Inferring Twitter User Locations With 10km Accuracy

KyoungMin Ryoo
Division of Web Science and Technology
KAIST
Korea
kmryu@an.kaist.ac.kr

Sue Moon
Department of Computer Science
KAIST
Korea
sbmoon@kaist.edu

ABSTRACT

Geographic locations of users form an important axis in public polls and localized advertising, but are not available by default. The number of users who make their locations public or use GPS tagging is relatively small, compared to the huge number of users in online social networking services and social media platforms. In this work we propose a new framework to infer a user's main location of activities in Twitter using their textual contents. Our approach is based on a probabilistic generative model that filters local words, employs data binning for scalability, and applies a map projection technique for performance. For Korean Twitter users, we report that 60% of users are identified within 10 km of their locations, a significant improvement over existing approaches.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications; J.4 [**Computer Applications**]: Social and Behavioral Sciences

Keywords

Microblog, location estimation, text mining

1. INTRODUCTION

In the past few years online social media have risen as a key venue for communicating with the public and monitoring public opinions. In order to weigh in the public opinions expressed on such social media as much as traditional poll results, the representativeness of the opinions has to be accounted for. Geographic location is one of the key factors in the representativeness. Yet, most users of online social media do not make their geographic location information public. For example, only 34% of Twitter users have meaningful location information in their profiles, and less than 1% of Twitter users tag their tweets with GPS locations [7, 8]. Frank *et al.* show that a large portion of tweets were generated near a user's home or workplace [5]. As most users in the US commute less than 20 miles a day and in Korea 12.5 km [10][1], we can interpret the main point of activities on Twitter as the representative locale of the user.

In this work we propose a new approach to infer a Twitter user's main point of activities. Previous work has investigated spatial correlation between web resources and geographic locations [3, 7]. From GPS-tagged tweets we extract the spatial correlation between words and GPS locations and refine the city-level granularity of previous work to 500 m distance bins. We use data binning to reduce computational cost. Also, computing the Euclidian distance from the longitudes and latitudes will cause distortion and we use map projection to convert bewteen coordinates of longitudes and latitudes and of the 3D Euclidian space. We verify the accuracy of our approach with large-scale data of Korean Twitter users. Our method estimates 74.9% of user locations correctly within 10 km of their main locations.

The rest of the paper is organized as follows. In Section 2 we review related work, and data collection methodology is in Section 3. Section 4 and Section 5 demonstrate the background of probabilistic model and inferencing geographic distribution of words, respectively. In Section 6 we present method of inferring user location. In Section 7, our algorithm is introduced. In Section 8, we present another method inferring user location using friends' words. We conclude in Section 9 with a brief discussion for future work.

2. RELATED WORK

Geographic locations of users on online social networking services are of paramount importance in marketing, advertising, and public opinion polling. Yet most users do not specify their towns of residence or use the GPS tagging feature. From the few users with annotated locations and GPS tagged status updates, inference techniques mine location information of unknown users [2, 3, 8, 11].

One set of location inference techniques relies on the social network of users. Sadilek *et al* examine the location information and the social network of users with annotated locales and predict the location of their friends using a dynamic Bayesian network [12]. Jurgens *et al* utilize reciprocal relationships on Twitter and estimate user locations. They report a success ratio of 74% on inferring users' locations

[1]http://kosis.kr/gen_etl/start.jsp?orgId=201&tblId=DT_201_00222&conn_path=I2&path=NSI

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2579236.

on Twitter even without the textual contents in social media [8]. The authors argue that a user's network in social media is a pertinent source of information for inferring user location. They also demonstrate that mixing multiple social media datasets have the potential to improve the accuracy and infer locations on another social network.

