Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)

A Collaborative Filtering Approach to Citywide Human Mobility Completion from Sparse Call Records

Zipei Fan,[1] Ayumi Arai,[1] Xuan Song,[1] Apichon Witayangkurn,[2]
Hiroshi Kanasugi,[1] Ryosuke Shibasaki[1]
[1]Center for Spatial Information Science, the University of Tokyo, Japan
[2]Asian Institute of Technology, Pathumthani, Thailand
{fanzipei, apichon, yok}@iis.u-tokyo.ac.jp, {arai, songxuan, shiba}@csis.u-tokyo.ac.jp

Abstract

Most of human mobility big datasets available by now, for example call detail records or twitter data with geotag, are sparse and heavily biased. As a result, using such kind of data to directly represent real-world human mobility is unreliable and problematic. However, difficult though it is, a completion of human mobility turns out to be a promising way to minimize the issues of sparsity and bias. In this paper, we model the completion problem as a recommender system and therefore solve this problem in a collaborative filtering (CF) framework. We propose a spatio-temporal CF that simultaneously infers the topic distribution over users, time-of-days, days as well as locations, and then use the topic distributions to estimate a posterior over locations and infer the optimal location sequence in a Hidden Markov Model considering the spatio-temporal continuity. We apply and evaluate our algorithm using a real-world Call Detail Records dataset from Bangladesh and give an application on Dynamic Census, which incorporates the survey data from cell phone users to generate an hourly population distribution with attributes.

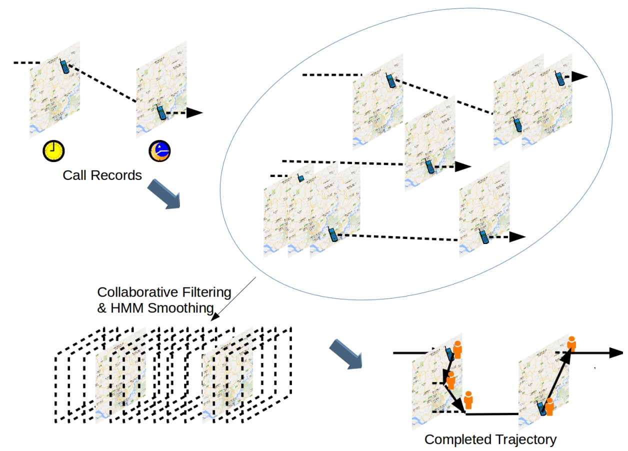

Figure 1: A collaborative filtering framework of human mobility completion. Observing sparse call records (top-left), we fuse the cues from the call records of on other days as well as other users' records in a collaborative filtering framework (top-right) to interpolate the blank time spans (bottom-left) to find out an optimal location sequence between sparse records (bottom-right).

1 Introduction

With the popularization of cell phones and the exploding growth of the human mobility big data available, the study of city has been vitalized by an accurate recording of human movements. Nowadays, human can be sensed nearly real-time via a variety of negligible-cost ways, which greatly deduces the granularity of our understanding of the city from yearly/monthly to hourly or even finer. This enables a variety of time-sensitive applications such as disaster management, epidemic prediction and transportation regulation.

In general, there are two categories of human mobility big data collection, namely passive recording and active recording. Passive recording means when the positioning device is on, the device is localized and the location is recorded periodically for every certain time interval. Although passive recording produces a more objective and denser data, it suffers from severe problems such as the limitation of devices, power consumption and privacy issues. In contrast, when the data is collected in an active way, the device is localized only when the user uses relevant functions, for example making a call or tweeting with geotag. Compared with passive recording, issues of device limitation, power and privacy can be significantly minimized, but at the expense of more bias (e.g. users make calls or tweets much more during daytime than nighttime) and sparsity of the data as shown on the top-left in Fig. 1(on average for each user, there are only about 3 records per day).

Thus, directly using active recording data to represent human trajectories would heavily biased towards their usage of cell phones, and a simple linear interpolation or routing algorithm for a long time span would be problematic due to the sparseness of the data. Most of the existing linear interpolation or routing algorithm assumes the user sets off from the starting point and arrives at the ending point by making a full use of the time span. This works well for a short term interpolation, however turns out to be problematic when the time span is hours or even days. Imagine one makes a phone call at midnight and makes the next one in the next morning at workplace, linear interpolation or routing algorithm would most probably make a completion that the user is moving whole night traveling to his workplace, ignoring the pattern of users' daily routines (he is most probably stay at home sleeping overnight). To counter this issue, in this paper, we propose a novel algorithm to complete the citywide human mobility from sparse call records in a Latent Dirichlet Allocation (LDA) based collaborative filtering (CF) framework, which incorporates the cues from the call records on this day, on other days and records from other users as well.

