Information Engineering Express
International Institute of Applied Informatics
2015, Vol.1, No.4, 33 – 42

Estimation of User Location and Local Topics Based on Geographical Distribution of Microblogging

Kazunari Ishida *

Abstract

This paper proposes a method for estimating microblogging user location to determine local topics of importance based on area-specific term co-occurrence. Geotagged information on social media has not previously been sufficient to determine local topics; however, the amount of information available on social media has continued to expand due to the widespread use of smartphones. Notably, the amount of information generated from regional cities is significantly smaller than that from metropolitan cities. Hence, we must estimate the location of each user in a regional city to obtain adequate local information for determining local topics. To extract this information, we define area-specific scores of terms and co-occurrences that are calculated using term frequency, as well as the average and standard deviation of the longitude and latitude of raw geotagged information.

Keywords: Geotagged information, Location estimation, Microblog, Term co-occurrence

1 Introduction

Social media gives us personal publishing and communicating media. For example, we can publish our personal experiences and opinions on blogs. Writing a blog post might require several tens of minutes depending on the length of the post. On the other hand, microblogging gives us a handy medium for publication and communication because of its short message format. For example a message on Twitter is limited to 140 characters. Due to widespread use of smartphones and high speed mobile communication infrastructure, we can write short messages on social media in timely manner. In addition, GPS enabled devices let us generate geotagged information, which has precise location information in terms of longitude and latitude. Hence, users of microblogging services are potential social sensors to sensing any kind of events or topics in specific local areas. It is possible to extract valuable local information from social media if there is sufficient geotagged information. However, of the amount of geotagged information available from social media is small, because many people are worried about the privacy risk of providing exact location information at any time. Fortunately, early adapters of location enabled messages on microblogging platforms continue to generate geotagged messages. These mes-

* Hiroshima Institute of Technology, Hiroshima, Japan

sages give us clues to identifying the locations of microbloggers. Users with estimated locations can help us to extract valuable local information from social media.

Extraction of local information generated by individuals is very important for managing place branding in cities, regions, and nations[†], because this information contains public opinion concerning these places. Opinion tends to be positive when public services and commercial zones are excellent and attractive. Hence, public administrations and businesses have to understand public opinion to adjust their services for their customers. Public opinion in specific geographic areas can be easily extracted from microblogging messages for place branding[‡]. However, research concerning place branding tends to collect information from a few official microblogging accounts concerning regional tourism centers and bureaus instead of the huge amount of individual microblogging accounts because of the scarcity of geotagged messages.

In recent years, many researchers have been trying to analyze location information on social media based on text classification technics, e.g. area-specific keywords and local words, which frequently appear in one local area but not in others. Employing text classification technics seems to promise ways to associate location with messages on microblogs. However, words are inherently ambiguous because they frequently have multiple meanings in natural language. In addition, we need to take into account the precision of location information when we try to identify interesting opinions and topics in local areas, because errors in location information can lead us to incorrect conclusions about the importance of various opinions and topics.

To improve the location estimation of microbloggers, this paper proposes a method for estimating the approximate location of microbloggers based on area-specific term co-occurrences, because such co-occurrences can represent multiple meanings of terms [1][2][3]. Based on microblog posts with estimated locations and precision of the estimation, term frequencies are re-calculated to reduce measurement errors on the importance of opinions and topics in local areas.

The rest of this paper is organized as follows: Related works are explained in section 2. Section 3 explains our dataset, defines three types of estimation method, and describes the evaluation of precision of the methods applied to a geotagged dataset which is a part of the entire dataset. We will show that the method with term co-occurrences is more precise compared to that with terms alone. In section 4, we apply the estimation method to the entire dataset to expand geotagged information. Employing the expanded information, we introduce weighted term frequency to measure the importance of opinions and topics in local areas to reduce measurement errors. We also provide several examples of local topics based on weighted term frequency. Section 5 concludes our work.

2 Related Works

In recent years, many researchers have been interested in analyzing location information from social media. Dalvi et al. [4] defined distance and language models using an expectation-maximization (EM) algorithm to match a tweet to an object in the real world. Choosing restaurants as their objects, these authors extracted 750,000 data records from Yahoo local data

[†] https://en.wikipedia.org/wiki/Place_branding
[‡] http://www.researchgate.net/publication/260363011

from December 2009 to January 2011. However, the present paper discusses topics without any geographical restriction.