Another approach is to take advantage of user-generated contents. Location inference of search engine queries and web pages has produced the idea of power and spread [4, 13], and Backstrom *et al.* refine it to build a probablistic model for spatial variation [1]. Hecht *et al.* produce a descriptive report on Twitter users' behavior. According to their paper, a low ratio of 34% Twitter users did not enter their actual geographic information on their profiles. They use a term-frequency-based Multinomial Naïve Bayes model on textual contents and estimate state-level user locations in the US [7]. Cheng *et al.* propose a probablistic framework similar to Backstrom *et al.*'s and refine the noisiness in tweet words by a local word classifier. They demonstrate that by filtering out non-local words, the estimation error is reduced from 1,773 miles to 539 miles. Also they employ smoothing to address the data sparseness and places "51% of users within 100 miles of their actual locations."

3. KOREAN TWITTER DATASET

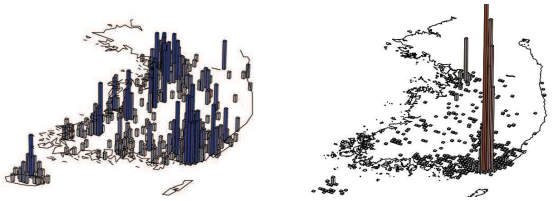
We have chosen Korean as the target langauge for this work. Most social media analyses have focused on English contents, and other languages have received relatively less coverage. As demographics, geography, and NLP (Natural Language Processing) tools all differ by the country and the language, we believe this work is interesting in its own right for designing and evaluating a location inference technique.

In order to find Korean users on Twitter, we used snowball sampling. Starting from two Korean celebrities with more than $100,000$ followers, we crawled those celebrities' followers, but limited to those who have at least one tweet written in Korean among their 200 most recent tweets. Using the Twitter API from June 2010 to April 2011, our crawl resulted in 615 million tweets and 3.3 million Korean user profiles. Our dataset consists of tweets, user profiles, and following-follower relationships.

Twitter provides two types of location information: the location field in the user profile and GPS tags of tweets. According to Hecht *et al.* most users leave the location field blank or do not write the formal location name [7]. Fewer users turn on the GPS tagging feature on their smartphones [8]. In our dataset of 614 million tweets, only 0.4% or 2.8 million tweets from $140,275$ users have the GPS tags. These users with GPS-tagged tweets form the ground truth in evaluating our tweet-based location inference.

With the GPS tagging on, a user is associated with multiple locations but a large portion of tweets come from the user's home or workplace [5]. In order to identify the single location of most representativeness to a user, we take the geometric median m, of all the GPS positions calculated as below and label it as the user's location. It can be the home, workplace, or some other location of frequent visits by the user.

$$m = \arg\min_{x \in L} \sum_{y \in L} distance(x, y)$$



(a) "story" (b) "Busan"

Figure 1: Two words "story" and "Buson" to demonstrate spatial locality of words. The area of the bar is 500x500 m^2 and the height represents the tweet frequency.

where L is a set of GPS locations of a particular user and *distance* is the physical distance between the two points.

In order to filter out those who often have travelled far or who have too few tweets with GPS tags from the center location, we limit to those who have at least 5 tweets within 15 km of their center location as Jurgens has done in [8]. Also, we sort out 826 social spammers using features as Lee *et al.* has done in [9]. The final tally is $22,525$ users. Of these users' tweets we apply the Korean Morpheme Analyzer (KKMA) [6] and extract $801,505$ words.

4. BACKGROUND

In this section, we present the basic idea of how we select local words in tweets. We examine the word's spatial locality in order to determine whether the word can be labeled as local. Words such as "time", "story", or "politics" do not show clear spatial locality because their use is not limited to a confined area, but is spread widely. On the other hand, words such as city names, names of local soccer teams, and regional dialects show spatial locality. Figure 1 shows the tweet frequencies of the terms "story" and "Busan" in bars of 500x500 m^2 grids over the map of Korea. The term "story" appears with similar frequencies at many locations, while the term "Busan" which is the second largest city at a diagonal opposite corner from Seoul has the peak frequency coinciding with the actual location of the city.

