A Gaussian Kernel Approach for Location Recommendations

Iury Nunes, Leandro Marinho
Federal University of Campina Grande, Brazil
iury.nunes@ccc.ufcg.edu.br, lbmarinho@dsc.ufcg.edu.br

Abstract. The Location Based Social Networks (LBSN) are systems where the user can share the locations he/she has visited with his/her friends. These systems have become very popular because of the increasing number of users with smartphones equipped with GPS, which allow them to be connected to the internet wherever they are and geotag their current locations. In this context, the task to recommend new places for the user to visit is very important in order to improve the user experience while using the system. Several recommendation models have been proposed to solve this problem. In all these works, the final model is a combination of several specialized recommender algorithms, being the most recurrent ones based on social networks, collaborative filtering, and geographic information. Although geographic information is a key dimension in LBSNs, there is no work so far providing an empirical comparison between the existing recommendation algorithms solely based on geographic information. In this paper, besides providing this comparison, we propose a new geographic-aware recommendation algorithm based on a gaussian kernel to infer the regions more likely to be visited by the user. We conduct experiments on real world datasets of two popular LBSNs: Gowalla and Foursquare. Our experiments show that our approach outperforms the state-of-the-art geographic-aware recommenders proposed by the literature.

Keywords: Recommender Systems, Location Based Social Networks, Gaussian Kernel Density Estimation, Context-Aware Systems

1. INTRODUCTION

The systems known as Location-Based Social Networks (LBSN) are those where the users are able to share their locations with their friends. Foursquare, as an example, is a LBSN where users endorse and share tips about visited points of interest (POI) through check-ins. LBSNs have gained great popularity in recent years, mostly because, nowadays, more people have access to smartphones equipped with GPS and mobile internet, which are essential for the usability of these systems.

The recommendation of locations in LBSNs is very important because it can improve the user experience. The welcome message of Foursquare, for example, reads "Find great places near you", which confirms the importance of this kind of service. There are several scenarios where the recommendation of locations is important for LBSN users, for example, suppose a user is visiting a new city and does not know which places are worth visiting. He might be interested, for example, in nightclubs, that are nearby his/her hotel and that have been visited by many of his/her faceb ook friends. In this paper we focus only on the geographic aspect of the recommendation, which in the previous example, would correspond to the problem of recommending nightclubs nearby the users’ location, which is inferred from his/her previous check-ins.

There are many issues that make location recommendations in LBSNs a challenging problem, being data sparsity the most severe. From all the available venues in LBSNs, the users tend to visit only a small portion of them. Although sometimes users may travel to different cities, most of the users’ visits occur on the cities they reside (or the surrounding cities). Even then, users only visit a small portion of the locations available in the cities they reside, specially if the city is large and touristic with many venues available.

To alleviate the problem of data sparsity, most of the existing approaches in the literature propose
a combination of several specialized recommender algorithms, being the most recurrent ones based on social networks, collaborative filtering, and geographic information [Ye et al. 2011; Cheng et al. 2012]. Although geographic information is a key dimension in LBSNs, there is no work so far providing an empirical comparison between the existing recommendation algorithms solely based on geographic information. Such comparison would shed some light on how different purely geographic-based approaches affect the recommendation quality, thus paving the way for more informative recommender algorithms that use geographic data as one of its components. We fill this gap by conducting a comparative investigation between purely geographic-based recommenders proposed by the literature.

The main contribution of this paper is a new geographic-aware recommender system based on Gaussian kernels that, besides presenting high accuracy, it is also very efficient since no parameter needs to be tuned. Our experiments were conducted in real world datasets from two popular LBSN: Gowalla and Foursquare; and the results show that our approach outperforms the state-of-the-art location recommenders solely based on geographic information. To summarize, the main contributions of this paper are:

—A new geographic-aware recommender system based on Gaussian kernels that outperforms the existing methods in two real-world LBSN datasets.

—A comparative study of several algorithms purely based on geographic information already proposed in the literature.

2. RELATED WORK

Many research works in this area have appeared recently indicating that the geographic proximity between venues and between venues and the users’ home have a great influence on the users’ decision of visiting a new location [Cheng et al. 2011; Cho et al. 2011; Noulas et al. 2011]. For example, locations nearby the locations the user is used to visit have a higher probability of being visited than the ones that are not. An analogous argument can be used with respect to the locations nearby the users’ home. In fact, [Cho et al. 2011] showed that users tend to center their check-ins around their homes and work places.