Bo et al. [5] predicted the location from which tweets had been sent based on text classification with area-specific keywords. These researchers employed administrative districts as area divisions and combined areas with small amounts of information. Based on these area divisions, the authors defined three types of words: (1) local words, (2) semilocal words, and (3) common words. Term frequency, area frequency, and information gain were all considered to be features of these words. This research combined areas with small amounts of information before processing data, while the present method allows combination to be conducted after processing for the purpose of discussion.

Cheng et al. [6] proposed an algorithm to estimate user location based on area-specific keywords. To select the keywords, these authors employed the word distribution model defined by Backstrom [7]. This research utilized keyword locality, while the present study improves the precision of location estimation by considering the co-occurrence of keywords in addition to individual keywords.

Ishida [8] estimated users' locations based on a local area mesh defined by the statistics bureau of the Japanese Ministry of Internal Affairs and Communications. Roller et al. [9] estimated users' locations with a language model and used an adaptive grid as the area division. Conversely, the present study employs administrative divisions due to their easy analytical interpretation.

Chandra et al. [10] estimated users' locations using local term frequency and retweet information. These authors used the locations from which the original tweets had been sent, as well as the locations from which successive retweets had been sent, to improve the precision of location estimation, while the present study employs the co-occurrence of terms.

Hong et al. [11] defined global area distribution, user distribution, global topic distribution, local topic distribution, global term distribution, local term distribution, global topic matrix, average location of latent area, and the covariance matrix of latent area using an EM algorithm to extract local topics. This research used term distribution, while the present study uses both term and co-occurrence distributions to increase the precision of location estimation. To estimate the locations of users, we focus on geotagged terms and co-occurrences from a microblog dataset.

3 Estimating User Location

Twitter users move about for various purposes, e.g. school, job, shopping, travel, and so on. However, each user tends to tweet around his or her home territory, e.g. hometown and workplace, even though they could tweet wherever they go. Hence, we assume that the number of tweets sent from the home territory of each user is relatively huge compared to that of tweets sent from other places. Based on this assumption, we are estimated each user's location. The assumption is acceptable for this research, because we need to estimate of users' locations in order to gauge local topics. In addition we only use Twitter for data source. We can also collect location information from Open Street Map and Wikipedia. However each data source has its own set of frequently used terms. According to our preliminary experimentation, there are few terms that overlap among the data sources, which means that we cannot extract location information from the other data sources with respect to terms and co-occurrences of terms. Hence, we decided to employ Twitter as the sole data source. Geotagged terms and co-occurrences from a microblog

dataset offer plentiful information for estimating user locations. Data extraction employed the following steps:

1. Extract geotagged tweets from the dataset.
2. Identify users who published geotagged tweets.
3. Extract all nouns from the geotagged tweets extracted in step 1.
4. Calculate the term frequency and average longitude and latitude of each term extracted in step 3. The average location determines the administrative address of the term. The standard deviations of the longitude and latitude are also derived to determine the area-specific score in step 5.
5. Calculate the area-specific score of each term on the administrative address with the standard deviations of longitude and latitude derived in step 4. The formula will be described in section 3.1.
6. Extract all tweets from all users identified in step 2.
7. Estimate the location of each user based on all of that user's tweets and the area-specific score calculated in step 5. All terms are extracted from all tweets of each user. Each term has area specific scores in multiple areas. The scores of the terms are accumulated for each area. As a result an area ranking list of each user is derived from each user's tweets based on the area-specific score calculated in step 5. The top area of the list is the estimated location of the user.
8. Evaluate the precision of the estimated location of each user based on the actual locations from which that user sent geotagged tweets.

We collected tweets using a sample of Twitter's public stream[§] from March 2011 to May 2014. The number of tweets in the resulting dataset was 347,742,872. The number of terms was 4,124,568,983, and the variety of terms was 58,994,705. The number of users was 17,251,905. The number of geotagged tweets was 1,132,580, representing 0.33% of the total tweets in the dataset. The number of users who published geotagged tweets was 311,812, representing 1.8% of total users.

