An innovative GPS trajectory data based model for geographic recommendation service

JunJie Xiong
Data mining
College of Computer, Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu, China
mail: junjie.sop@gmail.com

***Abstract*—Geographic services based on GPS trajectory data, such as location prediction and recommender services, have received increasing attention because of their potential social and commercial benefits. In this study, a Geographic Service Recommender Model (GSRM) is proposed, which loosely comprises three essential steps. Firstly, location sequences are obtained through a clustering operation on GPS locations. To improve efficiency, a programming model with a distributed algorithm is employed to accelerate the clustering. Secondly, in order to mine spatial and temporal information from the cluster trajectory, an algorithm (MiningMP) is designed. Last but not least, the next possible location to which the user will travel is predicted. An integrated framework of GSRM could then be constructed and provide the appropriate geographic recommendation service by considering location sequences as well as other related semantic information. Experiments were conducted based on real GPS trajectories from Microsoft Research Asia (182 users within a period of five years). The experimental results clearly demonstrate that our proposed GSRM model is effective and efficient at predicting locations and can provide users with personalized smart recommendation services in the follow- ing possible position with excellent performance in scalability, adaptability, and quality of service.**

***Categories and Subject Descriptors*—H.2.8 [Database Management]: Database ApplicationsData mining, Spatial database and GIS**

***Keywords*—distributed computing, location-based services, location prediction, recommender service, trajectory pattern**

I. Introduction

The pervasiveness of Global Positioning System (GPS) devices enables people to conveniently acquire personal geographical location information theoretically contains user geographical trajectories and also implies individual behavior and reg- ular social patterns Furthermore, recommender services can be provided to enable individuals to discover interesting locations based on the knowledge mined from other human behaviors [1], [2], [3], [4]

As we know, public places are visited frequently by individuals . It can be inferred that these individuals have similar interests and behavior, which theoretically can be mined from GPS trajectory data.

With the development of GPS devices and cloud computing technology, geographical position trajectories have been established during interactions with reality and cyberspace. A huge amount of trajectory data provides new research opportunities to help understand social behaviors and dynamic communities in different situations, including even hidden human behaviors. This condition makes all kinds of innovative applications possible. The following are two instances:

1. Most traveling users leave behind a series of GPS trajectories; thus, it is possible to understand users spatio-temporal patterns. For example, there are two local residents: User 1 and User 2. They frequently travel around Location A where there are some services (e.g. restaurants, cinemas, hotels). Based on their GPS trajectories, we can infer their preference so that we can provide quick services before User 1 and User 2 reach Location Aagain, which is of great value to the business owners because they can provide better recommendation services for con- sumers, such as the message of discount coupons.

2. Recommendation service is also of great value to individuals. If User 3 is a foreign tourist, then he/she may have no idea about the surrounding environment upon arriving at Location A. In this case, the experiences of local residents may be useful for User 3. Because User 1 and User 2 frequently travel around Location A, we could infer Location-A to be a hot spot with some candidate services that could be offered to User 3. Based on these two instances, a hot spot is critical to recommendation services because it represents a certain meaningful point. However, raw GPS trajectory data include only latitude, longitude and time without further semantic information, from which we cannot know whether one GPS point is a hot spot or whether one GPS point belongs to a specific meaningful region. Fortunately, the amount of GPS trajectory date is huge, from which we can discover new information. There are some patterns in these huge data, such as temporal and spatial patterns.

1. Temporal pattern: This means some chronological phenomenon . For example, A user often has similar GPS trajectories during his or her workdays. By observing a history of locations, one can infer hourly patterns of a day, e.g. going to the office around 9 a.m. on workdays.

2. Spatial pattern: Users tend to visit nearby hot places instead of places further away. This means that the distance between the current location and the next location has some influences on the users mobile behavior .

These two patterns provide a promising tool for analyzing users real-world behavior, which could potentially improve location-based services such as traffic forecasts and geographic recommendation services (Gao et al., 2012). In this article we focus on mining the above patterns, predicting the next

location and providing a corresponding geographic service recommendation. The remainder of this article is organized as follows. The related research is reviewed in Section 2. Section 3 explains the proposed model in detail; Section 3.1 elaborates the definitions and context of our study, Section 3.2 presents the overview of the model, Section 3.3 focuses on the clustering process, Section 3.4 describes the movement pattern mining, and Section 3.5 focuses on geographic recommendation services. Section 4 presents the experiments and the evaluation and, finally, conclusions are given in Section 5, followed by a discussion of future research directions for this topic. illustrated in Fig. 1, where is ...

Fig. 1. Illustration of a user-item bipartite networks. Here users are connected with the corresponding selected items.

