TEMMUS: A Mobility Predictor based on Temporal Markov Model with User Similarity

Felipe Araújo[1], Denis Rosário[1], Kássio Machado[2], Eduardo Cerqueira[1], and Leandro Villas[3]

[1]Federal University of Pará (UFPA) – Belém – Brazil

[2]Federal University of Minas Gerais (UFMG) – Belo Horizonte – Brazil

[3]University of Campinas (UNICAMP) – Campinas, Brazil

{felipearaujo, denis, cerqueira}@ufpa.br, kassiolsm@dcc.ufmg.br,

leandro@ic.unicamp.br

Abstract. *Location-Based Social Networks (LBSN) data contains spatial, temporal, and social features of user activity, providing valuable information that is currently available on large-scale and low-cost fashion via traditional data collection methods. In this way, LBSN data enables a system to predict user mobility based on spatial, temporal, and social features, which can be used in several areas, such as, device-to-device(D2D) communication, cache, and others. In addition, a Temporal Markov Chain (TMC) is a stochastic model used to model randomly changing systems, such as mobility prediction based on the spatiotemporal factor, for instance, location and day of the week. In this paper, we introduce the TEmporal Markov Model with User Similarity (TEMMUS) mobility prediction model. TEMMUS considers a TMC of variable order based on the day of the week (weekday or weekend) and the user similarity to predict the user's future location. The results highlight a higher accuracy of TEMMUS compared to three state-of-the-art Markov Chains predictors.*

1. Introduction

Location-Based Social Networks (LBSNs), such as Foursquare and Yelp, became popular to provide public data capable of mapping people by means of status, check-ins, and photos shared online, leading to a new urban computing era [Machado et al. 2016]. This is because the growth of mobile networks together with recent advances in localization techniques enhanced social networking services in urban computing, where LBSNs allow users to share their locations and location-related contents, such as geotagged photos and notes [Silva et al. 2019]. In this context, LBSNs users stopped being only consumers to become data producers, offering various research opportunities, such as mobility prediction and recommendation systems. In addition, location data bridges the gap between the physical and digital worlds, enabling a deeper understanding of users' preferences and behavior [Schipor et al. 2017].

LBSN data distinguish from traditional GPS data, *e.g.*, Call Data Records (CDR), mainly in social, geographic, and temporal resolutions, which can be used to model movement pattern and infer user similarity movements [Gao and Liu 2015]. Specifically, LBSN contains spatial, temporal, and social features of user activity, providing valuable information that is currently available on large-scale and low-cost fashion via any traditional data collection methods [Silva et al. 2014]. LBSN is an important tool in urban computing to provide urban data with social aspects, such as the user's preferences and routine. In this way, LBSN data enables us to understand user patterns, city dynamics and related social, economic, and cultural aspects based on social, spatial, and temporal user characteristics [Silva et al. 2019].

LBSN data can be used for capturing user mobility patterns to understand when and where a user commonly goes (location prediction), and exploiting user preferences and location profiles to investigate where and when a user wants to explore (location recommendation). In this sense, mobility prediction based on LBSN data helps several areas of the public and private sectors, such as observation patterns and urban planning [Machado et al. 2017]. For instance, mobility prediction can be used to improve Device-to-Device (D2D) communications in opportunistic networks, where user location is required to make mobile data offloading. In addition, mobility prediction can be used for proactive caching, alleviating back-haul traffic and mitigating latency caused by handovers [Abani et al. 2017].

The study of human movement patterns shows that people's actions are repetitive since they visit specific locations at a relatively fixed time every day [Yan et al. 2017]. In addition, people tend to visit the same places that his/her friends visited, enabling the investigation of social features [Silveira et al. 2016]. However, social information is less effective in predicting a user's repetitive mobility behavior compared to spatial and temporal information, since a user's repetitive mobility behavior is more affected by his personal interests than his friends' preferences [Gao and Liu 2015]. On the other hand, social correlation, called user similarity, can be considered to assist spatial and temporal information for mobility prediction. For instance, the check-ins sequence, *i.e.*, trajectory, from other users are correlated with the user's current check-in, where mobility prediction can be made considering the combination of all the predictions from friends. Hence, spatial, temporal, and social information plays a key role in enabling a highly accurate mobility prediction.

Several methods have been proposed for mobility prediction based on LBSN, where most of them use only the historical trajectories to identify user and group movement patterns. For instance, Markov Chain (MC) is one of the statistical models used in predictive analytics, which aims to find the probability of an event happen given n past events conforming to the order n of the model. Specifically, in the 1st Order MC (1-MC), the future location of a given user is only influenced by the actual location. However, 1st and 2nd Order MC may conflict with some human mobility behavior [Menz et al. 2018]. For instance, the higher the number of transitions "Home to Work" or vice-versa will outweigh any other transition for as many times as this behavior exceeds the next most possible transition regardless the day of the week. Consequently, in many cases, the assumption that "the next place that is going to be visited is only dependent on the current location" becomes false. Hence, it is important to build a Temporal Markov Chain (TMC) for mobility prediction, by considering different models for the weekdays and weekends.

