Capturing Geographical Influence in POI Recommendations

Shenglin Zhao[1,2], Irwin King[1,2], and Michael R. Lyu[1,2]

[1] Shenzhen Research Institute
The Chinese University of Hong Kong, Shenzhen, China
[2] Department of Computer Science & Engineering
The Chinese University of Hong Kong, Shatin, N.T., Hong Kong
{slzhao,king,lyu}@cse.cuhk.edu.hk

Abstract. Point-of-Interest (POI) recommendation is a significant service for location-based social networks (LBSNs). It recommends new places such as clubs, restaurants, and coffee bars to users. Whether recommended locations meet users' interests depends on three factors: user preference, social influence, and geographical influence. Hence extracting the information from users' check-in records is the key to POI recommendation in LBSNs. Capturing user preference and social influence is relatively easy since it is analogical to the methods in a movie recommender system. However, it is a new topic to capture geographical influence. Previous studies indicate that check-in locations disperse around several centers and we are able to employ Gaussian distribution based models to approximate users' check-in behaviors. Yet centers discovering methods are dissatisfactory. In this paper, we propose two models—Gaussian mixture model (GMM) and genetic algorithm based Gaussian mixture model (GA-GMM) to capture geographical influence. More specifically, we exploit GMM to automatically learn users' activity centers; further we utilize GA-GMM to improve GMM by eliminating outliers. Experimental results on a real-world LBSN dataset show that GMM beats several popular geographical capturing models in terms of POI recommendation, while GA-GMM excludes the effect of outliers and enhances GMM.

Keywords: Gaussian Mixture Model, Genetic Algorithm, Geographical Influence, Point-of-Interest, Location Recommendation.

1 Introduction

Point-of-interest (POI) recommendation is a significant service for location-based social networks (LBSNs). With the development of mobile devices and Web 2.0 technologies, many LBSNs like Foursquare and Gowalla emerge and attract many users. These LBSNs allow users to check in their locations, make friends, and share location-related information. In order to help users discover new interesting places in LBSNs, POI recommendations arise.

M. Lee et al. (Eds.): ICONIP 2013, Part II, LNCS 8227, pp. 530–537, 2013.

How to recommend a point of interest? User preference, social influence, and geographical influence are three aspects responsible for users' check-in activities [14,15]. Generally we derive user preference from user-based collaborative filtering, explore social influence based on users' social relationships, and model geographical influence from check-in locations' spacial features. And then we construct a POI recommendation system in the way of combining those three kinds of influence. The representative work is as follows. Ye et al. [15] propose a linear fused framework to combine them and Cheng et al. [2] propose a fused model to recommend a point of interest.

For POI recommendation in LBSNs, research about geographical influence is new and requires more attention, comparing with user preference and social influence. It is well-defined on how to derive user preference and social influence in a recommendation system [7,11]. Note that users' evaluations for items reflect their preferences and friends are inclined to share preferences. We derive user preference from user-based collaborative filtering and introduce social influence by containing similarity among friends. For POI recommendation system, we use collaborative filtering method to get user preference through treating location as item and check-in frequency as rating value, and we capture social influence by including friends' similarity in check-in locations [2,14,15,1]. In 2010, Ye et al. [14] first propose POI recommendation for LBSNs and utilize power law principle to model users' geographical influence. It is similar to a Gaussian model. Cho et al. [3] study user movement in LBSNs inspired by Gonzalez's discovery [4]. The study focuses on those users who frequently check in, since Gonzalez's discovery bases on call logs data that have strong periodic property. They propose a periodic mobility model (PMM) to capture user's geographical influence for location prediction in LBSNs. Experimental data select users whose check-in records are more than 10 times one day. Cheng et al. [2] propose multi-center Gaussian model (MGM) to capture geographical influence. This model assumes a user's check-in locations disperse around several centers and utilizes a greedy method to discover centers. It defines a district by a fixed distance and thus ignores discrepancy between users. How to capture geographical influence? Gaussian distribution based models perform well in previous studies but we still encounter problems in discovering centers accurately and eliminating the effect of outliers.

