The Where in the Tweet

Wen Li
Delft University of Technology
Delft, Netherlands
wen.li@tudelft.nl

Pavel Serdyukov
Yandex LLC
Moscow, Russia
pavser@yandex-team.ru

Arjen P. de Vries
CWI
Netherlands
arjen@acm.org

Carsten Eickhoff
Delft University of Technology
Delft, Neterlands
c.eickhoff@tudelft.nl

Martha Larson
Delft University of Technology
Delft, Neterlands
m.a.larson@tudelft.nl

ABSTRACT

Twitter is a widely-used social networking service which enables its users to post text-based messages, so-called tweets. POI tags on tweets can show more human-readable high-level information about a place rather than just a pair of coordinates. In this paper, we attempt to predict the POI tag of a tweet based on its textual content and time of posting. Potential applications include accurate positioning when GPS devices fail and disambiguating places located near each other. We consider this task as a ranking problem, i.e., we try to rank a set of candidate POIs according to a tweet by using language and time models. To tackle the sparsity of tweets tagged with POIs, we use web pages retrieved by search engines as an additional source of evidence. From our experiments, we find that users indeed leak some information about their accurate locations in their tweets.

Categories and Subject Descriptors: H.2.8[Database Management]Database applications-Data mining J.4 [Computer Application]Social and Behavioral Science

General Terms: Experimentation

Keywords: Twitter, location-based estimation, text mining, geotag

1. INTRODUCTION

Twitter is an on-line communication tool that is situated in between social networks and news media. It allows users to publish messages of up to 140 characters, so-called tweets [5]. Usually a tweet also contains various meta data including the profile of the author, the time of posting, location (coordinates) where the users sent the tweet. In March 2010, Twitter extended its API to provide more accurate geographical information for tweets. Users can specify their locations by tagging a predefined POI (Place of Interest) to their tweets, which includes place name, address, coordinates. However, this service is not yet widely used and there are few good place-aware applications on the market.

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.

To boost the awareness of geolocation related applications, in this paper, we investigate predicting POI of origin based on the textual and temporal information of a tweet. To those who are concerned more about their privacy, this paper aims to raise awareness of that they might leak some information about their location in their tweets.

The task, as we defined, treats places not solely as points located in space, but rather as tags implying the social function of that place. People associate social functions with a place based on why they go there and what they do there. It is our consideration of this semantics that makes our POI prediction more meaningful and better interpretable than mere geo-coordinates. As an example, consider a mall with a food court and a sports store. Both may occupy the same geo-coordinates (on different floors) or nearly indistinguishable geo-coordinates (contiguous in the same building), but for humans it is a relevant distinction whether a tweet is associated with an eating place or with a shopping place. The conventional perspective defines a place as a set of geo-coordinates and is inherently agnostic to this difference. Our work carefully avoids conflation of human-perceived places on the basis of geo-proximity.

In our task, we rank a candidate set of POIs by their relevance to a tweet. Our assumption is that tweets from one place usually follow a certain set of patterns, especially, in vocabulary that can be represented by Language Models. However, we are facing double-level sparsity problems. One aspect is that the terms in a tweet may be insufficient to characterize itself. The other aspect is that most POIs lack enough supporting tweets to build strong models. Therefore, we leverage web pages from search engines in order to mitigate the problem. In addition, we also explore the time dimension to boost the performance of prediction.

2. BACKGROUND AND RELATED WORK

Twitter is a microblog service provider with almost 175 million[1] registered users as of March, 2011. It limits its users to post up to 140 characters per message. Besides, meta data such as POI tags can be attached to tweets. With POI tags, users can share their location information in a more precise way, as POI tags contain more than just a pair of coordinates, e.g., place names and addresses. To support potential place-aware applications, in this paper, we undertake to predict POI tags of tweets by ranking them to

[1]http://www.businessinsider.com/chart-of-the-day-how-many-users-does-twitter-really-have-2011-3

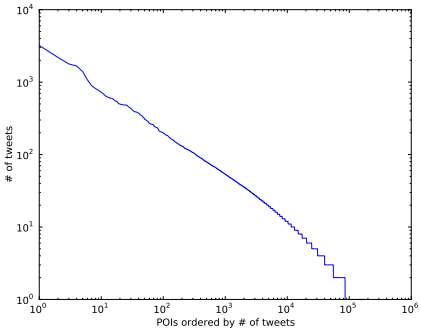

Figure 1: Tweet distribution among POIs (log-log scale)

show the relationship between users' tweets and the places of their origin.