In this paper, we propose the mobility predictor based on TEmporal Markov Model with User Similarity, called of TEMMUS. It ranks a set of possible locations to be visited by a given user based spatial and trajectory similarity. Specifically, TEMMUS takes into account variable order TMC depending on the day of the week and the similarity between users trajectories to rank a set of possible locations that a given user could be. We evaluate the performance TEMMUS using three real datasets, consisting of more than 200 thousand users' records over a period of 18 months. The results highlight a higher performance of TEMMUS compared to other state-of-the-art Markov Chains predictors, showing that the spatiotemporal and user similarity features influence on mobility predictions.

The remainder of this paper is organized as follows. In Section 2, we review relevant related work about mobility prediction based on LBSN data. In Section 3, we introduce the proposed TEMMUS mobility predictor. In Section 4, we describe our data collection procedures, evaluation metrics, and results. In Section 5, we introduce the conclusion and future work of this paper.

2. Related Work

[Abani et al. 2017] proposed a proactive caching strategy for reducing the latency of retrieving predictable content requests in a vehicular network. This proposal considers the individual strategy for mobility prediction since it is based on the history of the object itself. However, this approach is limited by the locations visited by the node, failing in predicting future locations of non-systematic objects due to the individuality of each object. [Nguyen and Giordano 2012] proposed a prediction-based routing algorithm, which considers both spatial and temporal contact dimensions. In this sense, the source knows when and where to start the routing process, which minimizes the network delay and overhead.

Existing mobility prediction models consider the historical record of users. For instance, [Jiang et al. 2016] proposed a method to extract the Region-of-Interest (ROI) from the historical data location. On the other hand, other authors consider not only the history of the user but also the spatial-temporal context to improve the accuracy of the mobility prediction model. For instance, [Wang et al. 2015] modeled the spatial and temporal activity preferences separately and combined them for preference inference. [Gao et al. 2013] proposed a general framework to exploit and model temporal cyclic patterns and their relationships with spatial and social data. The experimental results on two real-world LBSN data-sets that validate the importance of temporal effects in capturing user mobile behavior.

There are some researchers that study the social property on LBSNs to extract user movement and preference pattern. [Cheng et al. 2012] included the social information, and combined the geographical influence into a generalized matrix factorization framework to provide more accurate and efficient Points Of Interests (POI) recommendation. [Silveira et al. 2016] proposed a model to predict human mobility, called MobDatU, which considers data from mobile calls and LBSN data. MobDataU includes social interactions between users as an important factor to predict the next region. [Munjal et al. 2011] proposed SMOOTH, which is a simple and realistic model that leverage several known features of human movement to model human mobility.

Markov Model is one of the statistical models used in predictive analytics. In this way, [Chen et al. 2014] introduced three Markov-based models, namely, Personal Markov Model (PMM), General Markov Model (GMM), and Next Location Predictor

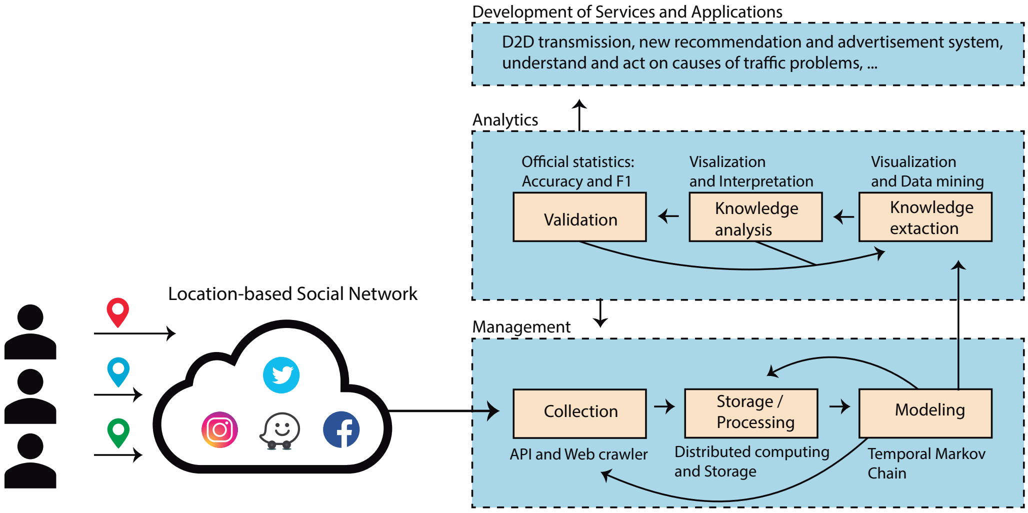
Development of Services and Applications
D2D transmission, new recommendation and advertisement system, understand and act on causes of traffic problems, ...
Analytics
Official statistics: Accuracy and F1
Validation
Visalization and Interpretation
Knowledge analysis
Visualization and Data mining
Knowledge extaction
Location-based Social Network
Management
Collection
API and Web crawler
Storage / Processing
Distributed computing and Storage
Modeling
Temporal Markov Chain

Figure 1. TEMMUS Overview

with Markov Modeling (NLPMM). PMM considers only the mobility of a specific user, *i.e.*, its own past trajectories, to build the mobility prediction model. On the other hand, GMM takes into account the collective aspects of the mobility, *i.e.*, considering not only the movement of a specific node but of all the nodes since they often share similar movement patterns. NLPMM combines PMM and GMM models using linear regression in order to explore the individual and collective aspects of mobility.