Collaborative filtering, which is widely used in recommendation system, is a method of making prediction about a user's preference or interest on items. The basic assumption is if users A and B share similar interests on some items (e.g. bought same goods), user A would probably have the similar opinion with B on other items (e.g. A may probably be interested in what B also buy). In this work, as shown in Fig. 1, we make a similar but more sophisticated assumption by taking the spatio-temporal factor into consideration, by making the "collaboration" not only with other users, but also with the records on different days from the same user. The former one gives a rich information of the typical routines of people in Dhaka and profiles the topic distribution of a day, while the later one extracts the geographical topic distribution that characterizes the user's important places such as home and work. In addition, we use the "filtering" results as a posterior over location and incorporate with Hidden Markov Model (HMM) [Rabiner, 1990] to smooth our interpolation results by exploiting the spatio-temporal continuity.

The contributions of this paper can be summarized as follows:

- We formulate the human mobility completion problem in a CF framework, thus transfer the emerging sparse trajectory interpolation problem into a well-studied recommender system problem.
- Considering about the particularity of human mobility completion, we propose a novel topic model that fuses the information of user important locations, topic distribution with respect to user, day and time-of-day.
- We apply and evaluate our algorithm on real-world data, and show potential applications based on this work.

2 Data

We use a Call Detail Records (CDRs) dataset in this work, one from a leading telecommunication company in Bangladesh. The data covers the records of all antennas (over 4 thousands) that are located in the Greater Dhaka in Bangladesh, from 10 million cell phone users within a period of four months. When fusing with other sources of data (census data, disaster or epidemic data), this dataset can used in a variety of urban computing scenarios such as disaster management, traffic regulation and epidemic control. To minimize the issues caused by the sparseness and bias of call records, we make a completion of the data as a key procedure for the preprocessing of the data.

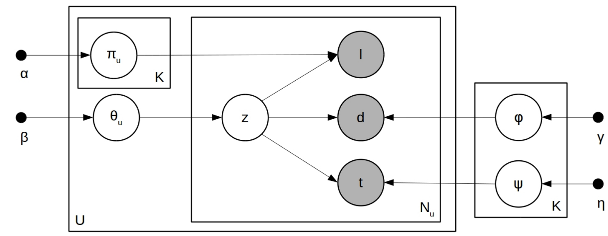

Figure 2: Graphical model for human mobility topic model.

3 Preliminary

We give our key definitions and explanations that are helpful before going deep into details.

Definition 1 (CDR dataset): The original dataset can be described by a set of 4-tuple:

$$X = \{x = (u, d, t, l)\}$$

where u, d, t, l are the user ID, the index of day, time-of-day (in our experiment, we discretize the time by hour) and the ID of the antenna of cell towers, which can be used as a rough estimation of the user's location. Looking up the database of the antenna, we can easily retrieve the coordinates of the antennas, namely the latitude $l.lat$ and longitude $l.lon$.

Definition 2 (Blank time span): a blank time span sp is defined as an interval:

$$sp = [(t_{start}, d_{start}), (t_{end}, d_{end})]$$

that lays between the two consecutive records (including the period before the first record and after the last record) from a single user. The collection of all the blank time spans is denoted as SP.

Note that blank time spans are what we intend to complete in this paper, and from a Bayesian network point of view, the latent locations l within each blank time span is conditionally dependent while independent for inter-blank time spans. This will help us simplify the model by decomposing the joint probability.

Definition 3 (Human mobility completion): A completion of human mobility C is an assignment of locations at each time slot within each blank time span. Observing the call records dataset X, we intend to find out the optimal completion by searching for the completion with the highest probability of $p(C|X)$.

4 Topic Distribution Inference

In this section, we detail our spatio-temporal topic model to infer the latent topic distribution over time-of-day, day and location. We make an analogy between "document", "vocabulary" in text mining and "one-day records from single user", "location/time-slot/sample day" respectively, and therefore discover the patterns underlain in the human mobility in a LDA model. Different from classical LDA model, the "word" in the document is sampled from three distributions rather than one, namely distribution over day, time-of-day and location. Notably, although people share some typical common daily routines (e.g. salaryman's life routine), human trajectories differ with each other for everyone has his own important locations (e.g. home/work location). To formulate this characteristics of human mobility, we put the topic distribution over location into the user plate in the graphical model, as shown in Fig 2.