To find activity centers more accurately and eliminate outliers, we propose two models—Gaussian mixture model (GMM) and genetic algorithm based Gaussian mixture model (GA-GMM) to capture geographical influence. In geographical perspective, whether one checks in depends on the locations' transport convenience—people prefer places that are nearer to their activity centers. Those frequent check-in places naturally form one's activity district. According to location's spacial clustering feature, we apply GMM to find one's activity district centers. However, outliers exist in the observed data that do harm to learn the model. How to eliminate the impact of outliers? Thang et al. [12] propose a genetic algorithm based EM algorithm to implement the trimmed likelihood estimate (TLE) method [10] to eliminate the outliers in mixture models.

We exploit this genetic based EM algorithm to train GMM. The genetic algorithm based GMM (GA-GMM) improves GMM and finds user's activity centers more accurately.

Our contributions are as follows. First, we propose GMM to automatically learn users' activity centers via exploring their check-in history records. Moreover, we enhance GMM by GA-GMM to intelligently eliminate outliers. Finally, we conduct experiments on a real-world LBSN dataset and demonstrate that the proposed models capture the geographical information better and improve the accuracy of POI recommendation.

The remainder is organized as follows. Section 2 introduces the related work. Section 3 demonstrates the two models: GMM and GA-GMM. Section 4 compares experimental results of different models. In the end, section 5 summarizes and outlines further work.

2 Related Work

In this part, we introduce related work in three aspects: POI recommendation in LBSNs, geographical influence capturing methods, and GA-GMM.

POI recommendation in LBSNs is a new research topic. POI recommendation is widely used in GPS-based mobile devices at first [5,6]. In 2010 Ye et al. [14] first propose POI recommendation in LBSNs. Further Ye et al. [15] point out that user preference, social influence, and geographical influence are three aspects responsible for recommending a point of interest and among them geographical influence is the most important. The representative work is as follows. Ye et al. recommend a point of interest through a linear fused framework combining user preference, social influence, and geographical influence [15]. Cheng et al. [2] propose a fused model to combine them to recommend a point of interest.

Study of geographical influence capturing methods is new for POI recommendation. In 2010 Ye et al. [14] first propose POI recommendation for LBSNs and arise a power law principle to capture geographical influence for POI recommendation. Earlier related work about geographical influence appears in the study of user movement pattern. Gonzalez et al. [4] build a model using call logs and discover that activities of an individual usually center around a small number of frequently visited locations. Based on this, Cho et al. [3] study the specific users frequently checking in and propose a periodic mobility model (PMM) to capture geographical influence for location prediction in LBSNs. Cheng et al. [2] employ multi-center Gaussian model (MGM) to capture the geographical feature of locations in the proposed fused POI recommendation model.

Genetic algorithm based GMM (GA-GMM) is a method to eliminate outliers when learning GMM. Trimmed likelihood estimate (TLE) method is adopted to eliminate outliers in some studies of mixture model analysis [10]. Thang et al. [12] first propose a genetic algorithm based method to implement the trimmed likelihood estimate method to train mixture models and demonstrate the performance through a genetic algorithm based GMM (GA-GMM). Wang et al. utilize the GA-GMM to process EEG signal and apply it on brain-computer interface [13].

3 Models

3.1 Gaussian Mixture Model (GMM)

Gaussian mixture model (GMM) [9] is the most widely used mixture model. We can formulize it as follows:

$$p(\mathrm{x}_i) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathrm{x}_i|\mu_k, \sum_k),$$

where $p(\mathrm{x}_i)$ denotes probability dense distribution of data x_i, μ_k indicates mean value, $\sum_k$ indicates covariance matrix for a base distribution, K denotes the number of base components, and π_k is the mixing coefficient.

We exploit GMM to capture geographical influence in POI recommendation. Each Gaussian distribution component represents an activity district and the mean value denotes the longitude and latitude of the district center. Centers may be his home, office, or some specific entertainment place. We assume places nearer to some center are geographically easier to arrive and people prefer those places.

