IPSJ SIG Technical Report

Vol.2009-DBS-149 No.17
2009/11/21

マイクロブロガーの移動履歴を用いた地域特性分析

藤坂達也[†]　李龍[†]　角谷和俊[†]

スマートフォンのようなモバイル端末の普及や Twitter に代表されるマイクロブログの流通により，アウトドア環境から日々の生活や社会的なイベントなどを短いメッセージとして伝える人々が爆発的に増加している．通常のブログと比べ，マイクロブログは場所や時間の制約に関わらず容易に情報を即時に発信できるため，ユーザらは様々な場所で頻繁にメッセージを発信している．これより，各地域から発信されたマイクロブログを解析することで，ユーザらの移動パターンを分析することが可能である．そこで，我々はモバイル端末から多くのマイクロブロガーらの移動履歴を分析することで地域空間における特徴的な傾向を発見することを目的としている．本稿では，ユーザらの特徴的な移動パターンを分析するために集中型・分散型の 2 つのモデルを提案し，Twitter から取得した実データを用いた実験を行うことにより地域ごとの特性の発見を行う．

Exploring Regional Characteristics Using the Movement History of Mass Mobile Microbloggers

Tatsuya Fujisaka[†]　Ryong Lee[†]　Kazutoshi Sumiya[†]

The explosive growth in the use of smartphones and social communities such as Twitter has enabled users to send numerous short messages detailing their daily lives and social events from outdoor places. Compared to conventional blogging, micro-blogging systems enable users to easily write and share their daily logs without any spatial or temporal restrictions. Such mass geo-tagged and time-stamped micro-blogs can inform us about social patterns, regardless of their scale, time, or significance. In this paper, we endeavor to discover characteristic patterns in urban areas from the movement histories of mass mobile micro-bloggers. In particular, some interesting movement patterns were frequently observed in urban areas using our two measures; aggregation and dispersion. We also present an experimental result to discover urban characteristics from real micro-blog data from Twitter.

1. Introduction

Recently, micro-blogging using mobile devices has attracted a great deal of interest; regardless of location, users can write† posts for micro-blogging sites such as Twitter [6]. Such outdoor posts naturally involve location information as well as a time-stamp. Of course, a key feature of such micro-blogging sites is that they generally permit short messages on personal updates, frequently providing little information that is of value to the public. However, the information is useful if we regard it as a set of mass movement histories. In fact, there has been explosive global growth in the numbers of mobile microbloggers willing to publish their life logs on micro-blogging sites without any privacy concerns. Some bloggers just would like to let their acquaintances know their movements or current location; other bloggers report on special events occurring around them, such as earthquakes or wildfires. It is clear that the availability of data on such a massive scale offers a framework for discovering interesting and useful patterns about social or natural incidents.

In this paper, we endeavor to discover characteristic patterns for urban behavior using micro-blog data gathered at a country level. In particular, we will focus on the movement histories of mass microbloggers. At the personal level, we can analyze lifestyle patterns in terms of geographic scope; from living or working areas to favorite restaurants or shops. Privacy is an important consideration, so we will only consider bloggers willing to share their life logs. The focus of this paper is to consider data gathered on a large scale; a data set of geo-tagged and time-stamped micro-blogs for a large urban area. This data can provide important information on the area; for instance, the amount of people who congregate there during the day can be identified from an analysis of how many microbloggers move in and out of the region. Furthermore, from mass movement flows we also can examine the bedroom suburbs that citizens commute from.

Obviously, a useful database for such a demographic study requires a sufficiently large population. A survey by Sysomos Inc. in June 2009 [1], showed that Twitter has experienced extraordinary growth over the past two years; it now has 11.5 million users, and would be the number one micro-blogging tool. The report also illustrated—through the explosive growth in users—that the most populous cities are New York, Los Angeles, Toronto, San Francisco and Boston. Actually, our platform found tweet messages from most countries, even for North Korea. Other micro-blogging sites such as Plurk, Jaiku, Prownce, and itenti.ca exist. There is a clear global trend: people are willing to post status updates, somewhat fearless of privacy and security issues, and are willing to disclose personal information such as their id, name,

[†] 兵庫県立大学環境人間学部
School of Human Science and Environment, University of Hyogo

age, and location. Such openness facilitates the analysis and use of data on such a massive scale for social networking and for demographic analyses. In order to achieve this goal, we first present a framework for gathering updates from micro-blogging sites that only support restricted data collection APIs. Moreover, we propose two analysis models to glean the characteristics of urban areas from the data gathered. We perform experiments using the two models and based on the micro-blog data.