Based on our analysis of the state-of-the-art, we conclude that LBSN data can be used for mobility prediction. Hence, it is essential to consider spatial, temporal, social, and user similarity information for mobility prediction based on LBSN data. However, to the best of our knowledge all of these key features have been provided in a unified mobility prediction model.

3. TEMMUS Model

In this section, we introduce the mobility predictor based on TEmporal Markov Model with User Similarity (TEMMUS). It takes into account TMC of a variable order based on the day of the week and the similarity between users trajectories to rank a set of possible locations that a given user could be. TEMMUS mainly consists of some components: *1)* training a variable order TMC for each user with its own trajectories (*i.e.*, individual pattern); *2)* computing the similarity between users based on the spatial locality and the users' trajectories to compute the probability of reaching each possible next location based on MC for each user and *3)* computing the future locations based only on the social feature for predictions with high Entropy (H) levels, *i.e.*, random mobility prediction. Each component is detailed in the following subsections.

Figure 1 introduces an overview of the prediction model. We focus on modeling and analytics. Specifically, it represents active and voluntary user participation, acting as a sort of social sensor, in a distributed process of sharing personal and also data about various aspects of the city in Web services, *i.e.*Foursquare. In order to build the mobility prediction model, we take into account a Temporal Markov Chain (TMC) of variable order based on the day of the week and the similarity between users trajectories to rank a set of possible locations that a given user could be.

3.1. System Model

Users could check-in at a physical place, and let their friends aware of this check-in by means of Foursquare[1], a popular LBSN. The check-in location indicates the current geographical status of a user in the real world and generates the local social networks of the user based on this location. Each check-in c is defined as a 3-tuple $\{u, l, t\}$, where u represents the user; l denotes the location and is defined by latitude and longitude coordinates; and t represents the timestamp. We denote the total set of check-ins as C and the set of check-ins for a specific user u as C_u. The trajectory seq is defined as a time-ordered sequence of check-ins, *i.e.*, locations, that the user just passed. For instance, for a sequence of length five (5), seq is equal to $\{l_1, l_2, l_3, l_4, l_5\}$, indicating that the user just checked-in at these locations in such order. In addition, we consider that two persons are able to interact, as soon as they checked-in at the same location in a one-hour time window period.

We formalize the mobility prediction problem for the user u, *i.e.*, next check-in location l_{n+1}, as follows. Given a user u whose current check-in is c (to venue l at time t), our aim is to rank the set of possible locations so that the next location l_{n+1} to be visited will be ranked at the highest possible location in the list. Hence, the mobility prediction problem is essentially a ranking task, where we compute a ranking score $\hat{r}$ for all venues in L. We consider an MC to solve the mobility prediction problem. Specifically, a Markov Chain is composed by a set of states and transitions, where a state corresponds to a sampling location l_n, while a transition $l_n \rightarrow l_{n+1}$ represents a mobility of a user u from a location l_n to a location l_{n+1}. The state transition might occur based on collective or individual users' patterns. The transition from a state to other has a probability denoted as $P(l_{n+1}|seq)$, meaning the probability of reaching the location l_{n+1} given the trajectory seq.

3.2. Temporal Markov Chain based on Individual Patterns

An often observable behavior of an LBSN user is the departure from home to work in the morning followed by a return to home in the afternoon [Menz et al. 2018]. However, considering the logic of a 1st and 2nd MC, the probability for the transition "Home to Work" and vice-versa increases every time the user follows this behavior regardless the day of the week. Hence, in many cases the assumption that "the next place that is going to be visited is only dependent on the current location" becomes false. One way to overcome this problem is building MC with higher orders, but the performance may not increase as would expect, once this may lead to long trajectories that are not directly related to the user next location, making the mobility prediction difficult.

Clustering user trajectories according to different aspects, such as day of the week, hour, etc, could improve the accuracy of the mobility prediction model. For instance, we must use only weekdays trajectories in order to predict the next location on a Monday. In this way, we build a TMC for the weekdays (from Monday to Friday) and weekends (Saturday and Sunday). We denote the set of trajectories for user u as S_u, and the number of times the user u has checked-in at location l_{n+1} after the sequence of locations seq as $|seq \rightarrow l_{n+1}|$, where $seq \in S_u$ and the size is defined by the order of the model. We compute the probability to reach each location that the user u visited after the trajectory

[1] http://4sq.com

of check-ins seq. Therefore, the probability $P(l_{n+1}|seq)$ of the user sign at location l_{n+1} in the next time given the sequence of locations seq is:

$$P(l_{n+1}|seq) = \frac{|seq \to l_{n+1}|}{\sum\limits_{i=1}^{N} |seq \to l_i| \; \forall \; l_i \in L}; \tag{1}$$