3.1 The first method

The first method of location estimation is based on the statistics and area-specific scores of the terms. We identified an administrative division for each geotagged term based on a database of relationships between administrative divisions and locations (longitude and latitude). We defined the area-specific score (1), which is defined based on term frequency (*tf*) and the standard deviations of longitude (*sx*) and latitude (*sy*). These statistics were calculated in step 4 of data extraction. Table 1 shows examples of the statistics. According to this formula, a term with small geographical dispersion and high frequency receives a high score. We can then estimate the location of each user by adding the scores of all terms which appeared in that user's tweets.

$$Score = tf \times \exp\left(-\sqrt{sx^2 + sy^2}\right) \quad \cdots (1)$$

[§] https://stream.twitter.com/1.1/statuses/sample.json

3.2 The second method

The second method is based on the first method but employs thresholds on term frequency and geographical dispersion to reduce estimation error due to noisy terms. A term can become noisy when it has extremely high frequency and a large geographical dispersion. Based on our preliminary experiment, we employed limits of 50,000 and 2.0 for the frequency and deviation, respectively.

Table 1: Term Frequency and Location

Term	Freq.	Avg. Lng.	Avg. Lat	Std. Dev. Lng.	Std. Dev. Lat.
Tokyo	34875	139.52	35.67	1.12	0.66
Kyoto	8257	135.86	35.03	0.87	0.44
Shinjyuku	6951	139.67	35.69	0.48	0.24
Earthquake	7997	138.74	36.51	3.63	2.75
Tsunami	230	138.64	36.35	3.90	3.32
…	…	….	…	…	…

3.3 The third method

The third method is based on term co-occurrence in tweets to reduce the effect of meaning variety on the usage of each term. A term may have multiple meanings, which increases error in location estimation. Co-occurrence between two terms can reduce this error by providing a context for usage. Term frequency and geographical dispersion limits were employed for only one term in a co-occurrence instead of both. If we had restricted both terms in the co-occurrence, we would have been unable to obtain sufficient geotagged co-occurrences, while if we had not restricted either term, we would have suffered from a computational complexity problem and noisy co-occurrences. We defined area-specific scores for each co-occurrence of term pairs using the same formula discussed in the first method. Table 2 shows examples of the co-occurrences and their statistics.

Table 2: Term Cooccurrence and Location

Term 1	Term 2	Area	Avg. Lat.	Avg. Lng.	Std. Dev. Lat.	Std. Dev. Lng.
Tsunami	Ibaraki	Ibaraki	36.50	140.62	0.48	0.30
Chiba	Tsunami	Chiba	35.65	140.31	0.24	0.43
Iwate	Tsunami	Iwate	39.32	141.76	0.43	0.16
Miyagi	Tsunami	Miyagi	38.43	141.31	0.99	0.62
Tsunami	Takahagi	Ibaraki	36.72	140.71	0.00	0.00
…	…	…	…	…	…	…

3.4 Evaluation of Estimated Location

We applied the three methods for estimating location to our dataset and evaluated the results in terms of error distance, i.e. the deviation between the estimated and actual locations. Figure 1 illustrates error distance and user distributions for the three types of estimation method. According to the results of the first method, a deviation between 250 and 300 km was most common for the estimated locations of users. The second method most frequently obtained deviations between 50 and 100 km. The third method most frequently yielded deviations of less than 50 km.

Figure 2 depicts cumulative user rate and the error distance for the three types of estimation method. We were able to estimate location for over 80% of users in the dataset with average deviations of 350, 300, and 100 km for the first, second, and third methods, respectively.

