The 5th International Workshop on Information Quality and Quality of Service for Pervasive Computing 2013, San Diego (22 March 2013)

Spatio-Temporal Provenance: Identifying Location Information from Unstructured Text

(Invited Paper)

Kisung Lee, Raghu Ganti, Mudhakar Srivatsa
IBM T. J. Watson Research Center
Email: kslee@gatech.edu, rganti, msrivats@us.ibm.com

Prasant Mohapatra
Department of Computer Science, UCDavis
Email: prasant@cs.ucdavis.edu

***Abstract*—Spatio-temporal attributes represent two aspects of physical presence - space and time - which are integral to human activities. Space-time markers of an entity in conjunction with correlation with other networks such as movements in social network, the road/transportation network encodes a wealth of *provenance* information. With the advent of mobile computing and cheap and improved location estimation techniques, encoding such information has become commonplace. In this paper, we will focus on deriving such *location provenance* information from unstructured text generated by social media. As social media such as Facebook and Twitter are integrated with mobile devices, information generated by individuals in these networks gets tagged with spatial markers. We can classify such markers into explicit and implicit tags, where explicit tags encode the spatial data explicitly by providing the accurate location attributes. On the other hand, a lot of social network data may not encode such information explicitly. Our hypothesis in this paper is that the unstructured textual data contains implicit spatial markers at a fine granularity. We develop algorithms to support this hypothesis and evaluate these algorithms on data from FourSquare to show that the spatial category information can be identified with an accuracy of over 80%.**

I. Introduction

The introduction of mobile smartphones and their integration with social networks has resulted in an explosion of social network data generated by these devices. For example, Facebook, Twitter, and FourSquare are social networks that are smartphone enabled, with FourSquare being explicitly focused on smartphones. The data generated by these devices is primarily unstructured text with additional sensor information, such as location (generated by GPS, WiFi), context (e.g., walking, running), and temperature of environment. As these mobile devices become popular and pervasive, their rich sensing capabilities provide a valuable source of information. For example, the GPS sensor provides exact location information of an individual. Such location sensor data when combined with other contextual information can enable novel applications. Examples of such applications include spatio-temporal localization of events, geo-spatial opinion mining, and automated geographical surveys.

Among these additional sensing capabilities, spatio-temporal attributes that represent two aspects of physical presence - space and time - are integral to human activities. These space-time markers in conjunction with correlation with other networks such as movements in social network, the road/transportation network encodes a wealth of *provenance*[1] information. We coin the term *spatio-temporal* provenance, motivated by the concept of provenance, to refer to the space-time tags (and the corresponding chronology) to a piece of information.

On the other hand, these space-time markers are also sensitive attributes that can violate an individual's privacy. Hence, space-time markers are not usually provided by individuals. For example, we analyzed over a month of Twitter data feeds (10% decahose feeds from Gnip [2]) in 2012 and observed that 99.1% of the tweets are not geotagged (i.e. no space marker available). Given the lack of explicit spatial tags in one of the largest unstructured text based social network feeds, this paper addresses the question of deriving location information from such unstructured text. In this paper, we tackle the problem of deriving spatial tags and leave the temporal inferences to future work.

We specifically address the following hypothesis in this paper - does unstructured text (provided by an individual) act as a good reference to geospatial markers? We focus on deriving fine-grained location information from unstructured text (such as tweets, tips in FourSquare). In the past [4], [5], it was shown that city and region level spatial information can be derived from unstructured text (Twitter) with an accuracy of about 100 miles (from the city of the tweet's origin). We believe that it is indeed possible to construct models at fine-grained locations (e.g., a city-block level) that can be used to derive the spatial provenance from unstructured text. In this paper, we take a first step toward this direction by building text models at fine-grained locations and show that these models are significantly "apart" from each other (using a KL-divergence metric). We evaluate these constructed models on data collected from FourSquare and show that more than 80% of the venues can be differentiated from each other based on textual information. We also build classifiers for identifying the type of location (e.g., food, travel, and outdoors) and show that these classifiers achieve up to 80% accuracy (on data collected from FourSquare).

This work was done when Kisung Lee was a summer intern at IBM. His primary affiliation is with College of Computing, Georgia Tech

[1]Provenance refers to the chronology of the ownership or location of the object/piece of information

978-1-4244-9529-0/13/$31.00 ©2013 IEEE

The rest of the paper is organized as follows, Section II describes the datasets we used for constructing and evaluating our text models. In Section III, we describe the method for constructing the text models (specific to fine-grained locations) and evaluate these models in Section IV. We describe related work in Section V and discuss various directions that we are going to pursue in Section VI. Finally, conclusions are presented in Section VII.