In this paper, Section 2 describes our initial motivation, and provides brief details of the research model. Section 3 presents full details of the model together with the framework for data gathering. Section 4 illustrates the experiment in order to examine movement histories and to establish the urban characteristics. Section 5 concludes this paper with suggestions for further work.

2. Exploring City Updates using the Micro-blog Histories of Mass Mobile Users

2.1 Micro-blog as a Geo-spatial Media

Geo-tagged micro-blogs can be written on most smart phones, including iPhones and Android-based cellular phones, equipped with GPS or cell station-based positioning functions. This small advancement, however, encourages users to publish status updates without intermitting location-based reporting due to manual position tagging. The effect of such a change showed its power in major social incidents such as the Mumbai terror attacks and the recent post-election protests in Iran.

Mumbai Blasts:Taj Hotel is a block from
my house! Hostages still inside;still burning;
smoke is pouring from windows; pics later

Figure 1: Pictures and messages from Mumbai reported through Twitter

The size restriction on the messages ensures that they can be written and uploaded quickly and in real-time, reaching a global audience in mere seconds. Using such a platform, it is easy to imagine a sudden burst of tweets from an unknown town where micro bloggers voluntarily broadcast up-to-date news on an unexpected accident as it happens. Their messages are likely to be relayed fast, to be unregulated, and more vivid than those offered by other media channels. In the Mumbai terror attacks, as shown in Figure 1, a local citizen provided a vivid report on the terror unfolding, using real-time messages and pictures posted on Twitter. The CNN news channel used the information on Twitter to break the news. In addition, people outside Mumbai used the platform to establish contact with acquaintances there. We can speculate that micro-blog updates appear in great volume throughout the duration of a crucial incident. Likewise, people have a wonderful tool for sharing their experiences. Such geo-spatial media is clearly a valuable information source, but it is often too raw to mine and establish any useful patterns.

2.2 Research Model

We present our research model for gathering micro-blog data and for establishing the patterns from mass movement histories with the following steps; 1) monitoring updates from micro-blogging sites in relation to geography; 2) clustering of posts gathered as a preprocessing step to measure movement patterns; and 3) extracting characteristic patterns from movement histories.

First, the development of a micro-blog observation system capable of collecting lots of geo-tagged posts from the region of interest was required. Using an open API supported by Twitter, we can request tweets for a region specified by a center and a search radius over 1-km. For each request, a maximum of 1,500 tweets from the previous week can be gathered. This is the only option in terms of geographic specification that is available from Twitter. Actually, the strict form of the query makes it inappropriate for surveying a large region in high resolution. In order to overcome this constraint, we propose a quad-tree based data gathering framework in Section 3. This framework is able to gather more high-resolution data, depending on the density of existing data, while ensuring the data is up-to-date.

The experimental data collected using our framework will be clustered to identify regions requiring detailed examination. For this, we simply applied the K-mean algorithm for the coordinate based on two-dimensional planar space. The clusters generated from the data over a one-week period will be used for further detailed examination with various granularities.

Fundamentally, each cluster has its own properties representing how many tweets or users there are. However, this primitive statistic cannot reveal the temporal status of each region. In order to examine the time each user remained in the region, we present two other statistics; an aggregation model will count how many new users enter the region, and a dispersion model will count those leaving the region. Based on these two simple measures, and on the data gathered for an area in Japan, we can discover some characteristics; illustrate high passing regions and analyze time periods when a region is more/less heavily populated.