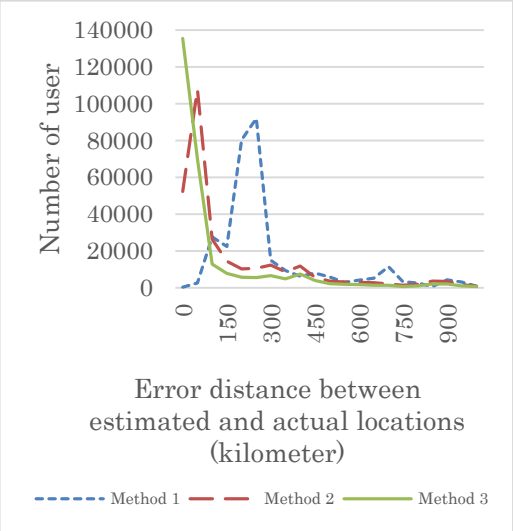

Figure 1: Error Distance and User Distribution

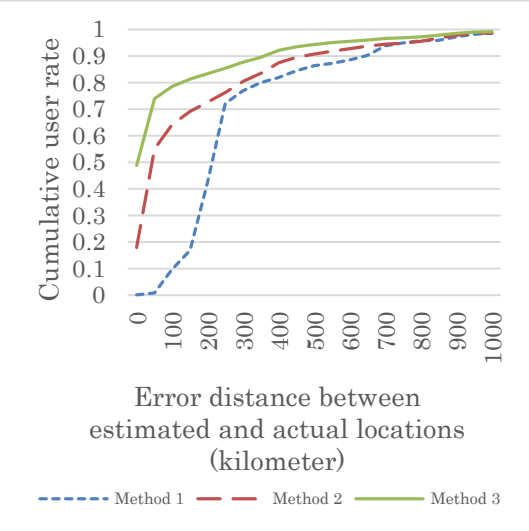

Figure 2: Distance and Cumulative User Rate

4 Extracting Local Topics with Location Estimation

We determined area-specific topics in terms of time series of term frequency for each administrative division. Employing the third method, we extracted and estimated the locations of 8,575,766 users, representing nearly 50% of the total users in the dataset. The number of geousers (users who published one or more geotagged tweets) was 331,812. Hence, we extracted a sample roughly 28 times larger than the number of area-specific users. To extract each time series of term frequency for each area, we employed Hadoop to handle the huge amount of tweets in the dataset. The following processing steps were employed: 1) Separate sets of tweets in terms of users with Hadoop. 2) Extract all term co-occurrences from each set of tweets for each user. 3) Estimate location based on the term co-occurrences. 4) Extract time series for each term based on users whose estimated locations are in the same administrative division. 5) Calculate the frequency of each term in each area with Hadoop.

4.1 Discussion of Information Expansion

To evaluate the benefit of location estimation, we compared the statistics for raw geotagged tweets and tweets with estimated information. The raw geotagged tweets contained 5,032,683 terms, and the average frequency of a term was 30.45. In contrast, the average frequency of terms in the tweets with estimated information was 613.51, over 20 times that of the raw geotagged tweets. Furthermore, the tweets with estimated information contained 49,309,065 terms that did not appear in the raw geotagged tweets. Hence, the variety of terms in the tweets with estimated information increased by nearly 10 times over that of the raw geotagged tweets.

4.2 Selection of Granularity of Local Area

To select a reasonable granularity for the administrative divisions, we calculated correlation coefficients between the population statistics provided by the Ministry of Internal Affairs and

Communication** and the estimated locations of users in the dataset. At the state or prefecture level in Figure 3, we obtained a high correlation coefficient of 0.823. However, at the city level in Figure 4, we obtained a low correlation coefficient of 0.381. We therefore chose to employ state or prefecture-level divisions.

