A Carpooling Recommendation System Based on Social VANET and Geo-social Data

Ahmed Elbery
Dept. of Computer Science
Virginia Tech
Falls Church, VA 22043
aelbery@vt.edu

Mustafa ElNainay
Dept. of Computer and Systems Engr.
Alexandria University
Alexandria, Egypt 21544
ymustafa@alexu.edu.eg

Feng Chen
Interdisciplinary Research Center
Carnegie Mellon University
Pittsburgh, PA 15213
fchen1@cmu.edu

Chang-Tien Lu
Dept. of Computer Science, Virginia Tech
Falls Church, VA 22043
ctlu@vt.edu

Jeffrey Kendall
Dept. of Computer Science, Virginia Tech
Falls Church, VA 22043
jdk34@vt.edu

ABSTRACT

Geo-social information can be utilized for user benefits in many applications. Social interaction in vehicular ad hoc networks (VANETs) is an important source for this type of information. In this paper, we first propose and describe a general architecture of the social VANET system (S-VANET) that supports social interaction through vehicular networks. Then, we present a new carpooling recommendation system that works as S-VANET application. The main objective is to recommend individuals to join their friends during trips or travels. The proposed recommendation system uses check-in history and home location to model users, and utilizes Fast Fourier transform to represent user check-ins and find the similarity between users. The system uses hierarchical clustering with weighted center of mass method to estimate the user home location.

Categories and Subject Descriptors

H.2.8 **[Database Management]:** Database Applications| Spatial databases and GIS

Keywords

Geo-Social Information; Vehicular Ad Hoc Networks.

1. INTRODUCTION

Social online networks are rich sources of data that can be analyzed and mined to get valuable information. This immense amount of information about users, their interests and relationships could be used in many applications. As wireless communications advance, new technologies are being imbedded such as GPS, such technologies provide location information about users. Integrating wireless communication into other technologies facilitates new services and applications. A Vehicular Ad-Hoc Network (VANET) exemplifies such integration and benefits. VANET provides a huge amount of information about user mobility patterns and trajectories. There are datasets that describe the user trajectories such as Microsoft data set [1] which records the trajectories of users using multiple transportation methods (Walk, Taxi, Car…..). However, due to privacy consideration, there is a lack for data that integrates user information and user mobility and location information.

SIGSPATIAL'13, Nov 05-08 2013, Orlando, FL, USA
ACM 978-1-4503-2521-9/13/11.
http://dx.doi.org/10.1145/2525314.2525327

By Integrating social networks and VANET to form a social VANET (S-VANET), we can get geo-social information about users' social profiles, interests, friendships and their mobility. This information can be utilized in wide spectrum of applications for enhancing and securing the vehicular communication, passing suggestion and recommendation systems for road congestion prediction and avoidance, as well as for city and road planning.

In this paper, a general framework for S-VANET is presented with its architecture and main components. Then we introduce a new recommendation system based on this S-VANET architecture. The main motivations include : (1) People prefer to work in communities, so it is better for individuals to accompany others when travelling or visiting specific venues. From this perspective, the recommendation system saves the users' time and effort in searching for friends who may join in visits. Finding those persons is a challenging task because it does not only depend on the friendship; it also depends on the interests of those individuals that mainly revealed from the user history and activities. The place characteristics are also important in this process. (2) Friendship locality: most of user friends are spatially close to each other. (3) Preference locality: users from a spatial region have common preferences different than users from other regions. (4) Travel locality: users tend to travel within a limited distance when visiting venues [2].

Combining 2, 3, and 4 together, we may conclude that close friends prefer traveling to the same nearby venues. Thus, when a user visits a venue, it is highly probable that other friends who are spatially close to him plan to visit the same places.

The contributions of this paper are as follows:

- **S-VANET framework**: a new framework is proposed to integrate the social networks and VANETs. The proposed integrated framework helps users form groups when travelling or visiting places of interest.

