 3.1 shows some examples of graph landmarks in the dataset. The top two user accounts are regarded as in-landmarks, the middle two user accounts are out-landmarks, and the bottom two user accounts are both in-landmarks and out-landmarks. These in- and/or out-landmarks tend to be local news accounts or commercial accounts, which are *bots* rather than *human accounts*. Specifically, most in-landmarks are local news accounts, which post about their local area. Although there are fewer out-landmarks, they tend to be commercial accounts. However, it should be noted that some user accounts can be regarded as both in-landmarks and out-landmarks. This type of graph landmark has a lot of followers and friends in a small region, indicating that it follows its followers back.

Our observations suggest that in-landmarks are authoritative user accounts that post

Table 3.1: Examples of graph landmarks

<i>User Name</i>	<i>Profile</i>	<i>Dominance Location</i>	<i>Degree (in : out)</i>	<i>Dispersion (in : out)</i>
denvernews	Denver-specific news from The Denver Post. ...	39.73, -105.0 (Denver,CO)	20,526 : 7,418	0.0013 : 20.36
BostonFire	Official Twitter Boston Fire. Spring starts outdoor grilling. ...	42.32, -71.09 (Boston,MA)	34,111 : 49	0.6523 : 56.60
HomeTheaterMI	Genesis Electronics is a family-owned business ...	42.39, -83.13 (Detroit,MI)	2,331 : 1,841	16.66 : 0.0067
alabamanews1	All Alabama News!	33.52, -86.81 (Birmingham,AL)	836 : 1,523	10.51 : 0.3815
komonews	The latest breaking news, traffic, and weather from Seattle ...	47.63, -122.3 (Seattle,WA)	29,989 : 2,102	0.0107 : 0.0029
OWHnews	Updates from Omaha.com and the Omaha World-Herald ...	41.26, -96.01 (Omaha, NE)	17,400 : 9,662	0.0306 : 0.0306

useful tweets about their dominance location. Thus, in-landmarks can provide useful information about local locations. This observation provides another motivating factor to utilize in-landmarks to extract useful local information, but it is beyond the scope of this thesis and will be examined in the future.

On the other hand, out-landmarks tend to be commercial accounts, including spammers, who want more followers in a small region. Although these graph landmarks do not post useful tweets, we can utilize them to infer home locations.

### 3.4 Proposed Method

This section proposes Landmark Mixture Model (LMM) to address the user location inference problem. LMM models both the dominance locations of graph landmarks and the home locations of users as continuous probability distributions. Section 3.4.1 formulates the model, and then Section 3.4.2 proposes a location inference method based on this model. Finally, Section 3.4.3 introduces the thresholds to adjust the trade-offs.

### 3.4.1 Model Formulation

According to the definition, graph landmarks have small dispersions, leading to clear dominance locations. Hence, LMM estimates all users' dispersions and dominance locations, and then regards users with small dispersions as graph landmarks. Note that dispersions and dominance locations can be calculated even for unlabeled users because only home locations of followers or followees are used to calculate those values.

**Dominance distribution.** Similar to several other studies that model the probability distribution over a geographical space [113] [114] [110], we model the dominance location as a Gaussian distribution. We call this distribution the *dominance distribution*. The underlying idea is that the Gaussian distribution has two parameters, mean and variance, which represent the dominance location and the dispersion, respectively. The value of the probability density at each location point indicates what fraction of followers (or followees) the corresponding user has at the point. A dominance distribution with a large variance (i.e., large dispersion) does not have a clear peak, indicating that the user has followers (or followees) in various locations. Consequently, the user cannot be regarded as a graph landmark.

Based on the above idea, we assign a Gaussian distribution  $N(\boldsymbol{\mu}_u, \boldsymbol{\Sigma}_u)$  for each user  $u$ . The mean parameter  $\boldsymbol{\mu}_u$  of 2-dimensional vector denotes the dominance location of user  $u$ , while the covariance parameter  $\boldsymbol{\Sigma}_u$  of  $2 \times 2$  matrix denotes the dispersion of  $u$ . Herein we assume that the shape of the dominance distribution is symmetric, that is,

$$\boldsymbol{\Sigma}_u = \begin{pmatrix} d_u & 0 \\ 0 & d_u \end{pmatrix}, \quad (3.1)$$

where the diagonal components are the dispersions. It should be noted that users have two types of dispersions. We assign two Gaussian distributions for each user: in-dominance distribution  $N(\boldsymbol{\mu}_u^{in}, \boldsymbol{\Sigma}_u^{in})$  and out-dominance distribution  $N(\boldsymbol{\mu}_u^{out}, \boldsymbol{\Sigma}_u^{out})$ .

**LMM.** Using the dominance distributions, LMM models the home locations of users as continuous probability distributions. Following a graph landmark provides a strong clue for inferring a user 's location because most of the graph landmark 's neighbors are in close proximity. On the other hand, following an ordinal user (i.e., a non-graph landmark) does not provide a good clue because the locations of neighbors of an ordinal user are geographically dispersed.

Based on this idea, LMM employs *Gaussian Mixture Model (GMM)* where the dominance distributions are mixed. Specifically, user  $u$ 's home location is modeled as a GMM where each Gaussian component is the dominance distribution of  $u$ 's neighbors.

The probability density at location  $l$  represents how likely user  $u$  lives in  $l$ . Hence, if a user 's location distribution has a clear peak at a specific locale, we can confidently state that identify the user 's home location.

We call this the *location distribution*. The location distribution is denoted as:

$$P_u(\mathbf{x}) = \sum_{v \in N_u^{out}} \pi_v^{in} N(\mathbf{x} | \boldsymbol{\mu}_v^{in}, \boldsymbol{\Sigma}_v^{in}) + \sum_{w \in N_u^{in}} \pi_w^{out} N(\mathbf{x} | \boldsymbol{\mu}_w^{out}, \boldsymbol{\Sigma}_w^{out}), \quad (3.2)$$

where  $N_u^{in}$  is the set of followers of  $u$ ,  $N_u^{out}$  is the set of followees of  $u$ , and  $\pi_v^{in}$  and  $\pi_v^{out}$  are the mixture weights. Mixture weights are defined as:

$$\pi_v^{in} \propto \log c_v^{in}, \quad (3.3)$$

$$\pi_v^{out} \propto \log c_v^{out}, \quad (3.4)$$

$$\sum_{v \in N_u^{out}} \pi_v^{in} + \sum_{w \in N_u^{in}} \pi_w^{out} = 1, \quad (3.5)$$

where  $c_v^{in}$  and  $c_v^{out}$  are in- and out-degree of user  $v$ , respectively.

The reason that we employ the logarithm of degree is that degrees of users in social graphs

follows the *power law*. In social graphs, some users have huge degree values, which requires us to moderate these values.

LMM does not explicitly differentiate graph landmarks from ordinal users in its location inference process. Instead, it imposes weights (i.e., mixture weights and variances) on all users to implicitly differentiate them. A Gaussian component with a small variance and large mixture weight, which corresponds to a dominance distribution of a graph landmark, strongly affects the shape of the overall location distribution. Consequently, our model mostly uses graph landmarks to determine a user's home location.

### 3.4.2 Parameter Estimation

Given a social graph  $G$ , we initially estimate the parameters of the dominance distributions for all users. Based on the maximum likelihood criteria, the parameters are estimated using the location points of users' neighbors as

$$\boldsymbol{\mu}_u^{in} = \frac{1}{|N_u^{in}|} \sum_{v \in N_u^{in}} \boldsymbol{l}_v, \quad (3.6)$$

$$\boldsymbol{d}_u^{in} = \frac{1}{2|N_u^{in}|} \sum_{v \in N_u^{in}} (\boldsymbol{l}_v - \boldsymbol{\mu}_u^{in})^2. \quad (3.7)$$

The parameters for the out-dominance distributions are estimated in the same manner.

However, we found that some of the neighbors are located far from the other neighbors, leading to noises that degrade the inference accuracy. To suppress the noise effect for Eqn. 3.6, we employ the median because the median is more robust against noises than the mean.

After the parameters of the dominance distributions are set, we can construct the users' location distributions. The estimated location distributions can be simply written as

$$\begin{aligned} P_u(\boldsymbol{x}) &= \sum_{v \in N_u^{out}} \pi_v^{in} N(\boldsymbol{x} | \boldsymbol{\mu}_v^{in}, \boldsymbol{\Sigma}_v^{in}) \\ &+ \sum_{w \in N_u^{in}} \pi_w^{out} N(\boldsymbol{x} | \boldsymbol{\mu}_w^{out}, \boldsymbol{\Sigma}_w^{out}). \end{aligned} \quad (3.8)$$

It should be noted that statistical inference methods (e.g., the EM algorithm [115]) are unnecessary because LMM simply mixes the dominance distributions. This substantially reduces the parameter estimation cost.

Based on this model, the user home location is inferred as the location with the largest probability density (i.e., mode point) as follows:

$$\hat{\mathbf{l}}_u = \arg \max_{\mathbf{x}} P_u(\mathbf{x}). \quad (3.9)$$

Finding the mode point of the mixture of Gaussian requires some kind of search algorithm such as grid search. However, finding the exact mode point is extremely costly. Hence, we adopt an approximate algorithm as follows. The algorithm firstly narrows down the candidate solutions to center points of Gaussian components. Then it sums up the probability density of all components at each candidate point, and selects the point with the largest sum of probability density as the mode point. The accuracy of approximation of this algorithm is adequate enough because 1) if each component is close to each other, then each candidate point is close to the exact solution, or 2) if each component is distant to each other, then there is little overlap, which means that the exact solution is nearly identical to one of the candidate points. The time complexity of this algorithm is  $O(k^2)$ , where  $k$  is the number of Gaussian components of GMM. Although this process still has a relatively high computational cost, we can reduce this by imposing degree threshold described in the next section.

### 3.4.3 Thresholds to Adjust the Trade-offs

LMM can adjust the trade-offs between precision and coverage, and between computational cost and coverage by imposing the *confidence threshold* and *degree threshold*, respectively.

**Confidence threshold.** The process of finding the mode point also gives the probability density  $p$  at that point, which indicates how likely the corresponding user 's home location

is at that point. If  $p$  is small, the confidence of this inference is low. To avoid making an unconfident inference, we impose the *confidence threshold* as follows:

**Definition 2 (Confidence threshold)** *If the probability density  $p_u$  of user  $u$ 's location distribution at the mode point is less than the predefined threshold  $p_0$ , the location of user  $u$  is not inferred.*

As the value of  $p_0$  increases, the precision increases, but the coverage decreases. On the other hand, as  $p_0$  decreases, the opposite is true. This trade-off is examined in Section 3.5.5.

**Degree threshold.** Because LMM does not explicitly discriminate between graph landmarks and ordinal users, it uses the dominance distributions of all users to infer the location. This causes a relatively expensive computational cost. Because graph landmarks provide the strong clues and have the wide coverage, we can infer the locations for a large segment of users using a small number of graph landmarks. To reduce the computational costs, we impose the *degree threshold*.

**Definition 3 (Degree threshold)** *If user  $u$ 's degree  $c_u$  is lower than the predefined threshold  $c_0$ , the dominance distribution of user  $u$  is not used for the inference.*

Because users with a low degree are not regarded as graph landmarks, they do not provide good clues for location inference. The degree threshold reduces the computational cost by eliminating the dominance distributions of these ordinal users in the location inference step.

Even if we exclude a significant fraction of users whose degrees are low, most users are connected to at least one graph landmark. This can be explained by the fact that the degree distribution of the Twitter social graph follows the *power law*<sup>1</sup> [18]. Based on *percolation theory*, in scale-free networks, the most vertices are connected even if a lot of vertices are removed as long as the degrees of removed vertices are small [116].

If the threshold  $c_0$  is large, then it is expected that both the computational cost and

---

<sup>1</sup>In fact, [18] reported that there are some Twitter users who have higher in-degrees than expected.



coverage decrease. On the other hand, if  $c_0$  is small, then a decrease in computational costs is small and the coverage remains high. This trade-off is verified in Section 3.5.6.

## 3.5 Experiments

This section describes the experiments to:

1. Compare the precision and the coverage of our proposed method to other existing methods.
2. Compare the results in different areas: the United States and Japan.
3. Compare the precision and the coverage of variations of our proposed method.
4. Examine the effects of two thresholds: the confidence threshold and the degree threshold.

Section 3.5.1 explains the experimental conditions, while Sections 3.5.2 - 3.5.6 describe the results.

### 3.5.1 Experimental Setups

**Dataset.** We used two Twitter dataset, namely, US dataset and JP dataset. US dataset is provided by Li et al. [110]. This dataset is composed of 3,122,842 Twitter users in the United States with 284,884,514 edges. Similar to previous studies, we geocoded users' location profiles into latitude and longitude pairs using the 2010 census U.S. gazetteer<sup>2</sup>. Specifically, we converted location profile texts in the form of  $[cityName, stateName]$  or  $[cityName, stateAbbreviation]$  into latitude and longitude pairs. As a result, we obtained 464,794 (14.9%) labeled users. Note that misreports of location in location profiles can degrade the location inference. However, Jurgens et al. [104] experimentally showed that

---

<sup>2</sup><http://www.census.gov/geo/maps-data/data/gazetteer2010.html>

Table 3.2: Datasets details.

	all users	labeled users	test users	edges	distinct locations
<b>US dataset</b> [110]	3,122,842	464,794	45,033	284,844,514	9,124
<b>JP dataset</b>	201,570	201,570	19,546	33,569,924	11,142

there are not so many misreports of location. So we believe users’ location profiles show their true home locations.

JP dataset is constructed in this research. This dataset is composed of 201,570 Twitter users in Japan with 33,569,924 edges. Location profiles of users in this dataset were also geocoded into coordinates of latitude and longitude using *Yahoo! geocoder*<sup>3</sup>. Note that all users in this dataset are limited to labeled users.

To evaluate the precision, we randomly divided the labeled users into a test set and a training set, where 10% were assigned to the test set and the rest were assigned to the training set. Details of these two datasets are shown in Table 3.2. Section 3.5.3 uses both US dataset and JP dataset, while Sections 3.5.2, 3.5.5, and 3.5.6 use only US dataset.

**Implementation.** We implemented our proposed method and other existing methods as described in Section 3.5.2. Our code is available at [http://github.com/yamaguchiyuto/location\\_inference](http://github.com/yamaguchiyuto/location_inference).

**Evaluation metrics.** We evaluated our method and existing methods using five metrics.

- *Precision*: The ratio of correctly inferred users versus all inferred users. If the error distance between the inferred and actual location is less than 160 km (100 miles), the inference is assumed to be accurate. This metric has been used in [97] [28] [110].
- *Coverage*: The ratio of inferred users versus all test users.
- *F-measure*: The harmonic mean of the precision and coverage.
- *Mean E.D.*: The mean error distance between the inferred location and the actual

---

<sup>3</sup><http://developer.yahoo.co.jp/webapi/map/openlocalplatform/v1/geocoder.html>

location.

- *Median E.D.*: The median error distance between the inferred location and the actual location.

In addition, we employ *accumulative precision plot* at various distances, which shows that  $y\%$  of users' error distances are within  $x$ km.

**Thresholds.** The confidence threshold  $p_0$  is varied in Section 3.5.5 to examine its effect on the trade-off between precision and coverage. The degree threshold  $c_0$  is used in Section 3.5.6 to examine its effect on the trade-off between coverage and computational cost. If the threshold values are not clearly specified,  $c_0$  is not used and  $p_0$  is set to 0.003, which achieves the best F-measure.

### 3.5.2 Comparison with Existing Methods.

Our method is compared to four graph-based methods including the state-of-the-art below:

- *UDI* is the state-of-the-art method proposed by Li et al.[110]. For UDI, we employ the *global prediction method*<sup>4</sup> as its inference method. Although their model can integrate user-generated contents and a social graph, we do not use user-generated contents because the objective of this experiment is to compare the performance of these graph-based methods.
- *Backstrom* is a method by Backstrom et al. [29], which considers the likelihoods of generated edges assuming that short-distance edges are more likely generated than long-distance ones.
- *Jurgens* is a method by Jurgens. [104], which is based on the label propagation [105].
- *Naive* is a naive method that infers user  $u$ 's location by simply calculating the medoid of locations of  $u$ 's neighbors. Note that this method uses both followers and friends

---

<sup>4</sup>The authors of [110] proposed two types of inference methods: global prediction method and local prediction method. The former achieved a higher accuracy than the latter.

as neighbors, which achieves better precision and coverage than the case of using only followers or friends.

These four existing methods are all based on the closeness assumption.

Table 3.3: Summary of the comparison of our method to existing methods (US dataset).

	<i>LMM</i>	<i>UDI</i>	<i>Backstrom</i>	<i>Jurgens</i>	<i>Naive</i>
<i>Precision</i>	<b>0.754</b>	0.594	0.513	0.530	0.445
<i>Coverage</i>	0.850	<b>0.926</b>	0.967	0.926	0.900
<i>F-measure</i>	<b>0.799</b>	0.724	0.671	0.674	0.596
<i>Mean E.D.</i>	<b>297,739</b>	542,483	703,271	596,022	616,196
<i>Median E.D.</i>	<b>3,804</b>	37,363	124,559	118,210	249,982

Table 3.4: Summary of the comparison of our method to existing methods (JP dataset).

	<i>LMM</i>	<i>UDI</i>	<i>Backstrom</i>	<i>Jurgens</i>	<i>Naive</i>
<i>Precision</i>	<b>0.679</b>	0.524	0.571	0.430	0.421
<i>Coverage</i>	0.850	<b>0.999</b>	0.980	0.982	0.961
<i>F-measure</i>	<b>0.754</b>	0.688	0.721	0.598	0.586
<i>Mean E.D.</i>	<b>171,567</b>	240,223	226,535	243,215	238,565
<i>Median E.D.</i>	<b>37,232</b>	149,963	41,946	175,667	18,1842

Figure 3.5 and Table 3.3 show the results of US dataset, and Figure 3.6 and Table 3.4 show the results of JP dataset. According to the two tables, our method achieves the best precision and F-measure. Our method improves precision by 17%-69% compared to other method and preserves about 85% of the coverage. UDI shows approximately the same results as the original paper in the US dataset. In addition, the mean E.D. and median E.D. of our method are substantially reduced compared to the other methods.

Figures 3.5 and 3.6 show the accumulative precision plot at various distances. Our method successfully reduces the error distances. Specifically, about a half of the users are located within a 1km error distance using our method in the result of the US dataset. Note that although in the result of the JP dataset, our method is inferior to Backstrom in the part of small error distance, our method outperforms Backstrom even in the JP dataset in regard to the precision metric and F-measure metric in Table 3.4.

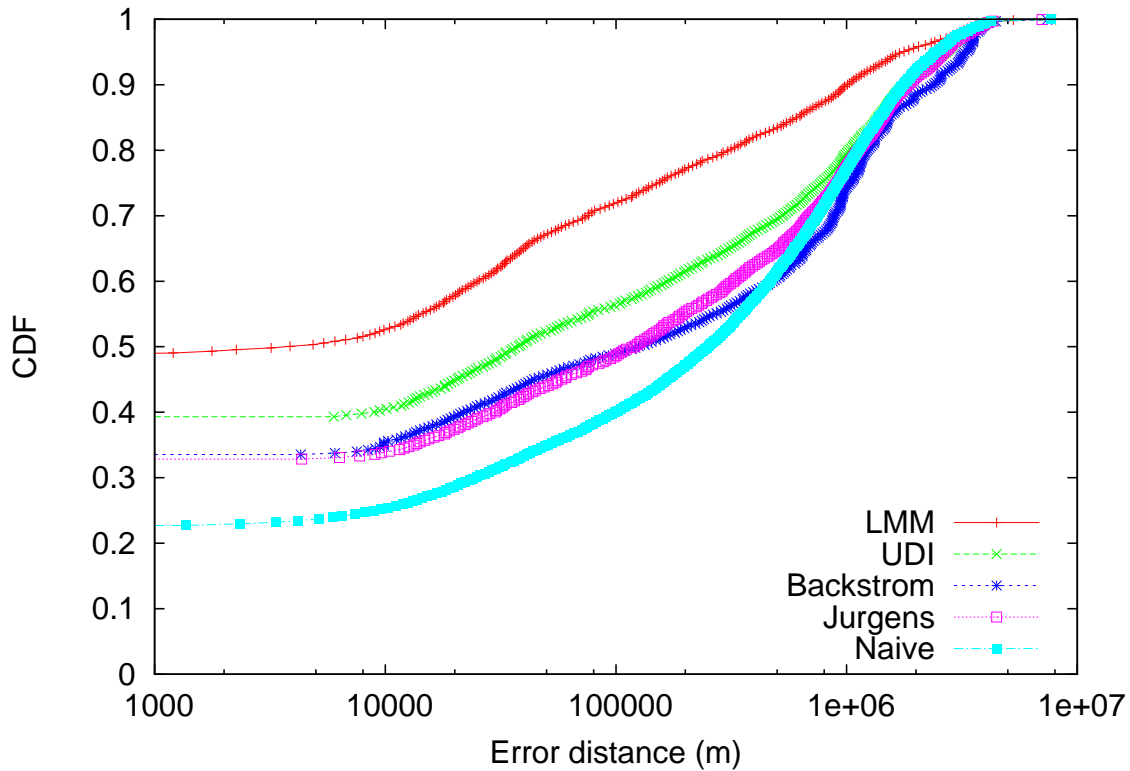


Figure 3.5: Accumulative precision of our method and existing methods at various error distances using US dataset. Our method successfully locates about a half of the users within a 1km error distance, and outperforms the other methods, including the state-of-the-art method.

These results indicate that our concentration assumption provides better clues than the closeness assumption. If we can find graph landmarks in social media, they make it possible to infer home locations of other users accurately.

The coverage of our method drops to a lower value because if the probability density at the mode point is lower than  $p_0$ , our method can decide not to infer the user location. The effect of this threshold is examined in Section 3.5.5.