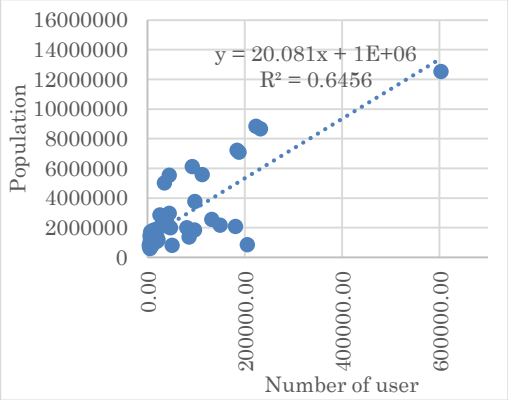

Figure 3: Users and Population in Prefectures

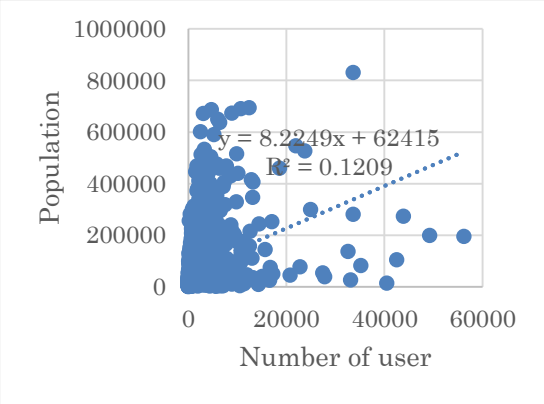

Figure 4: Users and Population in Cities

4.3 Extracting Estimated Time Series of Terms

To extract time series of each term based on users with estimated locations, we defined a weight for each user and a weighted term frequency. This estimation is not perfect; therefore, we must reduce the error for the extraction of terms in each area. The weight of each user is defined by the average area-specific score (*scave*) and standard deviations of longitude (*sx*) and latitude (*sy*), as shown in formula (2). *scave* is an average of area specific scores of all possible areas for a Twitter user estimated in step 7 of section 3. This definition ensures that estimated users with a high average score and small geographical dispersion are heavily weighted. Employing this weight, we also defined the weighted term frequency (*WTF*), as shown in formula (3).

$$Weight = scave \times \exp\left(-\sqrt{sx^2 + sy^2}\right) \quad \cdots (2)$$

$$\mathrm{W}TF = TF\left(1 - \exp\left(-\left(1 + Weight\right)\right)\right) \quad \cdots (3)$$

Based on weighted frequency, we aggregated the *WTF* of each term in each prefecture. Many place names were ranked highly in each area. For example, "Tokyo" was ranked 3rd (1,135,150), Shinjuku 11th (381,421), and Shibuya 12th (360,807) in Tokyo. In Osaka, "Osaka" was ranked 3rd (1,525,890) and "Umeda" 39th (135,327). Certain terms of dialect were also highly ranked, e.g., "honma" (really) at 2nd (2,413,362), "yakara" (because) at 48th (101,397), and "yakedo" (but) at 58th (90,295). Moreover, famous foods in an area often received high rankings, e.g., "Macdo" (Mc Donald's) at 115th (39,296) and "takoyaki" (ball-shaped snack) at 137th (32,216). In Hiroshima, the place names "Hiroshima," at 2nd (2,034,632), and "Kure," at 27th (173,125), were highly ranked. High rankings were also assigned to "sashimi" (raw fish), at 51st (80,569), and "okonomiyaki" (savoury pancake), at 53rd (79,038). Based on these rankings, we discuss area-specific topics related to area-specific foods. We also discuss natural disasters because they have occurred frequently over the past several years in Japan. In order to find the most discussed

** http://www.soumu.go.jp/menu_news/s-news/17216_1.html

area concerning each topic, we will find top five areas based on the fraction of each term in an area, because the fraction represents the importance of each topic in the area.

4.4 Local Topic: Okonomiyaki

Okonomiyaki is a Japanese savoury pancake containing a variety of ingredients. Okonomiyaki is mainly associated with the Kansai or Hiroshima areas of Japan, but is widely available throughout the country according to Wikipedia[††]. On one hand, the top five areas mentioning the topic "okonomiyaki" were Okayama, Hiroshima, Iwate, Hyogo, and Wakayama in terms of raw geotagged information. On the other hand, the top five areas mentioning the topic in terms of estimated data were Hiroshima, Okayama, Hyogo, Shimane, and Osaka. Because of the information expansion, Hiroshima rose to first place. According to the enterprise statistics for 2006 provided by the statistics bureau of the Ministry of Internal Affairs and Communication, Hiroshima is the top area by number of okonomiyaki restaurants. Concerning the Hiroshima area, statistics of the term "okonomiyaki" from the raw geotagged data and of the location-estimated data are summarized in Table 4. We can observe more than fourteen times the number of days in the estimated data compared to the raw geotagged data. We can also count *WTF* at almost two times the frequency. Change between fractions of *WTF* and *TF* concerning the term "okonomiyaki" were very different depending on the area, i.e. Hiroshima (+9.1%), Okayama (+0.6%), Hyogo (-0.3%), Shimane (-1.5%), and Osaka (+0.7%). To sum up, information expansion by location estimation and information adjustment by *WTF* contributed to making Hiroshima the most famous area concerning "okonomiyaki".