- **User home location estimation**: an approach for estimating user home location based on his check-ins and venues' weights, is presented.
- **Carpooling recommendation system**: a new recommendation system is developed to recommend users to join their friends in their travels. The system uses Fast Fourier Transform (FFT) to represent users' check-in history; the FFT is then used to calculate the similarity between users as well as to predict a user's future visits.

2. S-VANET SYSTEM OVERVIEW

S-VANET is a social communication system that tracks and connects social friends on the road, and shares the traveling information for social communities. Users can communicate to the system through Internet connection. The system can also share user location within his community.

2.1 S-VANET Components

The S-VANET system consists of three main components, including user, vehicles, and a social VANET server.

USER: a user can schedule a visit to a certain place on the site by identifying the place and the time. Then, the system can recommend the best roads from his location to the destination, and suggest some friends who can join him in this trip.

VEHICLE: A vehicle is a communication node that is equipped with a wireless communication device and GPS. The vehicle communicates to the server through either other vehicles or road side units. The vehicle coordinates are sent to the server periodically, so the server can track the vehicle and user location during travel.

SOCIAL VANET SERVER: the server has users' information including user profile, friends, privacy setting, and a summary of user location history. The system also tracks users in their travels and holds some details about their recent locations. The user interacts with the system through a web GUI. Integration with existing social network applications like Facebook and Twitter is a possibility for having users' information available.

2.2 S-VANET Architecture

The system architecture is shown in Figure 1, the user side consists of the user interface and the GPS module that periodically sends the vehicle coordinates to the system every *T* seconds.

In the system side, the tracking system receives the coordinates from different users and sends the information to the summarization module that stores the tracking information. The tracking system is responsible for detecting users' speeds and stops, and it controls time interval *T* to minimize the network overload and maintain a reasonable tracking accuracy.

The user schedules his visits on the scheduling subsystem that in turn communicates information to the recommendation system.

2.3 Illustrative Scenario

A user *U* is registered on the system, his profile includes mainly home location (latitude, longitude) and friend list. The system has a history of locations that *U* visited previously.

At time *T*, the user *U* scheduled a visit to venue *V* and selected a time range (T_1-T_2) in which he wishes to visit *V*.

Once the user has scheduled his visit, the system will search *U's* friends as well as friends of friends (*FoF*) and from them it selects the following users

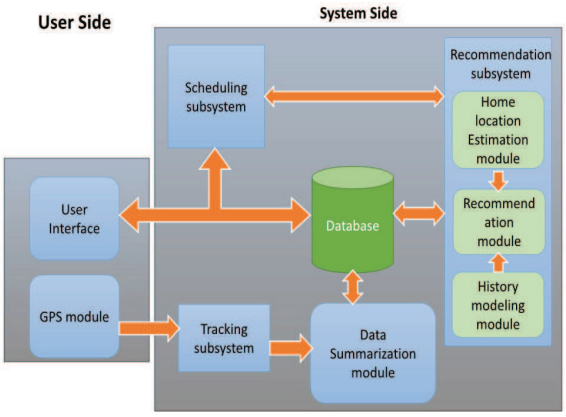

Figure 1: System Architecture

a) Users whose home locations are close to *U*
b) Users whose home locations are close to *V*
c) Users who visited the same location with *U*
d) Users who visited other locations with *U*
e) Users who are interested in the same location
   - Who previously visited the same location or
   - Who visited location of the same category as *V*.

For *a* and *b* users, the system needs the user home location coordinates. If the system allows users to register without identifying their home location, then it needs to estimate the home location for each of those friends.

To estimate the user's home location, we propose a simple mechanism that uses the user check-in history (the information available in the dataset). The system first hierarchically clusters these check-in locations and then selects a suitable number of clusters that are separated by sufficient distance, and for each cluster the system uses the weighted center of mass to estimate the home location. From this list (spatially close to U) the system finds the similarity between U's check-ins and check-ins for each of them. This similarity is used as a measure of closeness of interest between users. As user's interests become closer, they might join each other in their trips or visits, because the places reflect the user interest.