Cheng et al. [1] first looked into this kind of problem and proposed a method using local word identification to estimate a user's location at city-level. Later, Hecht et al. [3] pointed out the location information entered by users was not so accurate as researchers had thought before and proposed a machine learning method of predicting users' home city. Leuski et al. [6] investigated a similar topic based on the chat messages in an on-line game which resulted in a method of predicting events at given virtual locations. Cheng et al.'s and Leuski et al.'s works are based on coordinates which usually suffer from the problem of coarse positioning. Hecht et al.'s and Cheng et al.'s work are locating users' home city not their tweets of origins.

The sparsity of tweets was also pointed out by Cheng et al. [1] who adopted traditional smoothing method for text in tweets. However, the problem is more severe in our task because of fewer tweets tagged with a POI. As shown in Figure 1, the distribution of tweets follows Zipf's law. Only a few POIs (about 0.16%) are supported by more than 100 tweets and 93.11% of POI tags are used less than 10 times in our data set. Due to the limitation of tweet length, it is hard to build strong models for those impoverished POIs using traditional text classification methods, which are typically based on domains that offer numerous documents per category. Therefore, we need other sources to enrich our models for POIs. Sahami et al. [7] proposed a web-kernel similarity measurement which uses web search results to generate strong models for short text snippets, i.e., querying the snippet against search engines. However, our method queries place names instead of tweets themselves.

We also include the time dimension to our models. Like de Jong et al. did in [2], we assume tweeting at a POI is dependent on time. It is intuitive that places have their own hours for visitors, e.g., bars get crowded around midnight and parks are popular on weekends. Therefore, the time patterns found in the set of past tweets from a POI may contain reliable clues as to where new tweets come from.

3. METHODS

The prediction of POI tags can be seen as a ranking problem, i.e., to rank POI tags according to the tweet. We first build unigram language models for each candidate POI and also for the query tweet. Then, we rank those POIs by their KL-divergences with the tweet LM. Accordingly, the higher the reference POI in ranking, the better the ranking method performs.

Because of the sparse distribution of tweets among places, web pages from search engines are used as an external data source. It is reasonable that web pages regarding a place should contain rich information covering the kind of place it is or the kind of activities held there. For example, it may be likely that users at a cinema post tweets regarding the film they have just watched which is also listed on the home page of the cinema. However, it should be noted that posting a tweet from a place is quite different from publishing a web page. As revealed in [4], most tweets are talking about daily routines and what people are currently doing. Web pages regarding a place, on the other hand, aim at describing the functions of the place. Thus, the vocabulary use may be different from source to source. As a result, we decide to score the places separately by models from tweets and web pages and then combine the scores to obtain comprehensive rankings.

Linear combination of scores is an easy way of merging rankings. Since KL-divergence (our ranking score) scales differently with respect to different sources, we first normalize our ranking scores with respect to their own dimensions, i.e., map the scores to $[0, 1]$ and then linearly combine the scores for each POI. Generally, let $\mathbf{X}$ be the score matrix where x_{ij} is the score of POI_i given by the jth criterion. The normalized matrix is $\mathbf{X}' = [x'_{ij}]$ where

$$x'_{ij} = \frac{x_{ij} - \min_j x_{ij}}{\max_j x_{ij} - \min_j x_{ij}}.$$

Our ranking model generates a ranking score using a linear combination of contributions from the component dimensions, $\boldsymbol{\sigma} = \mathbf{X}'\boldsymbol{\lambda}$. Here, λ is a weight vector controlling the contribution of the different dimensions in final rankings. Since we are already dealing with a data sparsity problem, we focus our investigation on the performance that can be achieved without tuning the balance between dimensions. In our experiments, dimensions are all weighted equally. Then, we can rank POIs in a balanced manner taking multiple information source into account.

The time dimension of tweets is an interesting aspect to explore. As mentioned previously, some places have particular activity patterns related to their social functions. The time distributions of tweets from *Exit (Rock Club)* and *Runyon Canyon Park* suggest that the rock club is crowded at late night of working days and the park is more popular on weekends. Therefore, we propose a time model for POIs which combines the time distributions of tweets on different scales of periods. In this paper, we use three scales of periods: day, week and month. To measure the probability that a tweet was posted from the modeled POI at time t, the linear combination is again adopted. Let $\mathbf{P}_i(\mathbf{t}) = [P_i^{(d)}(t)\ P_i^{(w)}(t)\ P_i^{(m)}(t)]$ be the vector of probabilities that a tweet is posted at time t. The score from time models is then the linear combination of these values $s_i(t) = \mathbf{P}_i(t)\boldsymbol{\lambda}$.

To evaluate our ranking method, we use a modified precision curve, as rankings in our task are slightly different from those of other text retrieval systems, i.e., typically only a single label is considered as relevant one (the reference place). For this reason, the precision we choose for our evaluation

is the rate that the reference POI is ranked above pth place. We use curves to show the relationships between the rate and concrete settings of p. The larger the area under the curve, the better the ranking algorithm performs. All results are statistically significant ($p < 0.05$) tested by Wilcoson Signed-rank tests.