### 3.5.3 Comparison of Target Areas

Inference results are considered to vary with different target areas. Factors that potentially affect the results include:

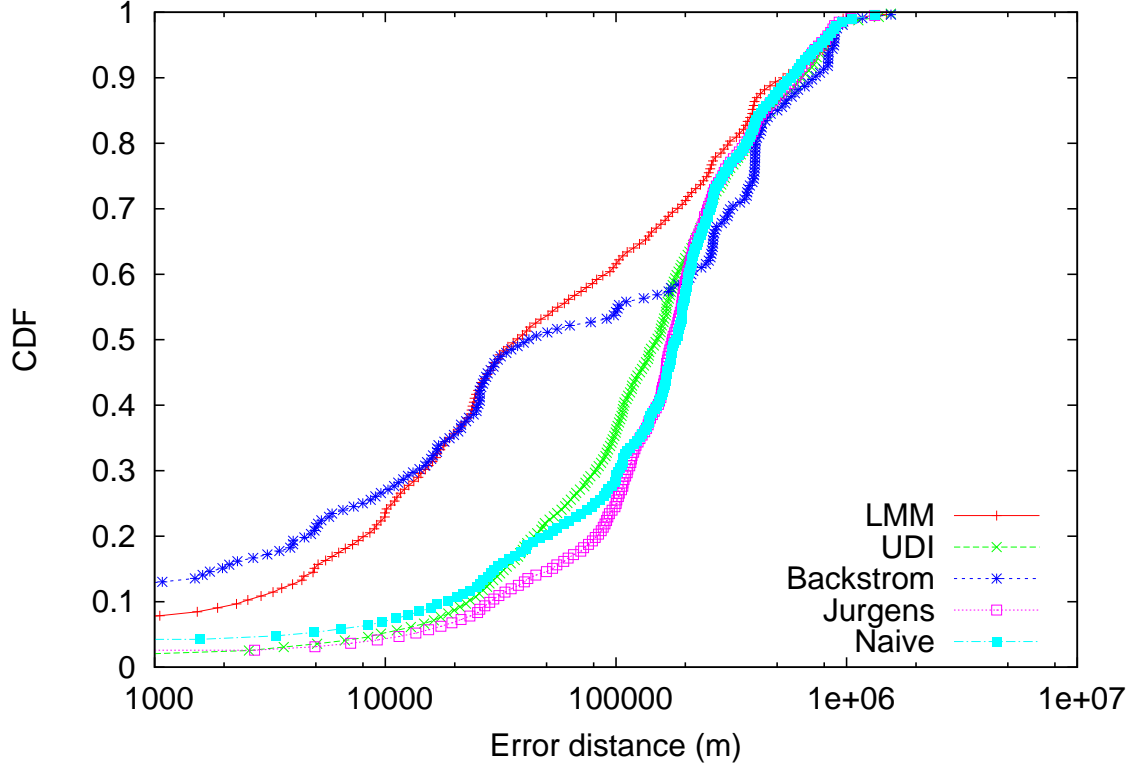


Figure 3.6: Accumulative precision of our method and existing methods at various error distances using JP dataset. Although the results of our method and Backstrom seem to be in the same level, our method outperformed other methods.

- the size of the target area (e.g., the United States and Japan),
- the skew of the population distribution,
- and the distribution of cities.

For the first factor, the error distance of the inference in a large target area (e.g., all over the U.S.) is naturally considered to be large. For the second one, it is thought that the inference result becomes better with the strong skew of population distribution. For example, in Japan, since approximately one third of population live in Kanto area, even a plain inference that locates all users in Kanto area achieves a small error distance. If this skew goes to the extreme situation where a half of population live in Kanto area, the

error distance of such plain method is further reduced. For the last one, which is similar to the second one, inference may become difficult if there are a lot of metropolises which are widely dispersed in the target area. For example, large cities in the U.S. are far from each other, while metropolises in Japan are concentrated near Tokyo and Osaka, which leads to the more accurate inference result in Japan. To examine the above discussion, in this section, we compared the results of US dataset and JP dataset.

Figures 3.5 and 3.6, and Tables 3.3 and 3.4 compares the results in two different areas. Comparing the results of US dataset and JP dataset, the maximum error distance of JP dataset is smaller than that of US dataset. This confirms the first factor discussed above.

On the other hand, in the part of the small error distance in Figures 3.5 and 3.6, the accuracy of US dataset is higher than that of JP dataset, which is considered because the number of distinct locations of US dataset is larger than that of JP dataset. It becomes difficult to infer the more precise home locations with the larger number of distinct locations.

### 3.5.4 Comparison of Variations of the Proposed Method.

To examine which part of the proposed method leads to the good results, five variations of our method are compared in this section.

- *LMM*: Our method described in Section 3.4.
- *LMM w/o m*: Our method without mixture weights. This method uses the same value for all mixture weights.
- *LMM w/o mv*: Our method without mixture weights and variances. This method uses the same value for all mixture weights, and regards variances of all the Gaussian distributions as 1.
- *Medoid*: A method that simply calculates the medoid of neighbors' dominance locations. This method does not use the dominance distribution.

- *Centoid*: A method that simply calculates the centroid of neighbors’ dominance locations. This method does not use the dominance distribution.

Table 3.5: Summary of the comparison of variations of our method (US dataset).

	<i>LMM</i>	<i>LMM w/o m</i>	<i>LMM w/o mv</i>	<i>Medoid</i>	<i>Centroid</i>
<i>Precision</i>	0.754	<b>0.757</b>	0.543	0.357	0.274
<i>Coverage</i>	0.850	0.846	<b>0.996</b>	<b>0.996</b>	<b>0.996</b>
<i>F-measure</i>	0.799	<b>0.800</b>	0.703	0.526	0.429
<i>Mean E.D.</i>	297,739	<b>292,917</b>	587,857	698,769	705,689
<i>Median E.D.</i>	3,804	<b>2,694</b>	75,885	413,459	455,695

Table 3.5 summarizes the results. Contrary to our expectations, LMM and LMM w/o m give approximately the same results. These two methods form almost the same curve in Figure 3.7, indicating that the mixture weights do not improve the precision. There are three reasons that the mixture weights do not work well. 1) Even if users have relatively small degrees, they provide some clues as long as they have small dispersions. 2) Most of the users have small degrees, which results in discarding a substantial part of the clues by imposing small weights. 3) Users with large degrees are weighted heavily regardless of their dispersions. Hence, we have to carefully design the mixture weight, and this is our future work.

The other three variations do not show good results, indicating that employing dispersion leads to good results. Moreover, comparing LMM w/o mv and Medoid indicates that the considering the dominance location as the probability distribution rather than just a location point positively influences the results. Although not all users provide clues, users with a small dispersion provide significant clues for location inference.

### 3.5.5 Effect of the Confidence Threshold

LMM can adjust the trade-off between precision and coverage by imposing the confidence threshold. This section shows the effect of the confidence threshold. Figure 3.8 shows the result by varying the value of threshold  $p_0$ . The x-axis denotes the value of  $p_0$ , and



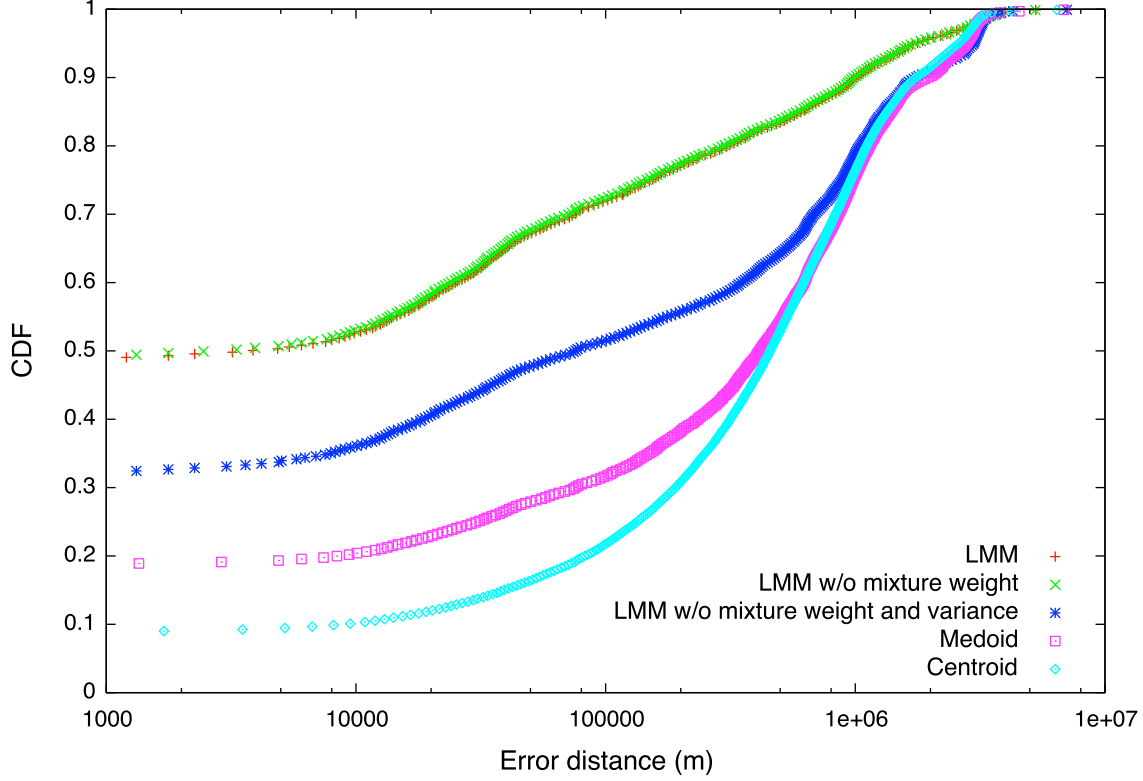


Figure 3.7: Accumulative precision for variations of our method at various error distances. Although considering the geographical dispersion improves the precision, considering the mixture weight does not. Medoid and Centroid do not work well because they simply use the dominant location as the points rather than as the distributions.

each line denotes the precision, coverage, and F-measure. As the value of  $p_0$  increases, the precision increases but the coverage decreases. The F-measure achieves the best score around  $p_0 = 0.003$ . Note that even if we do not impose the confidence threshold ( $p_0 = 0$ ), our method outperforms other methods for all these metrics.

If the probability density at the mode point is low, the overall probability distribution does not have a clear peak. In this case, we should not infer the home location of that user because there are insufficient clues to determine the location. Our method can select this option because it is based on the probability distribution.

If we require a high precision (e.g., in the case of sending disaster warnings), we can

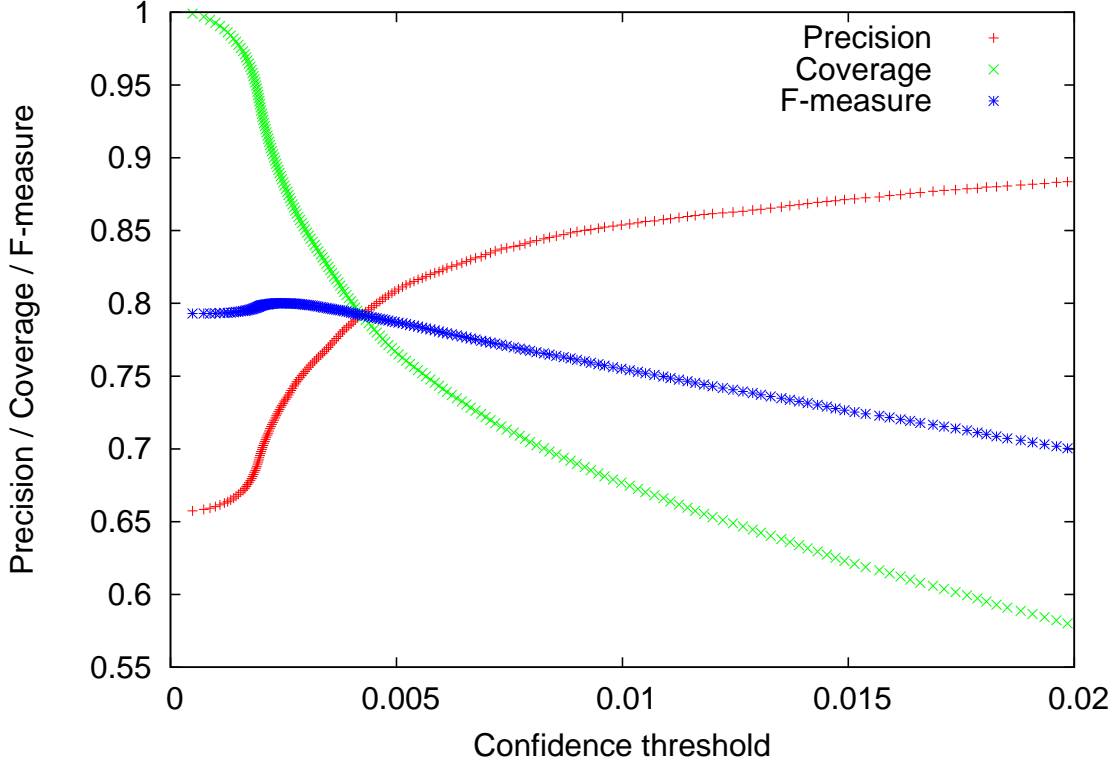


Figure 3.8: Effect of the confidence threshold. x-axis denotes the value of the threshold  $p_0$ . As the value of  $p_0$  increases, the precision increases but the coverage decreases. The F-measure achieves the best score around  $p_0 = 0.003$ , and the precision is about 0.88 at  $p_0 = 0.02$ .

achieve that by imposing the large  $p_0$  value. On the other hand, if we want a high coverage (e.g., local advertisements), we can get that by imposing no constraint, or small  $p_0$  value. The ability of this trade-off adjustment may expand the applications of users' home location profiles.

### 3.5.6 Effect of the Degree Threshold

LMM can also adjust the trade-off between the computational cost and coverage by imposing the degree threshold. Varying the value of threshold  $c_0$  demonstrates the effects of the cost, coverage, and precision. The confidence threshold is not imposed in this section.

Figures 3.9 and 3.10 show the results where the x-axis denotes the value of  $c_0$ . The ratio

of the utilized graph landmarks means the ratio of users satisfying the degree threshold (i.e., users with their degrees  $c_u > c_0$ ) versus all users. The average number of neighbors means the average number of each user’s neighbors whose degrees are larger than  $c_0$ , in other words, the average number of mixed components for each user’s location distribution, which dominates the computational complexity of our method. As for the computational time, the time of finding the mode point of LMM was measured.

As the value of  $c_0$  increases, the ratio of utilized graph landmarks decreases rapidly but the coverage remains high. Hence, we conclude that our method can infer locations of almost all users with only about 5% of users ( $c_0 = 100$ ). This means only 5% of users’ dominance distributions (i.e., mean and variance parameters) and following relationships need to be stored to infer almost all user locations.

In terms of computational cost, we can reduce the average number of neighbors to approximately 30%, preserving 85% of the coverage ( $c_0 = 200$ ). This means that because the computational complexity of our method is  $O(k^2)$  where  $k$  is the number of neighbors, the cost is reduced to about 10%, which is confirmed in Figure 3.10.

From  $c_0 = 0$  to 400, the precision remains about the same value or even decreases, but from  $c_0 = 500$  to 1000, it increases. Two factors may lead to such a behavior. If our method utilizes graph landmarks with high degree, good results are achieved because their small dispersion values are statistically significant. On the other hand, because only a small number of graph landmarks have a high degree, other users do not satisfy the degree threshold and are not used for location inference. Ignoring these ordinal users, which is most of the users, may degrade the performance.

### 3.6 Conclusion

Herein we introduce a novel concept of concentration assumption and graph landmarks and propose a Landmark Mixture Model (LMM) to address the user location inference problem. Graph landmarks have desirable features for location inference: strong clues and a

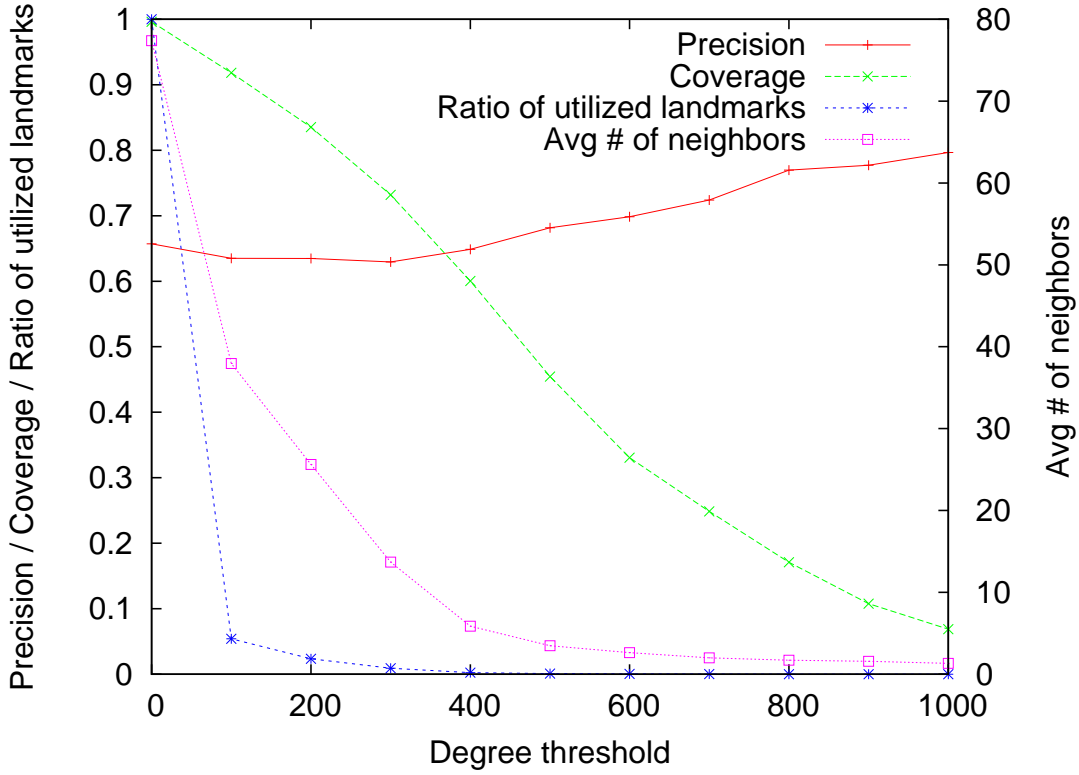


Figure 3.9: Effect of the degree threshold. x-axis denotes the value of the threshold  $c_0$ . As  $c_0$  increases, the ratio of utilized graph landmarks rapidly decreases, preserving the high coverage value. The decrease in the average number of neighbors denotes the reduction of the computational cost of our method. The precision remains high or even increases when we utilize a small number of graph landmarks.

wide coverage. LMM can adjust the trade-offs between precision and coverage, and between computational cost and coverage. This capability may expand applications employing users' home locations. The experimental results show that our inference method outperforms other existing methods, including the state-of-the-art method. The results also demonstrate that imposing the two thresholds allows our method to accomplish a high precision, reduce the computational cost, and preserve a high coverage.

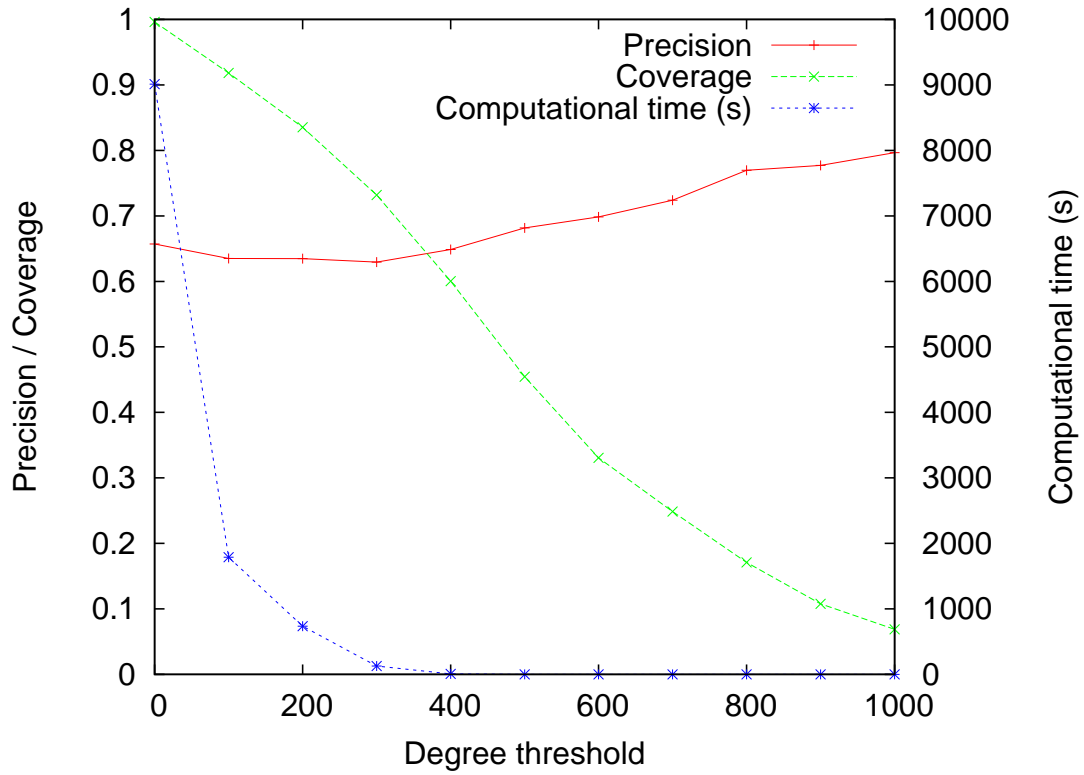


Figure 3.10: Effect of the degree threshold. x-axis denotes the value of the threshold  $c_0$ . As  $c_0$  increases, the ratio of utilized graph landmarks rapidly decreases, preserving the high coverage value. The decrease in the average number of neighbors denotes the reduction of the computational cost of our method. The precision remains high or even increases when we utilize a small number of graph landmarks.

## Chapter 4

# User Location Inference based on Local Events

People using social media transmit vast quantities of information in real time, which forms real-time information sources called social streams. By monitoring social streams, we can detect real-world local events (e.g., earthquakes, fires, etc.) because people all over the world tend to post messages about those local events instantly. In this chapter, we propose a method for user location inference using local events detected from social streams. Our method is based on the assumption that users who post about a local event likely to live close to the event. Specifically, the method first detects local events using messages posted by labeled users, and then infer home locations of unlabeled users who post about the detected event. Experimental results show that our method can properly detect local events and infer user locations more precisely than other existing location inference methods.

### 4.1 Introduction

A huge number of people now use various types of social media and transmits information as they like. For example, people can share their photos and movies, post texts describing their opinions, and review restaurants and books. Among these services, social media with

real-time features such as microblogs have a tremendous amount of texts sent from users in real-time. We focus on this streams of texts and call them *social streams*.

Many researches investigated real-time texts in microblogs and found that it is feasible to detect real-world events (e.g., earthquakes) utilizing social streams [68, 117, 73, 118, 71, 70]. For example, we can detect an earthquake in Tokyo by monitoring social streams because when it happens users in Tokyo immediately react and post texts like “Shaked!!!”. This kind of information helps us sense the real world because users frequently post about situations around them (described in Section 2.1).