The remainder of this paper is organized as follows. We give a review of related work in Section II. In Section III we introduce the proposed latent user interests and item topics model LITM in detail. Section IV describes our experimental setup, performance metrics, and discusses experimental results. Finally, the conclusions and future work are laid out in Section V.

II. Related Work

In order to mine the individual behavior and then provide a better recommendation, there are two challenges that need to be addressed. The first challenge relates to efficiently retrieving clusters from raw GPS trajectory data; the other challenge relates to discovering movement patterns based on the clustered result so that a smart recommenda- tion service can be developed.

Regarding the first problem, clustering analysis of the raw data and mining of meaningful information from unstructured datasets requires complex and intensive computing. There have already been many successful cases in the past decades. Ester proposed a DBSCAN algorithm that is able to discover clusters of arbitrary shape II. Arthur optimized the k-means clustering method by seeding the initial centers II. Because the two abovementioned general methods for clustering data may miss some significant places II, an improved algorithm was adopted to cluster GPS trajectory data into different geographic regions II. Guo proposed an algorithm for clustering massive spatial GPS points based on Delaunay triangulation, which exploited the spatial contiguity of GPS trajectory data (Guo et al., 2012). Andrienko used a density-based method to cluster GPS trajectory data, in which the clustered result could be visually analyzed (Andrienko et al., 2009). A space-time cube (STC) was utilized to mine the frequent stopping locations from pedestrian GPS trajectories II. However, visual pictures can become cluttered in STC for large amounts of trajectory data. So Demsar introduced the concept of 3D spacetime density of trajectories II. There have also been different distance measures (e.g. perpendicu- lar, parallel, and angle) that can be applied to grouping raw GPS trajectories II. However, these distance measures are incapable of determining user preference because little mean- ing exists among these raw data. The Stay Point method was efficiently used to extract meaningful information from raw data II and can help to find points of interest (POIs). A method of automatically building region with an O (n2) time complexity was proposed II. To understand the spatio-temporal patterns in mobility data consisting of origin-destination pairs, spatial clustering of massive GPS points was proposed with O (nlog n) time complexity (Guo et al., 2012). The clustering methods mentioned above belong to data-intensive computing methods that normally cost a large amount of computing time. The time of clustering compu- tation will increase rapidly with an increasing amount of data. In order to meet the computing demand of big data and decrease the time of clustering computing, inspired by the parallel computation II, we proposed a parallel clustering algorithm based on k-means, which often is considered to be of linear complexity in practice.

Regarding the second challenge, during the past several years, with the wide use of personal GPS devices, the study of extracting POIs and even obtaining user preferences has become more and more popular. A collaborative rec- ommendation model incorporating temporal, geographical, and social information was proposed by Yuan et al. (2013). A recent study showed that similar check-in behavior can help to improve the accuracy of recommendations (Gao and Liu, 2015). However, because of a lack of check-in and social information, we have only used temporal and geographi- cal information. To find the characteristic points of each GPS trajectory in a line segmentation process, some studies based on LBS (Location-Based Services) have also been explored successfully, such as grouping the trajectories, provid- ing location recommendations, and predicting movement based on frequent patterns (Zhang et al., 2001; Ashbrook Starner, 2003; Monreale, Pinelli, Trasarti, Giannotti, 2009). The methods of Zhang, Kao, Yip, and Cheung (2001) and Ashbrook and Starner (2003) were able to mine the sequential marshal pattern but are not appropriate here due to poor scalability to large GPS datasets. To understand spatio-temporal properties, a method based on T-pattern Trees was proposed, which may be used as a predictor of the next location with complex computing to build a tree (Monreale et al., 2009). Perego, Orlando, and Palmerini, (2001) proposed an enhanced Apriori algorithm for frequent patterns. Nevertheless, this algorithm is incapable of handling situations with a very long sequence. As for processing long sequences, Pei et al. (2001) proposed the PrefixSpan algorithm, which can reduce the mining time by exploring prefix-projections.

In contrast to these abovementioned studies, we aim to provide a location recommendation service for moving customers and taking into consideration spatial as well as temporal information from location sequences. In this research, the loosely coupled model Geographic Service Recommender Model (GSRM) is proposed to predict users preferences and provide appropriate recommender services. In brief, the proposed model has three advantages over these existing studies. Firstly, by exploring the spatial pattern, it can extract signifi-

cant spatial regions from raw GPS data with higher efficiency and accuracy using the k-means11 algorithm under the Hadoop platform. Second, by con- sidering the influence of temporal periodicities and time sequence, it employs user preferences and the PrefixSpan algorithm to calculate frequency patterns. Last, there is a linked list to connect POIs and clusters so that the next possi- ble location of moving users can be predicted with higher probability and better personalized recommendation services can be provided.