4. EXPERIMENTAL FRAMEWORK

For our experiment, we first collect a reasonably large set of tweets with POIs. To this end, the following strategy is used: 1) Retrieve an initial set of tweets from Twitter's stream API[2] and pick out those with POI tags. 2) Aggregate all the users who sent these tweets with POI tags and all the POIs. 3) Crawl tweets from the users and POIs gathered in the previous step. 4) Update the data set with new incoming POI-tagged tweets and 5) Repeat steps 2-5 to expand the dataset. Following this strategy, we collected about 31.6 million tweets by crawling from September 2010 to May 2011. However, there are very few tweets with POI tags and close observation suggests that most tweets with POI tags originate from Foursquare, an on-line geo-information sharing platform. Users check-in to places (i.e., post something with a POI tag on Twitter) to win titles or special treatment. This is reflected by tweets with the pattern "I'm at XXX. http://4sq.com/YYY". As it is trivial to predict the origin of tweets which have the POI names (XXX) in their text, we remove the part of text dedicated to Foursquare games. In the end, we have 700,288 tweets with POI tags from 177,817 POIs posted by 52,488 users in our dataset.

Second, we select four main cities in the USA, namely, Chicago, Los Angeles, New York and San Francisco, since they are home to many popular POIs. Then, we select the 10 most popular POIs per city, including shops, restaurants, parks, cafes and clubs. These POIs are all supported by between 100 and 400 tweets. To build language models for these POIs, a stemming tokenizer with a stop words filter from WHOOSH[3] is used in term extraction.

As stated previously, we use web pages from search engines as an additional source of evidence. For this, we query POI names against Microsoft Bing and gather the textual content of the top 30 returned web pages with HTML tags being filtered out.

5. RESULTS AND DISCUSSION

First, we investigate whether POIs are distinguishable through LMs built from tweets. We split the tweets from each POI into two equally large sets and build an LM for each set. Then, we compare the KL-Divergence between the LMs from the first part of tweets of each POI and that from the second part in a confusion matrix. The confusion matrix for Chicago is shown in Figure 2.

Significant differences can be found between places, which supports our assumption that language models are able to capture the differences in vocabulary. In other words, vocabulary used in most places is significantly more similar among tweets from the same POI than across POIs. However, language models are confused by some of the places, e.g., *AMC River East 21* and *Century Center Cinema* which are both cinemas. For another example, *Lakeview Athletic Club* and *Bally Total Fitness* are both fitness centers. This

[2]http://dev.twitter.com
[3]http://whoosh.ca/

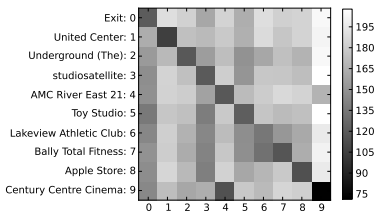

Figure 2: Confusion Matrix for POIs in Chicago

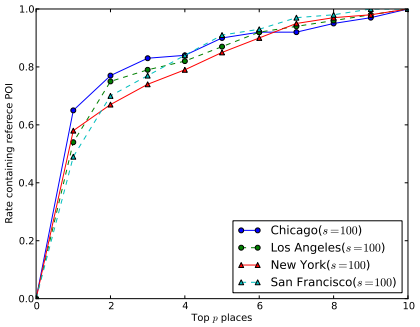

Figure 3: Ranking with rich models

observations leads to the insight that the POIs of tweets are activity dependent. That is, predicting the origin of tweets is actually predicting the place of activity reflected in the tweets.

To test our ranking methods, we conduct the following experiments in each of the four cities. 10 tweets from each POI are randomly selected to compose a test set and are queried against the ranking method based on the textual content of tweets. The results can be found in Figure 3.

The evaluation result seems to be positive in all four cities when we have enough tweets ($s = 100$) to build strong models for POIs. Nevertheless, looking into the tweets with which the ranking methods do not work well, we find: 1) A group of tweets exists whose content is short and not location-specific, such as "Thank you", "yaaaa". These tweets cannot be correctly tagged even by humans. 2) Words with a strong relationship to a place may not appear often enough when the data size is small, e.g., "swim" usually implies fitness, however, it only appears in 3 tweets in the dataset. 3) Tweets are time-related, i.e., words can show up for a while and then disappear, e.g., the title of a movie.