We describe the generative process as:

- Draw θ_u from $Dirichlet\left(\beta\right)$
- Draw $\pi_{u,z}$ from $Dirichlet\left(\alpha\right)$
- Draw ϕ_z from $Dirichlet\left(\gamma\right)$
- Draw ψ_z from $Dirichlet\left(\eta\right)$
- For each user u:
  - For each record x in X_u:
    * Draw a topic $z = k$ from multinomial distribution θ_u
    * Draw a location l from multinomial distribution $\pi_{u,k}$
    * Draw a time-of-day t from multinomial distribution ϕ_k
    * Draw a day d from multinomial distribution ψ_k

where α, β, γ, η are the Dirichlet priors. θ_u is the topic distribution of user u, while $\pi_{u,k}$ is the k-th topic distribution over location, ϕ_k and ϕ_k are time-of-day and day distribution of topic k respectively.

We make the inference of this model via Gibbs Sampling. Here, we omit the details on the produces of Gibbs sampling inference (for more details please refer to [Griffiths and Steyvers, 2004] while only give the conditional topic distribution over the latent topic assignment $z = k$.

$$\begin{aligned} p\left(z=k|Z_{\neg z},X\right) &= \theta_u\left(k\right)\cdot\pi_{u,k}\left(l\right)\cdot\phi_k\left(t\right)\cdot\psi_k\left(d\right) \\ &= \frac{n_{u,k}^{\neg x}+\beta}{n_{u,:}^{\neg x}+K\beta}\cdot\frac{n_{u,l,k}^{\neg x}+\alpha}{n_{n,:,k}^{\neg x}+L\alpha}\cdot \\ &\quad \frac{n_{u,t,k}^{\neg x}+\gamma}{n_{u,:,k}^{\neg x}+T\gamma}\cdot\frac{n_{u,d,k}^{\neg x}+\eta}{n_{u,:,k}^{\neg x}+D\eta} \end{aligned} \tag{1}$$

where $Z_{\neg z}$ is the collection of all the latent topic assignment except z, $n_*^{\neg x}$ is the number of records suffice the subscripts $*$ except record x, and K, L, T, D is the number of topics, locations, time slots and sample days respectively.

We visualize the results of inferred topic distribution over time-of-day and day in Fig. 3 and Fig. 4 respectively. From the variation of topic distribution, we can easily see that each topic describes a different temporal pattern in a day. For example, topic 5, which has a higher proportion during daytime while higher on weekdays, most probably describes a pattern that relates to "work". In contrast, topic 1 and topic 9, which takes up a dominating proportion during nighttime, most probably describes a pattern that relates to "home". Note that there is dramatic change of the topic distribution around Aug. 9th. On that day, people in Bangladesh celebrates Eid al-Fitr festivals, which is the most important religious festival to the people there.

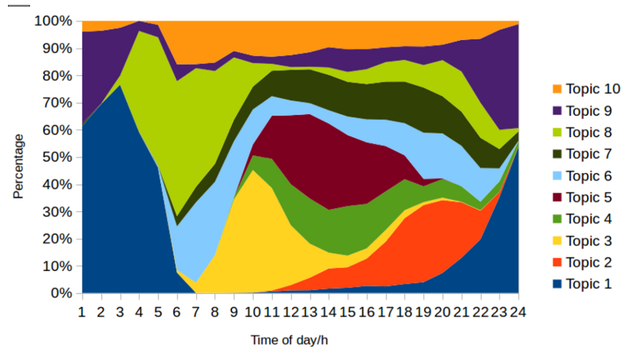

Figure 3: Topic distribution over "time-of-day".

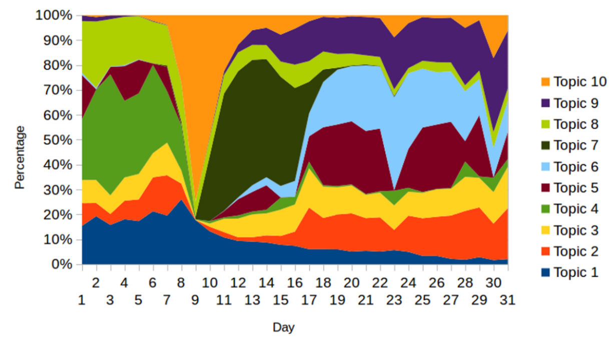

Figure 4: Topic distribution over "day".