[5], [6], [7].

III. Methodology

A. Notations

In this study, we first define a hot region as a Stay Point (Zheng Xie, 2011) and then cluster all Stay Points to extract public hot spots. The following definitions are used.

Definition 1. *(GPS Points): GPS Points P5fp1 ; p2 ; . . . ; pn g. Each GPS point pi 2 P contains the latitude pi :Lat, longitude pi:Lngt, and time stamp pi:T.*

Definition 2. *(GPS Trajectory): On a 2D plane (Figure 1), we can sequentially connect GPS points into a curve based on their time sequences and split this curve into a GPS Trajectory (Traj) if the time interval between consecutive GPS points is less than a certain threshold DT. Thus, Traj5p1 ! p2 ! ... ! pjPj, where pi 2 P;pi11:T¿pi:T and pi11:T2pi:T¡DT1 i¡jPj.*

Definition 3. *(Stay Point): A Stay Point denotes a geographic region in which a user stays over a time threshold Tthreh within the distance threshold Dthreh as radius. The set of Stay Points is denoted as SSet5fS1; S2; . . . ; Si; . . . ; Sng; n5jSSetj. As shown in Figure 1, a single Stay Point Si 2 SSet can be regarded as a virtual location characterized by a group of consecutive GPS points.*

Definition 4. *(Cluster): Suppose that Si is a Stay Point and there are n Stay Points in total, then a Cluster is defined from the finite set of the Stay Points as C5fS1; S2; S3; . . . ; Snng; nn5jCj, and the corresponding set of the clusters is recorded as CSet5fC1;C2;C3;...;Ckg;k5jCSetj*

Definition 5. *(Cluster Trajectory): A Cluster Trajectory is defined as CTraj5C1 ! C2 ! C3 ! . . . ! Ck , Ci 2 CSet and is a cluster that signifies a location.*

Definition 6. *(CTrajPrefix): Suppose that two CTraj trajectories are denoted by A5he1 e2 . . . en i and B5he1' e'2 . . . e'm i m n of the user, if ei'5eifori m21, then B is the CTrajPrefix of A.*

Definition 7. *(CTrajPostfix): Let A5he1 e2 . . . en i be a CTraj trajectory, and A' s CTrajPrefix be B5he1e2 . . . em21emim n, A's CTrajPostfix denoted as C5hem11em12 . . . eni w.r.t. CTrajPrefixB.*

Definition 8. *(SubCTraj): Given two CTraj trajectories are denoted by A5ha1a2 ...ani and B5hb1b2::bmi;m n, we consider B as a SubCTraj of A, i.e. B A, if there exist 1 j1¡j2¡...¡jm n such that b1 aj1 ;b2 aj2 ;...;bm ajm .*

Definition 9. *(Projection): Given two CTraj trajectories A and B such that B A, A' is denoted a Projection of A with respect to CTrajPrefix B if and only if (1) A' A; (2) B A'; (3) :8A"; B A" A' A" A' 146 A".*

Definition 10. *(Projected database): Suppose that sequence A is a sequential pattern of database D, A's projected database denoted as DjA, which is a set of postfix sequences and contains all postfix sequences with respect to CTrajPr efix A.*

These definitions will be illustrated in the following section.

- **latent user interest vector $\boldsymbol{u}_n$**: for each user n, we use K-dimensionality probability vector $\boldsymbol{u}_n$ to represent his interest distribution over the K topics;
- **latent item topic vector $\boldsymbol{v}_m$**: for each item m, we use K-dimensionality probability vector $\boldsymbol{v}_m$ to represent its topic distribution over the K topics.

Here a probability vector is a vector with non-negative entries that add up to 1, and that's why we can consider it as a distribution.

Based on definition of $\boldsymbol{u}_n(n = 1, 2, \cdots N)$ and $\boldsymbol{v}_m(m = 1, 2, \cdots, M)$, the occurrence probability of the edge between user n and item m, or the probability of user n select item m, can be calculated by the following:

$$\boldsymbol{u}_n^T \boldsymbol{v}_m = \sum_{k=1}^{K} u_{n,k} v_{m,k}, \tag{1}$$

where $u_{n,k}$ and $v_{m,k}$ are respectively the $k-$th elements of the vector $\boldsymbol{u}_n$ and $\boldsymbol{v}_m$. We define it like this because if we let $u_{n,k}v_{m,k}$ be the occurrence probability of the edge between user n and item m on the $k-$th topic, then the sum of the occurrence probability over all topics (i.e. Eq. 1) can represent the occurrence probability of the edge between user n and item m.