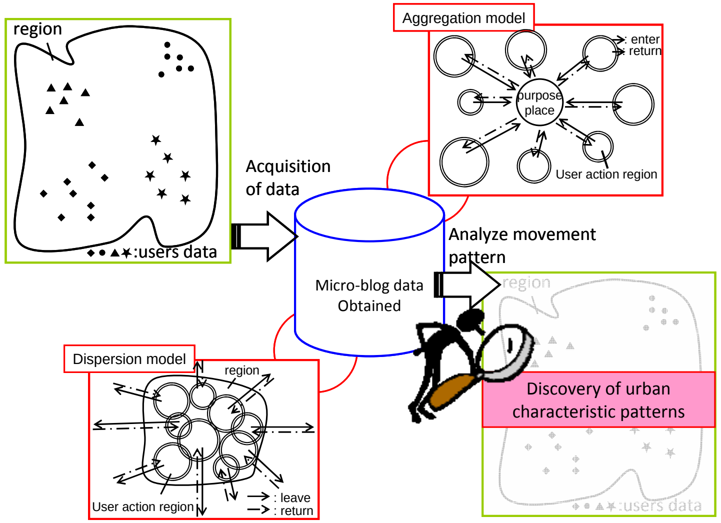


Figure 2: Process for Discovery of Urban Characteristic Patterns using Micro-blogging Sites

2.3 Related work

Micro-blogging is in a state of evolution, with many academic and practical issues. San et al. [3] examined the use of Twitter in relation to its impact on lifestyles and topical discussions. Iwaki et al. [2] considered the discovery of useful topics from micro-blogs. Both studies paid attention to the contents of messages and the link structure of follows among users. These researches mainly analyzed the trends in remarks and the discovery of tastes based on the context of the messages and link structures. Our study focuses on the location and time when users actually write micro-blogs, in order to discover regional characteristic patterns from movement histories. Moriya et al. [4] developed a system that estimates situation of a region from text, in relation to geographic information provided by blogs, and displayed the results on a digital map. This research is similar to our work in terms of social analysis being the point used to consider tendencies, but in our research we also consider habitual movement patterns and social events. As research analyzing the social movements of people, Otsuka et al. [5] revealed a relationship between the real world and web and analyzed how web behavior reflects the real world. This research analyzed the movement patterns of people using the web as a resource. Our study discovers regional characteristic patterns based on micro-blogs where people's movements are directly reflected in the data.

3. Geo-tagged Micro-blog Observation Model

3.1 Gathering Mass Micro-blog Data

In this section, we explain a method for obtaining micro-blog data. We developed a system that can obtain micro-blog data for a region designated for analysis. Figure 3 presents the procedure used in our study.

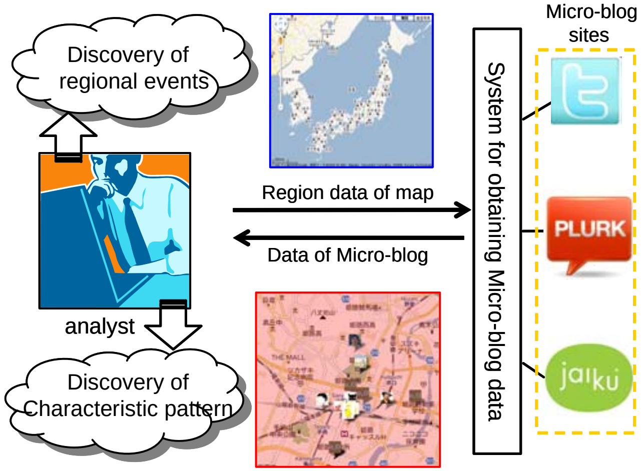

Discovery of regional events
analyst
Discovery of Characteristic pattern
Region data of map
Data of Micro-blog
System for obtaining Micro-blog data
Micro-blog sites
PLURK
jaiku

Figure 3: Outline of Micro-blogs Monitoring System

This system is able to retrieve the region a user wants to examine using a digital map. In other words, this system can receive various requests from analysts who want to obtain data on a large region, such as at a national level, or for a narrow region such as for a prefecture or city. The analyst first needs to identify the region for examination. The system then locates the geographic region, acccesses the micro-blog site and obtains data through the API [7].

We explain our preliminary experiments to reveal where most Twitter users are publishing their messages in Japan. Given the limitations of Twitter's open API, a naïve approach to examining such spatial distribution of human activities through Twitter would be to split the entire region of interest into 1 km grid cells, and to place a radar station in every cell. For a circular area with a 100 km radius, we would need $\lceil 100^2\pi \rceil$ cells to cover the area.