5 HMM based Completion

In a traditional collaborative filtering framework, the users' preference to a new item can be predicted using the topic distributions. Analogy to the recommendation system, user's probable location in the blank time span can also be predicted using the topic distributions we inferred in the previous section in a similar way. Following our definitions, we formulate the completion problem in a generative graphical model and decompose the joint probability as:

$$p\left(C|X\right) = \prod_{sp\in SP} p\left(C_{sp}|X\right) \tag{2}$$

where C_{sp} is the completion on blank time span sp. This decomposition is made from the conditional independence of blank time spans, as is said in definition 2. Then the completion on each blank time span can be formulated into the graphical model shown in Fig 5, the probability can be further decomposed into:

$$\begin{aligned} &p\left(C_{sp}|X\right) = p\left(l_0\right) \\ &\prod_{\tau} p\left(l_\tau|l_{\tau-1}\right)\sum_k p\left(\omega_\tau^k|l_\tau,\pi_u\right)p\left(X_{d_\tau},d_\tau,t_\tau,\pi_u,\phi,\psi|\omega_\tau^k\right) \\ &p\left(l_{T_{sp}}|l_{T_{sp}-1}\right) \end{aligned} \tag{3}$$

where T_{sp} is the length of the blank time span sp, τ is an iterative variable that iterates the blank time span sp, d_τ, t_τ are the day and time-of-day which τ corresponds to, and X_{d_τ} is the records on day d_τ. ω_τ is the latent topic distribution

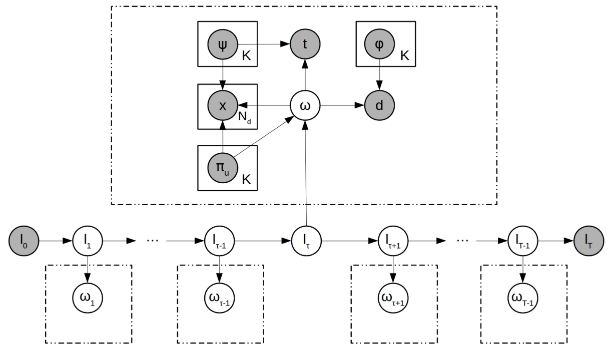

Figure 5: HMM for human mobility completion. We use the topic distributions extracted from previous section to estimate an emission probability of latent locations, and define a inter-antenna transition probability to characterize the enforce the spatio-temporal continuity.

of time slot τ and ω_τ^k is the distribution with respect to k-th topic. π_u, ϕ, ψ are the topic distributions estimated by previous section. To simplify the model, we assume that these distributions are the inherent property of X rather than the latent variables, therefore they are the observed variables in this completion phase. Note that in this decomposition, we also make an assumption that conditioned on $l_{\tau-1}$ and the topic distributions, l_τ is independent with the records on other days. We will explain this in more details in 5.1.

In the following subsections, we will introduce a further decomposition of the likelihood probability $p\left(X_{d_\tau}, d_\tau, t_\tau, \pi_u, \phi, \psi | \omega_\tau^k\right)$ and $p\left(\omega_\tau^k | l_\tau, \pi_u\right)$ in 5.1, and give the more details on the definition of transition matrix, the entries of which is $p\left(l_\tau | l_{\tau-1}\right)$, as well as the probability of the initial and ending locations in 5.2.

5.1 Emission Probability

When we observe some sparse call records of a user on one day, we can have some cues on what kind of day is to him/her and therefore estimate a location likelihood probability for each time slot in the blank time spans on this day. For example, when we observe a user go to work place on a day, then he/she would most probably also go to station to take public transportation rather than go to shopping mall during the daytime on the same day. Moreover, even there are no records on a day, we can still have some cues on the topic distribution on this day. For example, on Aug. 9th the particular religious festival for Bangladeshis, even a user does not make any phone call on that day, he is most probably having a similar "festival" pattern as others from the prior knowledge of topic distribution over day.