II. Datasets Used

The datasets that we use in this paper are primarily collected from FourSquare [1]. FourSquare is a mobile application that allows an individual to *checkin* at a particular *venue* and then leave a *tip* for that venue. A venue is described by its geographical coordinates (latitude/longitude), the type of the venue (e.g., restaurant, airport), and the name of the venue. A tip is a brief message (unstructured text) left by the individual who checked in at that venue (typically describing that venue or anything associated). FourSquare relies on the location determination of mobile devices to obtain the precise location of the venue.

A few examples of venues and their associated tips are presented in Table I.

We utilize the FourSquare's public API to query for venues and its associated properties (geolocation and tips). Since our goal in this paper is to be able to correlate the unstructured text and the corresponding location in order to be able to build models on the underlying language for location determination purposes, we obtain the tips (unstructured text) and geolocation of venues for our analysis. We note that the reason for utilizing FourSquare data is for validation purposes as they have explicit location and associated unstructured text references (i.e., tips). We obtain these data from the limits of New York city where the timespan of the tips are across four years, which is illustrated in Figure 2. The venues are pictorially depicted in Figure 1, where we observe that the venues are densely distributed in downtown Manhattan.

FourSquare categorizes each venue into one of nine categories based on the *type* of the venue. For example, if the venue is a restaurant or fast food place, it is categorized as *Food*. The number of venues per category is illustrated in Figure 3. We observe that *Food* is the predominant category.

III. Venue Specific Text Models

Our hypothesis in this paper is that different venues have different text models describing them. The intuition is that individuals at a given location leave a trail of text associated with that venue. For example, someone at the Times Square in New York City is likely to "leave" a tip or tweet about Times Square. As observed in Table I, the tips (from FourSquare) about New York Penn Station reflects the information about the venue itself. The references to the station are in different forms, *Seating area says Acela express*, *Really big station*, and *Amtrak train*. A key point is that such references (to the train station) are missing in other venues (Table I), which makes these references unique to the given location. One may

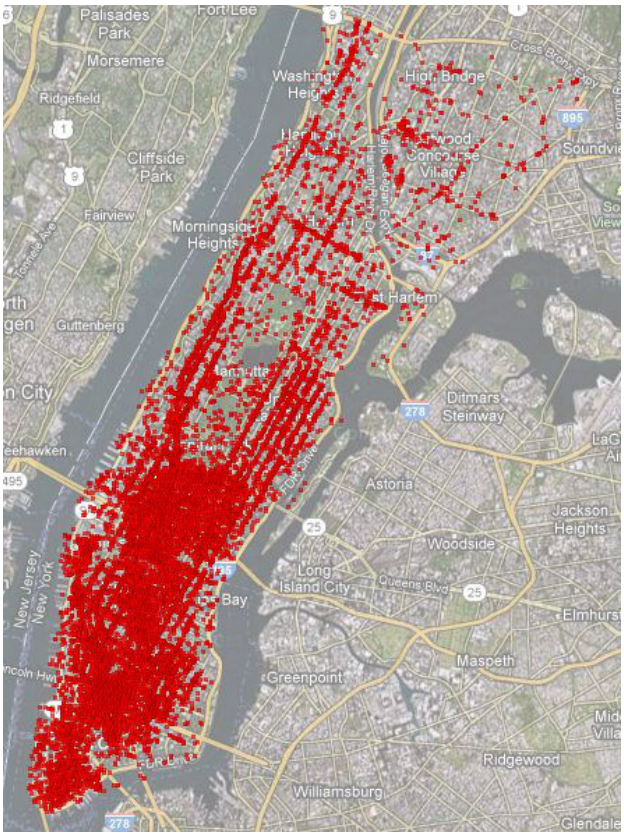

Fig. 1. Map of venues from FourSquare for Manhattan, New York City. A total of 20,034 venues were observed in Manhattan