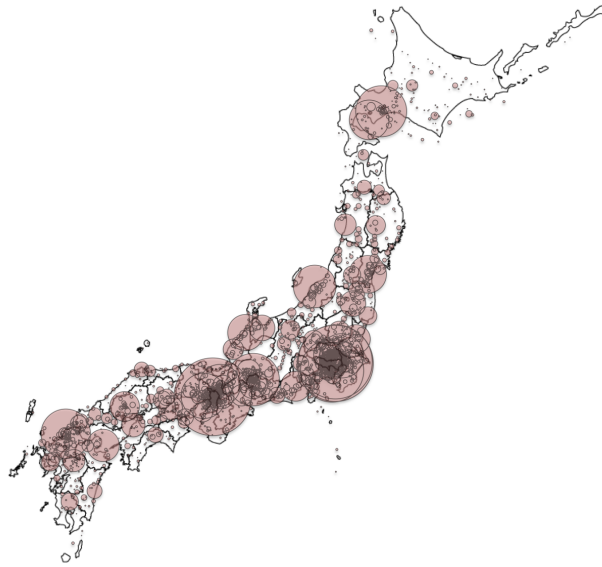
Recall our problem of inferring users’ home locations. In this chapter, we make the “reverse” idea of event detection and infer home locations based on the idea. For instance, it can be considered that a user who posted a text of “Shaked!!!” when an earthquake happened in Tokyo is likely to live in Tokyo. Figure 4.1(a) shows the geographical distributions of texts posted in Twitter in normal times<sup>1</sup>, while Figure 4.1(b) shows the distribution of texts that contain the word “earthquake” posted in Twitter when an earthquake happened in Hiroshima prefecture in Japan. The size of each circle indicates the number of tweets posted at the center of the circle. We can see from these figures that, when a local event occurs at a certain place, majority of posts related to the event tend to be posted by users whose locations are close to the event. In this way, when a user mentions a local event, we can infer his/her home location using the locations of the event. Note that in this chapter we focus on *local events* that have geographical locality, in other words, events that are associated with a certain area (e.g., earthquakes).

Contributions in this chapter are as follows:

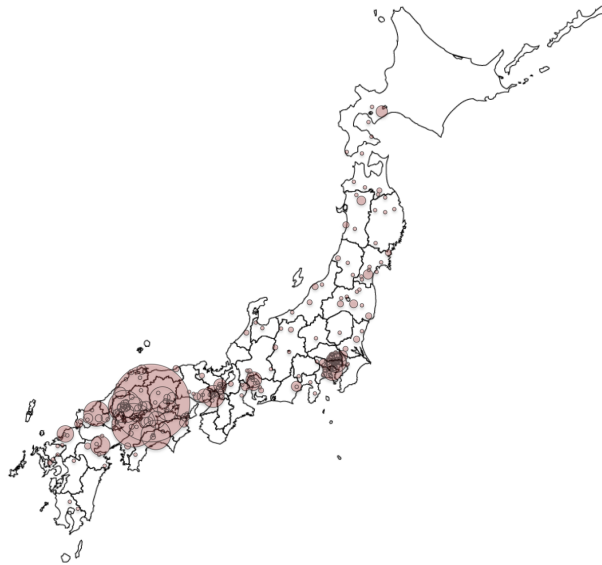
1. A method to infer users’ home locations based on the real-world local events is proposed.
2. Experimental results using Twitter dataset demonstrates the effectiveness of exploiting local events for the problem of inferring home locations.

---

<sup>1</sup>Precisely, it is a geographical distribution of home locations of users who posted tweets.



(a) Location distribution of tweets in Japan posted to Twitter.



(b) Location distribution of tweets containing the word "earthquake" posted when an earthquake occurred in the Hiroshima prefecture in Japan.

Figure 4.1: (a) The large part of messages are posted from metropolises such as Tokyo and Kyoto. (b) Unlike Figure 4.1(a), most tweets are posted from the Hiroshima prefecture, indicating that users who mention the event tend to have geographical proximity with the event.



Our method first detects real-world local events using messages posted by labeled users, and then infers home locations of unlabeled users who posted about detected events. This can be regarded as propagating labels of home locations from labeled users to unlabeled users.

According to the experimental results discussed in Section 4.4, the inference accuracy of our method was about 76%, which improved the accuracy of the existing methods by 33% and 121%. Besides, it was shown that more home locations can be inferred with more detected events, and more events can be detected with more location information.

The rest of this chapter is organized as follows. Section 4.2 states the problem addressed in this chapter and defines the terminology. Our method is described in Section 4.3 and evaluated in Section 4.4 with respect to the validity of detected local events, and the accuracy and coverage of location inference. Finally, Section 4.5 concludes this chapter.

## 4.2 Problem Statement

This section states the problems addressed within this chapter, and defines the terminology. Each *post* is defined as the triple consists of timestamp  $s$ , text  $t$ , and user  $u$ , and is denoted as  $p = (s, t, u)$ , which forms *social stream*  $SS = (p_1, p_2, \dots)$ . Each user  $u$  has home location  $l_u$ , which is set to *NULL* if the home location of the user is unavailable. All locations belong to a predefined location set  $L$ . *Event*  $e$  is defined as a set of posts satisfying the following conditions.

- For each post  $p_i \in e$ ,  $s_i$  belongs to the same predefined time window  $W$ .
- All texts  $t_i$  are adequately similar to each other.
- All locations  $l_{u_i}$  of user  $u_i$  are close to each other.
- The number of posts belonging to event  $e$  is larger than the predefined value.

For example, when a lot of posts containing the word "earthquake" are posted from the Hiroshima prefecture in a short period of time, these posts can be regarded as an event.

Using these definitions, a problem of *event detection* is stated as follows.

**Problem 3 (Local Event Detection)** *Detects a set of events  $E_j = \{e_1, e_2, \dots, e_n\}$  from a set of posts  $P_j$  within a time window  $W_j$ . An event is restricted to be a maximal set of posts satisfying above conditions. If no event is detected,  $E_j$  is an empty set.*

Each unlabeled user is assigned a discrete probability distribution  $P_u$  as:

$$0 \leq P_u(l) \quad (4.1)$$

$$\sum_{l \in L} P_u(l) = 1, \quad (4.2)$$

which is called *location distribution*. The goal of this chapter is to infer this location distribution based on detected local events. A set of events  $E_u$  that user  $u$  mentioned is denoted as:

$$E_u = \{e \in E \mid \text{mention}(u, e)\}, \quad (4.3)$$

where  $\text{mention}(u, e)$  is true if user  $u$  posted  $p \in e$ , and  $E$  denotes the set of all detected events. Using these notations, the main problem of event-based user location inference is defined as follows.

**Problem 4 (Event-Based User Location Inference)** *Given a set of events  $E_u$  that user  $u$  mentioned, infer the  $u$ 's location distribution  $P_u$  so that the mode point over the distribution  $\hat{l}_u = \arg \max_l P_u(l)$  is close to the true home location  $l_u$ .*

### 4.3 Proposed Method

In this section, we propose our Event-based Location Inference Method (ELIM). First our method receives a social stream as an input, and conducts event detection. The method then infers user home locations based on the detected events.

### 4.3.1 Local Event Detection

Our method initially detects local events using both a social stream and home locations of labeled users. Event detection consists of two steps: *content clustering* and *spatial filtering*. Content clustering groups the given posts using their contents, and spatial filtering removes clusters without geographical locality and outputs other clusters as events. The subsequent paragraphs describe the detail of each step.

**Content clustering.** Suppose a given set of posts  $P_j$  belongs to a certain time window  $W_j$ . The width of the time window is a parameter given by the time period, such as 10 minutes. Text  $t$  of each post  $p \in W_j$  is represented as a term vector  $\mathbf{v}(t)$  based on the *bag-of-words* model. The number of dimensions of the vector is the size of the vocabulary set. Each component of this vector is set to  $1/|T|$  if the corresponding term is contained in  $t$  of post  $p$ , otherwise, it is set to 0. Note that  $T$  is the set of terms contained in text  $t$ .

Posts are clustered based on their term vector  $\mathbf{v}(t)$ . The proposed method employs DBSCAN [119], which are robust to noises. That is, all posts are not necessarily clustered, and some of the posts are discarded as noise. This allows clusters to be extracted only when a lot of similar posts occur in a short time period. Euclidean distance  $dist_t(\mathbf{v}(t_i), \mathbf{v}(t_j))$  is adopted for the distance function between two term vectors in this method. DBSCAN takes two parameters *MinPts* and *Eps*, which are minimum number of posts which is allowed to construct an event, and maximal distance between posts in the same cluster.

The content clustering step outputs a set of clusters  $C_j = \{c_1, c_2, \dots, c_n\}$  for each time window  $W_j$ . Figure 4.2 illustrates the procedure of the content clustering step.

**Spatial filtering.** Spatial filtering receives a set of clusters  $C_j$  as an input, and outputs clusters with a geographical locality as events. As a measure of locality, we introduce *dispersion* described as Equation 4.4.

$$dispersion(c) = \frac{1}{|c|} \sum_{p_i \in c} dist_l(l_i, m_c) \quad (4.4)$$

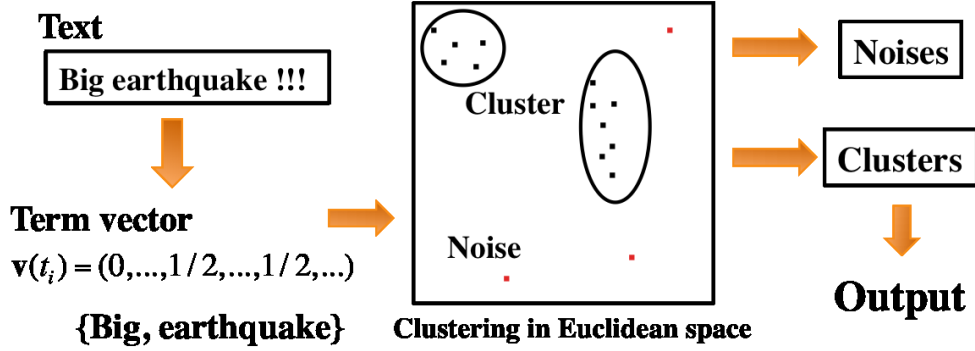


Figure 4.2: Procedure of the content clustering. Each text is represented as the term vector, and is clustered in Euclidean space. Posts in a sparse region are disposed of as noises.

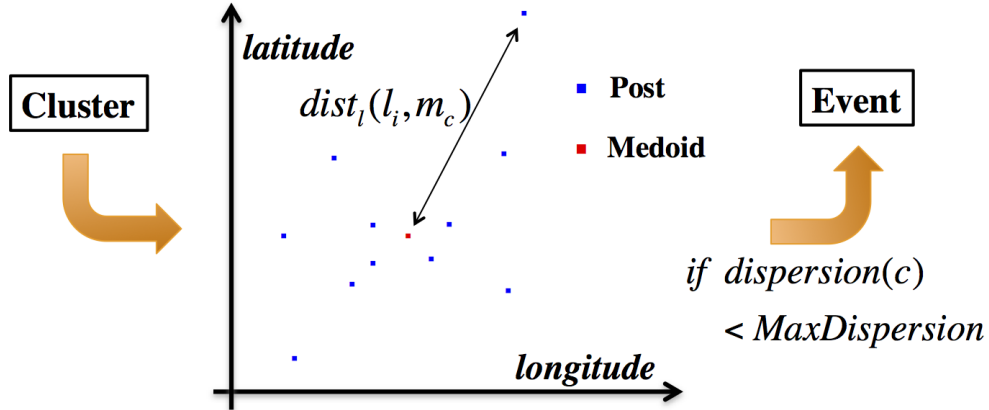


Figure 4.3: Procedure of the spatial filtering. Each post in a cluster is plotted in the Euclidean space. If the dispersion does not exceed the *Maxdispersion*, the cluster is regarded as an event.

where  $dist_l(\cdot, \cdot)$  denotes the Euclidean distance between two locations, and  $m_c$  is the *medoid* of locations in a cluster. The medoid is defined as the data point that minimizes the total distance between the data point and other points. A smaller dispersion means a larger geographical locality of the corresponding cluster. The spatial filtering outputs a set of events  $E_j = \{e_1, e_2, \dots, e_m\}$  that consist of clusters whose *dispersion* value is lower than the predefined parameter *MaxDispersion* (Fig. 4.3). Note that *MaxDispersion* indicates the maximal value of dispersion that is utilized for location inference.

---

**Algorithm 1** Local Event Detection

---

**Input:** set of posts  $P_j$  which belongs to time window  $W_j, Eps, MinPts$

**Output:** set of events  $E_j$

```
 $E_j \leftarrow \emptyset$ 
 $C_j \leftarrow \text{DBSCAN}(P_j, Eps, MinPts)$  // noises are discarded.
for all cluster  $c$  in  $C_j$  do
   $m_c \leftarrow \text{calculate\_medoid}(c)$ 
   $d_c \leftarrow \frac{1}{|c|} \sum_{p_i \in c} \text{dist}_l(l_i, m_c)$ 
  if  $d_c < \text{MaxDispersion}$  then
     $E_j \leftarrow E_j \cup c$ 
  end if
end for
return  $E_j$ 
```

---

**Event detection algorithm.** Algorithm 1 shows the algorithm for event detection. This algorithm receives a set of posts that belongs to one time window  $W_j$ , and then outputs a set of events  $E_j$ . The time complexity of this algorithm is dominated by that of DBSCAN.

#### 4.3.2 User Location Inference

ELIM infers location distributions of unlabeled users based on a set of all detected events  $E$ . Intuitively, users who mention event  $e$  when  $e$  occurs are likely to live at the location of  $e$ . Based on this notion, ELIM infers location distributions of unlabeled users who posted  $p \in e$ , so that the probability value of the location where  $e$  occurs becomes high. This makes the location point  $\hat{l}_u = \arg \max_l P_u(l)$  with the largest probability value approximates the true home location  $l_u$ . The following paragraphs detail the *location model* to infer the location and the MAP estimation of the distribution.

**Location model.** Given a set of possible locations  $L$ , location distribution  $P_u$  is denoted as a parameter vector  $\theta_u = (\theta_{u,1}, \dots, \theta_{u,|L|})$  as

$$P_u(l) = \theta_{u,l}, \sum_{l \in L} \theta_{u,l} = 1, 0 \leq \theta_{u,l} \leq 1. \quad (4.5)$$

ELIM aims to estimate this parameter vector  $\theta_u$ . The inferred location  $\hat{l}_u$  can be derived

from  $\theta_u$  as follows:

$$\hat{l}_u = \arg \max_{l \in L} \theta_{u,l}. \quad (4.6)$$

Parameter vector  $\theta_u$  is given by the MAP estimation as

$$\theta_u = \arg \max_{\theta} P(\theta|E_u). \quad (4.7)$$

This equation means that if user  $u$  mentions a set of event  $E_u$ , the parameter vector  $\theta_u$  can be inferred based on it.

**MAP estimation of location distribution.** MAP estimation can be transformed as follows based on Bayes' theorem:

$$\arg \max_{\theta} P(\theta|E_u) = \arg \max_{\theta} P(E_u|\theta)P(\theta). \quad (4.8)$$

Moreover, assuming that detected events are i.i.d (independent and identically distributed), then the likelihood function  $P(E_u|\theta)$  satisfies the following equation:

$$P(E_u|\theta) = \prod_{e \in E_u} P(e|\theta), \quad (4.9)$$

where  $P(e|\theta)$  denotes the likelihood function that event  $e$  is generated under the parameter  $\theta$ . Therefore, parameter vector  $\theta_u$  can be expressed as:

$$\theta_u = \arg \max_{\theta} P(\theta) \prod_{e \in E_u} P(e|\theta) \quad (4.10)$$

The following passages define the likelihood function  $P(e|\theta)$  and the prior distribution  $P(\theta)$ .

Let event  $e$  be represented as location vector  $\mathbf{v}(e) = (n_{e,1}, \dots, n_{e,|L|})$  based on the bag-of-words model. In this case,  $e$  is regarded as the set of locations of posts  $p \in e$ .  $n_{e,k}$  denotes the number of posts in  $e$  whose location is  $l_k$ . Using this location vector, event  $e$  is

generated by the multinomial distribution with parameter  $\theta$  as:

$$P(e|\theta) = \text{Multi}(\mathbf{v}(e); \theta) = \frac{N_e!}{\prod_k n_{e,k}!} \prod_k \theta_k^{n_{e,k}}, \quad (4.11)$$

where  $N_e = \sum_k n_{e,k}$ . Hence, ELIM adopts the multinomial distribution as the likelihood function of events.

Because the conjugate prior distribution of a multinomial distribution is the Dirichlet distribution, ELIM adopts the Dirichlet distribution for the prior distribution of parameter  $\theta$ .

$$P(\theta) = \text{Dir}(\theta; \alpha) = \frac{\Gamma(A)}{\prod_k \Gamma(\alpha_k)} \prod_k \theta_k^{\alpha_k - 1}, \quad (4.12)$$

where  $\Gamma(\cdot)$  denotes the gamma function,  $\alpha$  is the parameter of the Dirichlet distribution, and  $A = \sum_k \alpha_k$ .

Using the likelihood function and the prior distribution defined thus far, the MAP estimation can be written as<sup>2</sup>:

$$\begin{aligned} \theta_u &= \arg \max_{\theta} P(\theta) \prod_{e \in E_u} P(e|\theta) \\ &= \arg \max_{\theta} \text{Dir}(\theta; \alpha) \prod_{e \in E_u} \text{Multi}(\mathbf{v}(e); \theta) \\ &= \arg \max_{\theta} \text{Dir}(\theta; \alpha + \sum_{e \in E_u} \mathbf{v}(e)). \end{aligned} \quad (4.13)$$

By solving this equation, the desired parameter vector  $\theta_u$  is obtained as:

$$\theta_{u,k} = \frac{\sum_{e \in E_u} n_{e,k} + \alpha_k - 1}{\sum_{e \in E_u} N_e + A + |L|}. \quad (4.14)$$

**User location inference algorithm** Algorithm 2 shows the algorithm of user location inference. This algorithm takes a set of detected event  $E_u$  as an input, and then infers the

---

<sup>2</sup>Transformation from the second formula to the third one is based on the conjugate property of the multinomial distribution and the Dirichlet distribution

---

**Algorithm 2** User Location Inference Algorithm

---

**Input:** Set of events  $E_u$

**Output:** parameter vector  $\theta_u$

```
 $\theta_u \leftarrow \mathbf{0}$ 
for all event  $e$  in  $E_u$  do
  for all post  $p$  in  $e$  do
     $v \leftarrow \text{user}(p)$  // returns the user who posts  $p$ 
     $l \leftarrow \text{location}(v)$  // returns the location of user  $v$ 
    if  $u \neq v$  and  $l \neq \text{NULL}$  then
       $\theta_{u,l} \leftarrow \theta_{u,l} + 1$ 
    end if
  end for
end for
return  $\theta_u$ 
```

---

parameter vector of user  $u$ .

The time complexity of this algorithm is  $O(|E_u| \cdot |e|)$ . Although it seems relatively large, it does not matter because of the following reasons:

- The number of events  $|E_u|$  that user  $u$  mentions is small in most cases.
- The size of event  $|e|$  (i.e., the number of posts in event  $e$ ) is small in most cases.

## 4.4 Experiments

This section discusses the effectiveness of our method from three different perspectives.

1. *Validity of detected events*: whether or not the detected events consist of posts mentioning the same incident, and have geographical locality.
2. *Accuracy of location inference*: what fraction of users' home locations are correctly inferred.
3. *Coverage of location inference*: how many users' home locations are inferred regardless of the accuracy.



Section 4.4.1 describes the prototype system which implements our proposed method. Section 4.4.2 details the Twitter datasets for the experiments. Section 4.4.3 to Section 4.4.6 discuss the experimental results.

#### 4.4.1 Prototype System

Figure 4.4 illustrates our prototype system<sup>3</sup>, which implements the proposed method using the Twitter social stream as the input. The solid line frames the range of the system. Our system initially collects tweets using *Tweet Crawler*. Concretely, Tweet Crawler specifies keywords related to events that we want to detect, and collects tweets containing at least one of the keywords. For each collected tweet, *User Location Crawler* acquires location profiles of the user who posted the tweet. Because the location profile in Twitter is in textual form, *User Location Crawler* converts texts into the coordinates using Yahoo! Geocoder<sup>4</sup>. These coordinates of latitude and longitude are stored in *User DB*. For users who do not specify their locations, NULL values are stored in the User DB. Inferring these unknown locations is our objective.

The gathered tweets and user locations are fed into the *Event Detection* component. Events detected by Event Detection component are stored in *Event DB*, and then are fed into the *User Location Inference* component. The User Location Inference component receives events as inputs and infers user locations that are denoted as parameter vector  $\theta_u$ . In this prototype system, all hyperparameters  $\alpha_k$  are set to 1, in other words, the system uses the uninformative prior.

#### 4.4.2 Dataset

This section details the datasets gathered by Tweet Crawler and User Location Crawler. Five datasets are constructed using the collected tweets specified by five keyword sets in the period of Nov. 5-19, 2012 (Table 4.1). Note that all tweets are written in Japanese, and

---

<sup>3</sup>The code of the prototype system is available at <https://github.com/yamaguchiyuto/bagel>

<sup>4</sup><http://developer.yahoo.co.jp/webapi/map/openlocalplatform/v1/geocoder.html>

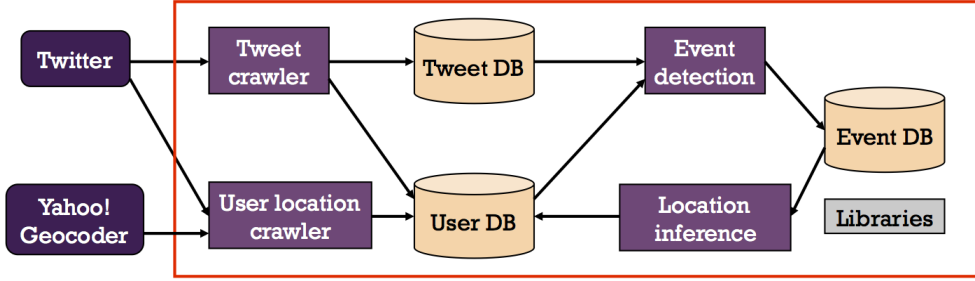


Figure 4.4: Overview of the prototype system. Two crawling components collect data that is used for event detection and user location inference are conducted.

Table 4.1: Dataset.

	<i>All</i>	<i>Earthquake</i>	<i>Weather</i>	<i>Tornado</i>	<i>Emergency</i>
<i># of users</i>	508,824	141,978	307,474	15,868	81,592
<i># of labeled users</i>	130,391 (26%)	36,613 (26%)	78,582 (26%)	3,412 (22%)	20,043 (25%)
<i># of tweets</i>	1,018,164	317,982	519,803	12,073	103,933
<i>keywords</i>	-	earthquake	heavy rain; flood; thunder; typhoon	tornado	patrol car; ambulance; fire truck; siren

keywords are specified also in Japanese. Each column shows one dataset: *All*, *Earthquake*, *Weather*, *Tornado*, or *Emergency*. *All* is composed of other four datasets. Each dataset contains about 25% labeled users. Since proper parameters for each dataset is different, different values are assigned to the parameters in different experiments and in different datasets. Specific parameter values are shown in each section.

#### 4.4.3 Validity of Detected Events

The experiment in this section confirms the validity of the events detected by our method with regard to two points:

- How large fraction of detected events to be regarded as valid.
- Whether the amount of known user locations affects the results.

Four datasets (*All* is excluded) were used in this experiment, and parameters were set to  $WindowSize = 600s$ ,  $MinPts = 15$ ,  $Eps = 0.2$ ,  $MaxDispersion = 200km$ . To verify the

influence of the amount of known user locations on the results, we compared cases where the percentage of labeled users to detect events are 100%, 50%, or 25%.