Following this intuition, we draw a graphical model in Fig. 5, and give a further decomposition:

$$p(X_{d_\tau}, d_\tau, t_\tau, \pi_u, \phi, \psi | \omega_\tau^k) \propto p\left(X_{d_\tau} | \omega_\tau^k, \psi, \pi_u\right) \cdot p\left(d_\tau | \omega_\tau^k, \phi\right) \cdot p\left(t_\tau | \omega_\tau^k, \psi\right) \tag{4}$$

These three terms, which are likelihood probability distributions over the latent topic distribution ω_τ, fuse the cues from three different aspects: the first one infers the topic distribution on the day d_τ using the records X_{d_τ} based on the topic distribution over location from his/her own, which is highly user-specific but unstable because probably there are no records on that day or the records are weak in cuing the topic distribution. This likelihood of observing X_{d_τ} by giving ω_τ with respect to topic k is calculated by:

$$p\left(X_{d_\tau} | \omega_\tau^k, \psi, \pi_u\right) \propto \prod_{x \in X_{d_\tau}} \psi_k\left(x.t\right) \cdot \pi_{u,k}\left(x.l\right) \tag{5}$$

the second one estimates the probability from drawing d_τ from a global topic distribution over day, which provides a less user-specific information but more reliable distribution to overcome the shortcoming when the information from the first term is weak or even missing. This probability is estimated by:

$$p\left(d_\tau | \omega_\tau^k, \phi\right) = \phi_k\left(d_\tau\right) \tag{6}$$

while the third term concerns about the time-of-day t_τ, which characterizes the topic variation within one day. Similar to Equation 6, this term can be calculated by:

$$p\left(t_\tau | \omega_\tau^k, \psi\right) = \psi_k\left(t_\tau\right) \tag{7}$$

Moreover, to complete an estimation of emission probability of HMM, we also need to define $p\left(\omega_\tau^k | l_\tau, \pi_u\right)$. Intuitively, this term characterizes a prior of topic distribution ω_τ for time slot τ giving the latent state (location) l_τ, and topic distributions over location π_u.

$$p\left(\omega_\tau^k | l_\tau, \pi_u\right) \propto \pi_{u,k}\left(l_\tau\right) \tag{8}$$

5.2 Inter-antenna Transition Matrix

To enforce the spatio-temporal continuity, we define a inter-antenna transition matrix which encodes the state transition probability in a HMM. The entries a_{ij} in the matrix is calculated as:

$$a_{ij} \propto \begin{cases} exp\left(-E \cdot dis\left(l_i, l_j\right)\right), & \text{for } i \neq j \\ 1 + \delta, & \text{for } i = j \end{cases} \tag{9}$$

where $dis\left(l_i, l_j\right)$ is the geographical distance between two antennas i and j, and E is a constant that controls the smoothness. δ is a regulation constant that enforcing a prior that transition are more reserved (more likely to stay in the same location rather than move around).

Note that different from classical HMM, we are making an interpolation for the blank time span, therefore the boundaries, namely, l_0 and $l_{T_{sp}}$, enforce a very strong prior on the inference of the optimal location sequence. We define $p\left(l_0\right)$ and $p\left(l_{T_{sp}} | l_{T_{sp}-1}\right)$ in Equation 3 as equal to 1 only when it matches our observation from the call records and 0 otherwise.

We have defined all the probability distributions in the HMM in the two subsections above, an exact inference of the latent location sequence is obtained by Viterbi algorithm, a dynamic programming method.

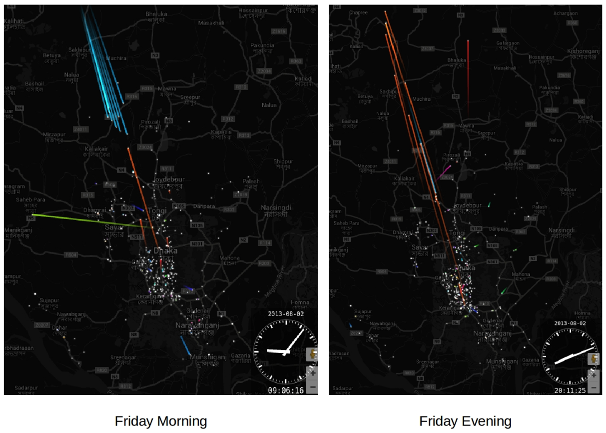

Figure 6: A visualization of human mobility completion result for CDR data in Bangladesh on Aug. 2, 2013, a usual Friday. On the left depicts people coming to Dhaka center city area in the morning while on the right people are leaving Dhaka downtown. Trajectory directions are indicated as hue.

