SemMobi: A Semantic Annotation System for Mobility Data

Fei Wu[†1] Hongjian Wang[†2] Zhenhui Li[†3] Wang-Chien Lee[‡4] Zhuojie Huang[§5]

[†]College of Information Sciences and Technology
[‡]Department of Computer Science and Engineering
[§]GeoVISTA Center
Pennsylvania State University, University Park, PA, USA
{[1]fxw133,[2]hxw186 , [3]jessieli}@ist.psu.edu, [4]wlee@cse.psu.edu, [5]seenhzj@gmail.com

ABSTRACT

The wide adaptation of mobile devices embedded with modern positioning technology enables the collection of valuable mobility data from users. At the same time, the large-scale user-generated data from social media, such as geo-tagged tweets, provide rich semantic information about events and locations. The combination of the mobility data and social media data brings opportunities for us to study the semantics behind people's movement, i.e., understand *why a person travels to a location at a particular time*. Previous work have used map or POI (point of interest) database as source for semantics. However, those semantics are *static*, and thus missing important *dynamic event information*. To provide dynamic semantic annotation, we propose to use contextual social media. More specifically, the semantics could be landmark information (e.g., a museum or an arena) or event information (e.g., sports games or concerts). The annotation method annotates words to each mobility records based on local density of words, estimated by Kernel Density Estimation model. The annotated mobility data contain rich and interpretable information, therefore can benefit applications, such as personalized recommendation, targeted advertisement, and movement prediction. Our system is built upon large-scale tweet datasets. A user-friendly interface is designed to support interactive exploration of the result.

Categories and Subject Descriptors: H.5.4 Information Systems:INFORMATION INTERFACES AND PRESENTATION(I.7)

Keywords: Density Estimation; Social Media; Mobility; User profile

1. INTRODUCTION

The rapid advancement in positioning technology enables us to collect large-scale mobility data from mobile users. Turning raw mobility into useful and interpretable knowledge has drawn increasing attention recently. Several systems have been proposed to mine patterns from movement purely using mobility data [5, 7].

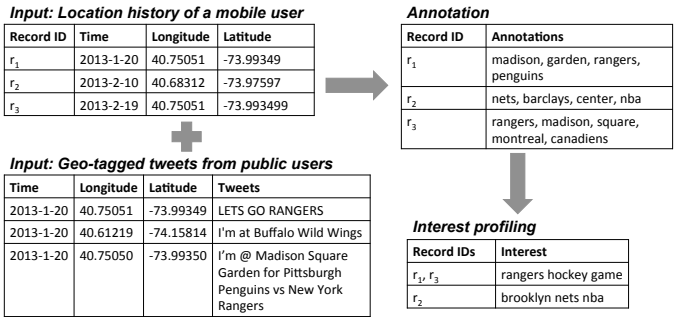

Figure 1: An example of mobility records annotation.

At the same time, popular location-based social media services, such as Twitter and Foursquare, provide platforms for mobile users to voluntarily generate enormous amount of spatiotemporel text. For instance, Twitter alone generates nearly 1 million geo-tagged tweets everyday. Such spatiotemporel data have demonstrated their value in a variety of tasks, such as event detection [4, 6], user profiling [3, 10], and city region profiling [2].

While the patterns from purely mobility data are useful to understand the inherent behaviours of a mobile user, they do not provide *contextual semantics* required for understanding the intended activities of the user. Previous work have considered using external information, such as landmarks, land-use categories, and POI (point of interest) database [9]. However, the semantics provided by those sources are *static*, i.e., a location will have the same semantics regardless of time. Therefore, they may fail to reveal the true purpose of a user's visit.

To provide dynamic semantic annotation, we propose to use contextual social media. More specifically, the semantics could be landmark information (e.g., a museum or an arena) and event information (e.g., sports games or concerts). The SemMobi system implements our recently developed annotation method, which has been accepted to WWW 2015 conference. The annotation method annotates words to each mobility records based on local density of the words, estimated by Kernel Density Estimation. The annotated mobility data contain rich and interpretable information, therefore can benefit applications, such as personalized recommendation, targeted advertisement, and movement prediction.

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742837.