The methodology of this experiment is described as follows. First, events were detected by the proposed method. Then the detected events were shown to five examinees, who are all graduate students majoring computer science. The examinees determined whether each event was valid or not, using the two guidelines below:

- 50% or more of tweets belong to a detected event mention one real event.
- Multiple users mention the identical event.
- The real event mentioned by these tweets has geographical locality.

In this experiments, we define the events as those that only users on-site could know. For example, a traffic accident can be an event, but it is no longer an event after it is broadcast in some kind of mass media. Based on the examinees' evaluations, the precision (i.e., the ratio of valid events) was calculated.

Figures 4.5 and 4.6 show the precision and number of detected events, respectively. The precision varies for each dataset. *Earthquake*, *Weather*, and *Tornado* have precisions between 0.7 and 0.8, whereas *Emergency* shows a low precision of 0.1. The results differ because the former three datasets have more frequent events which are mentioned by numerous users. On the other hand, *Emergency* dataset has small-scale events such as traffic accidents in multiple areas at the same time, leading to low accuracy because these events form multimodal distribution. Detecting events from such multimodal distribution is our future work.

By increasing the amount of the user locations used to detect events, the precision slightly increases and the number of detected events increases to some extent. Although the experiment did not use the inferred locations, the results confirm that as the percentage of known user locations increases, the better the event detection because the dispersion values can be more precisely calculated as the known user locations increases.

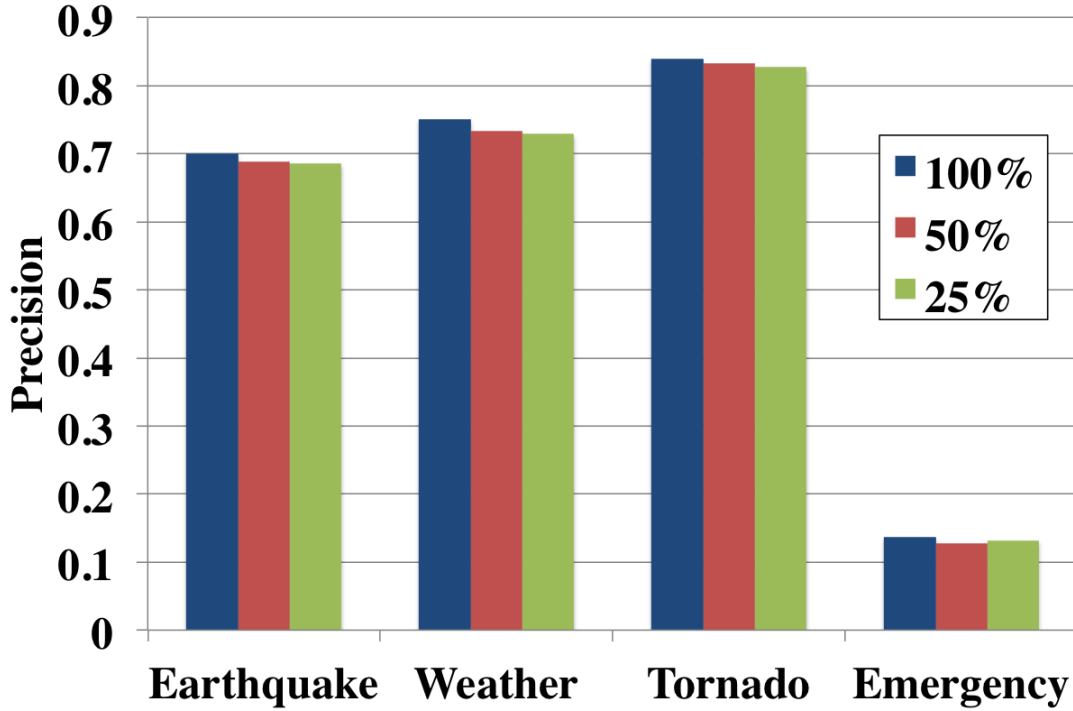


Figure 4.5: Precision of detected events. The precision of *Emergency* is low because few local events happened during this experiment in this dataset. As the number of known user locations increases, the precision increases slightly.

In this experiment, clusters with one or two labeled users are eliminated because the dispersion value cannot be calculated. Hence, the number of detected local events increases as the number of known user locations increases.

#### 4.4.4 Accuracy of User Location Inference

In this section, our proposed method is compared to following methods.

- *UDI* is the state-of-the-art method proposed by Li et al. [110]. For UDI, we employ the *global prediction method*<sup>5</sup> as its inference method.
- *Cheng* is a method proposed by Cheng et al. [28], which utilizes local words contained

---

<sup>5</sup>The authors of [110] proposed two types of inference methods: global prediction method and local prediction method. The former achieved a higher accuracy than the latter.

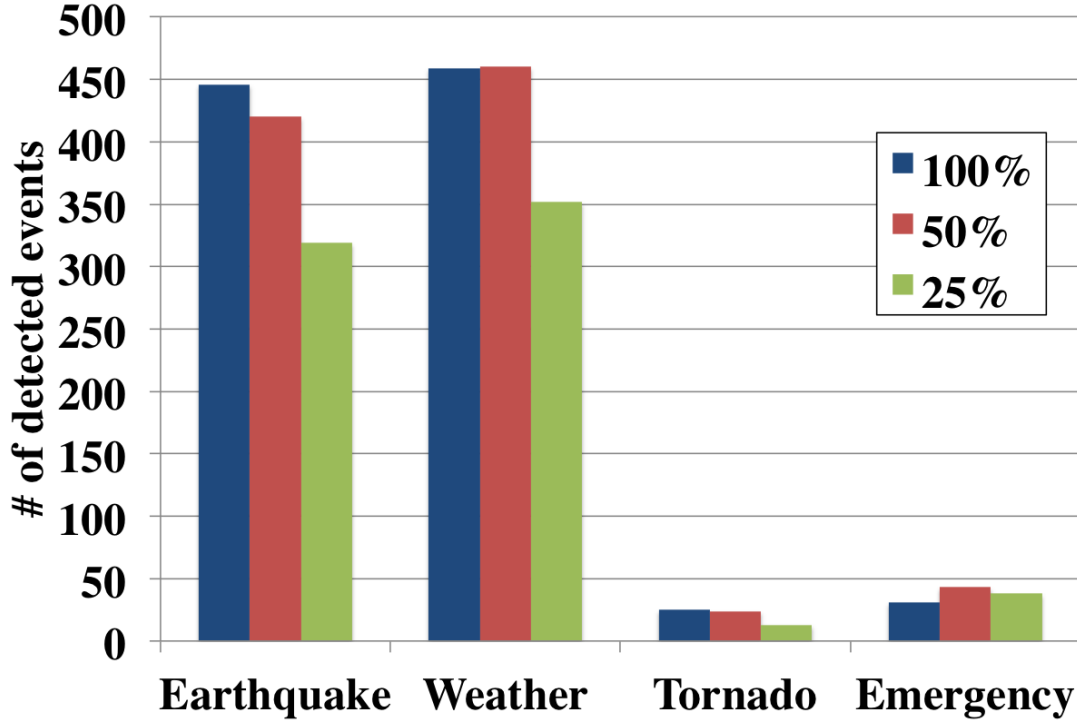


Figure 4.6: The number of detected events. The number of detected events considerably differs from each other. More events are detected as the number of known user locations increases.

in users' tweets.

- *Random* is a naive method that assigns randomly selected locations to all users.

Methodology of this experiment is described as follows. Using the *Weather* dataset, Proposed inferred the locations of labeled users, and evaluated its accuracy based on leave-one-out cross validation. Parameters were set to *WindowSize* = 600s, *MinPts* = 15, *Eps* = 0.5, and *MaxDispersion* = 150km. These values were fixed based on the parameter evaluation described in Section 4.4.6.

For the method of Cheng, training users and test users are needed. Hence we randomly sampled 1,000 labeled users for the test set, and 10,000 labeled users for the training set. Note that duplication of a test user and a training user is not allowed. For each training user, we collected 100 latest tweets, which forms the training tweet set. Cheng also collected

the latest 3,200 tweets from each test user, and inferred home locations of test users using test tweets.

UDI also uses above 1,000 test users, 10,000 training users, and 100 tweets for each training user. Toponyms are extracted from texts of tweets by Mecab<sup>6</sup>, a Japanese morphological tool. Extracted toponyms are then converted into coordinates of latitude and longitude by using OpenStreetMap<sup>7</sup>. Moreover, test users' followers and followees are also collected.

Figure 4.7 and Table 4.2 show the results. In Fig. 4.7, the horizontal axis indicates the error distance between the true home location and the inferred location, while the vertical axis indicates the precision within the corresponding error distance. In Table 4.2, *Mean E.D.* indicates the mean error distance, and *Median E.D.* indicates the median error distance. *Pre@160km* and *Pre@80km* indicate the precision where inference results with at most 160km or 80km error distances are regarded as correct inference, respectively. Proposed shows the highest precision and minimum error distances. Comparing the precisions of all the methods within a 160km error distance, Proposed shows about 34% and 122% improvements from UDI and Cheng, respectively.

The reason that Proposed more accurately infers locations is because during the experimental period heavy rains and thunderbolts occurred in various locations, which leads to more local events detected and high precision. On the other hand, Cheng shows a lower precision than in the original paper [28]. Cheng's original experiment was conducted in the United States, while this experiment was conducted in Japan. Finer-tunings (e.g., stop words) may improve the precision of Cheng's method because there are several settings (e.g., languages, test users) to be tuned.

#### 4.4.5 Coverage of User Location Inference

This section evaluates the coverage of the proposed method. We conducted an experiment to compare the results of the five datasets, and an experiment to compare the results with

---

<sup>6</sup><https://code.google.com/p/mecab/>

<sup>7</sup><http://wiki.openstreetmap.org/wiki/JA:Nominatim>

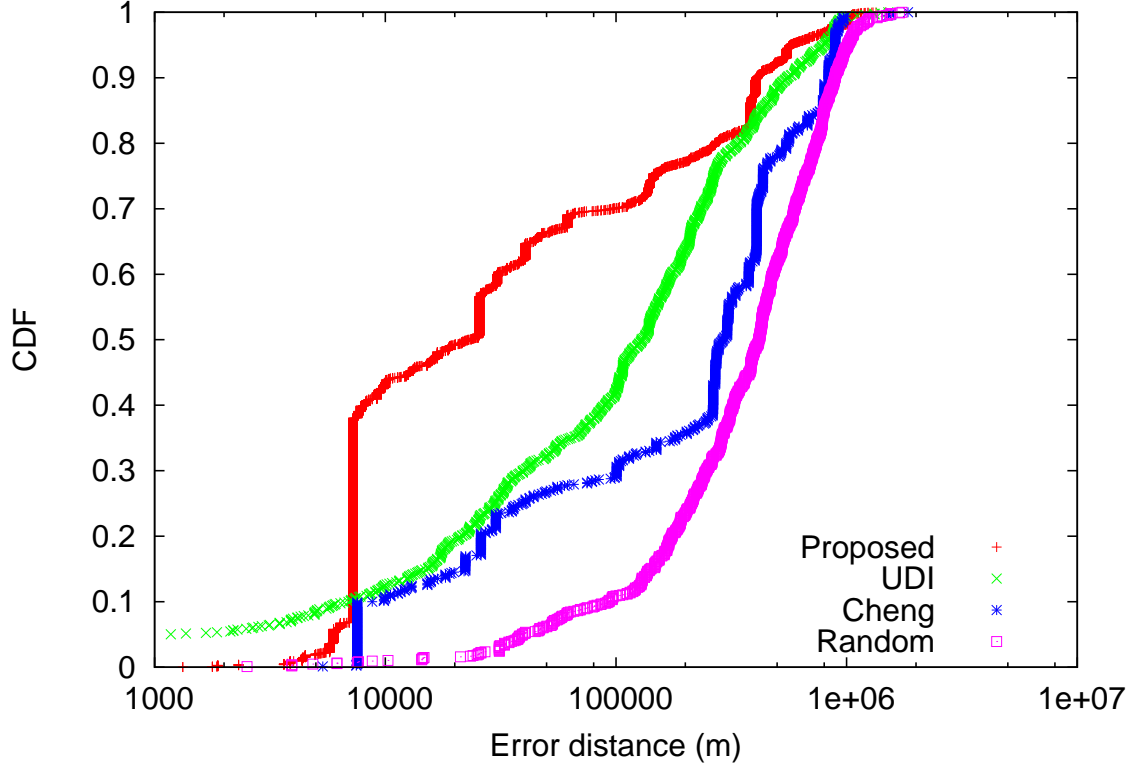


Figure 4.7: Accuracy of user location inference. Precision within 160km error distance of our method indicates improvements of 34% and 122% compared to the methods of UDI and Cheng, respectively.

the varied *MaxDispersion* values. The other fixed parameters were *WindowSize* = 600s, *MinPts* = 15, *Eps* = 0.5.

Figure 4.8 shows the trade-off between accuracy and coverage upon varying *MaxDispersion*. Horizontal and vertical axes represent coverage and accuracy, respectively. *MaxDispersion* was set to from 50km to 300km from the left point to the right. *NSF*, which is shown at the right-most point, means that the proposed method did not conduct spatial filtering; that is, *MaxDispersion* was set to infinity.

The coverage monotonically increases as *MaxDispersion* increases because the clusters with low locality are detected as events, and these detected events are used in the location inference step. In contrast, the accuracy tends to decrease as *MaxDispersion* increases

Table 4.2: Summary of the accuracies and error distances.

	<i>Proposed</i>	<i>UDI</i>	<i>Cheng</i>	<i>Random</i>
<i>Pre@160km</i>	<b>0.761</b>	0.570	0.344	0.170
<i>Pre@80km</i>	<b>0.696</b>	0.378	0.284	0.092
<i>Mean E.D. (m)</i>	<b>134,114</b>	208,699	336,489	457,666
<i>Median E.D. (m)</i>	<b>22,862</b>	129,249	290,465	416,106

because events with low locality do not provide sufficient clues for location inference. With regard to *All* and *Weather*, the accuracy also decreases when *MaxDispersion* becomes too small, indicating that interpretable events are not detected when *MaxDispersion* for the dataset is too small. Indeed, in these settings, most of detected events with a 50km dispersion do not seem to be valid.

Although the proposed method is superior to the other methods with regard to the accuracy, the proposed method cannot infer arbitrary users who do not mention any of the detected events. However, this difference would not matter to increase the number of known user locations. The number of labeled users can be increased with more events detected. Moreover, because our method and the other compared methods use different data, integrating them should achieve more accurate inference results.

#### 4.4.6 Effect of Parameters

We compared the results by varying *WindowSize* or *Eps*. The *Weather* dataset was used in this experiment, and the parameters were fixed to *WindowSize* = 600s, *MinPts* = 15, *Eps* = 0.5, and *MaxDispersion* = 150km when they were not changed explicitly.

**WindowSize.** Figure 4.9 compares the accuracy and coverage as *WindowSize* is varied from 60s to 3600s. The horizontal axis shows *WindowSize* values, while the vertical axis shows the accuracy and the coverage.

As *WindowSize* decreases, the accuracy decreases, but the coverage increases. A small time window may split events unsuitably, which leads to a decreased accuracy because the split events are less interpretable. The increase in coverage can be explained by the fact that



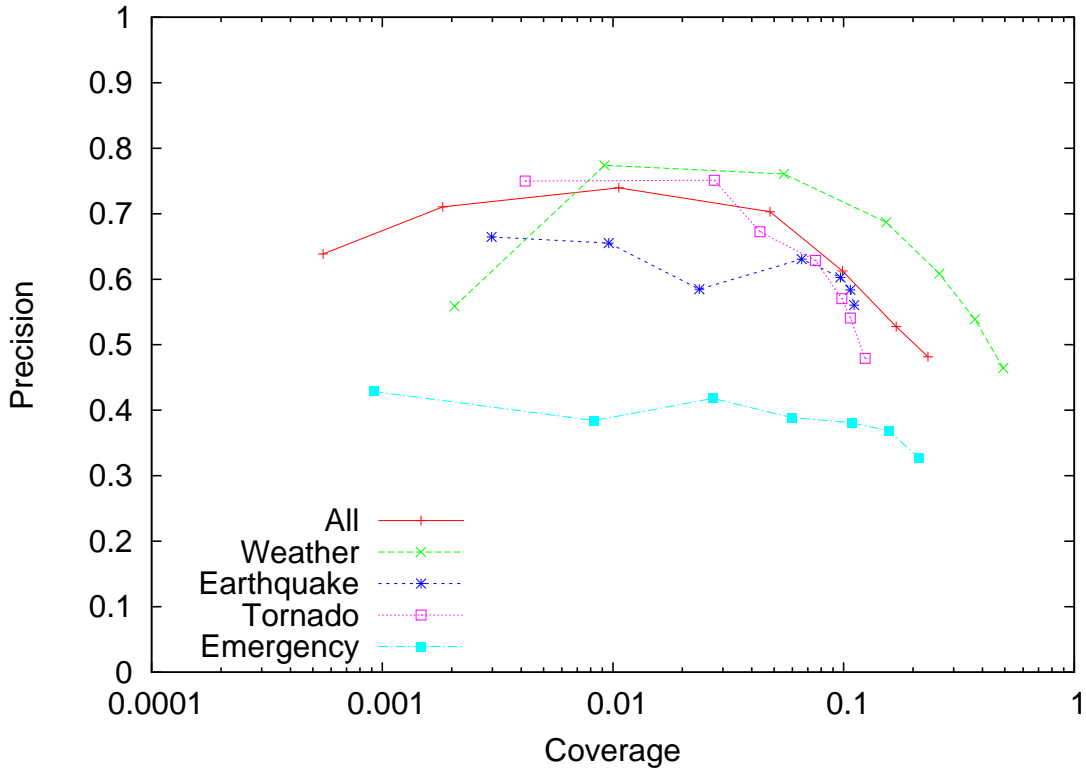


Figure 4.8: Efficiency and accuracy as *MaxDispersion* varies. Results indicate a trade-off between coverage and accuracy.

more small events were detected with the small window. On the other hand, the coverage decreases as *WindowSize* increases. Merging events in large windows cause the events to be less localized, and consequently, these events are removed by spatial filtering.

**Eps.** Figure 4.10 compares the accuracy and coverage as *Eps* is varied from 0.05 to 1. The horizontal axis shows the values of *Eps*, while the vertical axis shows the accuracy and the coverage.

Although the accuracy remains stable, the coverage decreases at both extremes. These observations indicate that with a small *Eps*, most of tweets are regarded as noises by DBSCAN, and the number of clusters becomes small, which results in a low coverage. Meanwhile, with a large *Eps*, most of the tweets are clustered into a few large clusters, and

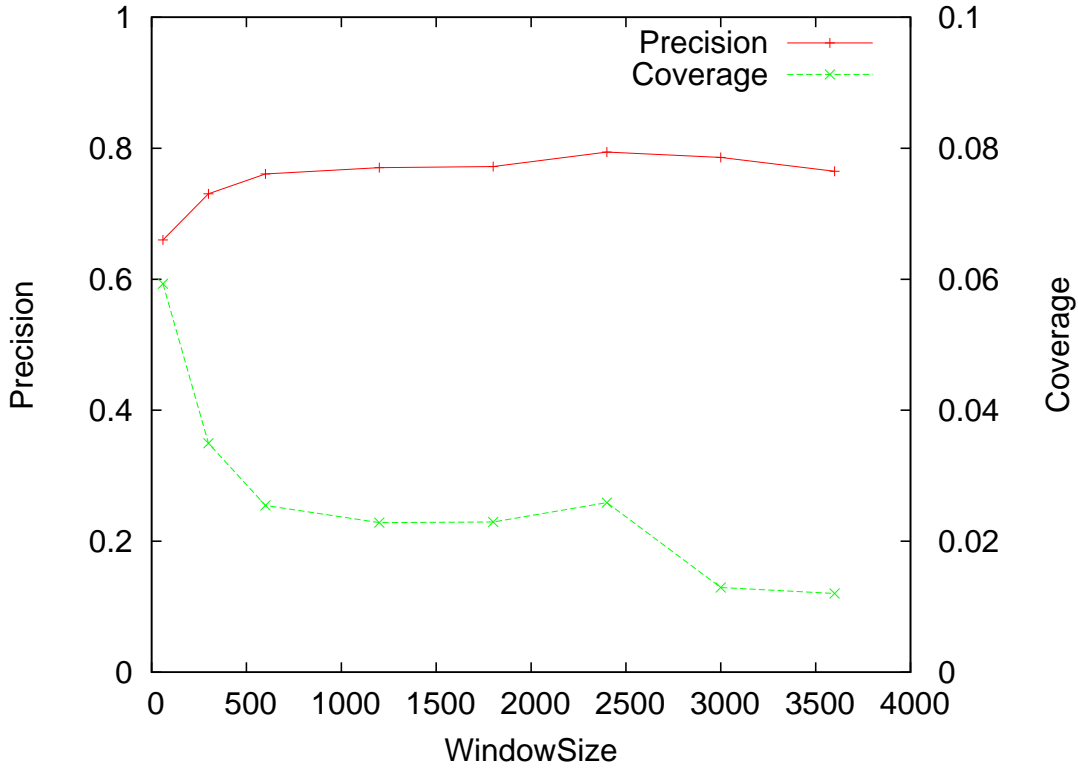


Figure 4.9: Coverage and accuracy with varied *WindowSize*. Increased *WindowSize* causes an increase in accuracy but a decrease in coverage.

are filtered out in spatial filtering due to its large dispersion, which also results in a low coverage.

## 4.5 Conclusion

In this paper, we proposed a method for user location inference based on real-world local events detected from social streams. The proposed method is based on the idea that users who post about an event with a geographical locality are likely to be close to the location of the event. Experiments, which used Twitter data, confirmed that the validity of detected local events, and showed the accuracy and the coverage of user location inference of our method.