B. Basic Model

Suppose both $\boldsymbol{u}_n(n = 1, 2, \cdots N)$ and $\boldsymbol{v}_m(m = 1, 2, \cdots, M)$ are known, the occurrence of edges between different user n and item m are independent according to the d-separation criterion [8] in probabilistic graphical models. So we can define the objective function as the occurrence probability of the whole bipartite network G as

$$P(G|U, V) = \prod_{n=1}^{N} \prod_{m=1}^{M} \left(\boldsymbol{u}_n{}^T \boldsymbol{v}_m\right)^{\sigma_{m,n}}, \tag{2}$$

where $U = \{\boldsymbol{u}_n | n = 1, 2, \cdots, N\}$, $V = \{\boldsymbol{v}_m | m = 1, 2, \cdots, M\}$.

Based on Alg. 1, we summarized the basic idea of solving the proposed model LITM (Eq. 1), which has been mentioned before in the current subsection. Line 3-4 means fixing V and optimizing U, Line 5-6 means fixing U and optimizing V, and Line 2 means repeating the two operations until convergence. At last we will get a local minimum of the objective function.

Algorithm 1: Basic idea for solving LITM

Input: $\sigma_{m,n} \in \{0,1\}, n = 1, 2, \cdots, N, m = 1, 2, \cdots, M; K \in \mathbb{Z}^+;$

Output: $\boldsymbol{u}_n, \boldsymbol{v}_m, n = 1, 2, \cdots, N, m = 1, 2, \cdots, M;$

Initialize $\boldsymbol{u}_n$ and $\boldsymbol{v}_m$;

while *not converged* **do**

  for $n = 1; n \leq N; n++$ **do**

    $\boldsymbol{u}_n = max_{\boldsymbol{u}_n} \sum_{m=1}^{M} \sigma_{m,n} \log\left({\boldsymbol{u}_n}^T \boldsymbol{v}_m\right)$ // *get the solution by Reduced Gradient method*(see Alg.1);

  for $m = 1; m \leq M; m++$ **do**

    $\boldsymbol{v}_m = max_{\boldsymbol{v}_m} \sum_{n=1}^{N} \sigma_{m,n} \log\left({\boldsymbol{u}_n}^T \boldsymbol{v}_m\right)$ // *get the solution by Reduced Gradient method*(see Alg.1);

return $\boldsymbol{u}_n, \boldsymbol{v}_m, n = 1, 2, \cdots, N, m = 1, 2, \cdots, M;$

IV. Experimental Evaluation

To evaluate our model, we conducted preliminary experiments. We performed the experiments using a server with a 16-core 2.6 GHz Intel Xeon processor with 32GB RAM, running Red Hat Linux 4.1.2-33.

A. Experimental DataSet

The Institute of Electrical and Electronics Engineers (IEEE, pronounced "I triple E") is a professional association with its corporate office in New York City and its operations center in Piscataway, New Jersey. It was formed in 1963 from the amalgamation of the American Institute of Electrical Engineers and the Institute of Radio Engineers. Today, it is the world's largest association of technical professionals with more than 400,000 members in chapters around the world. Its objectives are the educational and technical advancement of electrical and electronic engineering, telecommunications, computer engineering and allied disciplines.

V. Conclusion

This paper presents LITM, a .....

For the future work, we will .....

Acknowledgment

The authors would like to thank...

References

[1] T. Zhou, J. Ren, M. Medo, and Y.-C. Zhang, "Bipartite network projection and personal recommendation," *Physical Review E*, vol. 76, no. 4, p. 046115, 2007.

[2] T. de Paulo Faleiros and A. de Andrade Lopes, "Bipartite graph for topic extraction," in *Proceedings of the 24th International Conference on Artificial Intelligence*, pp. 4361–4362, AAAI Press, 2015.

[3] T. Hu, H. Xiong, and S. Y. Sung, "Co-preserving patterns in bipartite partitioning for topic identification.," in *SDM*, pp. 509–514, SIAM, 2007.

[4] X. Tang, M. Zhang, and C. C. Yang, "User interest and topic detection for personalized recommendation," in *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01*, pp. 442–446, IEEE Computer Society, 2012.

[5] R. Bell, Y. Koren, and C. Volinsky, "Modeling relationships at multiple scales to improve accuracy of large recommender systems," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 95–104, ACM, 2007.

[6] T. Hofmann and J. Puzicha, "Latent class models for collaborative filtering," in *IJCAI*, vol. 99, pp. 688–693, 1999.

[7] Y. Shen and R. Jin, "Learning personal+ social latent factor model for social recommendation," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1303–1311, ACM, 2012.

[8] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.