In *c* and *d* the system searches for those who previously accompanied *U* in visiting *V* or any other location. Those users probably wish to join *U* in his new visit. In *e*, the system finds users who are interested in the same location category as *V*.

For each of those selected users (*a*-*e*), the system estimates the probability that he might visit *V* in the specified time interval [T_1-T_2], and then selects those whose probability passes a specified threshold, or the highest *n* users.

The selected users are then recommended to join *U* in his visit to *V*. The user *U* can then choose some of them and send a join request asking them to join him in his visit to *V*. They can negotiate the time or how they will meet each other on their way to *V*. The system can also suggest the best road that they can follow.

3. RECOMMENDATION SYSTEM

The recommendation subsystem is responsible for finding users who might be wishing to join the active user in his scheduled visit. The recommendation system searches the active user friends and FoF and finds users who satisfy the above conditions listed in Section 2.3 (*a* through *e*). The system consists of three main

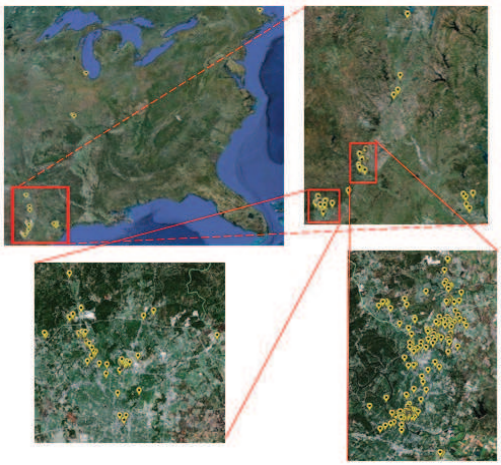

Figure 2: Spatial distribution for sample user's check-ins

components: home location estimation module, history modeling module, and recommendation module.

3.1 Estimating Home Location

The user home location is an important property because the system uses this location to estimate the distance between different users' locations as well as between users and venues.

Many approaches have been proposed in the literature to estimate the user's home., Cheng *et. al* proposed a probabilistic framework for estimating a Twitter user's city-level location based on the content of the user's tweets [3]. , Backstrom *et. al* estimate a user's location based on his Facebook friends [4]. , Cheng *et. al* used user check-ins and a recursive grid method to estimate the user home [5].

Our home location estimation is based on the analysis we made on the dataset in [5]. In this analysis, we have collected users who registered the home location coordinates and studied the check-in displacements from the registered home location. About 2600 user were collected with about 381,000 check-ins.

Figure 2 shows a distribution of a user check-ins. The figure shows that most user's check-ins are concentrated in two main areas (San Antonio and Austin) which are separated by about 120 KM. This distance indicates that the user may have homes in each area because it is not sustainable to travel this distance such number of times. We then have analyzed the user displacement distribution. As shown in Figures 3 and 4, more than 66% of the displacements are less than 20 KM, and more than 80% are less than 100 KM.

From the above analysis we can conclude that user might have multiple home locations, one in each area of interest. These areas should be sufficiently separated (i.e., in different cities). Thus, we need to locate user home in each area of interest.

To find the areas of interest for a user, we use the Hieratical Clustering. First, we find the pairwise distance between check-ins, and we link these points based on the centroid distance, as shown in Figure 6. From this linkage we find the clusters that are in different cities.

We decide that the clusters are in different cities based on the city size distribution. Figure 5 summarizes the sizes of the largest 250 cities worldwide [6]. According to the statistics in [6], the average city size is 1035 KM^2, and the average radius is 18.14 KM. To define different clusters, we used 22 KM distance because cities are not in regular circular form.