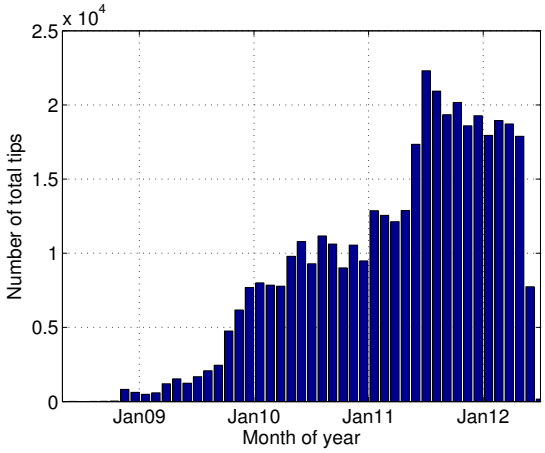

Fig. 2. Cumulative number of tips at venues in Manhattan, New York City over time

observe that there are many such train stations across the United States (or some other large geographical area), but we believe that city level information can be obtained from other sources of information such as the location fields from tweets or using other language models [4]. In this paper, we focus on developing algorithms that can capture the previously presented intuitions in mathematical models and show that these models can differentiate various venues. We also show

Venue	Tips
New York Penn Station	The **seating** area says *Acela express ticket* holders only but that's just for mornings
	Really big **station**, but they don't announce tee **train track** till few min before it **boards**. Not a lot of customer service in there either.
	Instead of waiting on line for your **Amtrak train**, take the stairs directly to the platform from the **NJT** level below.
Metropolitan Museum of Art	Take the elevator in the **European sculpture** and decorative **arts gallery** up to the top and grab a drink at the **roof garden cafe** and martini bar (open fro May through the fall)
	Everyone knows The **Met** is the city's most epic **museum**, with a vast collection from ancient to modern. dont' have to tell you that it is a must see. I love to twirl around the period rooms alone.
	It's tricky to navigate, and overwhelmingly humongous, but that's all part of the **Met's** charm. We love losing ourselves in the miles of corridors and ogline over the many world famous **treasures**.
Magnolia Bakery	Known for their **butter-cream cupcakes** and floral decor, it's a lovely place to grab one or two **desserts** for after dinner.
	Whoopie cookie is the freaking best thing I've ever tasted. Forget the **cupcakes**! They are too sweet, make sure u have water if u eat them
	Get the **red velvet mini cheesecake**, the **lemon bar**, and their **banana pudding**. Thank me later!

TABLE I
TABLE ILLUSTRATING VENUES FROM FOURSQUARE AND SOME EXAMPLES OF ASSOCIATED TIPS WITH KEYWORDS HIGHLIGHTED

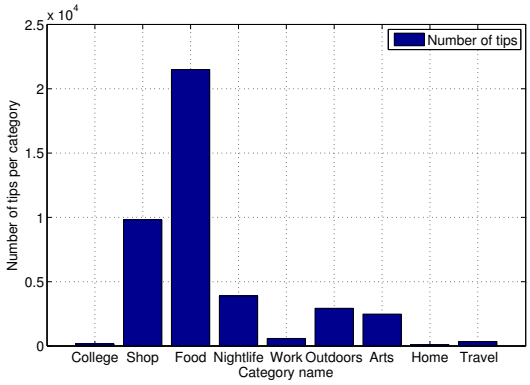

Fig. 3. Cumulative number of tips at venues in Manhattan, New York City by category

that these models can be used to accurately identify category-level venue information.

A trivial modeling method that relies on the keyword approach would be to consider all the keywords associated with a single venue and build a simple model that characterizes the frequency of the keywords at the given location. In more mathematical terms, for each venue i, there are N_i keywords, $K^i = k_1^i, k_2^i \ldots, k_{N_i}^i$ and each has frequency probabilities, $P_i = p(k_1^i), p(k_2^i), \ldots, p(k_{N_i}^i)$. This trivial solution has a clear problem - there will be many occurrences of simple keywords such as *the, food*, and *good* at various venues, which will have high probabilities (at each of these venues) and hence will fail to differentiate the models significantly.

In the past, extraction of location information from unstructured text (e.g., tweets) has been explored [4], [6], [5]. Although, all of these papers tackle the problem of extracting city-level locations and do not address fine grained venue location identification. Our hypothesis is that such fine grained location information can also be extracted, in addition to city-level locations. The challenge in modeling such fine grained locations is that of "determining" the appropriate language models and identifying the "correct" keywords for building such models. We take a two pronged approach, where we first predict location "tags" associated with a venue and then build venue specific models to predict the location. The location tag that we chose is the *category* of the venue, which provides details about the type of the venue and possibly narrowing the scope of search (in terms of predicting the venue). We use a predetermined set of eight categories (for category prediction, excluding *Residence* due to the non-availability of tips) defined by FourSquare and illustrated in Figure 3.