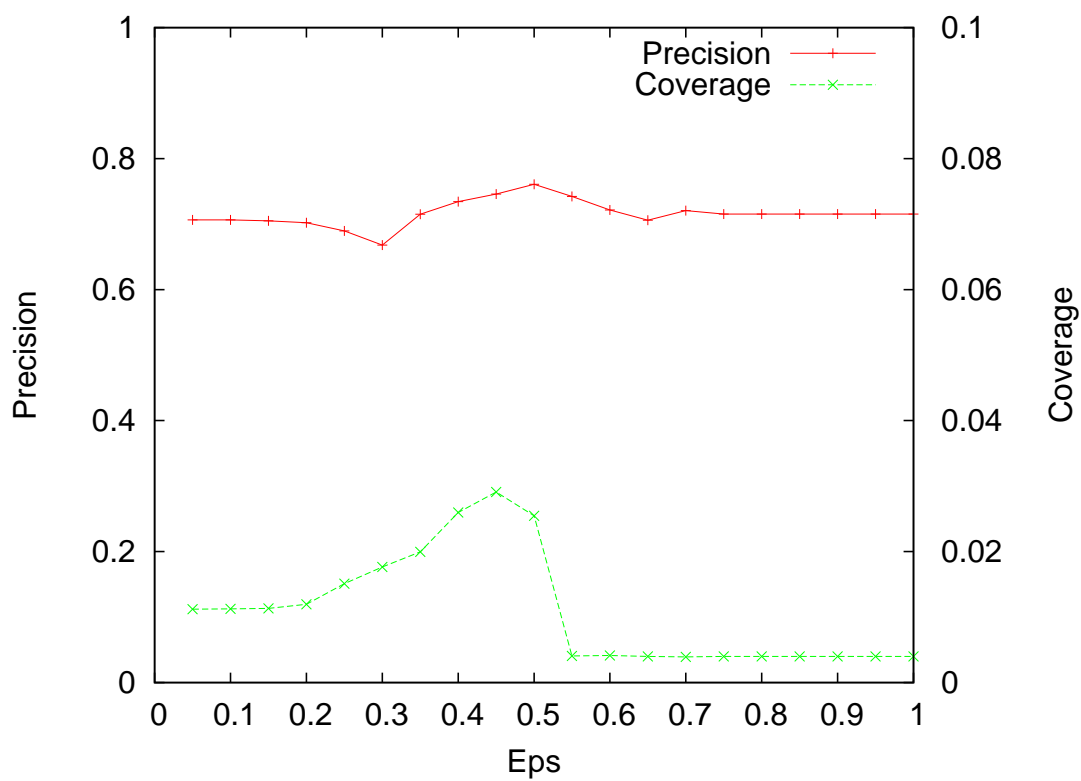


Figure 4.10: Coverage and accuracy with  $Eps$ . Although accuracy is not affected by  $Eps$ , the coverage decreases when  $Eps$  has an extremely value.

## Chapter 5

# Online User Location Inference based on Social Streams

Many works have proposed batch inference methods based on static user-generated contents. However, social media contents are generated in real-time, which creates data called social streams. Hence, we propose an online location inference method (OLIM) over social streams, which can update inference results using newly arrived user-generated contents. Our key idea is to exploit temporal features (temporally-local words; TL-words) in the contents, in addition to non-temporal features (statically-local words; SL-words) which are also used in other works. Specifically, the former features allow us to exploit the spatio-temporal correlation between users and locations in addition to the spatial correlation characterized by the latter. We propose a probabilistic user location model based on the distribution of SL- and TL-words and an online location updating scheme based on the model. The experimental results using a Twitter dataset show that our method has at least a 25% improvement in accuracy compared to existing methods, and can appropriately infer and update home locations, which reduces the inference errors over time.

## 5.1 Introduction

Recently, the location inference problem has been well studied. Most methods are based on static resources (i.e., social graphs or user-generated content). However, social media contents are generated in real-time in the form of content streams, called *social streams*. Such dynamic characteristics pose a new technical challenge (i.e., online location inference), where inference results are updated based on newly arrived user-generated content. For example, suppose that a user has generated only a small amount of contents at a point in time, which may lead to wrong inference because there are not enough clues to infer the home location of this user. However, if this user generates more contents after that, it is possible to more accurately infer the home location based on newly generated contents.

Updating inference results based on social streams has two main challenges: computational cost and storage cost. More precisely, prior works propose batch inference methods that repeat the entire inference process using all the old content as well as the newly arrived content when updating user locations. Obviously, these methods are not feasible to the scale of information in the social media era. Hence, herein we propose an online location inference method (OLIM) over social streams that can update inference results using only newly arrived user-generated content, which means that the proposed method does not require previous contents when updating the inference results. Consequently, both computational and storage costs are reduced.

Our online algorithm is based on *sequential rationality* of Bayesian inference, where the inference result of batch inference using whole observation and the result of online inference using an observation one by one are identical. Specifically, OLIM adopts distribution of labeled users' home locations as a prior, and updates it every time a new clue is obtained.

Moreover, OLIM exploits both non-temporal features (statically-local words; SL-words), which are also used in existing content-based methods, and temporal features (temporally-local words; TL-words). An SL-word is strongly correlated with a specific location regardless of time. Previous content-based methods use SL-words to exploit spatial correlations be-

tween users and locations. For example, the word “rockets” is an SL-word correlated with Houston, Texas [28] because the word tends to occur in the content generated by users in Houston. If the word “rockets” is used, previous methods can infer the user is likely to live close to Houston. In contrast, a TL-word is correlated with a certain location for a specific time period. For example, “tornado” can be regarded as a TL-word because when a tornado hits a certain location, users who live in the affected area are likely to generate contents containing the word. Hence, our proposed method can also exploit spatio-temporal correlations between users and locations via TL-words, improving the inference accuracy.

The proposed method is based on a probabilistic user location model over a geographical space, which denotes the probability that a user lives in a certain location. This model is continuously updated utilizing SL- and TL-words contained in the newly arrived contents.

Contributions of this chapter are as follows:

- We introduce a new concept of local words, *temporally-local words*, which are strongly correlated with a certain location in a specific time period.
- We propose an online location inference method (OLIM) that continuously updates the inference results based on statically-local and temporally-local words.
- Using a Twitter dataset, we conducted experiments, including a comparison to several existing methods including the state-of-the-art. Contributions of this chapter are as follows:

The experimental results showed that 1) OLIM successfully located about 51% of users within a 50km error distance, which was about a 25% improvement over existing methods, 2) the results demonstrated that OLIM can appropriately update the inference results, reducing the error distances over time, and 3) OLIM substantially reduced computational cost of existing methods in the setting of online inference.

The rest of this paper is organized as follows. Section 5.2 states the problem addressed in this paper and defines the terminology. The proposed method described in Section 5.3

is experimentally evaluated in Section 5.4. Finally, Section 5.5 concludes this chapter.

## 5.2 Problem Statement

This section defines the terminology and describes the problem addressed in this chapter.

**Social stream and time period.** Each *post* is defined via three elements; timestamp  $s$ , text  $t$ , and user  $u$ , and is denoted as  $p = (s, t, u)$ . We assume that posts are continuously received from *social stream*  $SS$  in chronological order. We define  $T_{(k)}$  as a set of posts in which timestamp  $s$  of each post  $p \in T_{(k)}$  belongs to the  $k$ -th predefined time period. For example, if the time period length is set to one hour increments beginning at 6 AM,  $T_{(1)}$  is the set of posts posted from 6 AM to 7 AM,  $T_{(2)}$  is the set of posts posted from 7 AM to 8 AM, etc.

Using these notations, our problem can be stated as:

**Problem 5 (Online Location Inference)** *Given a set of posts  $T_{(k)}$  within the  $k$ -th time period, for each unlabeled user  $u \in U^N$ , update the  $(k-1)$ th inferred home location  $\hat{\mathbf{l}}_{\mathbf{u}}^{(k-1)}$  to  $\hat{\mathbf{l}}_{\mathbf{u}}^{(k)}$  based on  $T_{(k)}$  so that the distance between  $\hat{\mathbf{l}}_{\mathbf{u}}^{(k)}$  and the true location  $\mathbf{l}_{\mathbf{u}}$  is reduced compared to the distance between  $\hat{\mathbf{l}}_{\mathbf{u}}^{(k-1)}$  and  $\mathbf{l}_{\mathbf{u}}$ .*

Because an unlabeled user's true home location  $\mathbf{l}_{\mathbf{u}}$  is unavailable, it is not guaranteed that the distance between the inferred home location and the true home location will decrease in the next time period.

## 5.3 Proposed Method

In this section, we propose an online location inference method (OLIM). Section 5.3.1 shows the overview of OLIM, and the following sections describe details of the OLIM processes.

### 5.3.1 Overview

We first explain our probabilistic model and then outline the offline process and the online process. The diagram of the OLIM is shown in Figure 5.1.

**Probabilistic model.** A user’s home location is modeled as a discrete probability distribution  $P_u(r)$  over a set of *regions*  $R$ , which is determined in the geo clustering step (Section 5.3.2).  $P_u(r)$  is called the *user distribution* over the region space. Using this model, we initially solve the location inference problem in the region space as:

$$\hat{r}_u = \arg \max_r P_u(r), \quad (5.1)$$

in which the most appropriate region for user  $u$ ’s home location is chosen.  $\hat{r}_u$  is the *home region* of user  $u$ .  $\hat{r}_u$  is then converted into the coordinates of latitude and longitude  $\hat{l}_u$  using the region center, which is also determined in the geo clustering step. That is, our solution goes through the region space (i.e., discrete probability space), and then determines the latitude and longitude coordinates.

**Offline process.** In the offline process, geo clustering divides the geographical space into multiple regions. The above-mentioned user distributions are defined over these regions. Geo clustering also produces the distribution of regions  $P(r)$ , called the *regular distribution*. The regular distribution can be regarded as the prior distribution of users’ home regions.

After geo clustering, statically-local words are extracted from a pre-stored set of posts using a *divergence metric* [97, 120] rather than a *dispersion metric* [28]. The extracted statically-local words are used in the online process to infer and update a user’s home location.

**Online process.** In the online process, a set of posts  $T_{(k)}$  is continuously received in the  $k$ -th time period. Temporally-local words from  $T_{(k)}$ , which are strongly correlated with a specific location in the  $k$ -th time period, are extracted based on the divergence metric. In



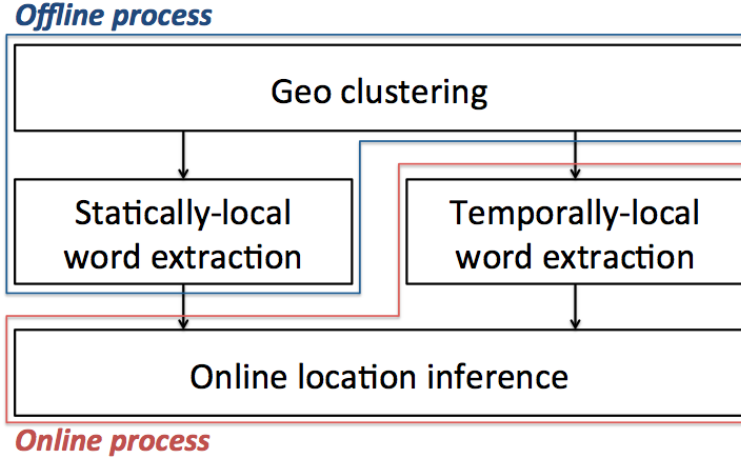


Figure 5.1: Diagram of OLIM.

each time period, user distributions are updated using both statically- and temporally-local words.

### 5.3.2 Geo Clustering

OLIM divides the geographical space into multiple regions using *GMM clustering* [121]. Specifically, home location points, which are denoted as latitude and longitude coordinates in geographical space, are clustered using GMM clustering. Then multiple Gaussian components, which are regarded as *regions* in this work, are obtained.

The distribution of users' home locations can be modeled as a mixture of Gaussian distributions because:

- Most users live in a city area, and the user density decreases as the distance from the city center increases.
- There are multiple cities.

Hence, in terms of the home location distribution, the geographical space can be divided into multiple regions by GMM clustering.

GMM over geographical space is denoted as:

$$P(l) = \sum_r P(r)P(l|r) = \sum_r \theta_r N(l|\mu_r, \sigma_r^2), \quad (5.2)$$

where  $r \in R$  is a region and  $P(r) = \theta_r$  is a regular distribution. Using home locations  $l_u$  of labeled users  $u \in U^L$ , we fit parameters of  $\theta_r$ ,  $\mu_r$ , and  $\sigma_r^2$  for all  $r$  by the *EM algorithm* [115]. The number of Gaussian components  $K$  is a predefined parameter, which is set by hand. The geographical space is divided into fine-grained and coarse regions for large and small  $K$  values, respectively. Hence the parameter  $K$  affects the complexity of the model in regard to the number of distinct locations, which is examined in Section 5.4.

**Example of a geo clustering result.** Figure 5.2 shows the result of geo clustering using Twitter users in the United States and in Japan. Each circle shows the center (i.e., mean parameter  $\mu_r$ ) of the corresponding region  $r$ . The size of the circle is proportional to the mixture weight  $\theta_r$  of the region  $r$ . Metropolises (e.g., New York and Chicago) are appropriately modeled as centers of Gaussian components in all settings (a), (b), and (c), whereas most small cities are ignored in (a) and (b), indicating that small cities have few users. Similarly, as for the result of Japan, large regions are located at metropolises (e.g., Tokyo).

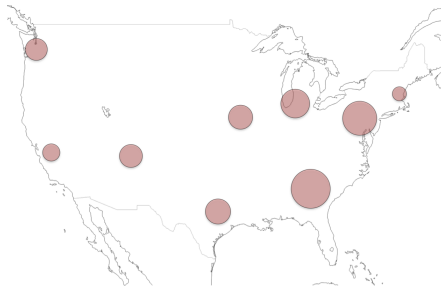
**Region location mapping.** Based on the result of geo clustering results, regions (i.e., categorical variables) and locations (i.e., coordinates of latitude and longitude) can be mutually transformed as:

$$f(l) = \arg \max_r P(r|l) \quad (5.3)$$

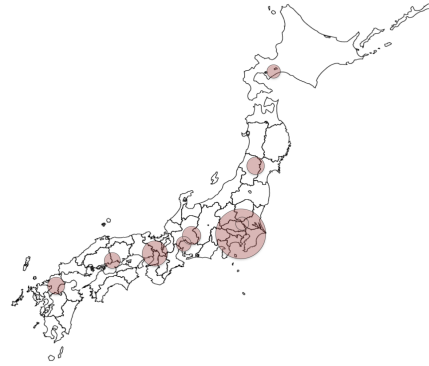
$$= \arg \max_r \theta_r N(l|\mu_r, \sigma_r^2),$$

$$g(r) = \mu_r, \quad (5.4)$$

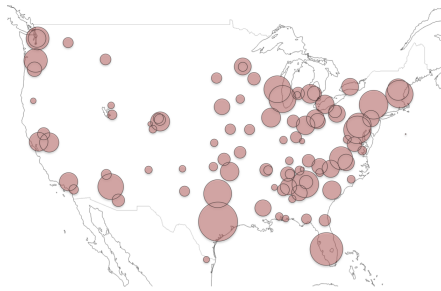
where  $f$  is a mapping from locations to regions, while  $g$  is a mapping from regions to



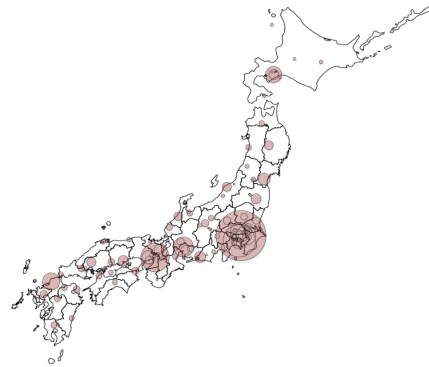
(a)  $K = 10$  (US)



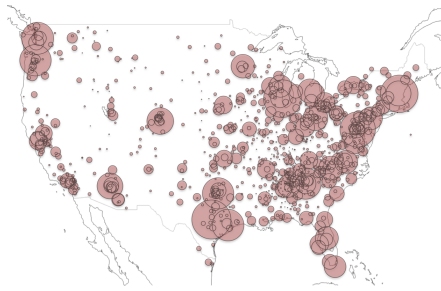
(b)  $K = 10$  (JP)



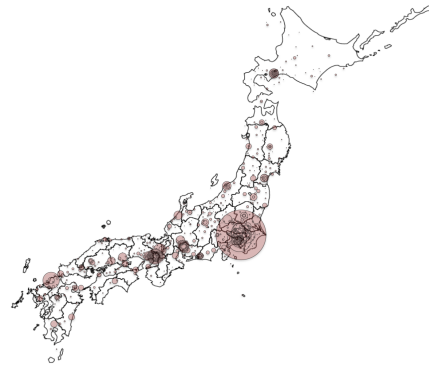
(c)  $K = 100$  (US)



(d)  $K = 100$  (JP)



(e)  $K = 1000$  (US)



(f)  $K = 1000$  (JP)

Figure 5.2: The result of geo clustering. Each circle shows the center of the corresponding region. The size of the circle is proportional to the mixture weight of the region. Large regions are located at metropolises (e.g., New York, Chicago, and Tokyo) both in the United States and in Japan.

locations. Hence, if the home region  $\hat{r}_u$  of user  $u$  can be determined, then the home location  $\hat{l}_u$ , which is the answer of the user locations inference, can be found using  $g$ .

### 5.3.3 Statically-Local Words Extraction

To extract statically-local words (SL-words), we employ a *divergence metric* rather than a *dispersion metric*. The dispersion metric assumes that local words tend to occur in one small region, and has been employed in [28]. However, this metric may miss local words that occur in multiple locations. For example, although a word “coast” can be a local word associated with locations near coasts, a dispersion metric misses this local word because this word tends to occur in more than one location. On the other hand, the divergence metric assumes that the geographical distribution of a local word differs from the regular distribution, and has been employed in [97, 120]. The divergence metric extracts “coast” as a local word because its distribution deviates from a regular one. To formulate the divergence metric, we define the *word distribution* and calculate the divergence between the word distribution and a regular distribution.

**Word occurrence matrix.** To define the word distribution, we introduce the *word occurrence matrix*, which shows how many times word  $w$  occurs in region  $r$ . Prior to SL-word extraction, we construct a set of posts  $T_{(0)}$ . SL-words are extracted from this set of posts rather than posts generated in real-time from social streams. We assume that there is a set of posts  $T_{(0)}$ , predefined vocabulary set  $V$ , and region set  $R$ , which is derived by geo clustering. For each word  $w \in V$ , a set of labeled users  $U_w$  is constructed from users who post  $p \in T_{(0)}$  which contains  $w$  as:

$$U_w = \{u \in U^L | u \text{ posts } p \in T_{(0)} \text{ which contains word } w\}. \quad (5.5)$$

Then we define the word occurrence matrix  $\mathbf{O}$  where each component  $o_{wr}$  is denoted as:

$$o_{wr} = |\{u \in U_w | f(l_u) = r\}|, \quad (5.6)$$

where  $f$  is the mapping from locations to regions.

**Word distribution.** Based on the word occurrence matrix, the word distribution  $P_w(r)$  is defined as the distribution over region set  $R$ :

$$P_w(r) = \frac{o_{wr}}{\sum_r o_{wr}}. \quad (5.7)$$

The value of  $P_w(r)$  denotes how likely word  $w$  occurs in region  $r$ .

**Divergence metric.** We assume that regular words (i.e., non-local words) are generated following the regular distribution, which means that words are more likely to occur in a region with a large  $P(r)$  value than in one with a small  $P(r)$  value. In contrast, local words do not follow a regular distribution. To evaluate this, we calculate the divergence between the word distribution and a regular distribution. KL-divergence is a popular divergence metric, which is defined as:

$$KL(P_w || P) = \sum_r P_w(r) \log \frac{P_w(r)}{P(r)}. \quad (5.8)$$

However, KL-divergence is susceptible to noises because it is defined by division; that is, if the probability value  $P(r)$  is too small, the net result of KL-divergence becomes too large. Hence, to diminish the effect of noises, we adopt the L2-distance between probability distributions for the divergence metric, which is defined as:

$$L2(P_w || P) = \sum_r (P_w(r) - P(r))^2. \quad (5.9)$$

Even if the probability value  $P(r)$  is too small, the result of the L2-distance does not become

too large because L2-distance is defined by subtraction.

If the L2-distance value  $L2(P_w||P)$  is large, word  $w$  is regarded as a local word, because the word distribution  $P_w$  differs significantly from the regular distribution  $P$ . Hence, the L2-distance and the following formula determine if word  $w$  is a local word:

$$L2(P_w||P) \geq d_{min}, \quad (5.10)$$

where  $d_{min}$  is a predefined threshold value. The KL-divergence and L2-distance are compared in Section 5.4.

**Local word extraction algorithm.** Algorithm 3 is the local word extraction algorithm. This algorithm takes tweet set  $T$ , vocabulary set  $V$ , region set  $R$ , regular distribution  $P$ , and threshold value  $d_{min}$  as inputs, and outputs a set of local words  $L$  and the occurrence matrix  $\mathbf{O}$ . This algorithm is comprised of three components. Lines 1 through 4 initialize the set of local words  $L$  and user sets  $U_w$ . Lines 5 through 9 construct the user set  $U_w$ . Lines 10 through 22 build the occurrence matrix  $\mathbf{O}$  and word distributions  $P_w(r)$ , and calculate L2-distances  $L2(P_w||P)$ .  $\mathbf{o}_w$  is a row vector of  $\mathbf{O}$  corresponding to word  $w$ . Time complexity of this algorithm is  $O(|T| \cdot |W| + |V| \cdot |R| \cdot |U_w|)$ , where  $W$  is the set of words contained in post  $p$ .

**A toy example.** Assume that the vocabulary set  $V$ , region set  $R$ , regular distribution  $P$ , and threshold value  $d_{min}$  are as follows<sup>1</sup>:

$$\begin{aligned} V &= \{\text{'RedSox'}, \text{'sports'}\} \\ R &= \{\text{'Atlanta'}, \text{'Boston'}, \text{'Chicago'}\} \\ P(A) &= 0.15, P(B) = 0.3, P(C) = 0.55 \\ d_{min} &= 0.05. \end{aligned} \quad (5.11)$$

---

<sup>1</sup>For simplicity, we denote each word and each region as its initial word.

---

**Algorithm 3** extractLWords()

---

**Input:**  $T, V, R, P, d_{min}$ **Output:** set of local words  $L$ , occurrence matrix  $O$ 

```
1:  $L \leftarrow \phi$ 
2: for all word  $w$  in  $V$  do
3:    $U_w \leftarrow \phi$ 
4: end for
5: for all post  $p_i$  in  $T$  do
6:   for all word  $w$  in text  $t_i$  do
7:      $U_w \leftarrow U_w \cup u_i$ 
8:   end for
9: end for
10: for all word  $w$  in  $V$  do
11:   for all region  $r$  in  $R$  do
12:      $o_{wr} \leftarrow 0$ 
13:   end for
14:   for all user  $u$  in  $U_w$  do
15:      $r \leftarrow f(l_u)$ 
16:      $o_{wr} \leftarrow o_{wr} + 1$ 
17:   end for
18:    $P_w \leftarrow \text{normalize}(\mathbf{o}_w)$ 
19:   if  $L2(P_w||P) \geq d_{min}$  then
20:      $L \leftarrow L \cup w$ 
21:   end if
22: end for
23: return  $L, O$ 
```

---

In addition, we have a set of posts  $T_{(0)}$ , which is stored in advance. Note that, in general, regions and real cities are not exactly the same.

First, we calculate the word occurrence matrix from  $T_{(0)}$ . This example assumes that the results of values  $o_{wr}$  are:

$$o_{RA} = 5 \quad o_{RB} = 30 \quad o_{RC} = 15 \tag{5.12}$$

$$o_{sA} = 20 \quad o_{sB} = 60 \quad o_{sC} = 120.$$

Based on this word occurrence matrix, we then calculate two word distributions as follows:

$$P_R(A) = 0.1 \quad P_R(B) = 0.6 \quad P_R(C) = 0.3 \quad (5.13)$$

$$P_s(A) = 0.1 \quad P_s(B) = 0.3 \quad P_s(C) = 0.6.$$

$P_s$  is similar to the regular distribution, while  $P_R$  differs from the regular distribution. This observation is evaluated by calculating the L2-distance:

$$L2(P_R||P) = 0.36 > d_{min} \quad (5.14)$$

$$L2(P_s||P) = 0.01 < d_{min}.$$

The L2-distance value of  $P_R$  is larger than the threshold, while that of  $P_s$  is smaller. Therefore, in this example, the word “RedSox” is extracted as a local word that is likely to occur in Boston.