4.1 Probabilistic Model

Recently, Backstrom *et al.* have proposed a generative probabilistic model that estimates a search query's physical location [1]. Their work is based on Yahoo!'s search query log. When a user issues a query, the search engine logs the query along with the user's IP address. If a query is *local* in nature, the query is likely to map to a single location near the IP's geolocation. If the query bears no strong relevance to a specific geographic location, then the query is not easy to be pinned down to a location. Backstrom *et al.* parameterize the query's geographic distribution with a focus and a dispersion and estimate them using a maximum likelihood approach.

Cheng *et al.* uses Twitter text contents to infer user locations at the city-level granularity [3]. A tweet is often a sentence or more with multiple words and just as in Backstrom's case not all queries or words have geographic relevance. Cheng *et al.* augments Backstrom's approach with classifiers in local word selection and smoothing.

Below we present a quick sketch of the probabilistic model that underline both approaches. The model posits that every word has a center, away from which the frequency decays fast. Let S_j and $\overline{S}_j$ be the set of tweets that contain the word j (or of queries indexed by j) and its complementary set, respectively. The distance d_{ij} is between the GPS tag of the tweet i and the center of the word j. Then, the likelihood function f is defined as:

$$f(C_j, \alpha_j) = \sum_{i \in S_j} log(C_j \times d_{ij}^{-\alpha_j}) + \sum_{i \in \overline{S}_j} \log(1 - C_j \times d_{ij}^{-\alpha_j}).$$

where a constant C_j represents the frequency of the word j at the center, and an exponent value α_j determines the dispersion of word j from the center. Backstrom *et al.* prove that $f(C, \alpha)$ is concave for both C and α, which guarantees $f(C, \alpha)$ to have exactly one local maximum over its parameter space. A large value of α determines a quick decay away from the center and thus represents high locality near the center.

5. GEOGRAPHIC DISTRIBUTION INFERENCE OF WORDS

At the end of Section 3 we are left with 801,505 words from 22,525 users' tweets for ground-truth building. First, as the center of a word, we use the center of mass of all the GPS locations of the word's tweets. Then we use the probabilistic model presented in Section 4.1 and compute the foci and dispersions of the words. When computing the focis and dispersions, we bin the distance between the GPS coordinates and the word's center by 500 m for computational scalability. In Table 1 we list the top 10 most frequently used words and their α values, latitudes and longitudes. Of the 10 listed words in Table 1 only one word, Gangnam, has α greater than 0.1. It refers to a a district in Seoul of about 40 sq km with half a million residents. Yet its geographic locality of use on Twitter is not confined.

Word	α	Latitude	Longitude
today	0.022	37.09506	127.24336
footprint	0.023	36.92209	127.38050
here	0.022	36.91559	127.38603
human	0.027	37.12846	127.22830
time	0.026	37.08744	127.23731
child	0.030	37.12474	127.21842
think	0.030	37.11521	127.22400
Gangnam	0.110	37.49431	127.03721
coffee	0.042	37.17903	127.20952
we	0.030	37.09761	127.24190

Table 1: Top 10 most used words, their α values, latitudes and longitudes

In Figure 2 we plot α versus words in decreasing order of α values. About 450 words have $\alpha > 0.4$. Once α drops below 0.3 beyond 1,000 ranked words, the decline is moderate until the very end. From the figure we conclude that using the top 1,000 words should provide enough information about spatial locality of textual contents.

We list the top 10 words with the highest α values and show them in Table 2. A quick look gives us the sense that the words are likely to be very local, as they all refer to

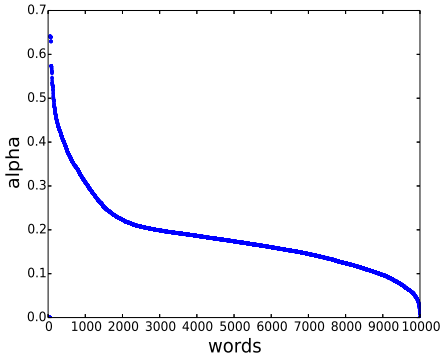


Figure 2: α versus words

Word	α	Latitude	Longitude
Resource	0.64	36.82286	127.18446
Seoul University Station	0.64	37.48034	126.95345
SKK University	0.54	37.33564	126.97723
Seoul Bus Terminal	0.52	37.50529	127.00784
Central-city (name of building)	0.42	37.50442	127.00385
Gimpo airport	0.42	37.56154	126.80487
Samsung C&T	0.38	37.45751	127.03466
Coex	0.37	37.51170	127.05890
Gangnam branch office	0.35	37.50370	127.01677

