Location Inference using Microblog Messages

Yohei Ikawa
IBM Research – Tokyo, IBM Japan, Ltd.
1623-14, Shimotsuruma, Yamato
Kanagawa, Japan
yikawa@jp.ibm.com

Miki Enoki
IBM Research – Tokyo, IBM Japan, Ltd.
1623-14, Shimotsuruma, Yamato
Kanagawa, Japan
enomiki@jp.ibm.com

Michiaki Tatsubori
IBM Research – Tokyo, IBM Japan, Ltd.
1623-14, Shimotsuruma, Yamato
Kanagawa, Japan
mich@jp.ibm.com

ABSTRACT

In order to sense and analyze disaster information from social media, microblogs as sources of social data have recently attracted attention. In this paper, we attempt to discover geolocation information from microblog messages to assess disasters. Since microblog services are more timely compared to other social media, understanding the geolocation information of each microblog message is useful for quickly responding to a sudden disasters. Some microblog services provide a function for adding geolocation information to messages from mobile device equipped with GPS detectors. However, few users use this function, so most messages do not have geolocation information. Therefore, we attempt to discover the location where a message was generated by using its textual content. The proposed method learns associations between a location and its relevant keywords from past messages, and guesses where a new message came from.

Categories and Subject Descriptors

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

General Terms

Experimentation

Keywords

Microblog, geolocation estimation, text analysis

1. INTRODUCTION

To sense and analyze disaster information from social media, there is high potential in microblogs used as social sensors. A microblog is a service in which users are able to broadcast short messages, generally less than 200 characters, much more easily compared to such social media as blogs and social networking services. Combined with the popularity of easy-to-use smartphones, people have begun sending messages more actively using microblog services, not only from home or office, but from trains or outdoors. Because of these factors, microblogs are more timely compared to other social media, and there have been attempts to utilize them as social sensors for disaster detection as well as location-based marketing and recommendations.

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

In this paper, we attempt to associate each microblog message with geolocation information. Identifying the location where a message was generated helps understand what incidents are occurring and allows the physical locations to be marked on maps. This solution is expected to be utilized in municipal emergency response centers.

Some microblogging services have a function that attaches latitude and longitude information for the present location to each message as a geotag with mobile devices equipped with GPS locators. However most users choose not to attach geotags to messages because of privacy concerns.

Therefore we attempt to determine the location where the message was generated without a geotag, using only its textual content. Our approach identifies the user location based on messages sent from location-related services, learning the associations between a location and its relevant keywords from past messages, and estimates where each new message was created by comparing the similarities between the keywords of the training data and those of the new messages. Experimental results show the location estimation is more accurate when trained for each user rather than when trained by all of the users.

2. RELATED WORK

Beginning with [4], there have been many studies on microblogs. From the point view of disaster management, microblog as a social sensor has recently attracted attention [7, 8, 9]. Combine microblogging messages with locations is helpful for understanding the impact of the disaster. In this section we focus on some research related to geolocation estimation for microblogs.

Cheng et al. [1] proposed a probabilistic framework for estimating a user's city-level location based on the content of microblog messages. Eisenstein et al. [3] introduced a cascading topic model that jointly identifies words with high regional affinity, geographically coherent linguistic regions, and the relationships between regional and topic variation. Hecht et al. [2] attempted state-level location estimation using a Multinomial Naive Bayes model to classify user locations. Our approach for location estimation is also based on textual content as in these approaches, but it differs since they estimate the user's residence whereas we estimate the location where the message was issued.

Kinsella et al. [5] attempted to predict the location of an individual message by building language models of locations using coordinates extracted from geotagged data. The purpose of estimating the location where the message was issued

is the same as our approach, but we do not use geotags for the location estimation.

3. MICROBLOGGING AND LOCATION SERVICES

In this section, we describe Twitter (http://twitter.com/), which is one of the most popular microblog services, and some location services that leverage our approach to identify each user's current location from the microblog messages.