It is worth noticing that most of the works on location recommendation propose some sort of ensemble of different recommender algorithms, each one specialized in some aspect of the data. [Ye et al. 2011] propose to combine a standard collaborative filtering algorithm with a social network-based recommender from the literature and a novel geographic-aware recommender. The final relevance score of a location is a simple linear combination of the scores of the individual recommenders. The main contribution of the paper is a novel mechanism of geographic influence based on a statistical model of how likely a user is to check-in to two places based on their distance. The geographic model is a distributional assumption (power law) fitted to real data sets from two popular check-in services. In [Cheng et al. 2012] a similar approach is proposed, but the collaborative filtering and the social recommenders are based on matrix factorization. The novelty of the paper lies on modeling the probability of a user’s check-in on a location as a Multi-center Gaussian Model.

Although the main contribution of the aforementioned works on location recommendations are the geographic recommender component, only the overall methods based on ensembles are compared, which do not say much about the performance of the geographic recommenders alone. In this paper, we propose a new model based on Gaussian kernels that outperforms the geographic recommender of the aforementioned works.
3. PROBLEM SETTING

The recommendation scenario occurs as follows: based only on the geographic coordinates of all locations a particular user has already visited, the recommender algorithm will suggest to him/her other locations he/she might like to visit but has not visited yet.

So, let $U$ be the set of all users of the LBSN and $L$ the set of all locations, represented only by the geographic coordinates. Then, the task of the geographic-aware recommender algorithm is to compute a scoring function $\hat{s} : U \times L \rightarrow \mathbb{R}$ that predicts the preference value for a given user $u \in U$ and a given location $l \in (L - L_u)$, where $L_u$ is the set of all locations visited by user $u$. Now, the topN recommendations for the user $u$ can be calculated as follows in equation 1

$$\text{topN}(u) = \arg \max_{l \in (L - L_u)} \hat{s}(u, l)$$

where $n$ is the number of locations we want to recommend for the user $u$.

4. GEOGRAPHIC-AWARE RECOMMENDER SYSTEMS

The main assumption of most of the geographic-aware recommender algorithms in the literature is that users tend to visit locations close to other locations he/she has already visited in the past. We can confirm this assumption observing the probability distribution of the distances between pairs of visited locations. As we can see in figure 1, for both datasets analyzed in this work, the probability distribution of the distances between visited locations resembles a power law distribution (especially for small distances), since it decreases linearly in log scale. However, notice that for larger distances, the density function oscillates heavily. One possible explanation for this phenomenon is that sometimes the users travel to long distances, and as a result, the users tend to create multiple regions of interest that are not necessarily close to each other. This observation is in line with the observations provided by other works that have made a similar analysis [Cho et al. 2011; Ye et al. 2011; Cheng et al. 2011; Cheng et al. 2012]. In the next subsections we will describe the most important geographic-aware recommendation algorithms proposed by the literature, that will serve as baselines to be compared against our approach, which is described in section 4.5.

4.1 Distance-Based

Assuming that the users indeed are more likely to visit locations that are near to the locations he/she has already visited, perhaps the most simple and intuitive geographic-aware recommender is the one which recommends based only on the distance. Formally, we could define a distance-based recommender as described in equation 2.
where function \( \text{dist} \) returns the physical distance between two locations and function \( \text{closest}(l, L_u) \) returns the nearest location in \( L_u \) to location \( l \). We multiply the distance by \(-1\) because the greater the distance the less is the chance that a user will visit that location.

One possible issue with this approach is that it is short sighted and look only the nearest venues. So, for instance, if a venue \( v \) is near from several locations the user has visited, we can assume that it is on the user’s region of interest. However, if there are locations that are closer to any location the user has visited (independently of how far they are from the other locations), then the distance-based approach would not recommend \( v \).

4.2 User Home

Recent work shows that the users tend to visit locations that are near to his/her home [Cho et al. 2011]. So, another very simple geographic-aware recommender would be the one which recommends the nearest locations to the user home. Formally, we define a user home based recommender as follows:

\[
\hat{s}(u, l) = -1 \times \text{dist}(\text{home}(u), l)
\] (3)

where function \( \text{home}(u) \) returns the location of the user \( u \) home. We can infer the homes of the users by discretizing the world into 25 by 25km cells and defining the home location as the centroid of check-ins in the cell with the largest number of check-ins [Cho et al. 2011].

A major limitation of this approach is that it is not flexible to recommend locations that are far from the user home. Suppose a user has a work that is far from his home, then he/she probably would have at least two regions of interest that are far from each other. Furthermore, the user home approach is not feasible to recommend locations when the users travel to cities that are far from their hometown.