Table 2: words with high α values

cities, station names, and universities, except for one word "Resource". The word happens to map to a Korean city by the name of Cheon-an. It has many companies that deal with scrap metal and recycling and thus the high α is justified. In [1] queries with a high value of α have shown great locality, while the contribution of C is less pronounced in comparison. In [3] they have chosen Bayesian classifiers and identify 3,183 words as local.

In order to evaluate how *local* they are in our case, we manually inspect the top 1,000 words and pick 712 words easily identifiable to be local. Those words refer to mostly cities, station names, and universities. For those words, we obtain their GPS coordinates from the Google Map API and compute the difference between the center from our approach and the Google Map coordinates. Figure 3 is the cumulative distribution of the differences. Among those 712 words, over 70% of the estimated centers fall within 10 km of their actual locations according to the Google Map. As most Korean cities are larger than 20 km in width, the accuracy lies at a finer granularity than the city level.

6. USER LOCATION INFERENCE

So far, we have used data from all of the 22,525 users in order to evaluate the accuracy of the geographic distribution inference of words. In this section we use the five-fold approach to build the geographic distributions of words and evaluate the quality of our user location inference method.

First we begin with the evaluation of our own method, in particular, decisions made at each step. There are two factors that contribute to the quality of the probabilistic model of local words: the vocabulary and coordinate trans-

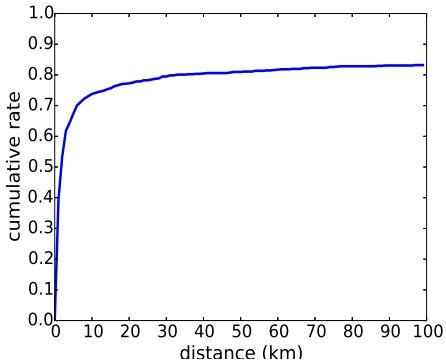


Figure 3: Cumulative distribution of the difference between words' centers and GPS locations from the Google Map

Method	Average error distance (km)	10km (%)	30km (%)
Baseline #1	73.0	0.002	0.10
Baseline#2	31.7	0.44	0.71
user's own words	**26.9**	**0.57**	**0.82**
friends' words	58.1	0.40	0.62
Cheng	57.6	0.35	0.73
Jurgens	57.4	0.33	0.66

Table 3: word location estimator results

formation. How much accuracy degradation do we see if we use less selective vocabulary of local words? How important is to compute the distance and the center of mass between coordinates in latitudes and longitudes?

In order to evaluate the importance of local vocabulary, we use the word distributions of all the words from a user in inferring the user location and call the method Baseline #1. It means we include not only the top $1,000$ words with high α values but all $801,505$ words. We state that it is the worst-case scenario for the case. Next, we use the top $1,000$ words for user location inference, but do not employ map project when computing the user location as the weighted center of mass of words. We label this method Baseline #2. When computing distance between two pairs of latitudes and longitudes, we use the haversine transformation. When computing the center of mass among sets of latitudes and longitudes, we have a choice among a straightforward numerical median (called Manhattan transformation) and the popular Transverse Mercator transformation, just to name a few. Not including the latter transformation in both baseline methods, we can evaluate the contribution from the transformation in our method's accuracy.