### 5.3.4 Temporally-Local Words Extraction

Temporally-local words (TL-words) are extracted in the online process, which can also be performed by Algorithm 3. As previously mentioned, a set of posts  $T_{(k)}$  is continuously received in the  $k$ -th time period from social stream  $SS$ . Hence, Algorithm 3 can extract TL-words using  $T_{(k)}$  instead of the pre-stored set of posts  $T_{(0)}$  as inputs.

Although the time complexity of Algorithm 3 seems relatively large, it is negligible in the TL-words extraction step because:

- The number of posts  $|T_{(k)}|$  is generally small unless the time period is too long.
- The number of users  $|U_w|$  is small because the number of posts is small.
- It is sufficient to enumerate the words that occur in posts  $p \in T_{(k)}$  rather than all words in  $V$ .



Table 5.1: Examples of temporally-local words.

	Description
Earthquake	Earthquakes happen in multiple locations across Japan.
Tornado	Several tornadoes hit the Kanto-area in Japan from 2012 to 2013.
Fireworks	Multiple places in Japan have firework festivals in the summer.
Power outage	Power outages are rare in Japan.
Flood	In 2012, several record rainfalls occurred in Japan.
Thunder	People tend to post when there is thunder.

Moreover, Algorithm 3 can be simply parallelized in regard to  $p \in T$  and  $w \in V$ , reducing the dominant component of the time complexity of this algorithm.

Table 5.1<sup>2</sup> shows examples of extracted TL-words from the Twitter dataset that is described in Section 5.4. These words are not SL-words because they are associated with a specific location only when the corresponding phenomena happen. Due to the temporary nature of the correspondence, existing methods cannot utilize these types of local words for location inference.

### 5.3.5 Location Inference Model

OLIM continuously infers and updates user locations based on newly arrived posts that contain SL- or TL-words. Given a set of regions  $R$  derived by geo clustering, we define the *user distribution*  $P_u^{(k)}(r)$  over  $R$ , which is inferred in the  $k$ -th time period. OLIM continuously infers and updates this distribution from  $P_u^{(k-1)}$  to  $P_u^{(k)}$  based on a set of posts  $T_{(k)}$  given in the  $k$ -th time period.

Based on the user distribution  $P_u^{(k)}$ , OLIM infers  $u$ 's home region as:

$$\hat{r}_u^{(k)} = \arg \max_r P_u^{(k)}(r). \quad (5.15)$$

The inferred home region is then converted into latitude and longitude coordinates using

---

<sup>2</sup>Note that the Japanese word that means power outage is just one word.

region location mapping as:

$$\hat{l}_u^{(k)} = g(\hat{r}_u^{(k)}). \quad (5.16)$$

$l_u^{(k)}$  is the  $k$ -th answer of the online user location inference problem stated in Section 5.2.

For simplicity of notation, we denote user distribution  $P_u^{(k)}$  using the parameter vector  $\theta_u^{(k)}$  as  $P_u^{(k)}(r) = \theta_{u,r}^{(k)}$ . Thus, OLIM aims to infer this parameter vector for each time period based on a set of tweets  $T_{(k)}$ .

**MAP Estimation of parameter  $\theta_u^{(k)}$**  For each time period, OLIM creates sets of local words  $SL_u^{(k)}$  and  $TL_u^{(k)}$  for each unlabeled user as follows:

$$SL_u^{(k)} = \{w \in SL \mid p_i \in T_{(k)}, u_i = u, w \in t_i\}, \quad (5.17)$$

$$TL_u^{(k)} = \{v \in TL^{(k)} \mid p_i \in T_{(k)}, u_i = u, v \in t_i\}, \quad (5.18)$$

where  $u_i$  is the user who posts  $p_i$ , and  $t_i$  is text of  $p_i$ .  $SL$  is the set of SL-words extracted in the offline process, and  $TL^{(k)}$  is the set of TL-words extracted in the  $k$ -th time period. Hence,  $SL_u^{(k)} \subset SL$  denotes the set of SL-words that user  $u$  posts in the  $k$ -th time period, and  $TL_u^{(k)} \subset TL^{(k)}$  denotes the set of TL-words that  $u$  posts in the  $k$ -th time period. Additionally, sets of local words posted by user  $u$  from 1st to  $k$ -th time period are defined as:

$$SL_u^{(1..k)} = \{SL_u^{(1)}, \dots, SL_u^{(k)}\}, \quad (5.19)$$

$$TL_u^{(1..k)} = \{TL_u^{(1)}, \dots, TL_u^{(k)}\}. \quad (5.20)$$

Based on these local words, parameter vector  $\theta_u^{(k)}$  is derived by the MAP estimation as:

$$\theta_u^{(k)} = \arg \max_{\theta} P(\theta \mid SL_u^{(1..k)}, TL_u^{(1..k)}). \quad (5.21)$$

Here we initially show the solution of the MAP estimation, and then demonstrate OLIM

can perform the online updating of the parameter vector based solely on newly arrived posts. Assuming that local words  $w \in SL_u^{(k)}$  and  $v \in TL_u^{(k)}$  for each time period are i.i.d. (independent and identically distributed), Eq. (5.21) can be transformed as:

$$\begin{aligned}\boldsymbol{\theta}_u^{(k)} &= \arg \max_{\boldsymbol{\theta}} P(\boldsymbol{\theta}) \prod_{s=1}^k P(SL_u^{(s)}|\boldsymbol{\theta})P(TL_u^{(s)}|\boldsymbol{\theta}) \\ &= \arg \max_{\boldsymbol{\theta}} P(\boldsymbol{\theta}) \prod_{s=1}^k \prod_{w \in SL_u^{(s)}} P(w|\boldsymbol{\theta}) \prod_{v \in TL_u^{(s)}} P(v|\boldsymbol{\theta}),\end{aligned}\tag{5.22}$$

where  $P(w|\boldsymbol{\theta})$  and  $P(v|\boldsymbol{\theta})$  denotes the likelihood function of obtaining word  $w$  and  $v$  under parameter vector  $\boldsymbol{\theta}$ , respectively. The following passages define the likelihood function  $P(w|\boldsymbol{\theta})$  and  $P(v|\boldsymbol{\theta})$ , and the prior distribution  $P(\boldsymbol{\theta})$ .

**Likelihood function.** Let  $\mathbf{o}_w^{(k)} = (o_{w,1}^{(k)}, \dots, o_{w,|R|}^{(k)})$  be a row vector in occurrence matrix  $\mathbf{O}^{(k)}$  in the  $k$ -th time period corresponding to word  $w$ . Based on the occurrence vector  $\mathbf{o}_w^{(k)}$ , SL-words and TL-words can be obtained from the multinomial distribution with parameter  $\boldsymbol{\theta}$  as:

$$P(w|\boldsymbol{\theta}) = M(\mathbf{o}_w^{(k)}; \boldsymbol{\theta}) = \frac{N_w^{(k)}!}{\prod_{r \in R} o_{wr}^{(k)}!} \prod_{r \in R} \theta_r^{o_{wr}^{(k)}},\tag{5.23}$$

where  $N_w^{(k)} = \sum_{r \in R} o_{wr}^{(k)}$ . As for SL-words, which are extracted from pre-stored posts  $T_{(0)}$ , the occurrence vector  $\mathbf{o}_w^{(0)}$  is used in this equation. Note that  $\theta_r^{o_{wr}^{(k)}}$  denotes the  $o_{wr}^{(k)}$ th power of  $\theta_r$ .

**Prior distribution.** Because the conjugate prior distribution of the multinomial distribution is the Dirichlet distribution, OLIM adopts the Dirichlet distribution for the prior distribution of parameter  $\boldsymbol{\theta}$ .

$$P(\boldsymbol{\theta}) = D(\boldsymbol{\theta}; \boldsymbol{\alpha}) = \frac{\Gamma(A)}{\prod_{r \in R} \Gamma(\alpha_r)} \prod_{r \in R} \theta_r^{\alpha_r - 1},\tag{5.24}$$

where  $\Gamma(\cdot)$  denotes the gamma function,  $\boldsymbol{\alpha}$  is the parameter of the Dirichlet distribution, and  $A = \sum_r \alpha_r$ . Intuitively, the user distribution should equal to a regular distribution if local words are not observed. Hence,  $\boldsymbol{\alpha}$  is set to  $\alpha_r = \gamma \cdot P(r)$ , where  $\gamma$  is a *sensitivity parameter*, which determines the importance of the prior distribution.

Using the likelihood function and the prior distribution defined thus far, the Eqn. (5.22) can be rewritten using the conjugate property of the multinomial distribution and the Dirichlet distribution as:

$$\boldsymbol{\theta}_u^{(k)} = \arg \max_{\boldsymbol{\theta}} D(\boldsymbol{\theta}; \boldsymbol{\alpha} + \sum_{s=1}^k (\sum_{w \in SL_u^{(s)}} \boldsymbol{o}_w^{(0)} + \sum_{v \in TL_u^{(s)}} \boldsymbol{o}_v^{(s)})). \quad (5.25)$$

By solving this equation, the desired parameter vector is expressed as:

$$\theta_{u,r}^{(k)} = \frac{\sum_{s=1}^k (\sum_{w \in SL_u^{(s)}} o_{wr}^{(0)} + \sum_{v \in TL_u^{(s)}} o_{vr}^{(s)}) + \alpha_r - 1}{\sum_{s=1}^k (\sum_{w \in SL_u^{(s)}} N_w^{(0)} + \sum_{v \in TL_u^{(s)}} N_v^{(s)}) + A + |R|}. \quad (5.26)$$

Consequently, OLIM can infer the  $k$ -th home region of user  $u$  as:

$$\begin{aligned} \hat{r}_u^{(k)} &= \arg \max_r \theta_{u,r}^{(k)} \\ &= \arg \max_r (\sum_{s=1}^k (\sum_{w \in SL_u^{(s)}} o_{wr}^{(0)} + \sum_{w \in TL_u^{(s)}} o_{wr}^{(s)}) + \alpha_r). \end{aligned} \quad (5.27)$$

The denominator of Eqn. (5.26) is omitted because it is not necessary to select region  $r$  that has the largest  $\theta_{u,r}^{(k)}$  value. Therefore, the inference result in the  $k$ -th time period is based solely on the following value:

$$B_{ur}^{(k)} = \sum_{s=1}^k (\sum_{w \in SL_u^{(s)}} o_{wr}^{(0)} + \sum_{w \in TL_u^{(s)}} o_{wr}^{(s)}) + \alpha_r. \quad (5.28)$$

### 5.3.6 Sequential Inference

$B_{ur}^{(k)}$  can be written as a recurrence formula:

$$B_{ur}^{(k)} = B_{ur}^{(k-1)} + \sum_{w \in SL_u^{(k)}} o_{wr}^{(0)} + \sum_{v \in TL_u^{(k)}} o_{vr}^{(k)}. \quad (5.29)$$

This equation guarantees that OLIM can update  $B_{ur}^{(k-1)}$  to  $B_{ur}^{(k)}$  by simply adding the occurrence vectors  $\mathbf{o}_w^{(0)}$  and  $\mathbf{o}_w^{(k)}$ . Consequently, values in the  $s(< k)$ th time period are not needed to update except for  $B_{ur}^{(k-1)}$  value. The  $B_{ur}^{(k)}$  value obtained by Eq. (5.29) is exactly equal to the one obtained by directly calculating Eq. (5.21).

Online updating has two advantages compared to batch inference:

- Computational costs are reduced because user locations can be updated using only newly arrived posts  $T_{(k)}$  in the  $k$ -th time period and the previous value of  $B_u^{(k-1)}$ .
- Storage costs are reduced because the whole sets of local words  $SL_u^{(1..k-1)}$  and  $TL_u^{(1..k-1)}$  do not need to be stored in order to update the  $(k-1)$ th result to the  $k$ -th result.

In social media where users continuously generate huge amounts of content, reducing these two costs is crucial to perform the online updating.

Algorithm 4 is the online location updating algorithm. The local words  $L$  ( $TL^{(k)}$  or  $SL$ ), occurrence matrix  $\mathbf{O}$  ( $\mathbf{O}^{(k)}$  or  $\mathbf{O}^{(0)}$ ), set of posts  $T$ , and  $\mathbf{B} = \{\mathbf{B}_u | u \in U^N\}$  are inputs, and the output is the updated value of  $\mathbf{B}$ , which is obtained by Eq. (5.29). The time complexity of this algorithm is  $O(|T| \cdot |W| \cdot |R|)$ , which is low because this algorithm uses simple vector addition. This algorithm is called for each time period in Algorithm 5 which is described in the next subsection.

### 5.3.7 OLIM Algorithm

The OLIM algorithm is shown in Algorithm 5. This algorithm is defined using *extractLWords()* in Algorithm 3 and *updateULocations()* in Algorithm 4. Inputs are social stream  $SS$ , set

---

**Algorithm 4** updateULocations()

---

**Input:**  $L, O, T, B$ **Output:**  $B$ 

```
1: for all post  $p_i$  in  $T$  do
2:   for all word  $w$  in text  $t_i$  of  $p_i$  do
3:     if  $w \in L$  then
4:        $B_{u_i} \leftarrow B_{u_i} + o_w$ 
5:     end if
6:   end for
7: end for
8: return  $B$ 
```

---

of pre-stored posts  $T_{(0)}$ , set of users  $U$ , vocabulary  $V$ , and parameters  $K, d_{min}, \gamma$ , and time period length  $pl$ . In line 1, *geoClustering()* performs geo clustering (Section 5.3.2), which returns the set of regions  $R$  and the regular distribution  $P$ . The set of SL-words  $SL$  and the occurrence matrix  $O^{(0)}$  are extracted from  $T_{(0)}$  by *extractLWords()* in line 2. In lines 3 to 5, the values of  $B_u^{(0)}$  are initialized by the hyper parameter  $\alpha$  of the Dirichlet distribution.

The online process of OLIM is depicted from lines 9 to 19. This algorithm continuously receives posts from the social stream  $SS$  in chronological order. In line 12, *extractLWords()* returns the set of TL-words  $TL_{(k)}$  and the occurrence matrix  $O^{(k)}$  in this time period using  $T_{(k)}$ . Then the values of  $B_u^{(k)}$  are updated based on  $SL$  and  $TL^{(k)}$  in lines 13 and 14. This online process updates  $B_u^{(k)}$  as long as the social stream continues. Note that all the values in the  $(k-1)$ th time period including  $B^{(k-1)}$  are discarded after it is updated to  $B^{(k)}$  because they are unnecessary for the subsequent inferences.

At any given point in time, we can infer user  $u$ 's home location based on  $B_u^{(k)}$ . Note that if there are no observations (i.e., no SL-words and TL-words) for user  $u$ , the inference is based solely on the prior, which is evaluated in Section 5.4.4 varying  $\gamma$  parameter.

## 5.4 Experiments

Here, three different experiments are described. In Section 5.4.2, OLIM is compared to other methods, including the state-of-the-art method. Section 5.4.3 verifies that online inferences

---

**Algorithm 5** OLIM

---

**Input:**  $SS, T_{(0)}, U, V, K, d_{min}, \gamma, pl$

```
1:  $R, P \leftarrow geoClustering(U^L, K)$ 
2:  $SL, \mathbf{O}^{(0)} \leftarrow extractLWords(T_{(0)}, V, R, P, d_{min})$ 
3: for all unlabeled user  $u$  in  $U^N$  do
4:    $\mathbf{B}_u^{(0)} \leftarrow \gamma \cdot P$ 
5: end for
6:  $k \leftarrow 1$ 
7:  $T_{(k)} \leftarrow \phi$ 
8:  $st \leftarrow$  timestamp of the first post
9: for post  $p_i$  from social stream  $SS$  do
10:   $T_{(k)} \leftarrow T_{(k)} \cup p_i$ 
11:  if  $s_i - st > pl$  then
12:     $TL^{(k)}, \mathbf{O}^{(k)} \leftarrow extractLWords(T_{(k)}, V, R, P, d_{min})$ 
13:     $\mathbf{B}^{(k)'} \leftarrow updateULocations(SL, \mathbf{O}^{(0)}, T_{(k)}, \mathbf{B}^{(k-1)})$ 
14:     $\mathbf{B}^{(k)} \leftarrow updateULocations(TL^{(k)}, \mathbf{O}^{(k)}, T_{(k)}, \mathbf{B}^{(k)'})$ 
15:     $k \leftarrow k + 1$ 
16:     $T_{(k)} \leftarrow \phi$ 
17:     $st \leftarrow s_i$ 
18:  end if
19: end for
```

---

can reduce inference errors based on the social stream over time, and Section 5.4.4 examines the effects of divergence metrics (i.e., KL-divergence or L2-distance) and the parameters of  $\gamma$ ,  $K$ ,  $d_{min}$ , and  $pl$ . Section 5.4.1 explains the experimental setups, and Sections 5.4.2 to 5.4.4 discuss the results.

### 5.4.1 Experimental Setups

**Dataset.** We used a Twitter dataset, which contains 201,570 location-known Twitter users in Japan, the latest 200 tweets of each user, and 33,569,924 follow edges among these users. Users in this dataset were randomly collected using Twitter API. Similar to previous studies, we geocoded these users' location profile texts into coordinates of latitude and longitude using *Yahoo! geocoder*<sup>3</sup>. If the user was incorrectly geocoded or the home location was outside of Japan, the results were discarded. For the same reason as Chapter 4, we believe

---

<sup>3</sup><http://developer.yahoo.co.jp/webapi/map/openlocalplatform/v1/geocoder.html>

that the users’ location profiles show their true home locations.

Tweets and follow edges were also collected by using Twitter API. Follow edges are used by other existing methods. For the evaluation, we randomly divided users into a training set, a validation set, and a test set, where 5% were assigned to the validation set, 5% were assigned to the test set, and the rest were assigned to the training set. The validation set is used for determining parameters  $K$ ,  $\gamma$ ,  $d_{min}$ , and  $pl$  in Section 5.4.4.

**Evaluation metrics.** The results were evaluated based on the *error distance*, which is the distance between the inferred location  $\hat{l}_u$  and the true location  $l_u$ . Concretely, the following metrics were used:

- **Error distance at p% (ED@p):**  $p$  is the percentile of error distances in ascending order. Note that ED@50 is equal to the median error distance.
- **Accumulative precision at d (AP@d):** The ratio of users whose error distance is less than  $d$ .

The average error distance was not used because the error distances do not follow Gaussian distribution.

**Compared methods.** The experiments compared six different methods<sup>4</sup>:

- **OLIM:** In the proposed method, the parameters are set to  $K = 100$ ,  $\gamma = 100$ ,  $pl = 3h$ , and  $d_{min} = 0.05$ , which are examined in Section 5.4.4. The vocabulary set  $V$  is composed of all nouns.
- **UDI:** Li et al.’s method [110] is the state-of-the-art method that we are aware of. UDI utilizes both social graphs and tweets. Venue locations are obtained from OpenStreetMap<sup>5</sup>.

---

<sup>4</sup>Our implementation is available at [http://github.com/yamaguchiyuto/location\\_inference](http://github.com/yamaguchiyuto/location_inference)

<sup>5</sup><http://wiki.openstreetmap.org/wiki/JA:Nominatim>



- *Cheng*: Cheng et al.’s method [28] leverages only SL-words. Local words are extracted based on the *dispersion metric*, which differs from the *divergence metric*.
- *Hecht*: Hecht et al.’s method [98] is based on the Naive Bayes classifier [99].
- *Kinsella*: Kinsella et al.’s method [95] does not adopt the concept of local words. It uses all words as clues.
- *NaiveC*: This is a very simple content-based approach, which initially extracts the venue names from tweets using Open Street Map, and then calculates the medoid of locations of the venues.

Table 5.2 shows the resource usage for each method.

Table 5.2: Resource usage.

	tweets	graph	gazetteer	SL-words	TL-words
OLIM	✓			✓	✓
UDI [110]	✓	✓	✓		
Cheng [28]	✓			✓	
Hecht [98]	✓			✓	
Kinsella [95]	✓				
NaiveC	✓		✓		

### 5.4.2 Comparison with Existing Methods

This experiment compared above described methods. Figures 5.3 to 5.5 and Table 5.3 show the results. In the following paragraphs, results of accuracy, coverage, and computational cost are discussed.

**Accuracy.** Figures 5.3 and 5.4, and Table 5.3 show the accuracy of compared methods. The x-axis shows the error distance and the y-axis shows the **AP@d** at the corresponding error distance. OLIM (TL+SL) utilizes both TL-words and SL-words, whereas OLIM (SL) utilizes only SL-words. These results indicate that OLIM achieves the best performance in

all evaluation metrics. OLIM (TL+SL) has a better accuracy than OLIM (SL), confirming that the temporal features are important for location inference.

Comparing OLIM (SL) and other methods, it is considered that geo clustering is effective for location inference. OLIM (SL), which employs geo clustering, has a less complex model (i.e., the number of possible locations is relatively small), especially when the  $K$  value is small, indicating that the complicated location inference problem is reduced to an adequate level. The effect of  $K$  value is also examined in Section 5.4.4.

**Coverage.** The ratio of inferred users from all the test users (i.e., coverage) is approximately 100% for all compared methods because home locations can be inferred as long as the user has at least one labeled friend or post with a local word or venue name. As for OLIM, it can infer home locations of exactly all test users based on the prior even if no observation is available.

**Computational cost.** Figure 5.5 shows the computational time of compared methods. This experiment measures the computational time of inferring 19,546 users' home locations consist of test users and validation users. Preprocessing of each compared method is excluded from the result. As for OLIM, the result shows the computational time of updating the result in one time period where the length of time period is set to 3 hours. The y-axes is logarithmic.