4.3 Linear Regression

Assuming the power law distribution of the check-ins, we can estimate the parameters of the power law through a simple linear regression as proposed in [Ye et al. 2011], and then we can use the distribution probability as the scoring function of a geographic-aware recommender. The linear regression may be applied as described in equation 4 to find the values of the parameters \( w_0 \) and \( w_1 \), and then we can apply the transformation also described in equation 4 to estimate the values of the parameters \( a \) and \( b \) of the power law distribution.

\[
\text{prob} = a \times \text{dist}^b \Rightarrow \log \text{prob} = w_0 + w_1 \times \log \text{dist} , \text{where } a = 2^{w_0} \text{ and } b = w_1
\] (4)

After computing the value of the parameter \( a \) and \( b \) of the power law distribution, we can define the probability that a given user \( u \) will visit a given location \( l \) by visiting the probabilities of the distances of location \( l \) to all the locations in \( L_u \). Formally, we define a linear regression based recommender as follows:

\[
\hat{s}(u, l) = \prod_{l_u \in L_u} a \times \text{dist}(l, l_u)^b
\] (5)
4.4 Multi-Center Gaussian Model

The idea that a user has a single home point that defines one region of interest, although intuitive, is very restrictive because the user may have more than one region of interest. To estimate several regions of interest for the user, the multi-center gaussian model as defined in [Cheng et al. 2012] proposes a clustering algorithm to find the clusters (or centers) that represent the regions of interest for the users. The clustering algorithm uses a greedy approach and works as follows. We iterate through the most visited locations and combine all other visited locations, which distance is less than \(d\) kilometers from the selected location, into a region. If the ratio of the total check-ins number of this region regarding the user’s total check-ins amount is greater than a threshold \(\theta\), we set these check-ins positions as a region and determine its center. Now that we have defined how to compute the centers \(C_u\) of each user \(u\), we can formally define a multi-center gaussian based geographic-aware recommender as follows:

\[
\hat{s}(u, l) = \sum_{c_u \in C_u} \frac{1}{\text{freq}(c_u)} \frac{\delta(l | \mu_c, \sigma_c)}{\sum_{c_i \in C_u} \delta(l | \mu_c, \sigma_c)}
\]

where \(\text{freq}(c_u)\) returns the number of check-ins that occurred in the region represented by the center \(c_u\); \(\alpha \in (0, 1]\) is introduced to maintain the frequency version property; and \(\delta(l | \mu_c, \sigma_c)\) is the probability density function of the Gaussian distribution, where \(\mu_c\) and \(\sigma_c\) correspond to the mean and variance of the region around the center \(c_u\), respectively.

4.5 Gaussian Kernel

Except for the distance-based approach, for all the approaches described in the previous subsections, we had to make some kind of pre-computation before we can use the geographic-aware recommender: either estimate the user homes, or discover the parameters of the check-ins probability distribution through a linear regression, or calculate regions of interest for each user through a greedy clustering algorithm. In this section, we propose a new geographic-aware recommender based on gaussian kernel that does not need any kind of pre-computation, furthermore, besides being computationally efficient, the gaussian kernel-based approach outperforms all the other geographic-aware recommender described in previous sections, as we will see in Section 5. We can formally define a gaussian kernel based geographic-aware recommender as follows:

\[
\hat{s}(u, l) = \sum_{l_u \in L_u} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} \text{dist}(l_u)^2}
\]

The high accuracy of this approach is achieved because the regions of interest of each user are inferred through the gaussian kernels formed around each location the user has visited. When we overlap the kernels of all the locations the user visited, we find his/her regions of interest that might have any kind of shape. An illustrative example of the gaussian kernels that are formed around each location (represented by a dot) can be seen at the figure 2a, where the color red represents a higher probability and blue represents a smaller probability. In figure 2b we can see an example of the overlap of the gaussian kernels forming two distinct regions of interest for the user. Differently from the multi-center gaussian model approach, where some locations are considered noisy data and do not participate in any user center, in the gaussian kernel approach proposed in this section, all the locations that the user has visited are considered to determine the several regions of interest for the users. This is very important because the data of the LBSN are generally very sparse.
5. EXPERIMENTS

In this section we conduct several experiments in different scenarios to evaluate the geographic-aware recommender approaches described in this paper with the goal to discover which one is more accurate.

5.1 Data Description

To conduct the experiments, we used datasets of two very popular LBSN: Foursquare and Gowalla. Both datasets were used and analyzed on recent works [Cho et al. 2011; Cheng et al. 2011]. As usual for recommender systems experiments, we focused on the dense part of the data, i.e., the check-ins of the users that had visited less than 10 locations as well as the check-ins on locations that were visited by less than 10 users were disregarded. Table I summarizes the characteristics of both datasets after this filtering. As we can see on Table I, although there is a smaller number of check-ins in the Gowalla dataset, the number of locations is larger if compared to the Foursquare dataset. Hence, the sparsity of the Gowalla dataset is larger than the sparsity of the Foursquare dataset.