How to recommend a point of interest through GMM? For a user, a location's geographical information ([longitude, latitude]) in his check-in history records represents data x_i. We recommend POIs through the following steps:

1. Learn the parameters of GMM,
2. Calculate candidate locations' probabilities fitting the trained model, and
3. Sort the candidate locations and recommend the top K locations.

3.2 Genetic Algorithm Based Gaussian Mixture Model (GA-GMM)

In order to eliminate the effect of outliers, we introduce a genetic algorithm based Gaussian mixture model (GA-GMM). Generally we could use maximum likelihood EM algorithm to learn GMM [9]. If we use Θ to denote the parameters, likelihood function could be represented as

$$p(X|\Theta)_{ML} = \prod_{i=1}^{n} p(\mathrm{x}_i|\Theta).$$

Further, if we use the logarithm form, we can denote the objective of maximum likelihood EM algorithm as follows:

$$\hat{\Theta}_{ML} = \arg\max \log p(X|\Theta)_{ML} = \arg\max \sum_{i=1}^{n} \log p(\mathrm{x}_i|\Theta). \tag{1}$$

This formula includes all observed data. Trimmed likelihood estimate (TLE)—that aims to to select the subset of data with maximum sum of likelihood values—is used to eliminate the outliers [10]. We can use a genetic algorithm to find the optimal subset and exploit maximum likelihood EM algorithm to

learn the parameters of GMM, as illustrated in Algorithm 1 [12]. In this case, the objective function could be represented as

$$\log p_{TLE}(X|\Theta) = \sum_{i=1}^{n} w_i \log p(\mathrm{x}_i|\Theta), \tag{2}$$

where $\forall i = 1, 2, ..., n, w_i \in \{0, 1\}$ and $\sum_{i=1}^{n} w_i = m$, m represents the number of valid data. When $w_i = 1$, it indicates that the corresponding data is chosen into the subset. Otherwise, the data is an outlier and should be discarded. Hence, the result is a subset of size m out of n original samples, which fits GMM most in terms of likelihood contribution.

As a genetic algorithm, GA-GMM contains properties of genetic algorithm—it includes encoding scheme, fitness function, and operators like crossover, mutation, and selection. We use the standard way to implement crossover and selection [8]. Encoding scheme, fitness function, and a self-defined mutation (Guided Mutation) are defined as follows.

Definition 1. *Encoding scheme. The chromosome is encoded into a binary string and each bit represents the existence of corresponding observed data. Each chromosome and its corresponding mixture model will be a possible solution to our problem.*

Definition 2. *Fitness function. The fitness score function is set as the trimmed logarithm likelihood of the corresponding GMM of a chromosome*—$\log p_{TLE}(X|\Theta)$.

Definition 3. *Guided Mutation. Guided Mutation ensures the chromosome in a population to mutate toward maximizing fitness score. It means we choose chromosome that has higher value fitting trained GMM.*

4 Experiment

4.1 Setup and Metrics

We prepare the data by cleaning and splitting. We filter locations of less than 10 visits. And then we split the dataset into three non-overlapping sets in sequence: a redundant set, a training set, and a test set. The test set keeps 10% of the whole data set. We test different cases in which the proportion of training data is 90% and 50% respectively. When training data set is 90%, there is no redundant data. When the training data set is 50%, redundant data is the former 40% data that will be discarded.

We evaluate the performance of different models in capturing geographical influence by the accuracy of POI recommendation that is measured by Precision and Recall. POI recommendation is to recommend the top-N highest ranked locations. However, the system should not recommend locations user has checked in. To evaluate the performance of POI recommendation, we use the Precision@N and Recall@N as the metrics that are standard metrics to measure the performance of POI recommendation [15]. Precision@N defines the ratio of recovered POIs to the N recommended POIs and Recall@N defines the ratio of recovered POIs to the size of test set.