Table 4: Estimated and Raw Data of Okonomiyaki in Hiroshima

	Estimated	Raw	Increase rate
Number of days	354	25	14.16
Average WTF	1.95	1.03	1.88
Sum of WTF	689.84	25.86	26.68

4.5 Local Topic: Tsunami

A huge earthquake hit Japan in 2011. A huge tsunami also hit the Pacific coastline of the Tohoku region of Japan. Many towns were destroyed by tsunami. More than nine thousand people went missing in Minamisanriku and almost one thousand bodies were recovered in towns by 14 March 2011[‡‡]. Hence, tsunami is one of the biggest calamities in Japan. On one hand, the top five areas mentioning the topic "tsunami" are Tochigi, Miyagi, Ibaraki, Saitama, and Gunma in terms of raw geotagged information. Tochigi, Saitama, and Gunma are inland or non-coastal. On the other hand, the top five areas mentioning the topic in terms of estimated data were Miyagi, Fukushima, Ibaraki, Tochigi, and Wakayama. Because of the information expansion, Miyagi gets the first place and some inland areas i.e., Saitama and Gunma, are excluded from the top five, although Tochigi remained in fourth place. Concerning Miyagi, statistics of the term "tsunami" from the raw geotagged data and from the location-estimated data are summarized in Table 5. We can observe more than seven times the number of days in the estimated data compared to the

[††] http://en.wikipedia.org/wiki/Okonomiyaki

[‡‡] http://en.wikipedia.org/wiki/2011_T%C5%8Dhoku_earthquake_and_tsunami#Tsunami

raw geotagged data. We can also count *WTF* at almost three times the frequency. Change between fractions of *WTF* and *TF* concerning term "tsunami" were also very different in those areas, i.e. Miyagi (-1.6%), Fukushima (+1.3%), Ibaraki (+0.4%), Tochigi (-0.5%), and Wakayama (+19.4%). The big increase in Wakayama might describe the Nankai Trough concern[§§]. Concerning Miyagi, Figure 5 and 6 illustrate trajectories of the term "tsunami" from raw geotagged and estimated data, respectively. The estimated data exhibits a huge spike on December 7, 2012, compared to the raw geotagged data. According to a news release of the Fire and Disaster Management Agency[***], a large earthquake (magnitude 7.4) occurred on this date around the offshore sea of Sanriku. The agency issued a tsunami warning to Miyagi Prefecture. Because of the huge earthquake and tsunami that had previously occurred on March 11, 2011, many people in the area were worried about similar disasters. The observed spike was successfully extracted from the huge dataset by the location estimation method. To summarize, information expansion by location estimation and information adjustment by *WTF* contributed to finding the frequency of the term "tsunami" in tweets from Miyagi.

Table 5: Estimated and Raw Data of Tsunami in Miyagi

	Estimated	Raw	Increase rate
Number of days	354	46	7.70
Average WTF	5.08	1.82	2.79
Sum of WTF	1799.16	83.69	21.50