Instead of such a naïve approach, we applied quad-tree based space splitting, where a space is recursively split into four rectangular areas of the same size until each space is larger than the minimum radius permitted in the query specification and the number of results will be under 1,500. In our case, the minimum radius was 1 km and the identifiable number was 1,500. Thus, in the quad-tree, if a node has a radius over 2 km and 1,500 results, it would be split repetitively. For practical processing and a simple discussion, the basic shape in this study is rectangular, as shown in Figure 4.

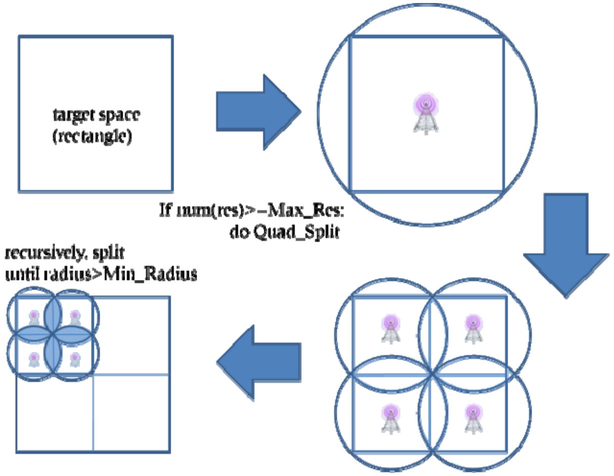


Figure 4: Adaptive Space Partitioning Based on the Density of Tweets

To survey the rectangular region, a circumcircle region enclosing the rectangle was used. The larger area increased by the circumcircle could damage the search results (due to four unintended coverings on each side of the rectangle). This method can help us to utilize general planar-based space indexing algorithms when asking for on-line APIs which usually only support a circular query. Furthermore, for exactness, we can also filter unwanted results outside the rectangle and inside the circumcircle by examining detailed coordinates for the results. The resulting number of radar stations and their coverage can be improved if the spatial distribution of the data is well agglomerated. As an example, Figure 5 shows a quad-tree targeting all regions in Japan for high-resolution tweets acquisition.

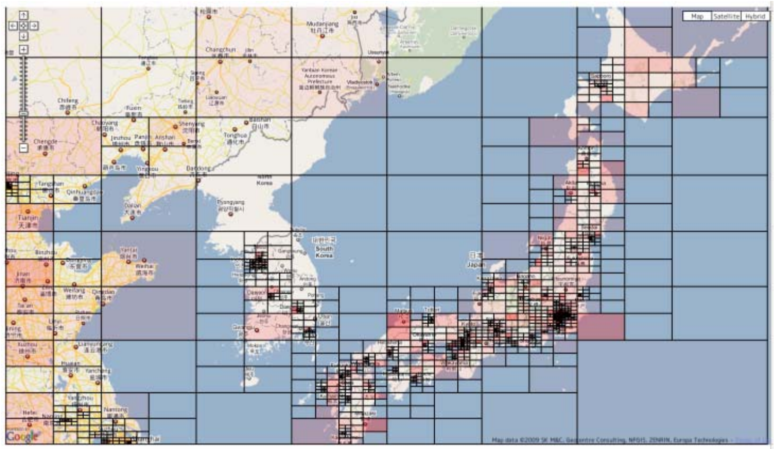


Figure 5: Geographic Distribution of Tweets (aggregated for 7.5 hours, at Aug. 11th, 2009.)

3.2 Clustering Mass Micro-blog Data

Before extracting the characteristics of urban areas from the mass micro-blog data gathered, we need to know the spatio-temporal distribution of the data. Here, we deal with a snapshot of the data from every gathering process; data captured over one week will be used. We are also able to expand such investigation terms for much longer periods. However, we will abide by Twitter policies which state that developers should not, on the grounds of privacy, log twitters for more than one week. From the one week data, we can consider different levels of granularity in the clustering. In the gathering stage we applied quad-tree based space splitting, but such data classification would be adequate for this kind of analysis work due to the limited possibility of granularity. Instead, we need to view the data from various angles; spatially, temporally or spatio-temporally at different levels of detail. Each post in the base data set has spatio-temporal attributes; location represented by latitude and longitude, or place names, date and time represented by universal time.