3.1 Twitter

Twitter is one of the largest microblog services with more than 100 million active users around the world. It sends a short message called a "Tweet" broadly or to specified users (up to 140 characters), or forwards a message sent by another user [6]. Many users of Twitter actively send messages everywhere they go such as home and office and even from restaurants or outdoors, since Twitter works well with smartphones. Therefore Twitter messages are quite timely compared to other social media such as blogs and social networking services.

Twitter also provides an option that attaches latitude and longitude information for the present location to a message in the form of a geotag for mobile devices equipped with a GPS sensor. However the number of messages with geotags is actually quite few to date, since geotags are only attached to messages when a user explicitly activates the geotag function of Twitter, and many users are concerned about their privacy.

3.2 Location Services

"Location services" is a generic term for services based on the user's present location based on mobile device equipped with GPS hardware, and among these services, we here focus on services for sending the present location broadly. Foursquare (https://foursquare.com/) and Loctouch (http://tou.ch/) are popular examples in Japan. In particular, Foursquare is one of the most popular location service with 10 million users worldwide (http://blog.foursquare.com/2011/06/20/holysmokes10millionpeople/).

Most of these location services work with Twitter. When a user submits a current location to the service, the location information is automatically spread via Twitter at the same time. Though many users are concerned that their current locations are always broadcast, a certain number of users are willing to publish their current locations and the number of users who use these location services is increasing.

Tweets from location services have specific formats for each service. The message formats of two leading location services are shown in Table 1. By pattern matching each tweet, the user's current location is tracked, since these location services generate tweets with specified message formats.

Table 1: Message formats generated from location services

Location Service	Message Format
Foursquare	I'm at [Location] ([Address]) [Comment] (@ [Location])
Loctouch (in Japanese)	[Location] にタッチ！ [Comment] @ [Location] にタッチ！

4. LOCATION ESTIMATION FOR MESSAGES

In our work, we are attempting to estimate where the message was issued using only its textual context without any geotags. Some of the messages expressing a user's current status contain the user's current location, such as messages from location services, but many of the messages do not contain specific location data. Therefore, we attempt to determine the location even if the message does not contain the user's current location information by learning associations between each location and its relevant keywords from past messages, and then estimating where each new message was generated.

The new location estimation method described in this paper has two steps: the training step and the location estimation step. It learns associations between each location and its relevant keywords from past messages during the training, and it estimates the location where a new message was issued in the location estimation step.

4.1 Training

First, the training dataset is selected from the past messages. On microblog services of the Twitter type, there are many messages communicating with friends and forwarding information sent by other users, but these messages are not always intended to report their current status at that time. For example, a message that is forwarded by another user who thinks it is interesting for many users is unrelated to the user's status. Such messages are excluded from the training dataset, since most of these messages do not have any relevance between the content and the user's current location

Next, our system classifies each message in the learning dataset into two types: a location that represents the user's current location, and an expression that does not represent the user's current location but only the user's current situation. A message generated from a location service is classified by its location type. By checking if each message matches specific formats such as shown in Table 1, location messages are with high probability eliminated from the training dataset. Also, it is possible to extract place names in a message by using Named Entity Recognition (NER) technology. However NER accuracy is less than the classification based on a location service, since even when a place name appears in a message the user is not necessarily at that location. Messages that are not locations are classified into the expression category.

Finally, for each expression message, a location message is associated with it, as long as the location message was created close to time t of the expression message. When multiple location messages exist in an interval t, the chronologically closest one is selected. Then a list of keywords that contain the expression message is generated, and the list is associated with the location. Through the training step, a set of keyword lists for each expression message associated with a location is generated.

4.2 Location Estimation

In the location estimation step, the location where a new message was issued is estimated based on the training results. A keyword list is generated for each new message by text analysis, and it is compared with each keyword list in the training result. The keyword list with the highest co-

sine similarity is selected from the training results, and the location associated with it is the estimated location.

5. EXPERIMENTATION

In our experiment, we evaluated the effectiveness of the location estimation method described in Section 4 for Twitter users who also use location services. We adopted two approaches in the training step in our method: training for each user, and training for all of the users. For comparison, we used a naive method as a baseline, where the estimated result is always the place from which the user has issued the largest number of messages in the past. In this experiment, we used data collected using public APIs provided by Twitter, Inc.