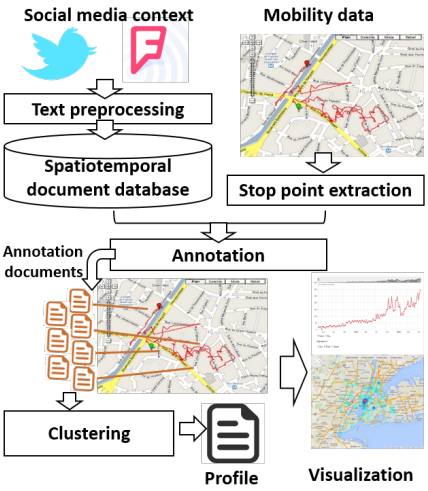

Figure 2: System architecture.

Figure 1 illustrates the annotation process, our system takes two input: (i) the location history of mobile user, represented by a sequence of location and the corresponding timestamps (i.e., $< time, location >$ pairs), and (ii) a set of documents from social media data with the time and location history. For each entry in the user's mobility history, SemMobi generates a list of ranked words that are most relevant to the purpose of the user's visit to that location at the particular time. We name this word list *annotation document*. For example, in Figure 1, the mobility record r_1 is annotated with words about a Rangers[1] game with Penguins[2] at Madison Square Garden. Given all the annotation documents, SemMobi generates an interest profile of that user. Note that SemMobi is designed to take raw mobility data of a mobile user that are not associated with any contextual information. The profile generated does not require any textual input to be collected from users. As Figure 1 illustrates, a profile that reveals the user interests (i.e., ice hockey and basketball) is generated. Furthermore, SemMobi also supports annotation of a region for aggregated-level analysis.

For interactive exploring the result, a Google Map[3] based interface is integrated into the system. In addition, SemMobi supports several visualization methods (e.g., time series and density map) for analysis of the social media data. We have tested our system on large-scale tweet datasets to ensure its practical use.

2. SYSTEM FRAMEWORK

Figure 2 illustrates the system architecture of SemMobi. Our system takes two inputs: (i) location history of a mobile user, and (ii) geo-tagged social media information. If the mobility data is collected from continuous tracking system, we will first extract the stop points of the trajectory [11]. The spatiotemporal documents gathered from social media service will be stored in a database after preprocessing. The documents are indexed by each day for efficient retrieval.

[1]Rangers is an ice hockey team based in New Yock city
[2]Penguins is an ice hockey team based in Pittsburgh
[3]http://www.google.com/maps/

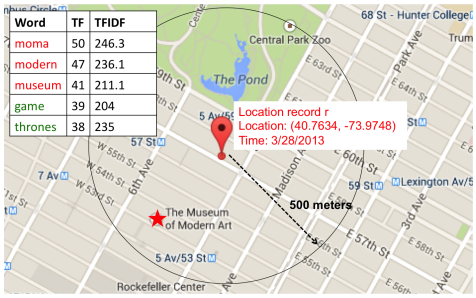

Figure 3: An example illustrating the problem of frequency based methods. The true user's intention of this location record is to attend the Game of Thrones event. But since MOMA is a more popular venue nearby, frequency-based methods will incorrectly use words "moma" and "modern" for annotation.

The semantic annotation component then operates on top of the processed data. We apply the annotation method from our recent work [8]. As a result, mobility records of a mobile user will be annotated with ranked lists of words, which we called annotation documents. Based on the annotation documents, SemMobi further applies clustering methods to generate a user profile. In addition, SemMobi provides two different forms of visualizations, i.e., time series and density map for each term. the users can look at the temporal trends of words to differentiate landmark related words and event related words. Showing spatial distribution helps the users to infer the region affected by a certain event.