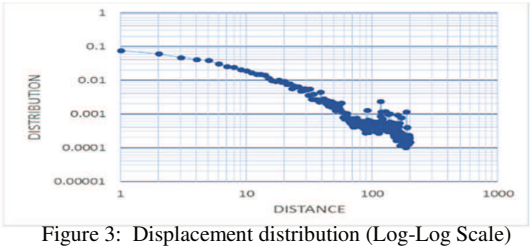

Figure 3: Displacement distribution (Log-Log Scale)

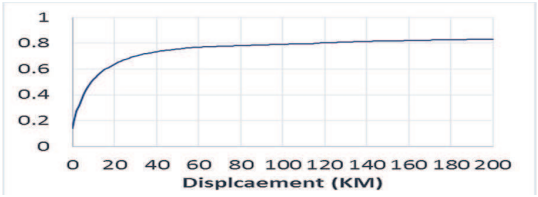

Figure 4: CDF for the displacement distribution

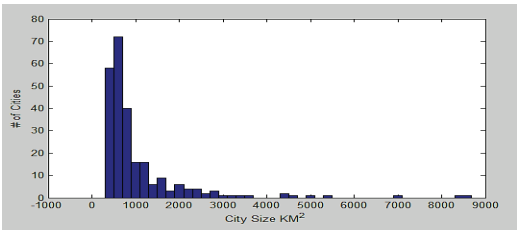

Figure 5: City Size Distribution

From the hierarchical clusters, we selected the dissimilarity level corresponding to about 22 KM, as shown in Figure 6. This dissimilarity level determines the number of clusters that are separated by this distance. However, there might be some small clusters. To remove such small clusters, we select only cluster whose size is greater than the average cluster size for the user.

After defining the clusters (areas of interest) for the user, we find the user home location in each of these areas using weighted central of mass. The central of mass locates the user home in the central point of his check-ins, but it suffers from splitting-the-difference. To avoid this drawback, we add the weights of the places to the check-ins. The weight of the place represents the importance of this place to the user. We used the ***tf.idf*** (Term Frequency – Inverse Document Frequency) as the weight after normalizing it to the [0, 1] range. The home coordinates *H* are calculated as

$$H = M - \sum_{i=1}^{n} w_i * (M - x_i)$$

where *M* is the center of mass, x_i is the coordinates of place in the i^{th} check-in, w_i is the weight of place

$R_i = Log(1 + tf) * \log(\frac{N}{df})$, $w_i = \frac{R_i - R_{min}}{R_{max} - R_{min}}$ *tf* is the number of check-ins to the place by the user, *df* is the number of users visited that place, and *N* is the total number of places. The rationale is that as user visits the place frequently; this place is probable more close

to his home. In the meantime as the place became popular to more (i.e., airport) users, this probability decreases.

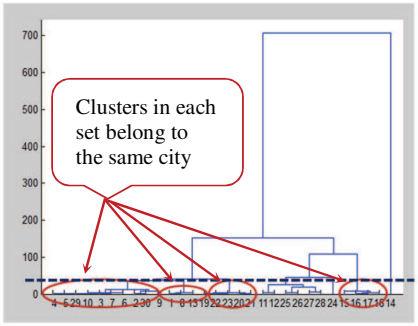

Figure 6: Hieratical Clustering of check-ins

3.2 Modeling User History

The Fast Fourier Transform is used to represent the user check-ins, the advantage of FFT is that it does not require many preprocessing steps (e.g., shifting and compression). In addition, using the magnitudes of the FFT coefficients, we can find whether the users visit the same place or near places. To demonstrate this concept, consider the example of three functions of the same values but in different orders X=[3; 7; 2; 1; 5] , Y=[1; 3; 5; 7; 2] and Z=[2; 1; 7; 5; 3] as shown in Figure 7. If we use the magnitude of the coefficients to represent the function, the data will be similar regardless of the order as shown in Figure 8. So using the magnitude of FFT, we can easily find the similarity between visited locations regardless of the visiting order.

For the calculation simplicity, the first 10 coefficients of the FFT are used and the similarity is estimated based on these coefficients. According to the power spectral density, these coefficients contain more than 90% of the information.