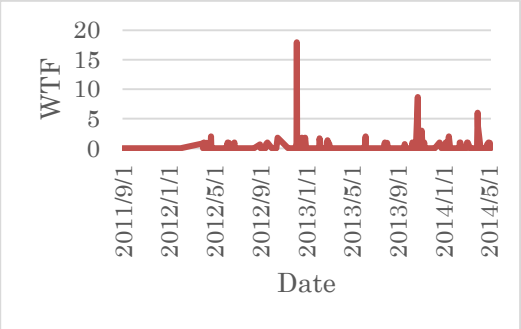

Figure 5: Raw Trajectories for tsunami

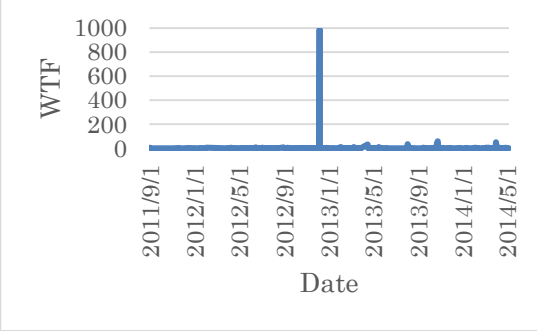

Figure 6: Estimated Trajectories for tsunami

5 Conclusions

This paper proposed a method for estimating the approximate locations of microblog users based on area-specific terms and term co-occurrence. To identify area-specific topics using a small amount of geotagged information, it is necessary to estimate the approximate locations of microbloggers for whom location is unknown. Each term extracted from geotagged microblogs is represented in terms of statistics, i.e., frequency of term use and the average and standard deviation of longitude and latitude for location. Based on these statistics, area-specific scores are defined for each term-area pair. The method thus estimates the approximate locations of microblog users by employing terms and co-occurrences. Based on these estimated locations, we deter-

§§ https://en.wikipedia.org/wiki/T%C5%8Dkai_earthquakes

***

http://www.fdma.go.jp/bn/%E4%B8%89%E9%99%B8%E6%B2%96%E3%82%92%E9%9C%87%E6%BA%90%E3%81%A8%E3%81%99%E3%82%8B%E5%9C%B0%E9%9C%87%28%E7%A2%BA%E5%AE%9A%E5%A0%B1%EF%BC%89.pdf

mined area-specific topics in terms of time series of term frequency for each administrative division.

References

[1] Ishida, K. and Ohta T., “An approach for organizing knowledge according to terminology and representing it visually,” IEEE Transactions on Systems, Man, and Cybernetics, Part C, Vol. 32, No. 4, 2002, pp. 366-373.

[2] Ishida, K., “Extracting Latent Weblog Communities: A Partitioning Algorithm for Bipartite Graphs,” Proceedings of the 2nd Annual Workshop on the Weblogging Ecosystem - Aggregation, Analysis and Dynamics in the 14th International World Wide Web Conference (WWW2005), Makuhari Messe, Chiba, Japan, May 10 - 14, 2005.

[3] Ishida, K, "Extracting Spam Blogs with Co-citation Clusters," Proc. Of the 17th International World Wide Web Conference (WWW2008), April 21 - 25, 2008.

[4] Dalvi N., Kumar R., and Pang B., "Object Matching in Tweets with Spatial Models," WSDM’12, Seattle, Washington, USA, February 8-12, 2012.

[5] Bo H., Cook P., and Baldwin T., "Geolocation Prediction in Social Media Data by Finding Location Indicative Words," Proceedings of COLING 2012: Technical Papers, pp. 1045-1062, COLING 2012, Mumbai, December 2012

[6] Cheng Z., Caverlee J., and Lee K., "A Content-Driven Framework for Geolocating Microblog Users," ACM Transactions on Intelligent Systems and Technology, Vol. 4, No. 1, Article 2, Publication date: January 2013.

[7] Backstrom, L., Kleinberg, J., Kumar, R., and Novak, J. “Spatial variation in search engine queries,” Proceeding of the 17th international conference on World Wide Web, WWW ’08, pages 357-366, Beijing, China, , April 21 - 25, 2008.

[8] Ishida K., "Extracting Geo-Social Information based on Geo-Tagged Social Media," 4th World Congress on Social Simulation (WCSS 2012), National Chengchi University, Taipei, Taiwan, September 4-7, 2012 .

[9] Roller S., Speriosu M., Rallapalli S., and Wing R., Jason Baldridge, "Supervised Text-based Geolocation Using Language Models on an Adaptive Grid," Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Jeju Island, Korea, 12-14 July 2012, pp. 1500-1510.

[10] Chandra S., Khan L., and Muhaya F. B., "Estimating Twitter User Location Using Social Interactions - A Content Based Approach," 2011 IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing, Boston, MA, October 9-11, 2011.

[11] Hong L., Ahmed A., Gurumurthy S., Smola A., and Tsioutsiouliklis K., "Discovering Geographical Topics In The Twitter Stream," WWW 2012, Lyon, France, April 16-20, 2012.