5.1 Data

Here is the data preparation for our evaluation. First using the Twitter APIs, we identified users who have sent messages with any geotags using Twitter APIs. Our approach does not use geotags for location estimation, but we selected these users to precisely evaluate the distance errors between the estimated and actual locations by calculating the distances between the pairs of points. We identified the users who had sent messages with geotags from a dataset called "Spritzer" that was obtained with the Twitter Streaming API (https://dev.twitter.com/docs/streaming-api/methods). According to the API documentation, the sampling rate of the dataset is about 1%. For these users, we obtained up to 3,200 messages that they had sent without sampling. Then we used the patterns shown in Table 1 to identify the location-service users. We identified 110 users from a dataset collected using the Twitter Streaming API from November 15 to 23, 2011. To exclude messages not expressing the user's current situation, we identified the Reply and Retweet messages using textual pattern matching, and excluded them from the target dataset.

Next, we classified the target dataset into training data and query data for the evaluation. After the training phase described in Section 4.1, any expression message with a geotag not associated with any location was used as query data. Here, the threshold time t to associate an expression and a location was 10 minutes in the learning step. The numbers of training and query data messages were 12,463 and 20,535, respectively. We estimated a location for each query data message, and evaluated the distance error by calculating the difference between the estimated result and the geotag attached to the query data.

5.2 Experimental Results and Discussion

We calculated the precision for three location estimation approaches: training for each user, training for all of the users, and the baseline. Here, precision means the accuracy of the location estimation when a distance error within 0.5 km, 1.0 km, 3.0 km, 5.0 km, 10 km, or 30 km is accepted. Figure 1 shows the relationships between distance error and precision.

In the figure, the precision of the training for each user is better than the other methods when the distance error is small, and the precision of the baseline approaches that precision as the distance error increases. The results show that the baseline can roughly estimate the user's residence location within a certain accuracy, but the location estimation based on the message body is efficient to estimate the user's current location more accurately. They also show that the precision of the training with all of the users is lower than half of the precision of the training for each user.

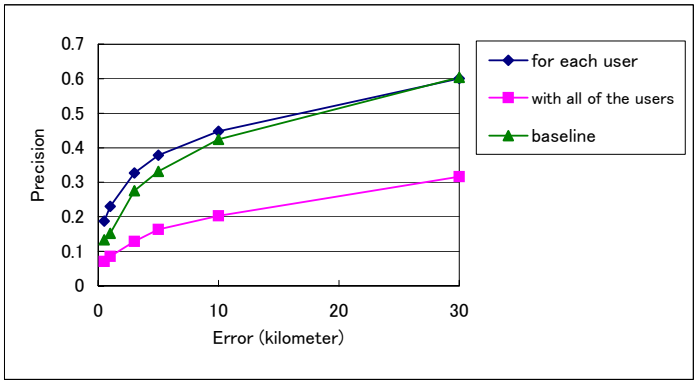

Figure 1: Relationships between distance errors and precision

Table 2 shows the relationships between the distance error and the recall in training for each user and for all of the users. Here, recall is the accuracy rate of the location estimation for all of the query data within the range of each distance error. Out of the 20,535 query messages, the numbers of attempted estimates were 7,658 for training for each user, and 16,380 for training with all of the users. The all-user training attempts to estimate more than twice the number of query messages than the each-user case, but the recall of the all-user case is slightly lower than the each-user case because all-user case fails to make an estimate for many of the messages.

One of the reasons why the precision is decreased by half in training by all of the users may be shortage of training data. However, the wide gap in these two approaches indicates that an approach to training by each user is efficient for this task of estimating the user's current location. For example, when a message "I'm working in the office." is sent, the user's current location is probably the user's office, but the approach of training with all of the users would cause failure for expressions associated with user's specific locations. In the contrast, the approach of training for each user makes it difficult to improve the recall. Therefore, a hybrid approach could enhance both the precision and recall.