3.3. Similarity

Users tend to visit specific locations at a relatively fixed time every day. In this way, we find users similar routines for mobility prediction. We computed the similarity from two perspectives: spatial similarity and trajectory similarity. For the first one, we calculated the relative frequency $RF(u,l)$ for a user u and a location l based on the number of times the user u visited the location l divided by the total number of visits. Therefore, $RF(u,l)$ is is given by:

$$RF(u,l) = \frac{|l|}{\sum\limits_{i=1}^{N} |l_i|}; \forall \; l_i \in L \tag{2}$$

After that, we built the matrix of relative frequencies (RF) for all users and locations, where a row indicates the user, and the column the location. Hence, the probability distribution of a user u is given by $RF(u)$, where it indicates the whole row of the matrix.

$$RF = \begin{bmatrix} RF(u_1,l_1) & RF(u_1,l_2) & RF(u_1,l_3) & \cdots & RF(u_1,l_N) \\ RF(u_2,l_1) & RF(u_2,l_2) & RF(u_2,l_3) & \cdots & RF(u_2,l_N) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ RF(u_M,l_1) & RF(u_M,l_2) & RF(u_M,l_3) & \cdots & RF(u_M,l_N) \end{bmatrix} \tag{3}$$

User similarity enhances the spatial and temporal information for mobility prediction since mobility from a user y could be correlated with a user x. In this sense, we computed the similarity between any two users x and y, denoted as $SRE(x,y)$, based on the Kullback-Lieber divergence (KL), also known as Relative Entropy. We considered similar users those whose SRE metric was above a given threshold ths. Therefore, $SRE(x,y)$ is given by:

$$SRE(x,y) = 1 - KL = 1 - \left[\sum_{i=1}^{N} RF(x,l_i) \times log \frac{RF(x,l_i)}{RF(y,l_i)} \right] ; \text{where } x \neq y \tag{4}$$

We determined the trajectory similarity based on the Adjusted Jaccard Index (AJI) for each user trajectory set S_u, where the index u indicates the user. For instance, let $seq_x \in S_x$ and $seq_y \in S_y$ be the trajectories of the users x and y, respectively. We computed the total relative abundances $V_x = \{p_1, p_2, \cdots, p_J\}$, for the user x, and $V_y = \{\pi_1, \pi_2, \cdots, \pi_K\}$ for the user y, where J and K are the number of unique locations. Then, the index AJI was computed considering the relative number of the sharing locations in these trajectories seq_x and seq_y (Eq. 5 and 6). Given that, we quantified the trajectory similarity using the intersections between users x and y in terms of a weighted average of all possible similarities in trajectories that belong to them, denoted as $SS(x, y)$. Therefore, $SS(x, y)$ is given by:

$$V_{xy} = |V_x| \times |V_y| \,; \forall \, S_x \cap S_y \tag{5}$$

$$AJI(seq_x, seq_y) = \frac{|V_{xy}|}{|V_x + V_y - V_{xy}|} \tag{6}$$

$$SS(x, y) = \frac{1}{|S_x| \times |S_y|} \sum_{seq_x} \sum_{seq_y} AJI(seq_x, seq_y) \tag{7}$$

After quantifying the similarities, we considered them as the weight (W) of the probability of reaching each possible next location taking into account the mean of the two kinds of similarities: spatial and trajectory. Therefore, the weight W and the vector of possible next location ($\hat{r}$) are given by:

$$W_{x,y} = \frac{SRE(x, y) + SS(x, y)}{2} \tag{8}$$

$$\hat{r}(x) = \underset{l \in L}{\operatorname{argmax}} \left\{ \sum_{y \in U} W_{x,y} \times P(l_{n+1} = l | seq_y) \right\} \tag{9}$$

3.4. Temporal Markov Chain based on Collective Patterns

In this last component, we classified our prediction according to the Shannon Entropy, also known as Entropy (H), to measure the uncertainty of our approach (eq. 10). In other words, the more uniform the vector $\hat{r}(x)$ is, the more randomness it has. Therefore, for those users whose next locations vector $\hat{r}(x)$ was above a threshold $thse$, we recalculated the probability of next locations using the collective patterns instead of the individual ones. In this sense, based on the argument that “users tend to visit specific locations at a relatively fixed time every day, and also they tend to visit the same places that his/her friends visited”, we built a TMC model leveraging the human interactions.

$$H = -\sum \hat{r}(x) log(\hat{r}(x)) \tag{10}$$

We considered that two or more users were able to interact since they have been in the same area/location in a one-hour time window. Consequently, for each user, we built a list of users that were in the same area following the specifications above, denoted as F. For instance, let x be a user who checked-in at a location l at timestamp t, hence, $c =< x, l, t >$. All users that checked-in at l within a period less than one hour are added in the F_x list, where the index x indicates that F is a list of the user x. Then, the vector containing the user possible next location, $\hat{r}(x)$, is given by:

$$\hat{r}(x) = \underset{l \in L}{\operatorname{argmax}} \{ P(l_{n+1} = l | seq_f) \forall f \in F \} \tag{11}$$

In summary, the probability $\hat{r}(x)$ of user x check-in next location is:

$$\hat{r}(x) = \begin{cases} \text{argmax}_{l \in L}[P(l_{n+1} = l | seq_f) \forall f \in F] \text{ if } H > thse \\ \text{argmax}_{l \in L}[\sum\limits_{y \in U} W_{x,y} \times P(l_{n+1} = l | seq_y)] \text{ otherwise} \end{cases} \tag{12}$$

4. Evaluation

In this section, we describe our data collection methodology and metrics used to evaluate TEMMUS.