3. METHODS

3.1 Mobility record annotation

Problem statement: For the problem of annotating mobility record, we are given: (i) one mobility record of a user $r_i = (loc_i^U, t_i^U)$, and (ii) external context data (sub-collection of all the social media data) as a set of spatiotemporal documents $\mathcal{D} = \{d_1, d_2, ..., d_m\}$. Each document d_j can be represented as a (W_j, loc_j^D, t_j^D) tuple, where W_j is the set of words in d_j, loc_j^D is a geographic coordinate, and t_j^D is a timestamp. We let $\mathcal{W} = \bigcup_{d_j \in \mathcal{D}_i} W_j$ denote the set of all words. Formally, given a mobility record r_i, to goal is to find an annotation document, $\mathcal{A}(r_i) = \{(w, s_w(r_i)) | w \in \mathcal{W}, s_w(r_i) > \theta\}$, where $s_w(r_i)$ measures the relevance of a word w to r_i.

Frequency-based approaches: A straightforward approach is to count the occurrences of each word (term frequency) near the given location record with respect to a distance threshold δ (specifying the neighbourhood). To avoid being overwhelmed by stop words, we can weight the term frequency by the inverted document frequency (i.e., tf-idf score). However, as shown in Figure 3, the frequency-based approaches may not reveal the true user intention, a proper distance threshold is hard to set. In this example, the user was attending the Game of Thrones (a TV series) exhibition, as inferred from the user's own tweets. The event location is very close to the Museum of Modern Art (MOMA). The highest ranked words are "moma" and "modern" by both methods, when setting $\delta = 500(m)$. The result is over-

whelmed by the popularity of MOMA, even though other words appear at closer places to the user.

Density-based approaches: Considering the density alleviates the problem of a hard threshold. One frequent-used model for spatial distribution is the Gaussian mixture model (GMM) [1]. Given all the occurrences of a word, we fit a GMM model and use the model to estimate the word density at the user's location. However, the true distribution of a word may not necessarily follow a mixture of Gaussian. In fact, the social media data generated by the crowd are constrained by the underlying city maps and natural landscapes, such as road networks, downtowns, lakes, and mountains, thus may deviate significantly from a Gaussian distribution. Furthermore, the number of components in GMM may very considerably across different words. For example, there could be 3 museums but 10 parks in a city.

To address the issues of GMM, we propose to use the density of the word occurrences as the measure and adapts Kernel Density Estimation (KDE) method to model the spatial density of the word occurrences. KDE is a non-parametric model for estimating density from sample points. The effectiveness of the KDE method in tackling the challenges has been demonstrated in our previous work [7]. Comparing with other methods, KDE provides a more local measure, i.e., word occurrences that are far from the user location has little effects and it naturally incorporating the effect of distance.

Following the kernel density model, we define the relevance score for a word w at r_i as follows:

$$s_w^{KDE}(r_i; L_i(w), H) = \frac{1}{|L_i(w)|} \sum_{loc_j^D \in L_i(w)} K_H(loc_i^U - loc_j^D),$$

and

$$K_H(x) = |H|^{-1/2} K(H^{-1/2}(x)),$$

where $K(x)$ is a kernel function of choice, $L_i(w)$ is a set of locations word w being mentioned, and H is a bandwidth matrix. We treat the two dimensions equally and use the same bandwidth h for both dimensions, i.e., $H = hI$ where I is the identity matrix. The estimated density will be more skewed when given a small h and more smoothed when given a large h.

It is known that KDE method is sensitive to the choice of bandwidth parameter h. There exist various ways of determining the Bandwidth parameter h, such as reference rule, cross validation, and adaptive bandwidth. SemMobi uses a fixed small $h = 10^{-4}$, which will return most reasonable result suggested by our previous work [8]. A small h applies larger penalties on points that are far from the user location, which results in a more local measure of the density.

3.2 Generating interests profile

SemMobi can generate an interest profile of the user by clustering similar annotation documents. The annotations are essentially a set of documents $\mathcal{D}_u = \{A(r_i)|\forall i\}$. Assuming that a user may express a limited number of interests K, such interests can be represented as clusters of the annotation documentations. For example, a New York City sports fan may have several annotation documents related to the sport events in New York City.