Figure 2 shows the distribution of the estimation results sorted in descending order of the correct distance errors within 10 kilometers. The figure shows that the failures for some users substantially reduce the accuracy, but it is possible to improve the accuracy to deal with these cases. in addition, messages that have no context about the user's

Table 2: Relationships between distance errors and recall

	Distance error (kilometer)					
Training Approach	0.5	1.0	3.0	5.0	10	30
for each user	0.07	0.09	0.12	0.14	0.17	0.22
with all of the users	0.06	0.07	0.10	0.13	0.16	0.25

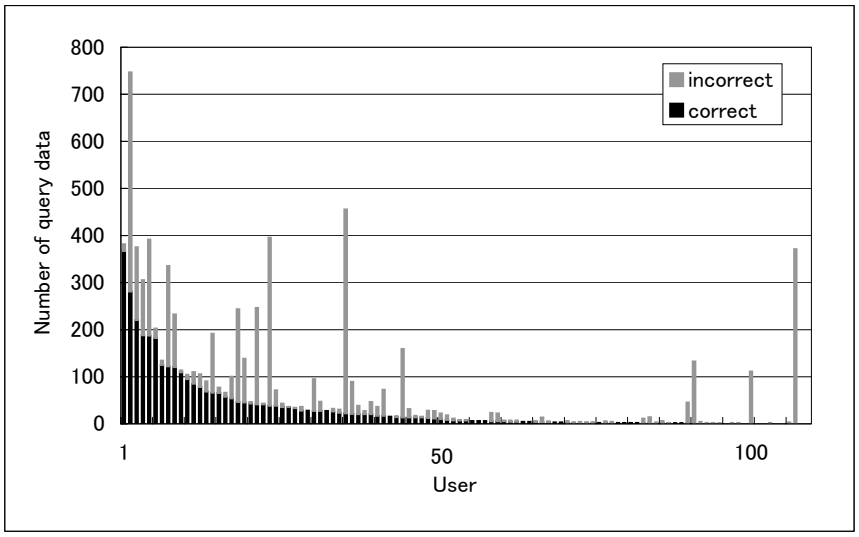

Figure 2: The distribution of estimation results in which the distance error is within 10 kilometers in the training for each user.

current location for accuracy improvement should not be used for estimating.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we examined two approaches for estimation of a user's current location using microblog messages without geotags. We attempted to predict the location by learning the associations between a location and its relevant keywords in past messages. In the experiment, we compared three location estimation approaches, and confirmed that the approach of training for each user has advantages in both precision and recall compared to the other approaches.

Future work includes other approaches for more specific solutions. Also, further experiments with larger datasets, and development of a hybrid estimation approach combining training for each user with training with all of the users should be studied.

7. REFERENCES

[1] Z. Cheng, J. Caverlee, and K. Lee. "You are where you tweet: A content-based approach to geo-locationg Twitter users". In Proceedings of CIKM, 2010.

[2] 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 CHI, 2011.

[3] J. Eisenstein, B. O'Connor, N. A. Smith, and E. Xing. "A Latent Variable Model for Geographic Lexical Variation". In Proceedings of EMNLP, 2011.

[4] A. Java, X. Song, T. Finin, and B. Tseng. "Why We Twitter: Understanding Microblogging Usage and Communities". In Proceedings of WebKDD/SNAKDD, 2007.

[5] S. Kinsella, V. Murdock, and N. O'Hare. "I'm Eating a Sandwich in Glasgow: Modeling Locations with Tweets". In Proceedings of SMUC, 2011.

[6] H. Kwak, C. Lee, H. Park, and S. Moon. "What is Twitter, a social network or a news media?". In Proceedings of WWW, 2010.

[7] T. Sakaki, M. Okazaki, and Y. Matsuo. "Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors". In Proceedings of WWW, 2010.

[8] K. Starbird, L. Palen, A. L. Hughes, and S. Vieweg. "Chatter on The Red: What Hazards Thread Reveals about the Social Life of Microblogged Information". In Proceedings of CSCW, 2010.

[9] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen. "Microblogging During Two Natural Hazards Events: What Twitter May Contribute to Situational Awareness". In Proceedings of CHI, 2010.