4.1. Data Collection Methodology and Evaluation Metrics

We used the regions of Tokyo, New York and Istanbul from Global-scale Check-in Dataset. Tokyo dataset contains about 1 million check-ins by 12592 users at 82258 venues over a period of 18 months (from April 2012 to September 2013). The region of New York contains 237268 check-ins of 6250 different users at 25049 locations against over 2.5 millions of check-ins of 31570 users and 111158 different venues collected in the same period from the region of Istanbul. This dataset has the following information: (i) User ID (anonimyzed); (ii) Latitude; (iii) Longitude; (iv) Timestamp/DateTime; (v) Venue ID.

In order to separate the dataset into train and test, we used a variation of the k-fold method, following the temporal aspects of the check-ins. Therefore, we defined k = 100, and the size of the train dataset was limited up to twice the test. For example, let D be a set of 101 time-ordered check-ins groups of equal size, where $D = [0, 1, 2, 3 \cdots 100]$. The train and test datasets are given by the following method:

$$\begin{gathered} TRAIN : [\,0\,]\, TEST : [1] \\ TRAIN : [0, 1]\, TEST : [2] \\ TRAIN : [1, 2]\, TEST : [3] \\ TRAIN : [2, 3]\, TEST : [4] \\ TRAIN : [3, 4]\, TEST : [5] \\ \vdots \\ TRAIN : [98, 99]\, TEST : [100] \end{gathered} \tag{13}$$

We can distinguish our prediction according to the type: classification or regression. In the first one, the output is a categorical class label. On the other hand, in the regression problem, the model learns a continuous function. It is common for classification models to predict a continuous value as the probability of a given example belonging to each output class. The probabilities can be interpreted as the likelihood or confidence of a given example belonging to each class. A predicted probability can be converted into a class value by selecting the class label that has the highest probability. In this paper, we return a vector containing the highest K predicted probabilities, indicating that the user can check-in at any of these K locations at next time.

	Day				
	1	2	3	4	5
Actual location	A	A	A	B	A
Predicted location	A	A	B	B	A
Correct	Yes	Yes	No	Yes	Yes

Table 1. Example of location prediction

For the first type of prediction, there are mainly three subsets of classifications: binary, multiclass and multilabel, each one containing its own subsets. For instance, the binary classification usually predicts the probability of a target variable to be Yes/No. On the other hand, the multiclass classification means a classification task with more than two classes; e.g., classify a set of images of fruits which may be oranges, apples, or pears.

In addition, it makes the assumption that each sample is assigned to one and only one label: fruit can be either an apple or a pear but not both at the same time. The multilabel classification problem also applies to tasks with more than two classes, however, the labels are not mutually exclusives, meaning that more than one label can be assigned at the same time. For instance, topics that are relevant to a document. A text might be about any of religion, politics, finance or education at the same time or none of these.

The tasks of multiclass as well as multilabel can be decomposed in multiple binary classification problems, where it can be categorized into One vs All (OvA) and One vs One (OvO). In the OvA strategy involves training a single classifier per class, with the samples of that class as positive samples and all other samples as negatives. In contrast, in OvO the classes are combined in pairs. For instance, if the multiclass problem has n classes, the OvO approach will be composed by $n(n-1)/2$ pairs, and then the label assigning stage is performed by majority voting. In this paper, we decompose our predicting problem into an OvA classifier.

In order to evaluate the classification performance, we compare different Markov-based methods using two metrics: accuracy and F1 metrics (eqs. 14 and 15). The former measures the number of correct predictions among the predictions made. However, while it is widely used in classification problems, for imbalanced datasets it can mislead. It occurs since a large number of classes may be wrongly classified. For instance, a predictor can be evaluated with a high overall accuracy score due to a great number of correct predictions made for a specific class while it presents a poor score for all others. Consequently, we compute the accuracy for each class and takes the average.

$$accuracy = \frac{\text{no. correctly predicted}}{\text{no. predictions}} \tag{14}$$

Following this same strategy, we compute the F1 metric. It is the weighted average of Precision and Recall, where the First is the ratio of correctly predicted positive observations to the total predicted positive observations while the Second is the ratio of correctly predicted positive to the total number of actually positive observations. For in-