We expect the there will be many researches looking at generating various types of clustering algorithms for the data based on the available properties. In this paper, we adopt the k-mean clustering method as a simple approach based only on location. We first generate k clusters using the entire one-week data; of course, each k cluster generated will be slightly different, since the algorithm starts with randomly selected k seeds, so that we cannot have the same results in repetitive experiments. However, once the clusters are generated, we will utilize them in the investigation process. We could actually obtain a desirable result from the generated clusters representing areas of high-density in major Japanese cities.

3.3 Analyzing Mass Movement Patterns

In this section, we explain our method for discovering places showing characteristic patterns by analyzing mass movement histories from micro-blog data. We also propose the use of aggregation and dispersion models for analyzing movement patterns.

3.3.1 Aggregation Model

This model seeks to characterize the urban area as shown in Figure 6 in relation to the aggregation pattern.

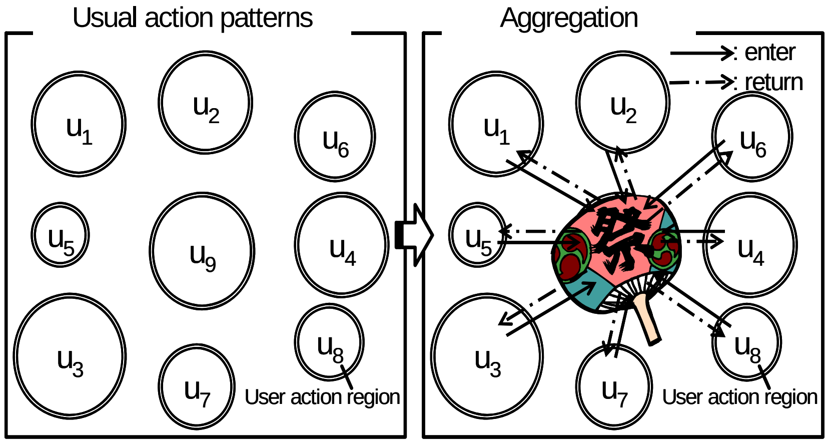


Figure 6: Aggregation Model

We consider that if people gather in a place and at a particular time, then an event will occur. We can imagine such events in urban areas during town festivals or morning commutes. However, there are two different examples of frequency. For example, a railway station can show such patterns all day, every day; while a town's yearly festival will be less frequent. The aggregation model is described in Figure 6; the left side shows the previously known clusters, and we can establish the changes occurring over time. From the two observations, we can figure out a common pattern of aggregation; the central clusters have a high-density.

3.3.2 Dispersion Model

We also need to consider the reverse of aggregation—the dispersion of data. For instance, office workers in downtown areas leave the area and return home. In addition, in a wildfire or during long national holidays like Christmas, a lot of inhabitants who write daily micro-blogs leave their typical daily location. Simple data distribution analysis cannot reveal the characteristic changes. The concept of dispersion is depicted in Figure 7; the clusters in the region of interest show a common pattern; most clusters will show a high dispersion level in the next observation.

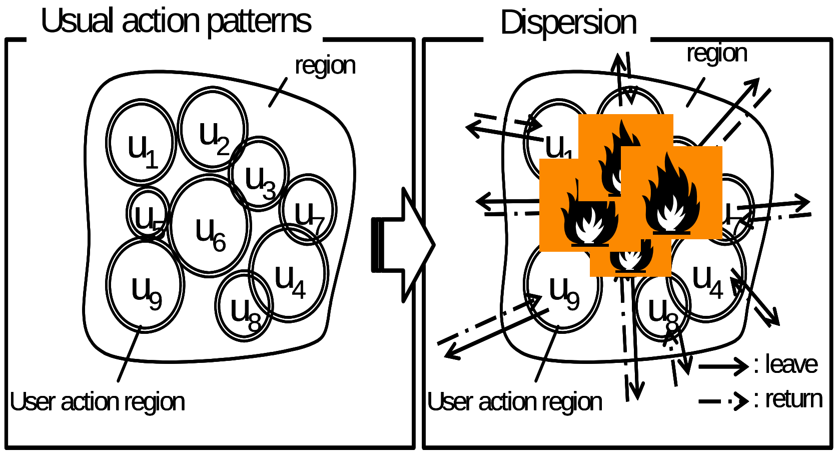


Figure 7: Dispersion Model

4. Experiment

4.1 Preprocessing Collected Data

We initially classified data as being "from_user" which is the unique id for the data received from Twitter. We then sorted each set according to time. We describe the data set we utilized in our experiment as follows:

$$M_{data} = \{0 \le i \le \#user \mid User[i].moving\_list\} \tag{1}$$

$$User[i].moving\_list = (m_0, m_1, \dots, m_k, \dots, m_n) \tag{2}$$

$$m_k = (t_k, loc_k, mesg_k) \quad n = the\ number of\ User[i]'s messege$$

M_{data} represents the data for micro-blog users during the data gathering process. *User[i].moving_list* denotes a user history, where the data are sorted in time order. m_k is a post; a user sent a post from a place (loc *k*), at time (t k), using texts (mseg k). That is, each post is a basic element representing each user's existence at a particular place and time. Finally, *n* describes the number of messages in the data set.

The data set is then clustered with the k-mean clustering algorithm. As described above, we fixed the clustering results utilizing the data for the full week, but investigated each cluster's updates as shown in Figure 8. Each regional cluster will have users coming in, staying and going out on the timeline. The size of the time slot is adjustable by the analyst's specification for investigation in different granularity.

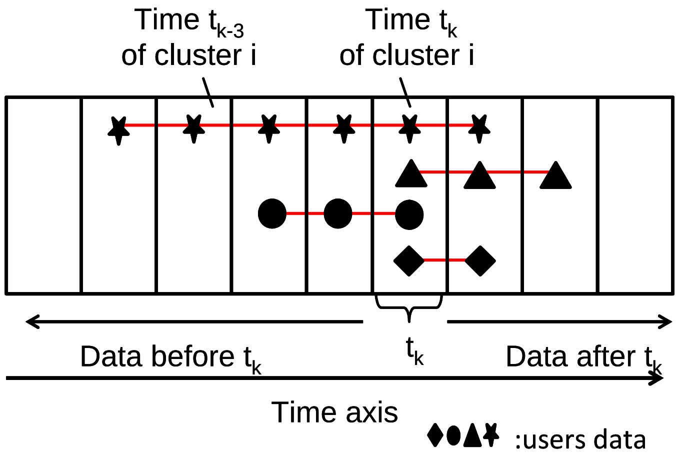


Figure 8: User-movement histories in a cluster

Based on cluster updates in the timeline, we are able to measure the two models as follows:

The degree of Aggregation
= The number of users entering into the cluster
/ The number of users in the cluster in t_k
The degree of Dispersion
=The number of users leaving for the cluster in t_k
/ The number of users in t_k before

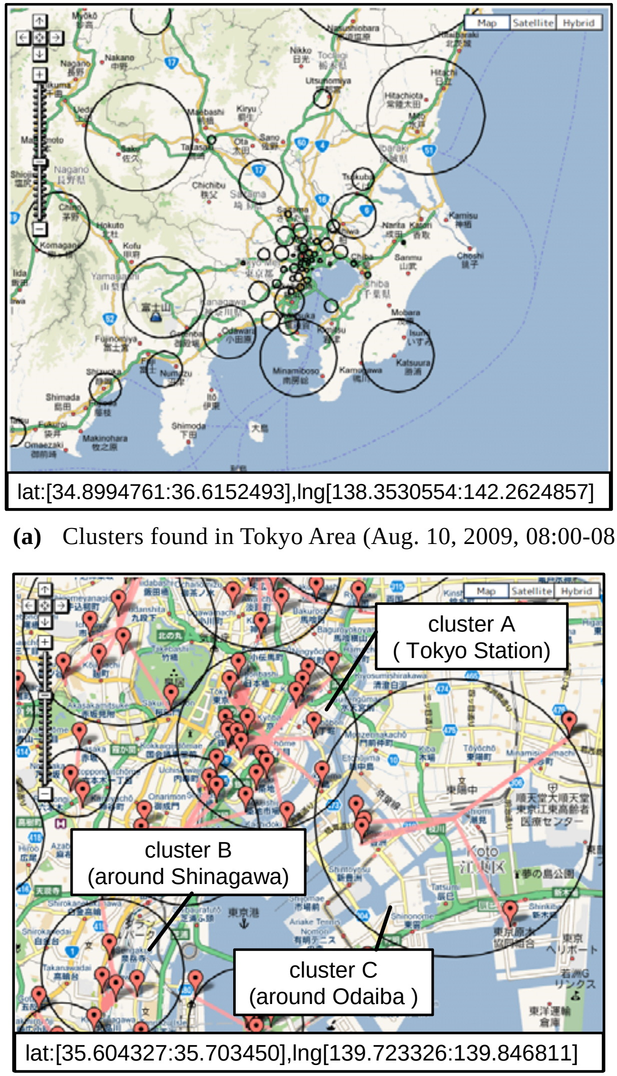
In the aggregation measurement, it is possible to judge whether a user entered the cluster at a particular time by analyzing data for the previous time period. Then, we calculate the degree of aggregation by the proportion of the number of entered users to the number of existing users in the time. Likewise, we calculate the degree of dispersion by the proportion of the number of disappeared users but disappeared in other clusters in the next stage to the number of users in the previous time.