From the result, computational time of OLIM was about 6 seconds, which is the shortest among all methods. Other methods spent more than 1,000 seconds, especially, Kinsella spent more than 100,000 seconds because Kinsella uses all words, including non-local words, contained in tweets. These results confirm that OLIM is feasible to online inference, while the other methods are not. Moreover, there is one more reason that only OLIM is feasible to online inference. The time complexity of OLIM does not depend on the accumulative number of tweets because it can infer home locations using only tweets in the current time period, indicating that the computational time of OLIM does not change, which is

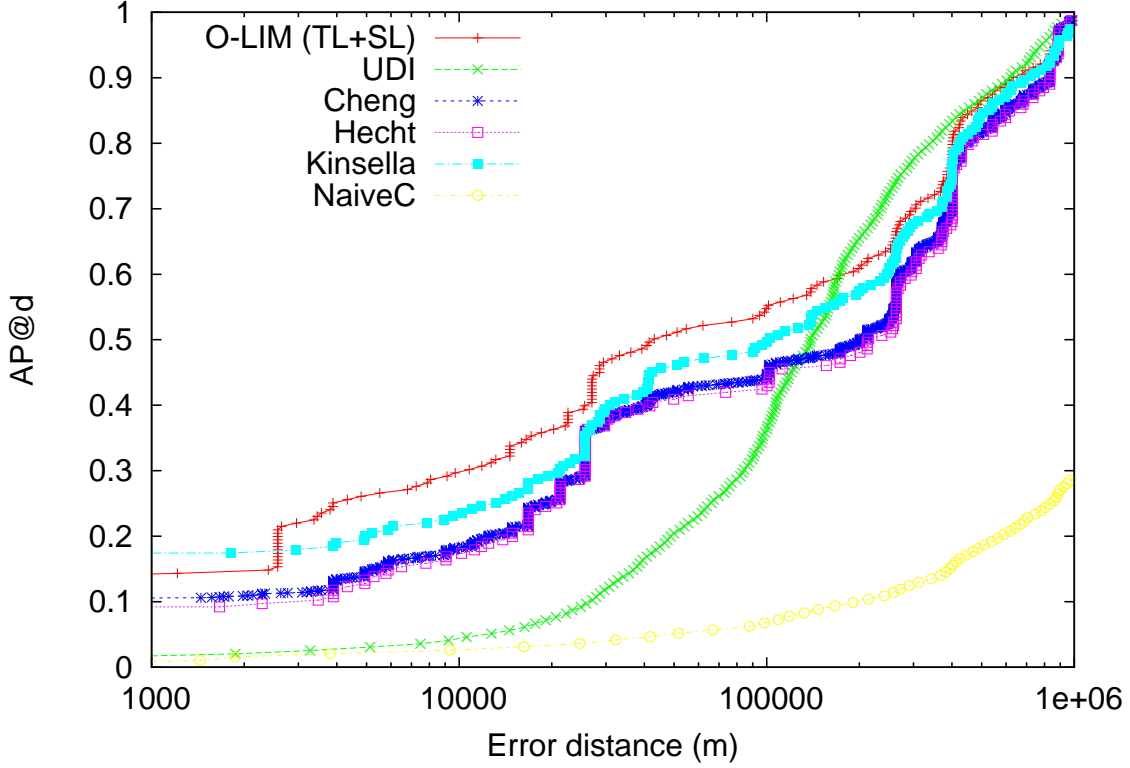


Figure 5.3: Comparison of different methods. A high value indicates a good result. The best result is from the proposed OLIM method.

6 seconds. On the other hand, the time complexity of other methods do depend on the accumulative number of tweets linearly, which means that the computational time increases with ever-increasing tweets.

#### 5.4.3 Error Reduction over Time

This experiment verified the ability of OLIM to reduce the error over time. In this experiment, all tweets in the dataset (excluding tweets posted by test users) were considered as a pseudo social stream, which provided each tweet in chronological order. Each time 5% of tweets were processed, OLIM inferred the home locations of all test users using the  $B^{(k)}$  value.

Figure 5.6 shows the results, where the x-axis denotes the percentage of tweets processed.

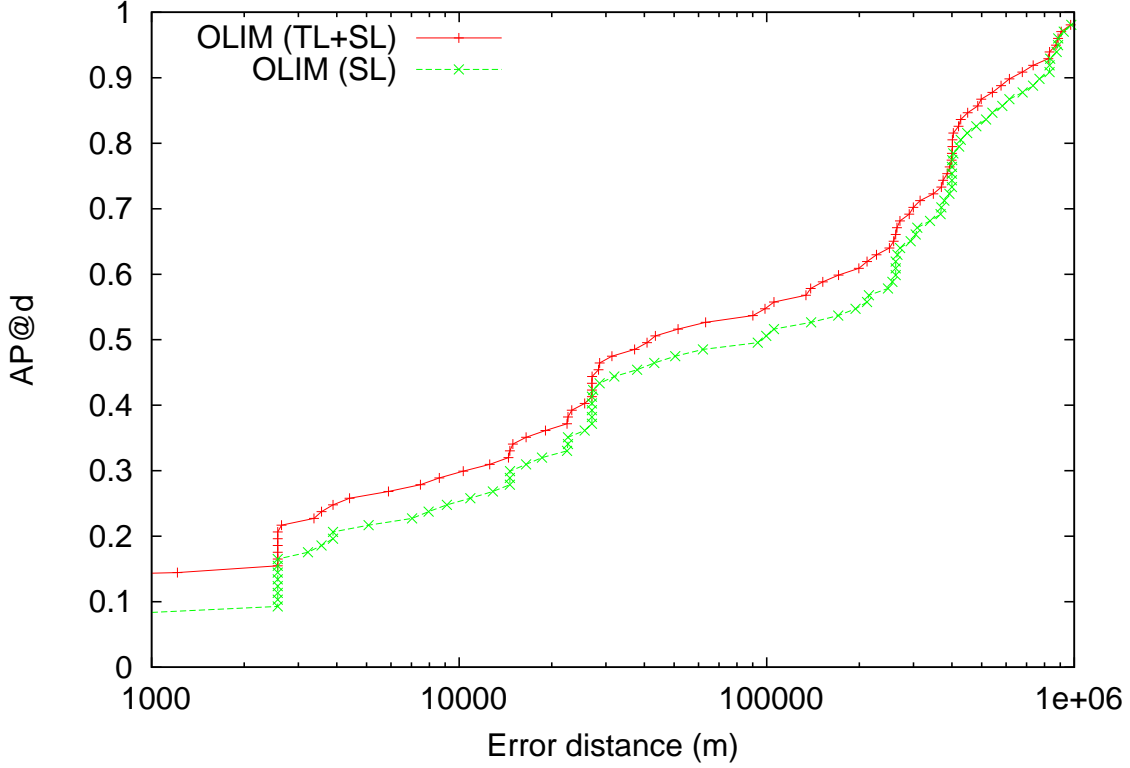


Figure 5.4: Impact of TL-words on location inference. TL-words can improve location inference.

The blue line and orange line show the results of  $ED@50$  of each method. Error bars denote the 95% confidence intervals. The red line and green line represent the accumulative number of SL-words and TL-words contained in processed tweets up to the corresponding time point in x-axis, respectively. The result shows that both versions of OLIM reduce the error distance as new tweets arrive. However, leveraging TL-words shows a greater improvement in the error distance.

The  $ED@50$  value of OLIM (SL) plateaus after about 50% of the tweets are processed, indicating that almost all test users reach consequent inference results. On the other hand, the  $ED@50$  value of OLIM (TL+SL) continues to decrease even after about 80% of the tweets are processed. These results indicate that users who are not likely to post SL-words (e.g., venue names) tend to post TL-words when some type of phenomena (e.g., tornadoes)

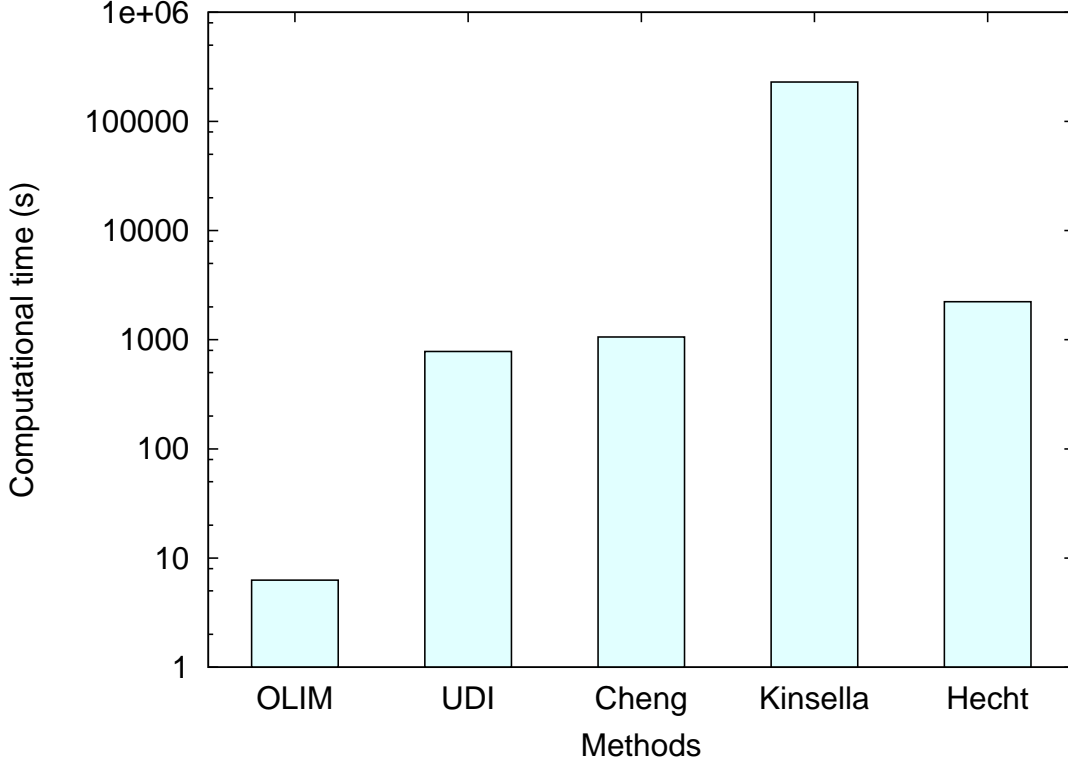


Figure 5.5: Computational time spent for five methods.

happens around them. Hence, the error distance should continue to decrease as OLIM (TL+SL) receives more tweets from the social stream.

In addition, it was shown that the number of TL-words is larger than that of SL-words, which supports the effectiveness of exploiting the new concept of TL-words. Hence, it is considered that OLIM can achieve the best performance than other methods that do not utilize TL-words.

#### 5.4.4 Effects of Parameters

This section validates parameters of OLIM by using the validation users. Figure 5.7(a) compares the effectiveness of the L2-distance and KL-divergence as a divergence metric. For the maximum performance, the parameter  $d_{min}$  should to be set to 0.05 for the L2-

Table 5.3: Comparison summary.

	ED@20	ED@50	AP@10km	AP@50km	AP@100km
OLIM (TL+SL)	<b>2,570</b>	<b>41,454</b>	<b>0.300</b>	<b>0.514</b>	<b>0.551</b>
OLIM (SL)	3,887	98,416	0.256	0.475	0.508
UDI [110]	48,651	138,862	0.043	0.205	0.362
Cheng [28]	12,483	200,851	0.183	0.420	0.441
Hecht [98]	14,894	212,333	0.174	0.409	0.427
Kinsella [95]	4,933	101,819	0.236	0.461	0.496
NaiveC	576,051	2,140,897	0.027	0.051	0.068

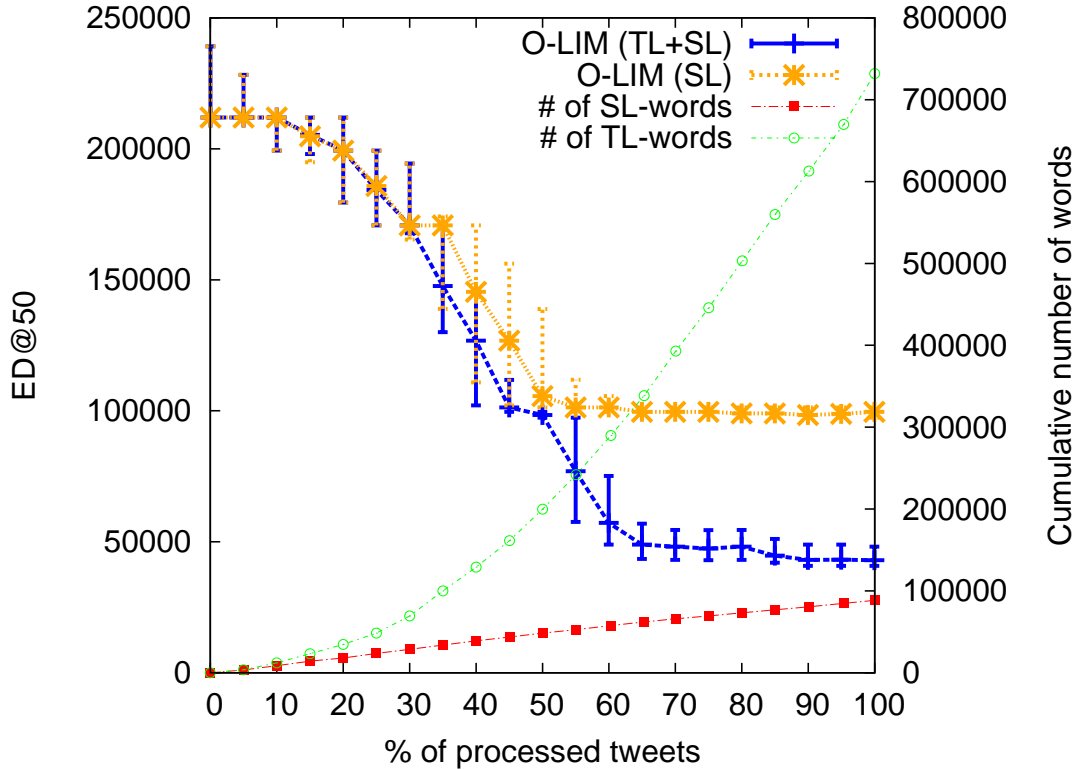


Figure 5.6: Error reduction over time. Online location inference can reduce the error distance as new tweets arrive.

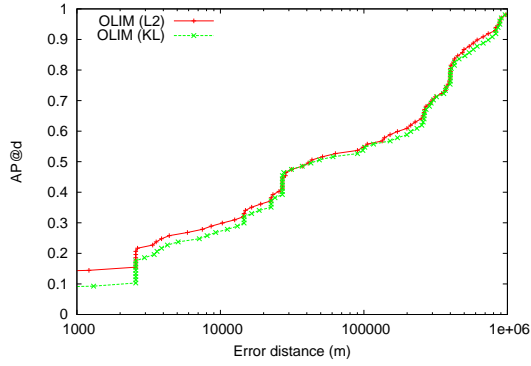
distance and 1.0 for the KL-divergence. Figures 5.7(b) and 5.7(c) show the results for various settings of the  $d_{min}$  parameters. At the appropriate setting, the L2-distance is a better divergence metric for inference accuracy than the KL-divergence. The L2-distance is suitable for thresholding because the result of L2-distance seems stable, while the result

of KL-divergence seems unstable.

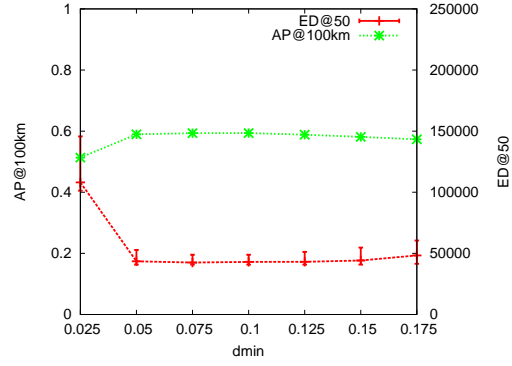
Figure 5.7(d) shows the results for various  $\gamma$  values. For a large  $\gamma$  value, the inference results are poor because OLIM places too much emphasis on prior, which means that observations are not used for location inference, whereas for a small  $\gamma$  value, OLIM relies heavily on observations and does not utilize prior, which reduces performance. Therefore, the  $\gamma$  value must be suitable for location inference, which is  $\gamma = 100$  in this experiment.

Figure 5.7(e) indicates that the number of possible regions, or the  $K$  value, significantly influences the inference result. If the  $K$  value is too small, the accuracy is poor because the geographical space is divided into too coarse regions. Similarly, if the  $K$  value is too large, the accuracy is poor because it is difficult to select the most appropriate region for each user from the large number of possible regions. Because the OLIM is sensitive to this parameter, the  $K$  value should be carefully selected for the target geographical space. The results of this experiment show that  $K = 100$  is suitable value in Japan.

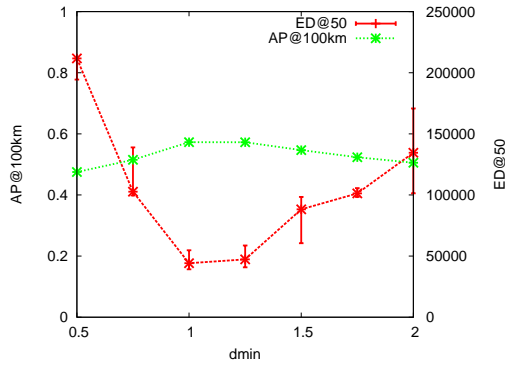
Figure 5.7(f) shows the results for various  $pl$  values. This parameter affects only the TL-words extraction step; as  $pl$  increases, TL-words are required to be local for a long time because the time period is long. On the extreme where  $pl$  is too large, TL-words are equal to SL-words because the words must always be local. For the other extreme where  $pl$  is too small, TL-words are not extracted because the time period is shorter than the duration of real-world phenomena. The results demonstrate that  $pl$  should be set between 2 to 5 hours.



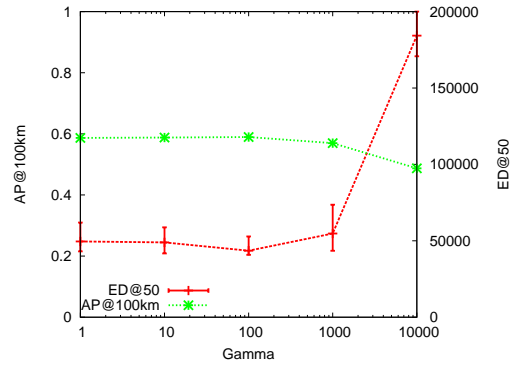
(a) L2 vs. KL



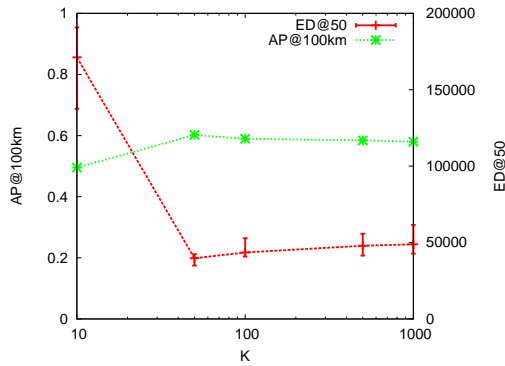
(b)  $d_{min}$  (L2)



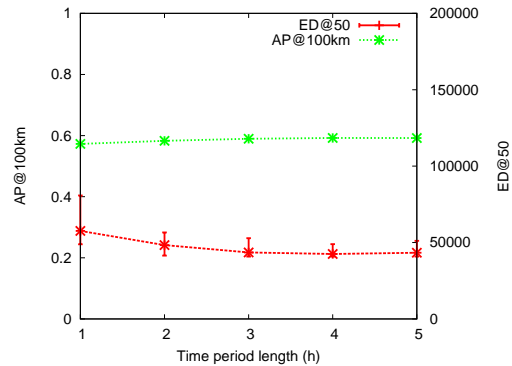
(c)  $d_{min}$  (KL)



(d)  $\gamma$



(e) K



(f) pl

Figure 5.7: Effects of parameters.



## 5.5 Conclusion

Herein we propose an online location inference method (OLIM) which can perform online inference leveraging two types of local words: statically-local words and temporally-local words. Because social media generates vast amount of contents in real-time, online inference is crucial because streams of user-generated contents provide new clues for location inference. In addition, our new concept of temporally-local words (e.g., earthquake and tornado), which are associated with a specific location for a certain time period, is promising for locations inference. The experimental results show that the proposed method outperforms existing methods, including the state-of-the-art method. Additionally, the results also demonstrate that OLIM can continuously update user locations, reducing inference errors over time.

## Chapter 6

# Conclusion and Future Work

This thesis explores the problem of user location inference in social media from three perspectives. In Chapter 3, we proposed a graph-based method which is based on the novel concentration assumption that states there are users with a lot of friends in a small area named graph landmarks. Experiments using two Twitter datasets in U.S. and in Japan showed that our method, named Landmark Mixture Model (LMM), outperformed other graph-based methods in both datasets. Compared to the existing graph-based methods, LMM does not depend on the conventional assumption that friends in social graphs tend to be close each other, and hence, expanded the applicability of graph-based location inference method to various types of social graphs.

Chapter 4 examined the effectiveness of utilizing real-world local events (e.g., earthquakes and typhoons) for the problem of inferring user home locations. Social media contents about real-world local events potentially have strong inference power for user home locations because social media users tend to post about an event when it happened near their home locations. According to the experimental results, our event-based method, named Event-based Location Inference Method (ELIM), outperformed existing methods in terms of inference accuracy.

In Chapter 5, we tackled the issue that existing inference methods do not deal with the

streams of contents that are posted in real-time in social media, proposing Online Location Inference Method (OLIM). This method can perform online inference based on streams of social media contents, which reduces the computational cost of inferring home locations in the situation of a growing number of social media contents.

Section 6.1 summarizes the contributions of this thesis by each chapter, and Section 6.2 states our future work.

## 6.1 Summary of Contributions

We summarize the contributions of this thesis as follows:

- We introduced the concept of *graph landmarks*, which are users having a lot of neighbors (e.g., friends, followers, and followees) in a small region. Graph landmarks were shown to be useful for user location inference (Chapter 3).
- We compared the characteristics of geographical features of Twitter social graphs in U.S. and in Japan by comparing the CDF plot of error distances of location inference, which demonstrated that it is relatively easy to infer user home locations in the area where the population is concentrated (Chapter 3).
- Temporal features were found to be effective to infer user home locations. Concretely, exploiting event-related tweets in Twitter achieved better results than using just static contents (Chapter 4).
- We devised an online inference algorithm, which enabled online location inference in the situation that the flood of contents are generated in real-time. In addition, this algorithm integrates static and temporal features of contents, achieving high accuracy and high coverage at the same time (Chapter 5).

We believe that this work not only explores what kind of information can be exploited for user location inference other than traditionally focused static contents and social graphs, but

also expands the possibility of analyzing social media contents in terms of location-related information to address the practical problems faced by us.

## 6.2 Future Work

We state our future work for each chapter and that of long-term view.

The future work for the work in Chapter 3 includes 1) refining our method to iteratively propagate graph landmark’s clues to increase the inference coverage and 2) examining the other applications of graph landmarks such as recommending graph landmarks’ tweets to travelers and searching local information utilizing graph landmarks.

For the work of Chapter 4, firstly, parameters need to be automatically tuned considering difference of scales of events (e.g., earthquakes and traffic accidents), and difference of frequency of events (e.g., earthquakes in Japan and France). Second, it is needed to detect the same category of events that occur in different places at the same time (i.e., events with multimodal distribution). It is considered that these types of events can be detected by clustering with regard to not only contents similarity but also geographical closeness.

Our future work of Chapter 5 is also twofold. First, we plan to improve the proposed method by integrating social graphs because propagating clues obtained by the newly arrived content in social graphs may improve the inference accuracy. Second, we plan to conduct an experiment to detect the move of users, which can potentially be detected by the online updating. For example, local words posted by a user who moved from Boston to Atlanta may change because the user becomes to be interested in Atlanta after the movement.