A variety of clustering methods, such as K-means and hierarchical clustering can be applied to cluster the annotation documents. In order to successfully reveal the interest of the mobile user, a proper distance measure should be chosen first. SemMobi uses ℓ_2-norm as the distance measure instead of other scale invariant measures, such as cosine similarity. The reason is that the absolute scores in a document are important in differentiating the event days and the normal days. For example, suppose we have annotation documents on two location records: $\mathcal{A}(r_1) = \{(nets, 150), (nba, 100), (others, 1)\}$ and $\mathcal{A}(r_2) = \{(nets, 10), (nba, 8), (others, 5)\}$, where record r_1 is on a game day. The cosine similarity will then consider these two documents as very similar: $Cos(\mathcal{A}(r_1), \mathcal{A}(r_2)) = 0.93$. Meanwhile, the Euclidean distance between the two documents in the previous example is large:
$Euc(\mathcal{A}(r_1), \mathcal{A}(r_2)) = 173.26$.

4. PREPROCESSING AND INDEXING

Text preprocessing. The textual content of spatiotemporal documents are first tokenized. As the raw social media data are often noisy (i.e., containing variations of the same word), SemMobi first aggregates similar word expressions by using Jaccard similarity. The most frequently used expression is selected to represent variations of a word. There are more sophisticated Natural Language Processing tools that can be applied to clean the social media data. We implement preprocessing step as a separated component, which can be changed easily.

Data indexing. The annotation is time-sensitive. Therefore, only a sub-collection of the entire spatiotemporal documents are needed for annotating one mobility record. Therefore, SemMobi partitions the entire collection by a time period and index each partition. We choose the time period as one day and use the date as index, as period for most events spans a day.

5. ABOUT THE DEMONSTRATION

In SemMobi, we have stored 50 million geo-tagged tweets from 2011 to 2013 over United States.

Annotating mobility history of a user. We start the demonstration by loading the mobility history of one mobile user into our system, as shown in Figure 4(a). After loading the trajectory, the processed mobility history is displayed on the map. The audience can click on the pins to view the annotation documents.

In this case, we examine the location history of user #901, who lives in Staten Island (based on his user profile) and visits New York City frequently. Figure 4(a) shows the top-5 words extracted by SemMobi. For the location record $\mathbf{r_3}$, SemMobi ranks the words "yankee", "stadium" and "yankees" as the top words with high relevance scores. We can infer that this user was attending Yankees' game. Similarly, we can infer that the user was attending Rangers' hockey game at Madison Square Garden for record $\mathbf{r_1}$. This user went to see the NBA game between San Antonio Spurs and Brooklyn Nets on 02/10/2013 at Barclays Center ($\mathbf{r_2}$) and the Game of Thrones exhibition on 3/28/2013 ($\mathbf{r_4}$). The location information of these records will be displayed on a map, which the audience can interact with.

User profiling generation. Given that the annotation documents are generated, SemMobi can apply clustering methods to obtain an interests profile of the user. A profile of the user will be shown in the box overlaid on the map, as shown in Figure 4(a). We can see the interests of user #901,