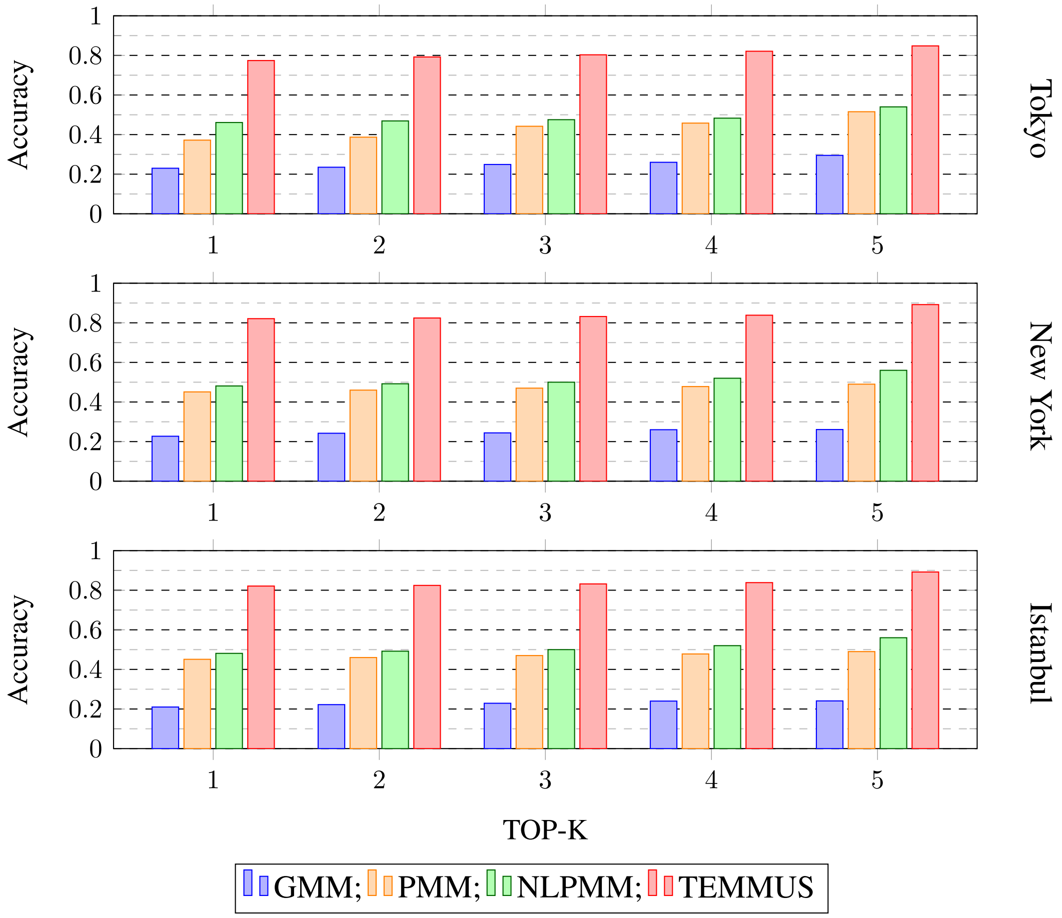

Figure 2. Accuracy according to the TOP-K predicted locations

stance, considering Table 1 that contains the set of predicted locations and the users' real locations for 5 different days, we have predicted that two locations are B (days 3 and 4). Therefore, there is a mistake on the third prediction, consequently, since there two days that are predicted as B (highlighted in red), and one location was really B, the precision is $1/2$ in this case while the recall is 1 (perfect).

$$F1 = 2 \times \frac{precision}{precision + recall} \tag{15}$$

4.2. Results

Figure 2 illustrates the accuracy metric according to the size of the next location vector for all the predictors and for all the scenarios: Tokyo, New York, and Istanbul. In this context, GMM has the worst performance considering. It occurs because this predictor uses only the collective patterns to compute the user's possible next location, highlighting the idea that the mobility behavior of a user is more influenced by his personal interests than the collective ones. In this sense, PMM considers only the individual patterns to predict the next location and outperforms the GMM's accuracy by 77.5%, for the list size equal to 3 (K=3) for the scenario of Tokyo. In addition, in this same scenario, the minor difference between GMM and PMM occurs for the lowest size of the list because PMM is more affected by the increase of the size if compared to GMM.