We observe that about 70% of the venues are of type "Food" and hence we build a first-level classifier using a poly-kernel SVM that separates a given tip into "Food" and non-Food categories (building a single classifier on all the eight categories results in poor performance). We build a similar poly-kernel SVM for the second stage classification as well, which predicts the specific category among the remaining seven categories. This algorithm is illustrated in Figure 4.

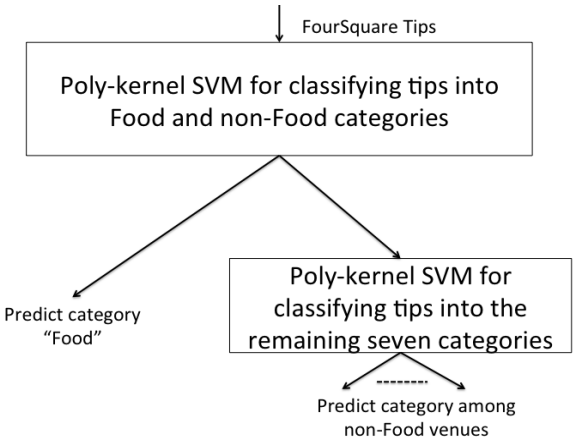

Fig. 4. Algorithm for classifying/predicting the category of the tip/unstructured text

The language model built at each venue follows earlier

approaches [4], [5] and is illustrated in in Figure 5. Our approach relies on smoothing the keywords by adapting the following steps: removing stop words (e.g., the, and) and then filtering infrequent words. On the filtered set of words, we build unigram language models, which will be the venue-specific location models.

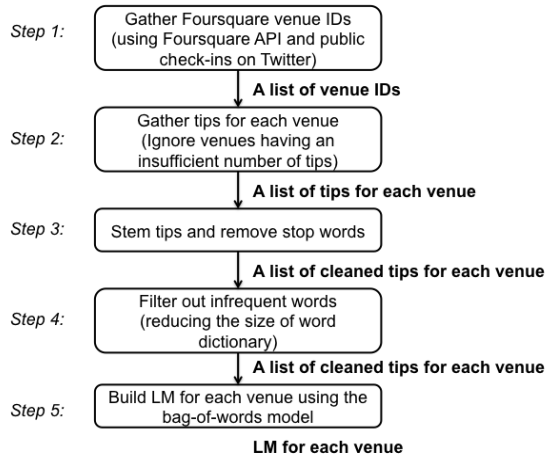

Fig. 5. Illustration of algorithm for construction of venue-specific language models

IV. Evaluation

This evaluation section is divided into two parts, the first evaluates the performance of the poly-kernel SVM classifier for category prediction and the second focuses on the difference between the text models (based on a KL-divergence metric [8]). We use FourSquare data for our evaluation as it provides exact venue information and associated tips (unstructured text).

We consider about 1066 venues in Manhattan sorted based on the number of tips available. In all our model building, we consider equal number of tips for each category and construct the model. Each tip is less than 200 characters in size. First, we built a poly-kernel SVM using 1400 tips across food and no-food categories and use a cross-validation approach to determine the performance of this model. The confusion matrix for this classifier on about 1400 tips (with equal number of food and non-food category tips) is shown in Table II. We observe from Table II that the classification accuracy is about 80%. We use this classifier to determine the accuracy across about 30,000 tips, which is about 78%, showing that the classifier is robust.

Food	Non-food
0.794	0.205
0.183	**0.817**

TABLE II
TABLE SHOWING THE CONFUSION MATRIX FROM THE CROSS-VALIDATION RESULTS ON THE POLY-KERNEL SVM FOR THE CLASSIFICATION OF TIPS INTO FOOD AND NON-FOOD CATEGORIES

We now evaluate the second stage classifier that is used to predict the exact category among the non-food cases. A similar poly-kernel SVM classifier is built for each of the categories and a cross-validation is performed. The number of tips considered for each category are 400 and we present the confusion matrix results in Table III. We observe from Table III that the overall accuracy has dropped to about 60% on an average. We also observe that the majority of the confusion is with the category *Professional*. We ran the classifier built on the remaining set of tips and observe that the average prediction accuracy falls down to about 46%, which is still much better than random prediction (about 14%).