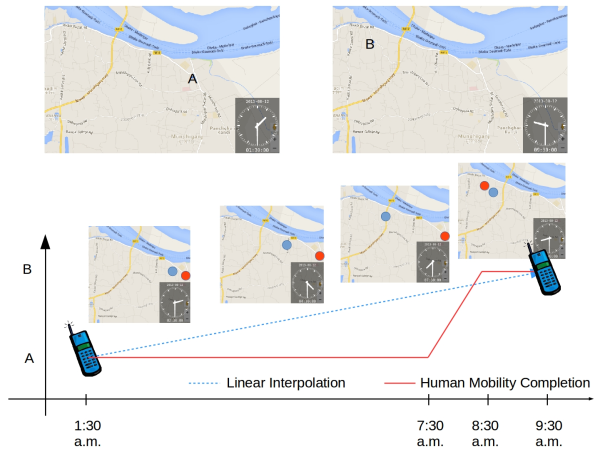

Figure 7: An example of completing a blank time span using the cues from topic model.

Figure 8: Two examples for the dynamic occupation map in Dhaka center city. On the left panel is in the afternoon while on the right panel is at the midnight. Different colors indicate different occupations of the users.

6 Experiments

We apply our algorithm to make an hourly human mobility completion on a real-world Call Detail Records dataset from Bangladesh. In this work, we do the testing and visualization on a subset of 10,000 unique ID from our dataset. As shown in Fig. 6, we visualize the completion result on a usual Friday. The hue of the trajectories indicate the movement direction of users. A blue color represents north-to-south direction, and a red color represents the reverse direction. From the left panel, we can see people are gathering to the center of Dhaka city in the morning (around 9 a.m.), while on the right panel, people are leaving the center of city in the night (around 8 p.m.).

In Fig. 7, we give an example of completing a blank time span between a call record made at around 1:30 a.m. and 9:30 a.m.. With no supportive information from other sources nor collaboration with records on other days and the other users in the city, the optimal completion would be a linear interpolation, as shown in the blue marker on the map and blue dashed line in the graph. However, it is unrealistic because it is extremely rare that one comes out from place A (where is probably referred to his home place) at 1:30 a.m. and walk slowly to the place B (which is probably referred to his work place) until 9:30 a.m.. The red marker and red line in the graph show our completion result makes a more realistic interpolation that the user stay at place A until 7:30 a.m. and move to place B at 8:30 a.m. and stay at place B thereafter, which is a fragment of a typical salaryman's daily routine.

6.1 Dynamic Census

An interesting application of human mobility completion is Dynamic Census [Arai, 2015]. Note that the trace of each cell phone user is more than a marker on a map, it carries personal attributes (e.g. gender, occupation, age etc.) to where he/she moves to and therefore makes regional statistics on these attributes varying with time. Traditional census, which is mainly conducted in a questionnaire form, can not be used for many time-sensitive applications. For example, during disaster time, people's temporary shelter allocation. In this work, we collect a diary survey data for mobile users in Bangladesh with personal attributes and use a random forest based approach in [Arai *et al.*, 2015] to infer the users' attributes in our CDR dataset based on their calling behavior.

In Fig. 8, we give an example of a dynamic occupation map in Dhaka center city at two different times (one in daytime, another at midnight). Red points are the users we infer as housewife, green ones are working males, and the blue ones are other types that do not include in these two major categories. As we can see, in the center of the city, there are obviously more working males in the daytime than those during midnight. With the help of this Dynamic Census map, the government will have a better understanding of the population distribution with respect to different attributes, and make more effective decisions by having a comprehensive considering of the population distribution of each type of people.

Methods	Avg. Error (km)	Error $<$ 100m
Ours	1.122	78.96%
Linear	3.909	60.05%
Nearest	3.595	64.37%

Table 1: Comparison of different methods of human mobility completion. Our method outperforms other interpolation algorithms.

6.2 Evaluation

In the evaluation part, we randomly select about 10% as ground truth and apply our algorithm on the rest 90% data to make a completion. We use two baseline methods, namely the linear interpolation, which assumes a linear movement between the adjacent two records, and nearest record interpolation, which uses the location of the temporally-closest record. The geographical distance between our completion and ground truth is calculated and as one metric to measure the accuracy of our completion. Concerning that there are some outliers that for unknown reasons a user travels hundreds km away, which is unpredictable from the CDR data alone but contributes a large part of average error, we use another metric that compares the number of "correct completions", the error of which is below 100m. As listed in Table 1, we can see that our algorithm outperforms the baseline methods and reaches an accuracy level that is good enough for regional statistics, considering the accuracy level of positioning with antenna (accuracy level of about $10^2 - 10^3$m).

7 Related Work

Urban computing: Recent years, urban computing has become a hot topic attracting the researchers from a wide range of fields such as urban planning, data science and ubiquitous computing. The huge demand and economical profit of urban computing have nourished a number of interesting researches, including but not limited to city noise level estimation [Zheng *et al.*, 2014], bike sharing station placement optimization [Chen *et al.*, 2015] and human mobility prediction during disaster time [Song *et al.*, 2014] and big event [Fan *et al.*, 2015].