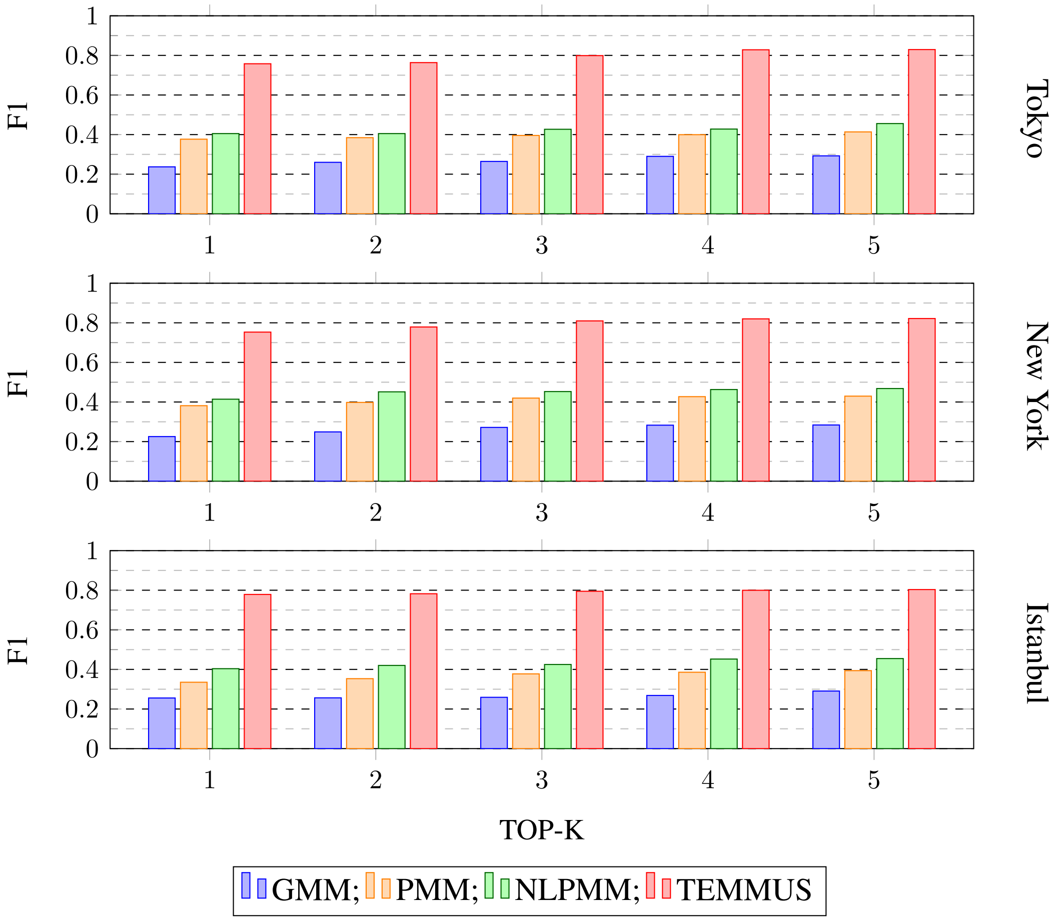

Figure 3. F1 according to the TOP-K predicted locations

The combination of the individual and collective patterns through a linear regression is used in the NLPMM predictor. Therefore, it shows a significant improvement if compared to the other approaches mentioned above. However, it is important to note that this model gives more importance for the individual patterns than the collective patterns using a factor α. In this context, we use the α-value that most fit for each scenario, considering the highest accuracy average. For instance, for the New York region, we use $alpha = 0.7$, indicating that the individual pattern has a weight of 70% while the collective 30%.

Our approach performs the best among all the predictors in the three scenarios due to the use of spatial, temporal and social aspects to predict user next location. The accuracy metric achieves an improvement of more than 2 times – considering the relative percentages – compared with the worst predictor, GMM. Considering the Second best predictor, NLPMM, our model outperforms up to 70% as long as it uses a TMC with user similarity to compute the prediction.

Similarly, Figure 3 presents the F1 metric according to the size of the next location vector for all the scenarios. In this context, TEMMUS performed best among all the predictors and all regions: Tokyo, New York, and Istanbul. More specifically, it improves almost twice the F1 score of the worst predictor. For example, for the region of Tokyo, TEMMUS outperforms GMM by 195%. Our approach is followed by NLPMM, PMM, and GMM as the best average results. It occurs since a lot of predictability is encoded in the sequence order of place visits, but a significant share of predictability is also encoded in a temporal order of visitation pattern, implying that the models can be improved by the inclusion of temporal information (as in our proposed model).

Similarly, the inclusion of the social factor to compute the next location improves the performance of our model. This makes the model dynamic because generates different predictions at different times. Also, it enables predicting a much broader range of locations, which is a big contributor to higher recall scores, implying a much higher predictive power. Additionally, the model also makes conceptually more meaningful predictions. For example, TEMMUS does not predict work location as the next location during the weekends as long as it is clear from the temporal factor that this kind of location is not frequently visited during the weekends.

From the results, we can conclude that the proposed method is a promising solution for predicting human mobility from LBSN. By exploiting both trajectory and spatial similarities into a ranking-classification approach, based on historical visiting information and user trajectory similarity, it inherited the benefits of both the methods, exhibiting high accuracy and F1 score rates. Nevertheless, it is worth noticing that the proposed similarity metrics (spatial and trajectory) resulted to be effective in predicting user future location.

5. Conclusion

In the paper is presented an approach to predict human mobility by exploiting LBSN data. TEMMUS predicts the future positions visited by a user at a specific time by exploiting her typical mobility behavior and the ones of other users in the given geographic region. More specifically, we identified a mobility pattern being as similar as possible to the current user trajectory and routine. Then, we computed the top-k future possible positions.