Office	Shop	Nightlife	Travel	Arts	College	Outdoor
0.67	0.065	0.01	0.02	0.025	0.155	0.055
0.215	**0.58**	0.035	0.025	0.045	0.09	0.01
0.18	0.055	**0.54**	0.015	0.17	0.015	0.025
0.18	0.085	0.055	**0.455**	0.065	0.085	0.075
0.19	0.07	0.175	0.03	**0.455**	0.04	0.04
0.31	0.07	0.005	0.04	0.02	**0.515**	0.04
0.24	0.065	0.01	0.015	0.055	0.07	**0.545**

TABLE III
TABLE SHOWING THE CONFUSION MATRIX FOR THE CLASSIFICATION OF TIPS INTO VARIOUS CATEGORIES USING THE POLY-KERNEL SVM APPROACH

Thus far, we have shown that the unstructured text can be used to tag "location" attributes such as the type of the place, which can significantly narrow the *search* scope for the venue location depending on the category. We now evaluate the power of prediction of the language models presented in the previous Section. The metric we use to quantify the predictive power of the language models built is the KL-divergence metric, which measures the non-symmetric difference between two probability distributions. We compute the KL-divergences among the 1066 venues and compute the difference between the KL-divergence of the given venue and the smallest KL-divergence among the remaining venues. The corresponding cumulative distribution function on these differences is illustrated in Figure 6. We observe that if the difference is negative, it would imply that a classifier that can "predict" the venue location does not exist, whereas a positive difference indicates the existence of such a classifier. We observe from Figure 6 that most (about 80%) of the venues have positive KL-divergence difference, suggesting that it is indeed possible to predict the venue locations accurately based on the language models.

V. Related Work

We divide this related work section into three parts, the first looks at deriving fine grained location information from keywords, the second summarizes papers that look at deriving city-level location information from keywords, and finally we look at web-based location determination.

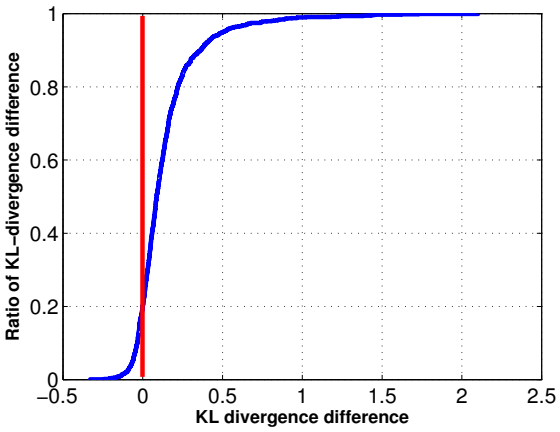

Fig. 6. CDF plot of the difference between the KL-divergence for the given venue and the minimum KL-divergence across all venues (except the given venue)

A. Fine-grained Locations

As far as the authors are aware, this is the first piece of work that tackles the problem of determining fine-grained location information based on keywords. This paper is the first step in this direction and explores the hypothesis of generating unique language models (at a given location). Related to the FourSquare data analysis, in [5], [11], checkins of individual users are obtained from Twitter and are analyzed across a period of nearly four months. These checkins were analyzed to identify hotspots in space and time. Our work on the other hand examines building differentiating language models from the FourSquare venue data. In [8], POIs are ranked based on relevance to a tweet. POI models are built based on past tweets, provided enough tweets at that POI. In contrast, we do not rely on tweets to build models, which can potentially be very sparse [4].

B. City-level Locations

The closest related work and the motivation for our paper is presented in [4]. In the paper, language models are built at a city-level and used to predict such location information from tweets. It was shown that 51% of the Twitter users can be placed within 100 miles of their actual location. City-level topic modeling using tweets was explored in [6]. In the paper, geographical location models were built using a sparse additive generative model, and it was shown that topics at a city level can be tracked provided a large number of tweets are available. In contrast, we aim at being able to determine language models at a venue level, which is far more fine-grained than a particular city.

C. Web-based Locations

A description of various challenges posed for *geoparsing* are presented in [7]. Gelocation of individuals based on their social relationships is presented in [3]. Facebook geolocations are utilized to obtain locations of individuals lacking such information and it is shown to outperform IP-based location determination. In contrast, we are building models from the unstructured text shared in social networks.

An algorithm for identifying geographical intent such as location of city associated with web searches was proposed in [12]. Language models were built on keywords which were associated with individual cities. In contrast, our goal in this paper is to understand the correlation across different datasets for identifying fine grained locations.