We present the long-term goal. In the information explosion era, we tackle the volume of data and the variety of data. For the problem of the data volume, efficient inference algorithm need to be devised to address the growing size of graph and the flood of user-generated contents. To deal with the variety of data, we plan to investigate attributes of vertices on several kind of graphs such as web graphs, biological networks, and so forth.

There has appeared several kinds of graphs where vertices have various types of attributes. For example, it is worth inferring the habitat of species by analyzing food chain networks.

# Bibliography

- [1] Andreas M. Kaplan and Michael Haenlein. Users of the world, unite! the challenges and opportunities of social media. *Business horizons*, 53(1):59–68, 2010.
- [2] *MySpace*. <http://myspace.com>.
- [3] *Facebook*. <http://facebook.com>.
- [4] *Flickr*. <http://flickr.com>.
- [5] *YouTube*. <http://youtube.com>.
- [6] *Ustream*. <http://ustream.tv>.
- [7] *Delicious*. <http://delicious.com>.
- [8] *Twitter*. <http://twitter.com>.
- [9] *Weibo*. <http://weibo.com>.
- [10] *Foursquare*. <http://foursquare.com>.
- [11] Paul Heymann, Georgia Koutrika, and Hector Garcia-Molina. Can social bookmarking improve web search? In *Proceedings of the international conference on Web search and web data mining*, pages 195–206. ACM, 2008.
- [12] Anlei Dong, Ruiqiang Zhang, Pranam Kolari, Jing Bai, Fernando Diaz, Yi Chang, Zhaohui Zheng, and Hongyuan Zha. Time is of the essence: improving recency ranking

- using twitter data. In *Proceedings of the 19th international conference on World wide web*, pages 331–340. ACM, 2010.
- [13] Ioannis Konstas, Vassilios Stathopoulos, and Joemon M. Jose. On social networks and collaborative recommendation. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 195–202. ACM, 2009.
  - [14] Stanley Milgram. The small world problem. *Psychology today*, 2(1):60–67, 1967.
  - [15] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of ‘ small-world ’ networks. *nature*, 393(6684):440–442, 1998.
  - [16] Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42. ACM, 2007.
  - [17] Jure Leskovec and Eric Horvitz. Planetary-scale views on a large instant-messaging network. In *Proceedings of the 17th international conference on World Wide Web*, pages 915–924. ACM, 2008.
  - [18] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.
  - [19] Mike Thelwall. Homophily in myspace. *Journal of the American Society for Information Science and Technology*, 60(2):219–231, 2009.
  - [20] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

- [21] Miller McPherson, Lynn Smith-Lovin, and James M. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444, 2001.
- [22] Sitaram Asur and Bernardo A. Huberman. Predicting the future with social media. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 1, pages 492–499. IEEE, 2010.
- [23] Andranik Tumasjan, Timm Oliver Sprenger, Philipp G. Sandner, and Isabell M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, 10:178–185, 2010.
- [24] Michael J. Paul and Mark Dredze. You are what you tweet: Analyzing twitter for public health. In *ICWSM*, 2011.
- [25] Sarah Vieweg, Amanda L. Hughes, Kate Starbird, and Leysia Palen. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1079–1088. ACM, 2010.
- [26] Nilesch Dalvi, Ravi Kumar, and Bo Pang. Object matching in tweets with spatial models. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 43–52. ACM, 2012.
- [27] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.
- [28] 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.



- [29] Lars Backstrom, Eric Sun, and Cameron Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on World wide web*, pages 61–70. ACM, 2010.
- [30] Jeffrey McGee, James A. Caverlee, and Zhiyuan Cheng. A geographic study of tie strength in social media. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 2333–2336. ACM, 2011.
- [31] Daniel Gruhl, Ramanathan Guha, Ravi Kumar, Jasmine Novak, and Andrew Tomkins. The predictive power of online chatter. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 78–87. ACM, 2005.
- [32] Gilad Mishne and Natalie S. Glance. Predicting movie sales from blogger sentiment. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 155–158, 2006.
- [33] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.
- [34] Wenbin Zhang and Steven Skiena. Improving movie gross prediction through news analysis. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology- Volume 01*, pages 301–304. IEEE Computer Society, 2009.
- [35] Mahesh Joshi, Dipanjan Das, Kevin Gimpel, and Noah A. Smith. Movie reviews and revenues: An experiment in text regression. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 293–296. Association for Computational Linguistics, 2010.

- [36] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, 11:122–129, 2010.
- [37] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.
- [38] Xue Zhang, Hauke Fuehres, and Peter A. Gloor. Predicting stock market indicators through twitter “ i hope it is not as bad as i fear ” . *Procedia-Social and Behavioral Sciences*, 26:55–62, 2011.
- [39] Eduardo J. Ruiz, Vagelis Hristidis, Carlos Castillo, Aristides Gionis, and Alejandro Jaimes. Correlating financial time series with micro-blogging activity. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 513–522. ACM, 2012.
- [40] Huina Mao, Scott Counts, and Johan Bollen. Predicting financial markets: Comparing survey, news, twitter and search engine data. *arXiv preprint arXiv:1112.1051*, 2011.
- [41] Michael D. Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. Political polarization on twitter. In *ICWSM*, 2011.
- [42] Michael D. Conover, Bruno Gonçalves, Jacob Ratkiewicz, Alessandro Flammini, and Filippo Menczer. Predicting the political alignment of twitter users. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 192–199. IEEE, 2011.
- [43] Nello Cristianini and John Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

- [44] Marco Pennacchiotti and Ana-Maria Popescu. Democrats, republicans and starbucks aficionados: user classification in twitter. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 430–438. ACM, 2011.
- [45] Lu Chen, Wenbo Wang, and Amit P. Sheth. Are twitter users equal in predicting elections? a study of user groups in predicting 2012 us republican presidential primaries. In *Social Informatics*, pages 379–392. Springer, 2012.
- [46] Ingmar Weber, Venkata Rama Kiran Garimella, and Asmelash Teka. Political hashtag trends. In *Advances in Information Retrieval*, pages 857–860. Springer, 2013.
- [47] Panagiotis Takis Metaxas, Eni Mustafaraj, and Daniel Gayo-Avello. How (not) to predict elections. In *Privacy, security, risk and trust (PASSAT), 2011 IEEE third international conference on and 2011 IEEE third international conference on social computing (SocialCom)*, pages 165–171. IEEE, 2011.
- [48] Daniel Gayo-Avello. "i wanted to predict elections with twitter and all i got was this lousy paper"—a balanced survey on election prediction using twitter data. *arXiv preprint arXiv:1204.6441*, 2012.
- [49] Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2008.
- [50] Aron Culotta. Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the first workshop on social media analytics*, pages 115–122. ACM, 2010.
- [51] Vasileios Lamos and Nello Cristianini. Tracking the flu pandemic by monitoring the social web. In *Cognitive Information Processing (CIP), 2010 2nd International Workshop on*, pages 411–416. IEEE, 2010.

- [52] Harshavardhan Achrekar, Avinash Gandhe, Ross Lazarus, Ssu-Hsin Yu, and Benyuan Liu. Predicting flu trends using twitter data. In *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on*, pages 702–707. IEEE, 2011.
- [53] Tetsuro Takahashi, Shuya Abe, and Nobuyuki Igata. Can twitter be an alternative of real-world sensors? In *Human-Computer Interaction. Towards Mobile and Intelligent Interaction Environments*, pages 240–249. Springer, 2011.
- [54] Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. Twitter catches the flu: Detecting influenza epidemics using twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1568–1576. Association for Computational Linguistics, 2011.
- [55] *Influ-kun*. <http://mednlp.jp/influ/>.
- [56] Vasileios Lamos and Nello Cristianini. Nowcasting events from the social web with statistical learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(4):72, 2012.
- [57] Bertrand De Longueville, Robin S. Smith, and Gianluca Luraschi. Omg, from here, i can see the flames!: a use case of mining location based social networks to acquire spatio-temporal data on forest fires. In *Proceedings of the 2009 International Workshop on Location Based Social Networks*, pages 73–80. ACM, 2009.
- [58] Andrew Crooks, Arie Croitoru, Anthony Stefanidis, and Jacek Radzikowski. # earthquake: Twitter as a distributed sensor system. *Transactions in GIS*, 2012.
- [59] Takeshi Sakaki, Fujio Toriumi, and Yutaka Matsuo. Tweet trend analysis in an emergency situation. In *Proceedings of the Special Workshop on Internet and Disasters*, page 3. ACM, 2011.
- [60] Kate Starbird, Leysia Palen, Amanda L. Hughes, and Sarah Vieweg. Chatter on the red: what hazards threat reveals about the social life of microblogged information.

- In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 241–250. ACM, 2010.
- [61] Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. Twitter under crisis: Can we trust what we rt? In *Proceedings of the first workshop on social media analytics*, pages 71–79. ACM, 2010.
  - [62] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684. ACM, 2011.
  - [63] Yan Qu, Chen Huang, Pengyi Zhang, and Jun Zhang. Microblogging after a major disaster in china: a case study of the 2010 yushu earthquake. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, pages 25–34. ACM, 2011.
  - [64] Mai Miyabe, Asako Miura, and Eiji Aramaki. Use trend analysis of twitter after the great east japan earthquake. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work Companion*, pages 175–178. ACM, 2012.
  - [65] Fujio Toriumi, Takeshi Sakaki, Kosuke Shinoda, Kazuhiro Kazama, Satoshi Kurihara, and Itsuki Noda. Information sharing on twitter during the 2011 catastrophic earthquake. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 1025–1028. International World Wide Web Conferences Steering Committee, 2013.
  - [66] Tye Rattenbury, Nathaniel Good, and Mor Naaman. Towards automatic extraction of event and place semantics from flickr tags. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 103–110. ACM, 2007.

- [67] Ling Chen and Abhishek Roy. Event detection from flickr data through wavelet-based spatial analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 523–532. ACM, 2009.
- [68] Hila Becker, Mor Naaman, and Luis Gravano. Beyond trending topics: Real-world event identification on twitter. In *ICWSM*, 2011.
- [69] Ryong Lee and Kazutoshi Sumiya. Measuring geographical regularities of crowd behaviors for twitter-based geo-social event detection. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks*, pages 1–10. ACM, 2010.
- [70] 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 20th ACM international conference on Information and knowledge management*, pages 2541–2544. ACM, 2011.
- [71] Alan Ritter, Oren Etzioni, Sam Clark, et al. Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1104–1112. ACM, 2012.
- [72] Heather S. Packer, Sina Samangooei, Jonathon S. Hare, Nicholas Gibbins, and Paul H. Lewis. Event detection using twitter and structured semantic query expansion. In *Proceedings of the 1st international workshop on Multimodal crowd sensing*, pages 7–14. ACM, 2012.
- [73] Chung-Hong Lee. Mining spatio-temporal information on microblogging streams using a density-based online clustering method. *Expert Systems with Applications*, 39(10):9623–9641, 2012.
- [74] Maximilian Walther and Michael Kaisser. Geo-spatial event detection in the twitter stream. In *Advances in Information Retrieval*, pages 356–367. Springer, 2013.

- [75] Elizabeth M. Daly, Freddy Lecue, and Veli Bicer. Westland row why so slow?: fusing social media and linked data sources for understanding real-time traffic conditions. In *Proceedings of the 2013 international conference on Intelligent user interfaces*, pages 203–212. ACM, 2013.
- [76] Sílvio S. Ribeiro Jr, Clodoveu A. Davis Jr, Diogo Rennó R. Oliveira, Wagner Meira Jr, Tatiana S. Gonçalves, and Gisele L. Pappa. Traffic observatory: a system to detect and locate traffic events and conditions using twitter. In *Proceedings of the 5th International Workshop on Location-Based Social Networks*, pages 5–11. ACM, 2012.
- [77] Axel Schulz and Petar Ristoski. The car that hit the burning house: Understanding small scale incident related information in microblogs. In *AAAI Technical Report WS-13-04*, 2013.
- [78] Makoto Okazaki and Yutaka Matsuo. Semantic twitter: analyzing tweets for real-time event notification. In *Recent Trends and Developments in Social Software*, pages 63–74. Springer, 2011.
- [79] Bella Robinson, Robert Power, and Mark Cameron. A sensitive twitter earthquake detector. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 999–1002. International World Wide Web Conferences Steering Committee, 2013.
- [80] Mark A. Cameron, Robert Power, Bella Robinson, and Jie Yin. Emergency situation awareness from twitter for crisis management. In *Proceedings of the 21st international conference companion on World Wide Web*, pages 695–698. ACM, 2012.
- [81] Orkut Buyukkokten, Junghoo Cho, Hector Garcia-Molina, Luis Gravano, and Narayanan Shivakumar. Exploiting geographical location information of web pages. In *WebDB (Informal Proceedings)*, pages 91–96, 1999.

- [82] Junyan Ding, Luis Gravano, and Narayanan Shivakumar. Computing geographical scopes of web resources. In *Proceedings of the 26th International Conference on Very Large Data Bases*, pages 545–556. Morgan Kaufmann Publishers Inc., 2000.
- [83] Kevin S. McCurley. Geospatial mapping and navigation of the web. In *Proceedings of the 10th international conference on World Wide Web*, pages 221–229. ACM, 2001.
- [84] Einat Amitay, Nadav Har’El, Ron Sivan, and Aya Soffer. Web-a-where: geotagging web content. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 273–280. ACM, 2004.
- [85] Wenbo Zong, Dan Wu, Aixin Sun, Ee-Peng Lim, and Dion Hoe-Lian Goh. On assigning place names to geography related web pages. In *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, pages 354–362. ACM, 2005.
- [86] Mário J. Silva, Bruno Martins, Marcirio Chaves, Ana Paula Afonso, and Nuno Cardoso. Adding geographic scopes to web resources. *Computers, Environment and Urban Systems*, 30(4):378–399, 2006.
- [87] Luis Gravano, Vasileios Hatzivassiloglou, and Richard Lichtenstein. Categorizing web queries according to geographical locality. In *Proceedings of the twelfth international conference on Information and knowledge management*, pages 325–333. ACM, 2003.
- [88] Lars Backstrom, Jon Kleinberg, Ravi Kumar, and Jasmine Novak. Spatial variation in search engine queries. In *Proceedings of the 17th international conference on World Wide Web*, pages 357–366. ACM, 2008.
- [89] David J. Crandall, Lars Backstrom, Daniel Huttenlocher, and Jon Kleinberg. Mapping the world’s photos. In *Proceedings of the 18th international conference on World wide web*, pages 761–770. ACM, 2009.



- [90] Pavel Serdyukov, Vanessa Murdock, and Roelof Van Zwol. Placing flickr photos on a map. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 484–491. ACM, 2009.
- [91] Qiaozhu Mei, Chao Liu, Hang Su, and ChengXiang Zhai. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *Proceedings of the 15th international conference on World Wide Web*, pages 533–542. ACM, 2006.
- [92] Michael D. Lieberman and Jimmy Lin. You are where you edit: Locating wikipedia contributors through edit histories. In *ICWSM*, 2009.
- [93] 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.
- [94] Liangjie Hong, Amr Ahmed, Siva Gurumurthy, Alexander J. Smola, and Kostas Tsoutsoulis. Discovering geographical topics in the twitter stream. In *Proceedings of the 21st international conference on World Wide Web*, pages 769–778. ACM, 2012.
- [95] Sheila Kinsella, Vanessa Murdock, and Neil O’Hare. I’m eating a sandwich in glasgow: modeling locations with tweets. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pages 61–68. ACM, 2011.
- [96] Swarup Chandra, Latifur Khan, and Fahad Bin Muhaya. Estimating twitter user location using social interactions—a content based approach. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 838–843. IEEE, 2011.
- [97] 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)*, pages 111–118. IEEE Computer Society, 2012.
- [98] Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H. Chi. Tweets from justin bieber’s heart: the dynamics of the location field in user profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 237–246. ACM, 2011.
  - [99] Andrew McCallum, Kamal Nigam, et al. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Citeseer, 1998.
  - [100] Yohei Ikawa, Miki Enoki, and Michiaki Tatsubori. Location inference using microblog messages. In *Proceedings of the 21st international conference companion on World Wide Web*, pages 687–690. ACM, 2012.
  - [101] Axel Schulz, Aristotelis Hadjakos, Heiko Paulheim, Johannes Nachtwey, and Max Muhlhäuser. A multi-indicator approach for geolocalization of tweets. In *Seventh International AAI Conference on Weblogs and Social Media*, 2013.
  - [102] Satyen Abrol and Latifur Khan. Tweethood: Agglomerative clustering on fuzzy k-closest friends with variable depth for location mining. In *Social Computing (Social-Com), 2010 IEEE Second International Conference on*, pages 153–160. IEEE, 2010.
  - [103] Clodoveu A. Davis Jr, Gisele L. Pappa, Diogo Rennó Rocha de Oliveira, and Filipe de L. Arcanjo. Inferring the location of twitter messages based on user relationships. *Transactions in GIS*, 15(6):735–751, 2011.
  - [104] David Jurgens. That ’s what friends are for: Inferring location in online social media platforms based on social relationships. In *Seventh International AAI Conference on Weblogs and Social Media*, 2013.
  - [105] Olivier Chapelle, Bernhard Schölkopf, Alexander Zien, et al. *Semi-supervised learning*, volume 2. MIT press Cambridge, 2006.

- [106] Dominic Rout, Kalina Bontcheva, Daniel Preoțiuc-Pietro, and Trevor Cohn. Where's@ wally?: a classification approach to geolocating users based on their social ties. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, pages 11–20. ACM, 2013.
- [107] Mark S. Granovetter. The strength of weak ties. *American journal of sociology*, pages 1360–1380, 1973.
- [108] Jeffrey McGee, James Caverlee, and Zhiyuan Cheng. Location prediction in social media based on tie strength. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 459–468. ACM, 2013.
- [109] Adam Sadilek, Henry Kautz, and Jeffrey P. Bigham. Finding your friends and following them to where you are. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 723–732. ACM, 2012.
- [110] Rui Li, Shengjie Wang, Hongbo Deng, Rui Wang, and Kevin Chen-Chuan Chang. Towards social user profiling: unified and discriminative influence model for inferring home locations. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1023–1031. ACM, 2012.
- [111] Rui Li, Shengjie Wang, and Kevin Chen-Chuan Chang. Multiple location profiling for users and relationships from social network and content. *Proceedings of the VLDB Endowment*, 5(11):1603–1614, 2012.
- [112] Leo Breiman. *Classification and regression trees*. CRC press, 1993.
- [113] Sergej Sizov. Geofolk: latent spatial semantics in web 2.0 social media. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 281–290. ACM, 2010.

- [114] Zhijun Yin, Liangliang Cao, Jiawei Han, Chengxiang Zhai, and Thomas Huang. Geographical topic discovery and comparison. In *Proceedings of the 20th international conference on World wide web*, pages 247–256. ACM, 2011.
- [115] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.
- [116] Réka Albert, Hawoong Jeong, and Albert-László Barabási. Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382, 2000.
- [117] Anqi Cui, Min Zhang, Yiqun Liu, Shaoping Ma, and Kuo Zhang. Discover breaking events with popular hashtags in twitter. In *CIKM*, pages 1794–1798, 2012.
- [118] Chenliang Li, Aixin Sun, and Anwitaman Datta. Twevent: Segment-based event detection from tweets. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 155–164. ACM, 2012.
- [119] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, volume 96, pages 226–231, 1996.
- [120] Bo Han, Paul Cook, and Timothy Baldwin. Geolocation prediction in social media data by finding location indicative words. In *COLING*, pages 1045–1062, 2012.
- [121] Christopher M. Bishop and Nasser M. Nasrabadi. *Pattern recognition and machine learning*, volume 1. springer New York, 2006.

# Reference Papers

## Journal Paper

- Yuto Yamaguchi, Yohei Ikawa, Toshiyuki Amagasa, and Hiroyuki Kitagawa, “ User Location Inference using Local Events in Social Media, ” Information Processing Society of Japan Transactions of Databases, Vol. 6, No. 5 (TOD 60), pp.23-37, 2013 (in Japanese with English Abstract).

## Conference Paper

- Yuto Yamaguchi, Toshiyuki Amagasa, and Hiroyuki Kitagawa, ”Landmark-Based User Location Inference in Social Media,” The 1st ACM Conference on Online Social Networks (COSN 2013), pp.223-234, Boston, USA, October 7-8, 2013.

# Other Papers

## Journal Papers

- Yuto Yamaguchi, Toshiyuki Amagasa, and Hiroyuki Kitagawa, "Tagging Users based on Twitter Lists," International Journal of Web Engineering and Technology, Vol. 7, No. 3, pp.273-298.
- Yuto Yamaguchi, Tsubasa Takahashi, Toshiyuki Amagasa, and Hiroyuki Kitagawa, "Ranking Twitter Users Based on Information Propagation Graph Analysis," Information Processing Society of Japan Transactions of Databases, Vol. 4, No. 2 (TOD 50), pp.142-157, 2011 (in Japanese with English Abstract).

## Conference Papers

- Yuta Sakakura, Yuto Yamaguchi, Toshiyuki Amagasa, and Hiroyuki Kitagawa, "A Local Method for ObjectRank Estimation," The 15th International Conference on Information Integration and Web-based Applications & Services (iiWAS 2013), pp.92-101, Vienna, Austria, December 2-4, 2013.
- Yuto Yamaguchi, Toshiyuki Amagasa, and Hiroyuki Kitagawa, "Recommending Fresh URLs using Twitter Lists," The 7th International AAAI Conference on Weblogs and Social Media (ICWSM 2013), pp.733-736, Boston, USA, July 8-10, 2013.
- Yuto Yamaguchi, Toshiyuki Amagasa, and Hiroyuki Kitagawa, "Tag-based User Topic

Discovery using Twitter Lists,” The International Conference on Advances in Social Network Analysis and Mining (ASONAM 2011), pp. 13-20, Kaohsiung City, Taiwan, July 25-27, 2011.

- Takahiro Komamizu, Yuto Yamaguchi, Toshiyuki Amagasa, and Hiroyuki Kitagawa, ”FACTUS: Faceted Twitter User Search using Twitter Lists,” The 12th International Conference on Web Information System Engineering (WISE 2011), pp. 343-344, Sydney, Australia, October 13-14, 2011.
- Yuto Yamaguchi, Tsubasa Takahashi, Toshiyuki Amagasa, and Hiroyuki Kitagawa, ”TURank: Twitter User Ranking Based on User-Tweet Graph Analysis,” The 11th International Conference on Web Information Systems Engineering (WISE 2010), pp. 240-253, Hong Kong, China, December 12-14, 2010.

## **Nonreferred Domestic Conference Papers**

- Yuto Yamaguchi, Toshiyuki Amagasa, and Hiroyuki Kitagawa, ”A User Tagging Method using Twitter Lists,” The 3rd Forum on Data Engineering and Information Management (DEIM 2011), A1-1, Izu, Japan, February 27 - March 1, 2011 (in Japanese).
- Yuto Yamaguchi, Tsubasa Takahashi, Toshiyuki Amagasa, and Hiroyuki Kitagawa, ”Twitter User Ranking by Link Structure Analysis,” The 72nd National Convention of Information Processing Society of Japan, ”1-751”-”1-752”, Tokyo, Japan, March 8, 2011 (in Japanese).