In order to compare the performance of our method to that of others, we select two studies. we select studies of Jurgens *et al.* and Cheng *et al.* because they are the latest study of a location estimator in Twitter and the most analogous study to ours, respectively. We apply all of the methods to the same data (Korean Twitter data that we crawled) and compare our method's performance to that of the others. Table 4 shows average error distance and the performance of word location estimators. "baseline1" is an estimator calculating each user's center of mass using all words and not implementing map projection; Also, "baseline2" is an estimator using only local words and not implementing map projection. The baseline1 estimator placed only 0.002% of users within 10km and 0.10% of users within 30km. The results inevitably show a low performance because not all the words are location-related. On the other hand, estimators "user's own words" and "friends' words" show higher performance compare to the baseline estimator. The two estimators use only local words and calculate the center of mass with map projection. Performance gain was about 57% within 10km compared to baseline1 method. This means filtering local words and implementing map projection apparently improve performance.

Our Method	Average error distance (km)	10km (%)	30km (%)
baseline1	73.0	0.2	10
baseline2	31.7	44	71
user's own words	**26.9**	**0.57**	**0.82**
friends' words	58.1	40	62
Other Methods	Average error distance (km)	10km (%)	30km (%)
Cheng	57.6	35	73
Jurgens	57.4	33	66

Table 4: word location estimator results

Also, average error distance of "user's own words" estimator is about a half comparing to that of other studies. Figure 4 shows the performance of our methods ("user's own words") and other studies. As shown in Figure 4, our methods outperform the others in all distance sections. The method proposed by Jurgens *et al.* placed 33.3% of users within 10 km, while the method of Cheng *et al.* and ours placed 34.6 and 56.7% of users at the same distance section, respectively. Within 30 km, the rates were 65.8, 73.1, and 81.9%, respectively. These gaps were maintained until the distance section reached 100 km. Also, our method scored over 90% within 70 km.

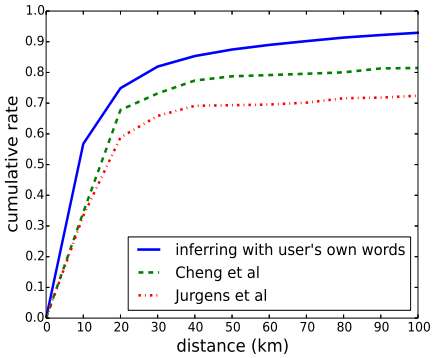


Figure 4: Performance of methods. Our method outperforms the others in all distance sections

Comparing to the work of Cheng *et al.*, our work shows smaller granularity. As their work used per-city word distributions for calculating probability, ours considers each tweet as a calculating point of the probability. U.S. has tens of thousands of cities and method of Cheng *et al.* has less then one hundred thousand points, while ours have 2,783,271

points. In addition, although performance of Cheng *et al.* is higher than that of Jurgens *et al.* in all distance section, the average error distance of these two methods are similar. This tells us error distances using method of Jurgens *et al.* are relatively smaller than that of Cheng *et al.* in large distance section.

7. SUMMARY OF OUR ALGORITHM

At the end of Section 3 we are left with 801,505 words from 22,525 users' tweets for ground-truth building. First, as the center of a word, we use the center of mass of all the GPS locations of the word's tweets. Then we use the probabilistic model presented in Section 4.1 and compute the foci and dispersions of the words. When computing the focis and dispersions, we bin the distance between the GPS coordinates and the word's center by 500 m for computational scalability.

Then we pick the top 1,000 words with the highest α values. In [1] queries with a high value of α have shown great locality, while the contribution of C is less pronounced in comparison. In [3] they have chosen Bayesian classifiers and identify 3,183 words as local.

In order to evaluate how *local* they are in our case, we manually inspect the top 1000 words and pick 712 words easily identifiable to be local. Those words refer to mostly cities, station names, and universities.

Next, we take the probablistic generative model from [**?**] for the spatial variation of queries, and compute the focii and dispersions for each word [3].

for identifying local words. We filter the top 1000 local words among 2,783,271 Korean tweets. Our method consider all GPS-tagged tweets as a point for calculating probability $C_j \times d_i^{-\alpha_j}$ in likelihood function. We multiply the probability $C_j \times d_i^{-\alpha_j}$ to the likelihood function f_j if a tweet i have a word j and $1-C_j \times d_i^{-\alpha_j}$ otherwise. Since millions of GPS-tagged tweets are used to optim ize likelihood function, summing all the log transformed probability up to likelihood function would be computationally expensive. Instead, we put the calculated distances into 500-meter intervals, called bins. In this way, we can reduce computational cost without losing granularity. Computational cost of our method will be depend on the size of bins. We next calculate the center of mass for finding the center of each word. All geotagged tweets containing a particular word act as a part of the weight.