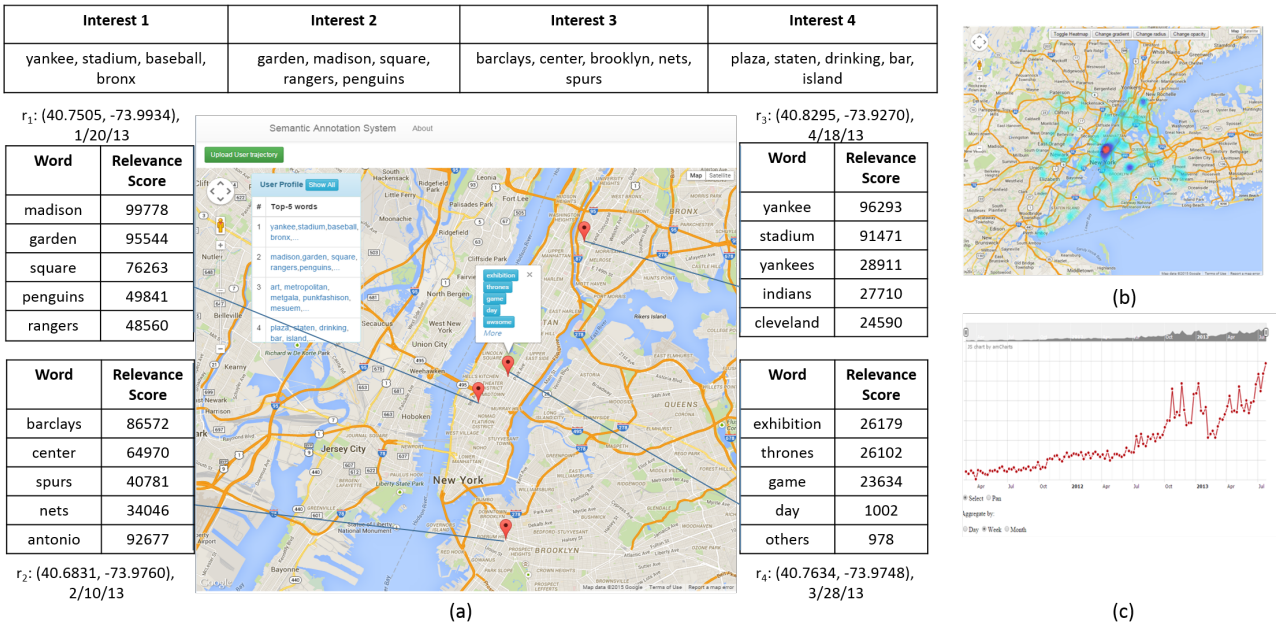

Figure 4: (a) Demonstration and user interface of SemMobi. (b) Time series of a selected word. (c) Density map of a selected word.

i.e., baseball, ice hockey, and drinking. Note that SemMobi only takes the mobility records from the user. The contextual information are from social media.

Exploring the result. The audience can further click on each word in the annotation documents or the user profile to view the time series (as shown in Figure 4(b)) and density map (as shown in Figure 4(c)). The time series can help audience to further differentiate words about landmarks and words about events. Intuitively, words related to landmarks should has a more consistent frequency pattern over the time, while words related to events may express a temporal burst. The density map suggests the area affected by some event. The audience can also verify the relevance of the annotation on via the density map.

6. SUMMARY

SemMobi demonstrates the semantic annotation of mobility data. We apply our recently developed annotation method [8] that is based on Kernel Density Estimation. Our system is built on large-scale social media data. It also incorporates several visualization tools for exploring the result. SemMobi provides useful tools for researchers to explore individual mobility records with semantics.

7. REFERENCES

[1] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proc. ACM KDD*, 2011.

[2] J. Cranshaw, R. Schwartz, J. I. Hong, and N. M. Sadeh. The livehoods project: Utilizing social media to understand the dynamics of a city. In *ICWSM*, 2012.

[3] M. Kosinski, D. Stillwell, and T. Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805, 2013.

[4] R. Li, K. H. Lei, R. Khadiwala, and K.-C. Chang. Tedas: A twitter-based event detection and analysis system. In *Data Engineering (ICDE), 2012 IEEE 28th International Conference on*, pages 1273–1276. IEEE, 2012.

[5] Z. Li, M. Ji, J.-G. Lee, L.-A. Tang, Y. Yu, J. Han, and R. Kays. Movemine: Mining moving object databases. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data (SIGMOD'10)*, pages 1203–1206. ACM, 2010.

[6] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proc. WWW*, 2010.

[7] F. Wu, T. K. H. Lei, Z. Li, and J. Han. Movemine 2.0: Mining object relationships from movement data. *Proceedings of the VLDB Endowment*, 7(13), 2014.

[8] F. Wu, Z. Li, W.-C. Lee, H. Wang, and Z. Huang. Semantic annotation of mobility data using social media. In *ACM International Conference on World Wide Web*, 2015, to appear.

[9] Z. Yan, D. Chakraborty, C. Parent, S. Spaccapietra, and K. Aberer. Semantic trajectories: Mobility data computation and annotation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 4(3):49, 2013.

[10] Q. Yuan, G. Cong, Z. Ma, A. Sun, and N. M. Thalmann. Who, where, when and what: discover spatio-temporal topics for twitter users. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 605–613. ACM, 2013.

[11] Y. Zheng, X. Xie, and W.-Y. Ma. Geolife: A collaborative social networking service among user, location and trajectory. *IEEE Data Eng. Bull.*, 33(2):32–39, 2010.