Algorithm 1. Genetic-based Expectation Maximization Algorithm

1. t=0;
2. Initialize $P_0(t)$;
3. **for** $t = 1 : G$ **do**
4. $P_1(t) \leftarrow$ perform several cycles of EM on $P_0(t)$;
5. $P_2(t) \leftarrow$ Guided Mutation in $P_1(t)$;
6. $fScore_2 \leftarrow$ evaluate $P_2(t)$;
7. $P_0(t)^{'} \leftarrow$ selection and crossover to generate offspring from $P_2(t)$;
8. $P_1(t)^{'} \leftarrow$ perform several cycles of EM on $P_0(t)^{'}$;
9. $P_2(t)^{'} \leftarrow$ Guided Mutation in $P_1(t)^{'}$;
10. $fScore_2^{'} \leftarrow$ evaluate $P_2(t)^{'}$;
11. $P_3(t) \leftarrow$ selection from $[P_2(t), P_2(t)^{'}]$;
12. $iBest \leftarrow$ best individual from $P_3(t)$;
13. **if** $iBest$ satisfies convergence condition **then** break;
14. $P_0(t+1) \leftarrow P_3(t)$;
15. $t = t + 1$;
16. Perform EM on $iBest$ until convergence;

4.2 Dataset

We use the Gowalla data records from February 2009 to September 2011. We select 3836 active users' records to experiment. We define active users as users whose check-ins are more than 1000 times and experience of using Gowalla is more than 1 year. After removing locations with less than 10 visits, all check-ins of active users include 183,667 different locations. We illustrate statistics of the data in Table 1, where "C." represents the check-in times of a user and "T." represents the time span (unit is day) from first check-in to last check-in.

Table 1. Data statistics

Min. C.	Max. C.	Avg. C.	Min. T.	Max. T.	Avg. T.
1,001	50,243	2,505	366	968	593

4.3 Results

We compare the POI recommendation performance of GMM and GA-GMM with Gaussian model (GM) and Multi-center of Gaussian model (MGM) [2] when training data set is 90% and 50% respectively.

Gaussian model (GM) [4] is a baseline model used in [3]. It models human movement as a stochastic process centered around a single point.

Multi-center Gaussian model (MGM) [2] is a latest model. It uses a fixed distance to define a district. When check-ins in a district are more than a threshold, the mean of all check-ins is the center. It utilizes a greedy method to find the district and requires no overlapping between two districts.

We illustrate experimental results in Fig. 1. GMM outperforms GM and MGM; further GA-GMM improves GMM. Hence, GA-GMM could better capture the geographical influence. In the experiment we set the number of centers in GMM and GA-GMM as 2 for simplicity, since Cho et al. pose that the check-in behaviour comprises two states in [3]. We set the radius of region in MGM as 1 kilometer and the threshold as 10% (that means the ratio of check-ins in one district is at least 10% of all his check-ins).