After that, we quantified this similarity using the Entropy measure. For those users whose location prediction entropy score was high or none user was similar, we recomputed the top-k next possible locations using the human physical interaction approach. Therefore, we considered that users who were in the same geographic region able to interact. Moreover, we evaluated our model using the LBSN datasets collected in Tokyo, New York, and Istanbul cities from April 2012 to September 2013. Additionally, our approach achieved high accuracy and F1 rates, enabling the application of this model in D2D transmissions, where the location is required.

Acknowledgement

This study was financed in part by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. The author Denis Rosario would like to thank the National Council for Scientific and Technological Development (CNPq) for the financial support through Grant 431474/2016-8.

References

[Abani et al. 2017] Abani, N., Braun, T., and Gerla, M. (2017). Proactive caching with mobility prediction under uncertainty in information-centric networks. In *Proceedings of the 4th ACM Conference on Information-Centric Networking, ICN 2017, Berlin, Germany, September 26-28, 2017*, pages 88–97.

[Chen et al. 2014] Chen, M., Liu, Y., and Yu, X. (2014). Nlpmm: A next location predictor with markov modeling. In Tseng, V. S., Ho, T. B., Zhou, Z.-H., Chen, A. L. P., and Kao, H.-Y., editors, *Advances in Knowledge Discovery and Data Mining*, pages 186–197, Cham. Springer International Publishing.

[Cheng et al. 2012] Cheng, C., Yang, H., King, I., and Lyu, M. R. (2012). Fused matrix factorization with geographical and social influence in location-based social networks. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, AAAI'12, pages 17–23. AAAI Press.

[Gao and Liu 2015] Gao, H. and Liu, H. (2015). *Mining human mobile behavior with location-based social networks*. Morgan & Claypool.

[Gao et al. 2013] Gao, H., Tang, J., Hu, X., and Liu, H. (2013). Modeling temporal effects of human mobile behavior on location-based social networks. In *22nd ACM International Conference on Information & Knowledge Management*, CIKM '13, pages 1673–1678. ACM.

[Jiang et al. 2016] Jiang, J., Pan, C., Liu, H., and Yang, G. (2016). Predicting human mobility based on location data modeled by markov chains. *Fourth International Conference on Ubiquitous Positioning, Indoor Navigation and Location Based Services (UPINLBS)*, pages 145–151.

[Machado et al. 2016] Machado, K., Boukerche, A., Cerqueira, E., and Loureiro, A. A. (2016). Long-term spatiotemporal analysis of social media for device-to-device networks. In *IEEE Global Communications Conference (GLOBECOM)*, pages 1–6. IEEE.

[Machado et al. 2017] Machado, K., Boukerche, A., Cerqueira, E., and Loureiro, A. A. (2017). A socially-aware in-network caching framework for the next generation of wireless networks. *IEEE Communications Magazine*, 55(12):38–43.

[Menz et al. 2018] Menz, L., Herberth, R., Luo, C., Gauterin, F., Gerlicher, A., and Wang, Q. (2018). An improved method for mobility prediction using a markov model and density estimation. In *2018 IEEE Wireless Communications and Networking Conference, WCNC 2018, Barcelona, Spain, April 15-18, 2018*, pages 1–6.

[Munjal et al. 2011] Munjal, A., Camp, T., and Navidi, W. C. (2011). Smooth: a simple way to model human mobility. In *Proceedings of the 14th ACM international conference on Modeling, analysis and simulation of wireless and mobile systems*, pages 351–360. ACM.

[Nguyen and Giordano 2012] Nguyen, H. A. and Giordano, S. (2012). Context information prediction for social-based routing in opportunistic networks. *Ad Hoc Networks*, 10(8):1557 – 1569.

[Schipor et al. 2017] Schipor, O.-A., Wu, W., Tsai, W.-T., and Vatavu, R.-D. (2017). Software architecture design for spatially-indexed media in smart environments. *Advances in Electrical and Computer Engineering*, 17(2):17–23.

[Silva et al. 2014] Silva, T. H., Melo, P. O. S. V. D., Almeida, J. M., and Loureiro, A. A. F. (2014). Large-scale study of city dynamics and urban social behavior using participatory sensing. *IEEE Wireless Communications*, 21(1):42–51.

[Silva et al. 2019] Silva, T. H., Viana, A., Benevenuto, F., Villas, L., Salles, J., Loureiro, A., and Quercia, D. (2019). Urban computing leveraging location-based social network data: a survey. *ACM Computing Surveys*, pages 1–37.

[Silveira et al. 2016] Silveira, L. M., de Almeida, J. M., Marques-Neto, H. T., Sarraute, C., and Ziviani, A. (2016). Mobhet: Predicting human mobility using heterogeneous data sources. *Computer Communications*, 95:54 – 68. Mobile Traffic Analytics.

[Wang et al. 2015] Wang, Y., Yuan, N. J., Lian, D., Xu, L., Xie, X., Chen, E., and Rui, Y. (2015). Regularity and conformity: Location prediction using heterogeneous mobility data. In *21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15. ACM.

[Yan et al. 2017] Yan, X.-Y., Wang, W.-X., Gao, Z.-Y., and Lai, Y.-C. (2017). Universal model of individual and population mobility on diverse spatial scales. *Nature Communications*, 8(1):1639.