To sum up, user location in Twitter is determined as described in Algorithm 1. Line 1∼8 describe calculating the center of each word; line 9∼16 describe calculating C and α of each word; line 17∼24 describe calculating the center of each user.

With algorithm 1, we infer location of a user who has at least one local word. However, since not all the users in Twitter use local words in their tweets, our method cannot be applied to users who do not use any local words. Thus, we consider another method by taking into account users' friends to alleviate our limitation. We infer location of users who do not have any local words with their friends' local words (only friends follow the user back). In this case, the accuracy of the estimator may decrease because the locations of particular user's friends are not the same as those of the users. Also, Twitter users often follow people who are not physically nearby. The results are demonstrated in section 5.

Algorithm 1 Find location of users

INPUT: U,W and S_j : set of users, words, tweets containing the word j
OUTPUT: Location of users

Require: tweet i contains g_i and at least one word
1: g_i : GPS formatted location
2: **for** $j \in W$ **do**
3: **for** $i \in S_j$ **do**
4: $c_i \leftarrow m(g_i)$ // $m(g_i)$: map projection function
5: **end for**
6: $c_j \leftarrow \frac{\sum_{i \in S_j} c_i \times freq_{ij}}{\sum_{i \in S_j} freq_{ij}}$
7: $g_j \leftarrow m'(c_j)$ // $m'(c_j)$: reverse of function m
8: g_j : the center of the word j
9: **for** $i \in S_j$ **do**
10: $f_j(C_j, \alpha_j) = f_j(C_j, \alpha_j) + \log(p_i)$
11: **end for**
12: **for** $i \in \overline{S}_j$ **do**
13: $f_j(C_j, \alpha_j) = f_j(C_j, \alpha_j) + \log(1 - p_i)$
14: **end for**
15: Find C_j and α_j that maximize $f_j(C_j, \alpha_j)$
16: **end for**
17: $L \leftarrow$ the top 1000 words which have high α value
Require: user k has at least one local word
18: **for** $k \in U$ **do**
19: **for** $j \in L_k$ **do**
20: $c_j \leftarrow m(g_j)$
21: **end for**
22: $c_k \leftarrow \frac{\sum_{j \in L_k} c_j \times freq_{jk}}{\sum_{j \in L_k} freq_{jk}}$
23: $g_k \leftarrow m'(c_k)$ // g_k : center of the user k
24: **end for**

8. FRIENDSHIP-BASED LOCATION INFERENCE

In order to infer the location of users who have none of the top 1,000 local words, we resort to their social network. In this work we define a friend of a user with whom the follow-following relationship reciprocally.

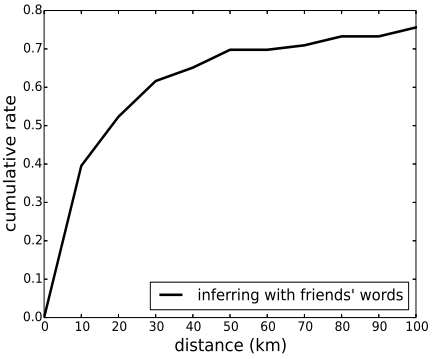


Figure 5: Performance of inference using friend's local words. This method uses friendship information in twitter.

Additionally, in order to estimate the location of users who do not have any local words, we apply our method to them with only their friends' local words("friends' words in Table 4 and Figure 5). Inferring locations with friends' local words also enables acceptable performance, although the performance is lower than that of inferring with user's own words, as we surmised. In comparison with the method of Jurgens *et al.*, which also used users' relationship with friends for estimating user location, the performance of the estimator using friends' local words falls behind only within the range of 15 km to 50 km. However, we note that the method of Jurgens *et al.* covers almost all users in the network, while ours only predicts users who use at least one local word in their tweets including their friends' tweets. Future work will consider a fusion method that reflects the topology of users on a social network graph to select local words for estimators.