Human mobility big data: The popularization of cell phone and other devices that leaves electrical traces (e.g. IC card) has made it possible to collect a citywide or even nationwide human mobility data in real-time and thus consider "human as sensors" [Wang *et al.*, 2014] to have a dynamic perspective of a city. The most widely used human mobility data are GPS data [Fan *et al.*, 2015], CDR data [Blondel *et al.*, 2012], location-based social network data [Kurashima *et al.*, 2013] and IC card data [Liu *et al.*, 2014]. Note that most of these datasets suffer the problem of sparseness and bias, and some pioneered works give some inspirations on this work but still cannot solve this problem perfectly. [Zheng *et al.*, 2012] makes an interpolation of the low-frequency sampling trajectory via a routing selection model learned from the data. However, the time interval of this work is at most a few minutes so that only routing is taken into consideration, while in our work the blank time span can be hours or even days, which makes modeling users' routine the key factor to be taken into consideration.

Collaborative filtering: Collaborative filter is a family of algorithms that are massively used in recommender system. In a recommender system, collaborative filtering involves automatically estimating the similarity among users and items and predicting users' preference on a new item based on the similar users' attitude. Some widely known collaborative filtering algorithms are svd++ [Koren, 2008], SoRec [Ma *et al.*, 2008] and LDA [Wang and Blei, 2011]. [Zheng *et al.*, 2010], [de Spindler *et al.*, 2006] and [Lian *et al.*, 2015] have introduced collaborative filtering into the human mobility big data and make location recommendation to a cell phone user that may interest him/her. However, although we use collaborative filtering frame in this work, the objective of location recommendation and completing a citywide human mobility from sparse records are different. Novelty is an essential factor to the location recommendation (it does not need to recommend home place for the user), while in our work, the patterns underlain in the users' daily routines are the key factors to make a proper completion.

8 Conclusion

In this work, we propose a collaborative filtering approach to complete sparse human mobility data, and successively apply our algorithm on a real-world dataset from Bangladesh.

We note several limitations of our work: first, for simplicity, we do not use map information in this work. The road network, point-of-interest distribution as well as land use can enforce a strong prior on the travel-time it may cost for traveling from one antenna to another. In addition, the owner bias can hardly be reduced in this work. For example, children are less likely to have a cell phone compared with grownups. An extension on incorporating with census data would be a promising solution to this issue. We will extend our model to counter these limitations in our future works.

9 Acknowledgment

This work is partially supported by JST, Strategic International Collaborative Research Program (SICORP); GRENE-ei; Grant-in-Aid for Young Scientists (26730113) of Japans Ministry of Education, Culture, Sports, Science, and Technology (MEXT); Microsoft Research collaborative research (CORE) program. We thank the Mobile Network Operator and volunteers who provided the data for this research.

References

[Arai *et al.*, 2015] Ayumi Arai, Apichon Witayangkurn, Teerayut Horanont, Xiaowei Shao, and Ryosuke Shibasaki. Understanding the unobservable population in call detail records through analysis of mobile phone user calling behavior: A case study of greater dhaka in bangladesh. In *2015 IEEE International Conference on Pervasive Computing and Communications, PerCom 2015, St. Louis, MO, USA, 23-27 March, 2015*, pages 207–214. IEEE Computer Society, 2015.

[Arai, 2015] Ayumi Arai. *Dynamic Census: Estimation of Demographic Structure and Spatiotemporal Distribution of Dynamic Living Population by Analyzing Mobile Phone Call Detail Records*. PhD thesis, University of Tokyo, 2015.

[Blondel *et al.*, 2012] Vincent D. Blondel, Markus Esch, Connie Chan, Fabrice Clérot, Pierre Deville, Etienne Huens, Frédéric Morlot, Zbigniew Smoreda, and Cezary Ziemlicki. Data for development: the d4d challenge on mobile phone data. *CoRR*, abs/1210.0137, 2012.

[Chen *et al.*, 2015] Longbiao Chen, Daqing Zhang, Gang Pan, Xiaojuan Ma, Dingqi Yang, Kostadin Kushlev, Wangsheng Zhang, and Shijian Li. Bike sharing station placement leveraging heterogeneous urban open data. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, pages 571–575, New York, NY, USA, 2015. ACM.