As mentioned, the sparsity may affect ranking performance in our task. Therefore, we shrink the number of tweets s used for model training, so that we can compare the performance of the ranking method at different popularity levels. The results of this modified setting can be found in Figure 4, which shows great degenerations of performance when s drops. Especially, for the case $s = 10$, the ranking behaves like random selection. Therefore, we turn to our additional source of information, i.e., web pages from search engines. We rank the candidate POIs by two sets of models, i.e., one from tweets and the other from web pages and then linearly combine the scores to generate a comprehensive ranking.

In order to show the performance of our web-enriched ranking method, we compare it with rankings based on only tweets and only web pages. The results in Figure 5 indi-

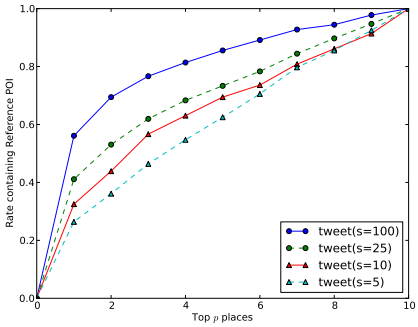

Figure 4: Ranking at different popularity levels

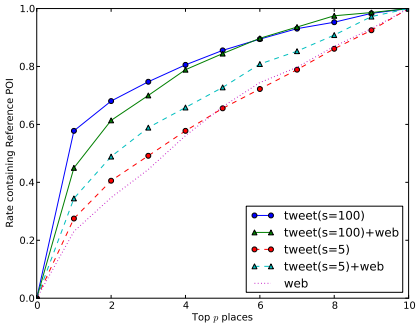

Figure 5: Web-enriched rankings

cate web pages can help increase the performance of ranking when POIs are supported by fewer tweets ($s = 5$) but will harm it somehow in the case that POIs have enough supporting tweets ($s = 100$). For example, *Lakeview Athletic Club* which is ranked at lower positions by the methods based on only tweets is ranked higher by the web-enriched method. An opposite example can be found on *Nokia Theatre* which is better ranked by pure tweets, as most of tweets from the theatre are talking about shows there while the web pages from the search engine include only big flash objects (its home page), another related theatre (*Best Buy Theatre*), and introduction (Wikipedia). Therefore, careful tuning of the weighting parameters is needed to suppress bad effects from noisy resources.

The time dimension also boosts performance when there are few supporting tweets for modeling candidate POIs. Figure 6 shows the evaluation of the time-enriched ranking method compared with other combinations of modeling techniques. As we can see, the time model can boost performance on both rich POIs and web-enriched POIs. However, the time model is also affected by sparsity problems, therefore it cannot substantially increase ranking performance. By looking into the POIs that the model does not work well with, we find that some POIs like *In & Out Burger* and *Best Buy* are always busy and cannot be characterized by the time model.

6. CONCLUSION AND FUTURE WORK

In this paper, we have shown the viability of applying a ranking approach to the prediction of the POIs of tweets' origin. Using a language modeling method, we can achieve good performance given enough POI-tagged tweets for building models. For those POIs not associated with sufficient tweets, using web-enriched models can significantly boost ranking performance. However, the web-smoothing ranking method shows its own limitation when ample number of tweets are available, presumably since the detrimental contribution of mismatched web vocabulary overwhelms the positive contribution of the tweet vocabulary. As to the time dimension, in spite of the sparsity problems, time models are more stable in boosting the performance. In general, we can conclude that POI-tagged tweets have strong relationships with their place of origin.

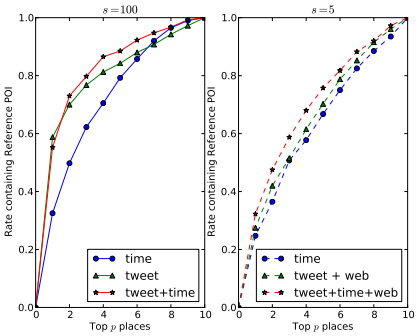

Figure 6: Ranking with comprehensive models

In the future, we will examine whether there are enough data available to effectively train the $\boldsymbol{\lambda}$ parameters balancing the dimensions. We will further explore other features to boost the performance of place prediction. For example, users' friendships may also imply the place they like going to.

7. REFERENCES

[1] Z. Cheng, J. Caverlee, and K. Lee. You are where you tweet: a content-based approach to geo-locating twitter users. CIKM '10.

[2] F. de Jong, T. Westerveld, and A. de Vries. Multimedia search without visual analysis: The value of linguistic and contextual information. *Circuits and Systems for Video Technology, IEEE Transactions on*, 17(3).

[3] 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. CHI '11.

[4] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. WebKDD/SNA-KDD '07.

[5] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? WWW '10.

[6] A. Leuski and V. Lavrenko. Tracking dragon-hunters with language models. CIKM '06.

[7] M. Sahami and T. D. Heilman. A web-based kernel function for measuring the similarity of short text snippets. WWW '06.