4.2 Experimental Results

We obtained the micro-blog data using the micro-blog monitoring system. We obtained data from the target region's latitude [31.1846:45.7875] and longitude [116.2792:148.3813]. The number of pieces of data was 359,709 and the number of users was 8,139. Next, we only extracted data which had latitude and longitude in the micro-blog post. As the real data, there are data which are attached name of prefecture and city, but these data don't have a trust to be able to decide freely by user. Therefore, we decided not to use such data. We performed experiments to identify the data satisfying the requirements. The results show the following:

(data satisfying the requirements/ all data) * 100 = (4131/8139) * 100 = 51(%)

Then, we performed a clustering of the micro-blog data by location. We utilized the SciPy library of python language for the k-mean clustering with a setting of K=100. We visualized a result of clustering actually performed for an area around Tokyo with the Google map as shown in Figure 9 (a) and (b).


Map
Satellite
Hybrid
lat:[34.8994761:36.6152493],lng[138.3530554:142.2624857]
cluster A
(Tokyo Station)
cluster B
(around Shinagawa)
cluster C
(around Odaiba)
lat:[35.604327:35.703450],lng[139.723326:139.846811]

(a) Clusters found in Tokyo Area (Aug. 10, 2009, 08:00-08:59)

(b) Clusters in a zoomed region around Tokyo Station

Figure 9: Clusters found with the data

In Figure 9 circles express the clusters found. It is clear that many small clusters are formed in the capital of Japan (Tokyo) and big clusters appear in the region around Tokyo Station. This result clearly shows the most populated regions in Tokyo area.

In the next experiment, we calculated the degrees of aggregation and dispersion in some clusters to discover character patterns. We decided that the number of clusters is 50 and performed the clustering method and observed three clusters as shown in Figure 9 (b). Moreover, we made calculations for the morning and evening. We show the results in the following Tables.

Table 1: The estimated results of aggregation or dispersion during 8:00-8:59, Aug. 10

Cluster	The number of users (7:00-8:59)	Aggregation (%)	Dispersion (%)
A. Tokyo Station	16	63.6	50.0
B. Sinakawa	17	100.0	100.0
C. Odaiba	9	77.8	0.0

Table 2: The estimated results of aggregation or dispersion during 18:00-18:59, Aug. 10

Cluster	The number of users (17:00-18:59)	Aggregation (%)	Dispersion (%)
A. Tokyo Station	17	90.0	87.5
B. Sinakawa	18	63.6	60.0
C. Odaiba	17	75.0	81.8

We calculated the degree of aggregation and dispersion for users in a one hour unit. In other words, when we want to examine time t_k, data for one hour before that is t_{k-1}, and we calculated the results using the formula described above.

In Cluster A around Tokyo Station, the degrees of aggregation and dispersion were very high in both time periods. This region has a big railway station with large people flows at key commuting times. In Cluster B at Sinagawa, with a bullet train station, the degrees of aggregation and dispersion were only high in the morning. Actually, the region is populated by many offices, so that in the morning many office workers gather in the region and few people leave. As a result, this region showed the high occurrence only in the morning. In Cluster C around Odaiba, a famous tourist spot with the Fuji broadcasting center, amusement parks, and a number of shopping malls, the degree of dispersion was 0.0 in the morning. But, the degree of dispersion in the evening was very high. From this result, we can assume that this area is not one where people live and is just a place for people to visit.