3.3 User Recommendations

After finding the users home locations and modeling the users, the system collects the active user friends and friends of friends in L. To select users who are spatially close to him or close to the designated venue V, the distance between the friend and the active user d_{lU}, and the distance between the friends and the venue d_{lV} are calculated. For each friend l the system selects the minimum of them as $d = \{ d_l : d_l = min(d_{lV}, d_{lU}) \forall\ l \in L\}$. Then, the system calculates the similarity $S = \{S_l \forall\ l \in L\ \}$ between the active user check-ins and those for each of his friends. This similarity is calculated as the distance between the FFT coefficients calculated by the user modeling module. Then, the system selects friends who are interested in visiting the designated venue, similar venues (of the same category) or venues near the designated venue. To do this task, the system calculates P_l, which is the probability that a friend l might visit $V \pm \delta V$ in the scheduled time. δV is the closeness distance that is assumed to be a constant of 500 meter. This represent a circle of diameter 1000 meter centered at V. These ranks are then combined together to find the recommended users. The final rank $R = \alpha\ d_l + \beta\ S_l + \gamma\ P_l$, the three factors α, β and γ are constants between 0 and 1 that represent the importance of these parameters d_l , S_l, and P_l, respectively. Once the user U schedules a visit to venue V, the recommendation system works as follows:

1. $F = \{f_i\}$; A list of all friends of U,
2. $\forall\ f_i$ Finds FoF_i ; The friends of f_i,
3. $L = F\ \cup FoF$; Compile F and all FoF_i into L; a list of all friends and friends of friends,
4. $L = unique(L)$; remove repeated users,
5. $\forall\ l \in L$ finds H_l; The home coordinates for user l and finds $d_l = min(d_{lV}, d_{lU})$; the minimum of the distances between l and U's location and between l and V,
6. $S = \{S_l \forall\ l \in L\ \}$; The similarity between U and each $l \in L$,
7. $P = \{P_l \forall\ l \in L\ \}$; The probability that l may visit $V \pm \delta V$,
8. $R = \alpha\ d_l + \beta\ S_l + \gamma\ P_l$; the final rank.

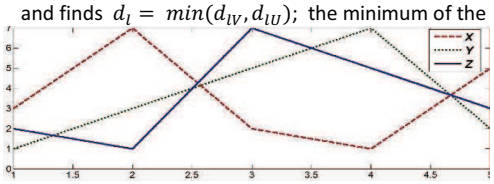

Figure 7: X, Y, and Z functions

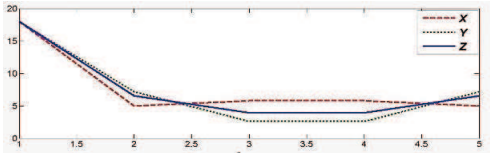

Figure 8: Inverse FFT of the magnitudes of the X, Y, and Z

4. SUMMARY

In this paper, a system that integrates social networks and VANET communication is proposed (S-VANET). S-VANET can offer many services to users such as car routing and road congestion prediction. S-VANET can collect large amounts of information related to users' profile, interests, friendship and mobility that enable novel applications to the location aware services. One of the S-VANET applications, the carpooling recommendation system, is proposed that help users plan their trips and recommend friends for carpooling. The system utilizes Fast Fourier Transform to represent user check-ins and find the similarity between the users. Moreover, hierarchical clustering with weighted center of mass method is proposed to estimate the user home location coordinates.

5. REFERENCES

[1] 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.

[2] J. Bao, Y. Zheng, and M. F. Mokbel. Location-based and preference-aware recommendation using sparse geo-social networking data. In Proceedings of the 20th International Conference on Advances in Geographic Information Systems, pages 199-208. ACM, 2012.

[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] Z. Cheng, J. Caverlee, K. Lee, and D. Z. Sui. Exploring millions of footprints in location sharing services. ICWSM, 2011, pages 81-88, 2011.

[5] M. Research. Geolife gps trajectories, Aug. 2012.

[6] C. Statistics. The largest cities in the world by land area, population and density, Jan. 2007.