Other related work includes building data models to infer locations from social media [10], geographical topic modeling (of large regions, such as cities) based on GPS-tagged documents [13], and moving object data mining [9].

VI. Discussion and Future Work

We have shown thus far that building textual models for various venues (location specific) that can differentiate these venues is indeed possible. In this Section, we will discuss the implications of our hypothesis and the experiments that we performed.

First and foremost, can we improve the textual models? In this work, we have considered unigram language models for capturing the textual model (at a given location/venue). But, this is prone to problems as it fails to capture words that occur together, for example the word *Lion* may indicate a zoo, but the words *Lion King* may indicate Broadway theater (in New York City). Unigram language models capture only single differentiating keywords and fail to capture multiple occurrences. But, it is not immediately apparent if bigram or N-gram language models would indeed improve the performance either. Initial experimental results suggest that there are cases when bigram models outperform unigram models and vice-versa. Hence, one possibility is to combine the unigram and bigram language models to capture different variations in the keywords (at different venues).

One of the goals of our work in building such text models at fine grained locations is to be able to predict accurate locations of individuals who are generating the text. In particular, we are interested in predicting the location information of tweets from Twitter. Twitter, a source of unstructured text social media generates valuable information in the form of individuals' opinions. Twitter introduced geo-location tags for tweets recently, but the adoption of which has been very minimal. In fact, previous work [4] (and our experiments corroborate it) has shown that only about 1% of them had geo-tags associated with them. Hence, the location as an attribute is missing in most of the tweets. Based on the results in this paper, we believe that fine-grained location prediction of individual tweets is indeed possible. Our immediate future work is to build classifiers using the models generated in this paper for determining accurate location information of tweets. Such location prediction techniques will enable new applications such as automated geographical surveys, geo-spatial opinion mining, and locating events in space.

VII. Conclusions

In this paper, we have addressed the hypothesis of - is it possible to predict the location of an individual based on

unstructured text information? We collected FourSquare data, a smartphone application that allows individuals to checkin and leave a tip at the given location to address this hypothesis. We first built models to categorize the type of location an individual is currently present at, which are shown to have an accuracy between about 50%-80%. We then constructed language models for predicting fine-grained location/venue information (based on the unstructured text) and showed that these models have a KL-divergence difference greater than 1 for 80% of the cases, suggesting that a classifier to predict locations (based on the unstructured text) exists. We concluded this paper by discussing various directions that we are currently pursuing to predict locations accurately.

REFERENCES

[1] Foursquare. http://www.foursquare.com.
[2] Twitter decahose. http://gnip.com/twitter/decahose.
[3] L. Backstrom, E. Sun, and C. Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of WWW*, pages 61–70, 2010.
[4] Z. Cheng, J. Caverlee, and K. Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of CIKM*, pages 759–768, 2010.
[5] Z. Cheng, J. Caverlee, K. Lee, and D. Sui. Exploring millions of footprints in location sharing services. In *Proc. of International AAAI Conference on Weblogs and Social Media*, 2011.
[6] L. Hong, A. Ahmed, S. Gurumurthy, A. J. Smola, and K. Tsioutsiouliklis. Discovering geographical topics in the twitter stream. In *Proceedings of WWW*, pages 769–778, 2012.
[7] J. L. Leidner and M. D. Lieberman. Detecting geographical references in the form of place names and associated spatial natural language. In *Proceedings of SIGSPATIAL*, pages 5–11, 2011.
[8] W. Li, P. Serdyuko, A. P. de Vries, C. Eickhoff, and M. Larson. Poster: The where in the tweet. In *Proceedings of CIKM*, pages 2473–2476, 2011.
[9] Z. Li et al. Movemine: Mining moving object databases (system demo). In *Proceedings of SIGMOD*, pages 1203–1206, 2010.
[10] M. Naaman. Geographic information from georeferenced social media data. In *Proceedings of SIGSPATIAL*, pages 54–61, 2011.
[11] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil. An empirical study of geographic user activity patterns in foursquare. In *ICWSM*, 2011.
[12] X. Yi, H. Raghavan, and C. Leggetter. Discovering users' specific geo intention in web search. In *Proceedings of WWW*, pages 481–490, 2009.
[13] Z. Yin, L. Cao, J. Han, C. Zhai, and T. Huang. Geographical topic discovery and comparison. In *Proceedings of WWW*, pages 247–256, 2011.