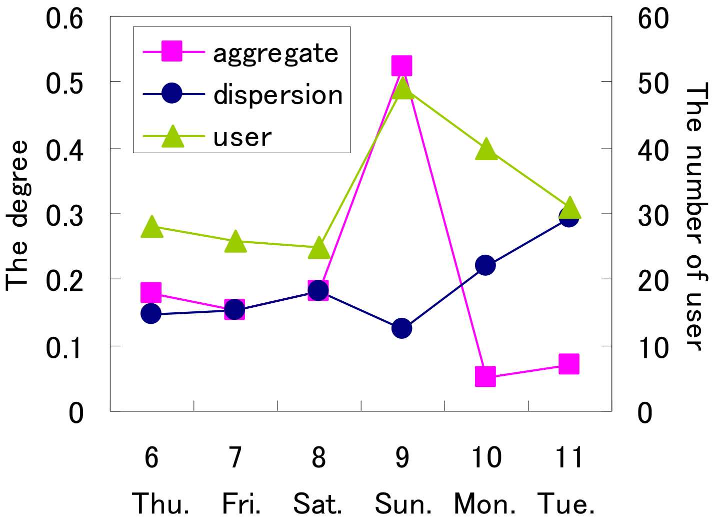


Figure 10: Daily change in aggregation and dispersion

In addition, we changed the timeslot size in the clustering to examine the characteristics of Cluster C, and observed the patterns of microbloggers there. As a result, we could identify temporally significant features found in our experimental data. As shown in Figure 10, the degree of aggregation on the 9th was very high, and the degree decreased from the day and the degree of dispersion increased. In relation to the aggregation, we can speculate that at the weekend, people who usually write micro-blogs from other places actually visited the region. Meanwhile, the reason for the decreasing patterns after the 9th was that it was just before a big holiday in Japan—'Bon' vacation—with many people leaving for other regions, probably heading for their home towns, instead of coming to work or visiting amusement parks in this region.

5. Conclusion

In this paper, we introduced an urban analysis system based on micro-blog data. Obviously, the explosive growth in the publication and transmission of such data makes it an important source for discovering social phenomena or events. We gathered tweets at a nation-wide level and revealed the characteristics of urban areas based on two proposed investigation models. In our future work, we will examine further various measurable patterns and their causes and effects using better clustering methods in different levels of granularities.

6. Acknowledgments

This research was supported in part by a Grant-in-Aid for Scientific Research (B)(2) 20300039 from the Ministry of Education, Culture, Sports, Science, and Technology of Japan.

REFERENCES

[1] "Inside Twitter: An In-depth Look Inside the Twitter World". Sysomos. 2009-06-10. Retrieved on 2009-06-23.

[2] Yusuke IWAKI, Adam JATOWT, and Katsumi TANAKA. Supporting finding read-valuable articles in micro-blogs. DEIM Forum 2009 A6-6, (2009)

[3] Akshay Java, Xiaodan Song, Tim Finin, Belle Tseng. Why we twitter: understanding microblogging usage and communities, International Conference on Knowledge Discovery and Data Mining. Proceedings of the 9th WebKDD and 1st SNAKDD 2007 workshop on Web mining and social network analysis, San Jose, California, pp 5665, (2007).

[4] Keita MORIYA, Shiori SASAKI, and Yasushi Kiyoki. A Dynamic Creation Method of Environmental Situation Maps Using Text Data of Regional Information. DEIM Forum 2009 B1-6 , (2009)

[5] Shingo OTSUKA, Masao TAKAKU, Masaru KITSUREGAWA, and Nobuyoshi MIYAZAKI. A Study for Analysis of User Behavior in The Free Magazine Site for Women. DEIM Forum 2009 B8-4, (2009)

[6] Twitter: http://twitter.com/

[7] Twitter open API: http://apiwiki.twitter.com/Twitter-Search-API-Method%3A-search