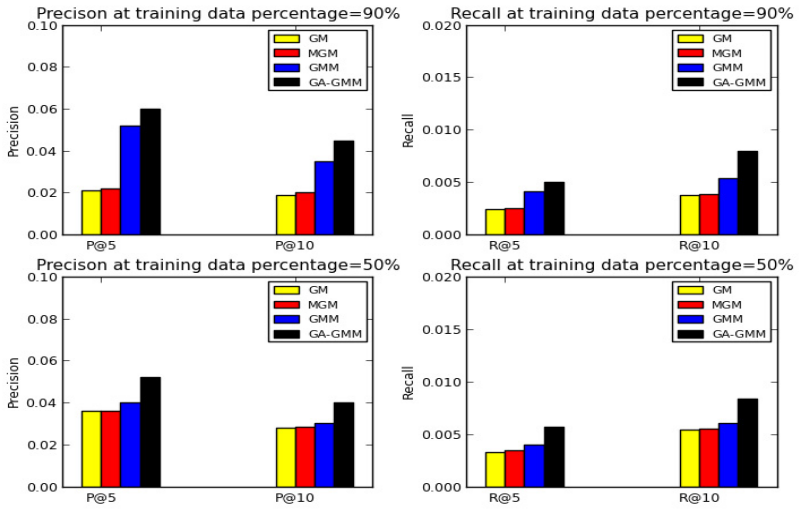

Fig. 1. Comparison of different models

5 Conclusions and Further Work

We apply GMM and GA-GMM to capture geographical influence in POI recommendation. According to experimental results, we draw conclusions as follows. 1) GMM outperforms the baseline model GM and the latest model MGM. 2) GA-GMM eliminates the outliers of data and improves GMM. It discovers the activity centers more precisely, which increases the accuracy of POI recommendation.

There are two aspects of further work. One is to establish a sophisticated POI recommendation system through adding the influence of user preference and social relationship. In this paper, we focus on how to model the geographical feature of check-in activities. For a POI recommendation system, we still have to consider more features of LBSNs such as user preference and social relationship. The other is to improve the efficiency. We learn each user's model separately. That provides opportunities to implement a parallel version.

Acknowledgments. The work described in this paper was partially supported by the Shenzhen Major Basic Research Program (Project No. JC201104220300A) and the Research Grants Council of the Hong Kong Special Administrative Region, China (Project Nos. CUHK413212, CUHK415212).

References

1. Cheng, C., Yang, H., Lyu, M.R., King, I.: Where you like to go next: Successive point-of-interest recommendation. In: IJCAI, Beijing, China (2013)
2. Cheng, C., Yang, H., King, I., Lyu, M.R.: Fused matrix factorization with geographical and social influence in location-based social networks. In: AAAI 2012: Proceedings of Twenty-Sixth Conference on Artificial Intelligence (2012)
3. Cho, E., Myers, S.A., Leskovec, J.: Friendship and mobility: user movement in location-based social networks. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1082–1090. ACM (2011)
4. Gonzalez, M.C., Hidalgo, C.A., Barabasi, A.L.: Understanding individual human mobility patterns. Nature 453(7196), 779–782 (2008)
5. Horozov, T., Narasimhan, N., Vasudevan, V.: Using location for personalized poi recommendations in mobile environments. In: International Symposium on Applications and the Internet, SAINT 2006, p. 6. IEEE (2006)
6. Kang, E.-y., Kim, H., Cho, J.: Personalization method for tourist point of interest (POI) recommendation. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) KES 2006. LNCS (LNAI), vol. 4251, pp. 392–400. Springer, Heidelberg (2006)
7. Koren, Y., Bell, R., Volinsky, C.: Matrix factorization techniques for recommender systems. Computer 42(8), 30–37 (2009)
8. Melanie, M.: An introduction to genetic algorithms, Cambridge, Massachusetts London, England, Fifth printing, vol. 3 (1999)
9. Murphy, K.P.: Machine learning: a probabilistic perspective. The MIT Press (2012)
10. Neykov, N., Filzmoser, P., Dimova, R., Neytchev, P.: Robust fitting of mixtures using the trimmed likelihood estimator. Computational Statistics & Data Analysis 52(1), 299–308 (2007)
11. Ricci, F., Shapira, B.: Recommender systems handbook. Springer (2011)
12. Thang, N.D., Lihui, C., Keong, C.C.: An outlier-aware data clustering algorithm in mixture models. In: 7th International Conference on Information, Communications and Signal Processing, ICICS 2009, pp. 1–5. IEEE (2009)
13. Wang, B., Wong, C.M., Wan, F., Mak, P.U., Mak, P.I., Vai, M.I.: Gaussian mixture model based on genetic algorithm for brain-computer interface. In: 2010 3rd International Congress on Image and Signal Processing (CISP), vol. 9, pp. 4079–4083. IEEE (2010)
14. Ye, M., Yin, P., Lee, W.C.: Location recommendation for location-based social networks. In: Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 458–461. ACM (2010)
15. Ye, M., Yin, P., Lee, W.C., Lee, D.L.: Exploiting geographical influence for collaborative point-of-interest recommendation. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 325–334. ACM (2011)