[de Spindler *et al.*, 2006] Alexandre de Spindler, Moira C. Norrie, Michael Grossniklaus, and Beat Signer. Spatio-temporal proximity as a basis for collaborative filtering in mobile environments. In Moira C. Norrie, Schahram Dustdar, and Harald C. Gall, editors, *Proceedings of the CAISE*06 Workshop on Ubiquitous Mobile Information and Collaboration Systems UMICS '06, Luxemburg, June 5-9, 2006*, volume 242 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2006.

[Fan *et al.*, 2015] Zipei Fan, Xuan Song, Ryosuke Shibasaki, and Ryutaro Adachi. Citymomentum: An online approach for crowd behavior prediction at a citywide level. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, pages 559–569, New York, NY, USA, 2015. ACM.

[Griffiths and Steyvers, 2004] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1):5228–5235, April 2004.

[Koren, 2008] Yehuda Koren. Factorization meets the neighborhood: A multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 426–434, New York, NY, USA, 2008. ACM.

[Kurashima *et al.*, 2013] Takeshi Kurashima, Tomoharu Iwata, Takahide Hoshide, Noriko Takaya, and Ko Fujimura. Geo topic model: Joint modeling of user's activity area and interests for location recommendation. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 375–384, New York, NY, USA, 2013. ACM.

[Lian *et al.*, 2015] Defu Lian, Yong Ge, Fuzheng Zhang, Nicholas Jing Yuan, Xing Xie, Tao Zhou, and Yong Rui. Content-aware collaborative filtering for location recommendation based on human mobility data. In Charu Aggarwal, Zhi-Hua Zhou, Alexander Tuzhilin, Hui Xiong, and Xindong Wu, editors, *2015 IEEE International Conference on Data Mining, ICDM 2015, Atlantic City, NJ, USA, November 14-17, 2015*, pages 261–270. IEEE, 2015.

[Liu *et al.*, 2014] Yanchi Liu, Chuanren Liu, N.J. Yuan, Lian Duan, Yanjie Fu, Hui Xiong, Songhua Xu, and Junjie Wu. Exploiting heterogeneous human mobility patterns for intelligent bus routing. In *Data Mining (ICDM), 2014 IEEE International Conference on*, pages 360–369, Dec 2014.

[Ma *et al.*, 2008] Hao Ma, Haixuan Yang, Michael R. Lyu, and Irwin King. Sorec: Social recommendation using probabilistic matrix factorization. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 931–940, New York, NY, USA, 2008. ACM.

[Rabiner, 1990] Lawrence R. Rabiner. Readings in speech recognition. chapter A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, pages 267–296. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1990.

[Song *et al.*, 2014] Xuan Song, Quanshi Zhang, Yoshihide Sekimoto, and Ryosuke Shibasaki. Prediction of human emergency behavior and their mobility following large-scale disaster. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 5–14, New York, NY, USA, 2014. ACM.

[Wang and Blei, 2011] Chong Wang and David M. Blei. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 448–456, New York, NY, USA, 2011. ACM.

[Wang *et al.*, 2014] Dong Wang, Md Tanvir Amin, Shen Li, Tarek Abdelzaher, Lance Kaplan, Siyu Gu, Chenji Pan, Hengchang Liu, Charu C. Aggarwal, Raghu Ganti, Xinlei Wang, Prasant Mohapatra, Boleslaw Szymanski, and Hieu Le. Using humans as sensors: An estimation-theoretic perspective. In *Proceedings of the 13th International Symposium on Information Processing in Sensor Networks*, IPSN '14, pages 35–46, Piscataway, NJ, USA, 2014. IEEE Press.

[Zheng *et al.*, 2010] Yu Zheng, Xing Xie, and Wei-Ying Ma. Geolife: A collaborative social networking service among user, location and trajectory. *IEEE Data Eng. Bull.*, 33(2):32–39, 2010.

[Zheng *et al.*, 2012] Kai Zheng, Yu Zheng, Xing Xie, and Xiaofang Zhou. Reducing uncertainty of low-sampling-rate trajectories. In Anastasios Kementsietsidis and Marcos Antonio Vaz Salles, editors, *ICDE*, pages 1144–1155. IEEE Computer Society, 2012.

[Zheng *et al.*, 2014] Yu Zheng, Tong Liu, Yilun Wang, Yanmin Zhu, Yanchi Liu, and Eric Chang. Diagnosing new york city's noises with ubiquitous data. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '14, pages 715–725, New York, NY, USA, 2014. ACM.