In figure 3, we can see the geographic location of the check-ins in both datasets. It is clear, as figure 3 shows, that the regions with the largest number of LBSN users are USA and Europe. One of the possible explanations for this phenomenon is the fact that due to the high economical standards of these regions, more people have access to smartphones and similar devices.

5.1.1 Cities Analysis. Besides the scenario where the geographic-aware recommender might recommend locations from anywhere around the globe, there is another scenario where the location recommendations can be very important to help the users to discover new places to visit: the scenario...
A Gaussian Kernel Approach for Location Recommendations

<table>
<thead>
<tr>
<th>City</th>
<th>Check-ins</th>
<th>Users</th>
<th>Locations</th>
<th>Sparsity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Philadelphia</td>
<td>21,604</td>
<td>2,817</td>
<td>380</td>
<td>97.99%</td>
</tr>
<tr>
<td>Houston</td>
<td>24,760</td>
<td>2,764</td>
<td>468</td>
<td>98.00%</td>
</tr>
<tr>
<td>London</td>
<td>31,305</td>
<td>3,025</td>
<td>548</td>
<td>98.00%</td>
</tr>
<tr>
<td>Los Angeles</td>
<td>64,494</td>
<td>6,317</td>
<td>1,274</td>
<td>98.20%</td>
</tr>
<tr>
<td>New York</td>
<td>184,760</td>
<td>12,005</td>
<td>3,073</td>
<td>99.50%</td>
</tr>
</tbody>
</table>

Table II: Global Data Set Statistics

where we know exactly which city the user is currently visiting. In such scenario, it does not make sense to recommend places that are not located in the city of interest.

To conduct experiments in the scenario where we know exactly which city the user is interested in, we created five different datasets filtered from the cities with the largest number of check-ins from the foursquare dataset. The reason why we have chosen to evaluate the algorithms only on foursquare for this scenario is because the foursquare dataset is less sparse than the gowalla dataset. The chosen cities were: Philadelphia, Houston, London, Los Angeles and New York. The characteristics of each city are summarized in Table II.

5.2 Evaluation Protocol

To perform the comparison between the geographic-aware recommenders described in Section 4, we split the data into two distinct partitions: the training partition and the testing partition. From all the data available in each dataset, we randomly marked off 10% of users’ visited locations for the testing partition, all the remaining 90% check-ins were used for training. This process of separating randomly the data into training and testing was repeated 10 times for each dataset in order to avoid taking conclusions from biased data. Finally, we had 10 training and testing partitions for the seven datasets: the foursquare and gowalla global datasets and the five cities, described in Section 5.1.1, datasets.

The chosen metric to evaluate the performance of the algorithms was the Normalized Discounted Cumulative Gain (NDCG). The NDCG is a metric that takes into account the ranking of the recommendation, i.e., a relevant item recommended at the top of the recommendation list will have a greater value than a relevant item recommended in lower positions of the list. For all datasets, we computed the NDCG confidence interval with 95% of confidence.

6. RESULTS

The results of the experiments on the foursquare and gowalla global datasets are exposed in figure 4. As we can see, the Gaussian Kernel approach outperforms, with 95% of confidence, all the other approaches in both LBSN datasets. Another interesting result is that the value of the NDCG is higher in the foursquare dataset if compared to the gowalla dataset. One possible explanation for this phenomenon is the fact that the gowalla dataset is more sparse than the foursquare one.

6.1 Results for Cities

The results of the experiments on the cities datasets are depicted in Figure 5. In the scenario where we know which city the user is interested in, the Gaussian Kernel approach also outperforms, with 95% of confidence, all the other approaches in all cities considered. It is important to notice that, differently from the global scenario, the linear regression approach performed worse than the multi-center gaussian model approach. One possible explanation for this is that a power law might not be a good fitting for the check-ins probability distribution in this scenario, in contrast to the global scenario (scenario considered in the original paper), since the users can not visit locations very far from each other.
7. CONCLUSIONS

In this work, we first provide the first comparative study between several location recommenders, based solely on geographic information, using data from two popular location-based social networks: Gowalla and Foursquare. More importantly, we proposed a new geographic-aware recommender based on gaussian kernels. Our experiments showed that such approach outperformed the other approaches proposed in the literature. For future work, we will investigate other aspects of the LBSN domain, such as the temporal and social context. Furthermore, we also plan to investigate the best approach to combine all sub-models into one single recommender that takes into account several contexts inherent to the domain.

REFERENCES