To compare spatial distributions of words in different linguistic culture, We are also interested in investigating language dependency of our method. In addition, the size of a country can significantly influence the performance of our method since the distortion of map projection depends on the size of the total area. Thus, future work will consider applying our method to other countries which have different language, culture, and territory size.

9. CONCLUSIONS

A large proportion of Twitter users deliberately leave out their location information, incorrectly fill their location information on the profile, or disable the GPS function on their devices. Yet people tweet about movies they watch, restaurants they visit, and views they enjoy, insinuating their whereabouts. In this paper we propose a user location inference method for Korean Twitter users. Based on GPS-tagged tweets, we first build geographic distributions of words, and then compute the user location as a weighted center of mass from the user's words. Binning the distance in 500 m unit makes our approach computationally scalable. Converting latitudes and longitudes to 3D Euclidean-space coordinates also saves much time in computing the center of mass among GPS tags of words. In comparison with two other approaches, our method show improved accuracy and places 56.7% of users within 10 km of their main locations.

According to Frank *et al.* People move about but mostly "spend the vast majority of their time near two locations" [**?**]. In order to take multiple centers of activities into consideration, we should first cluster GPS tags and then for each cluster investigate the geographic distributions of text contents. We leave this for future work.

It remains to develop a hybrid estimator for higher precision of predictions including network-wise information of the social ties of users. Also, we are interested in applying our method to other countries which have different language, culture, and territory size from our dataset.

10. ACKNOWLEDGMENTS

This research was funded by the MSIP(Ministry of Science, ICT & Future Planning), Korea in the ICT R&D Program 2013

11. REFERENCES

[1] L. Backstrom, J. Kleinberg, R. Kumar, and J. Novak. Spatial Variation in Search Engine Queries. In *Proceedings of the 17th International Conference on World Wide Web*, pages 357–366. ACM, 2008.

[2] L. Backstrom, E. Sun, and C. Marlow. Find Me If You Can: Improving Geographical Prediction with Social and Spatial Proximity. In *Proceedings of the 19th International Conference on World Wide Web*, pages 61–70. ACM, 2010.

[3] Z. Cheng, J. Caverlee, and K. Lee. You Are Where You Tweet: A Content-Based Approach to Geo-Locating Twitter Users. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 759–768. ACM, 2010.

[4] J. Ding, L. Gravano, and N. Shivakumar. Computing Geographical Scopes of Web Resources. 2000.

[5] M. R. Frank, L. Mitchell, P. S. Dodds, and C. M. Danforth. Happiness and the Patterns of Life: A Study of Geolocated Tweets. *Scientific reports*, 3, 2013.

[6] S. goo Lee. KKMA : A Tool for Utilizing Sejong Corpus Based on Relational Database. *Journal of KIISE : Computing Practices and Letters*, 16(11):1046–1050, 2010.

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

[8] D. Jurgens. That's What Friends Are for: Inferring Location in Online Social Media platforms Based on Social Relationships. In *Seventh International AAAI Conference on Weblogs and Social Media*, 2013.

[9] K. Lee, J. Caverlee, and S. Webb. Uncovering Social Spammers: Social Honeypots+ Machine Learning. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 435–442. ACM, 2010.

[10] A. K. F. Melanie A. Rapino. Mega Commuters in the U.S. In *Association for Public Policy Analysis and Management Conference*, 2013.

[11] T. Pontes, M. Vasconcelos, J. Almeida, P. Kumaraguru, and V. Almeida. We Know Where You Live: Privacy Characterization of Foursquare Behavior. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 898–905. ACM, 2012.

[12] A. Sadilek, H. Kautz, and J. 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.

[13] L. Wang, C. Wang, X. Xie, J. Forman, Y. Lu, W.-Y. Ma, and Y. Li. Detecting Dominant Locations from Search Queries. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 424–431. ACM, 2005.