Towards the Inference of Travel Purpose with Heterogeneous Urban Data

Chuishi Meng, *Member, IEEE*, Yu Cui, *Member, IEEE*, Qing He, *Member, IEEE*, Lu Su, *Member, IEEE*, and Jing Gao, *Member, IEEE*

Abstract—In people's daily lives, travel takes up an important part, and many trips are generated everyday, such as going to school or shopping. With the widely adoption of GPS-integrated devices, a large amount of trips can be recorded with GPS trajectories. These trajectories are represented by sequences of geo-coordinates and can help us answer simple questions such as "where did you go". However, there is another important question awaiting to be answered, that is "what did/will you do", i.e., the trip purpose inference. In practice, people's trip purposes are very important in understanding travel behaviors and estimating travel demands. Obviously, it is very challenging to infer trip purposes solely based on the trajectories, because the GPS devices are not accurate enough to pinpoint the venues visited. In this paper, we infer individual's trip purposes by combining the knowledge from heterogeneous data sources including trajectories, POIs and social media data. The proposed Dynamic Bayesian Network model (*DBN*) captures three important factors: the sequential properties of trip activities, the functionality and POI popularity of trip end areas. In addition, we propose an efficient method with local candidate pools to identify POIs from geo-tagged social media messages, and learn the POI popularities from nearby social media data. Moreover, trip data is usually imbalanced across different activities. This data imbalance problem can cause serious challenges because the ***DBN*** model could be biased by those "popular" class labels. Considering this challenge, we propose an ensemble DBN method with sampling technique (*eDBN*) which results in more accurate inference. Furthermore, real-world trip data are continuously collected on a daily basis. The batch model would result in unnecessary computation because historical data need to be revisited. We handle this problem by proposing an incremental DBN method (*iDBN*) which is both effective and efficient. Extensive experiments are conducted on real-world data sets with trajectories of 8,361 residents and the 6.9 million geo-tagged tweets in the Bay area. Experimental results demonstrate the advantages of the proposed method on correctly inferring the trip purposes.

Index Terms—Dynamic Bayesian network, trajectory, social media, point of interest, trip purpose

1 Introduction

In this big data era, we are able to collect a large amount of human trajectories because of the ubiquitous adoption of GPS-integrated devices. For instance, smart phones can track real-time trajectories and geo-tagged social media messages can also reveal users' trajectories. On the surface, these human trajectories record people's daily trips with a sequence of geo-coordinates. But in essence, they reveal people's activities or trip purposes, e.g., "shopping" or "eat out". With these abundant trajectory data in hand, we are curious to ask "what did/will you do when you arrive at one place?". This is the trip purpose inference problem.

The inference of people's trip purposes has many benefits for the whole society. First, people's trip purposes can help government officials understand travel behaviors and estimate travel demands which will lead to better city planning and investment decisions. In addition, it can provide customers with more accurate recommendations and better services, and this recommendation can be made even before users start their trips.

However, trip purpose inference is very challenging, because the GPS records are not accurate enough to pinpoint the venues visited, let alone revealing the purposes of the trip. Although there have been some attempts on this research topic [1], [2], [3], existing methods overlooked several important properties such that makes them less practical. First, existing methods do not take fully advantage of the heterogeneous data, such as trajectories, POIs and social media. Most of them only utilize one or two of these data sources. In addition, people's activities usually follow certain patterns, and there are intrinsic relationships among the sequence of activities. Take the trips shown in Fig. 1 as an example. Parents may drop off their children at school before going to work, and people would eat in a restaurant after shopping. Similar activity patterns commonly reside across different users. As a result, modeling the whole sequence of activities can help us infer the past or future activities. Moreover, although nearby POI information is useful to infer the trip purposes [2], [4], their geo-graphical distribution cannot reveal how much do people like the venues. Obviously, not all the POIs attract equal attentions, and some of them are usually more popular than the others.

- *C. Meng is with the JD Intelligent Cities Business Unit and JD Intelligent Cities Research, Beijing, China. E-mail: meng.chuishi@jd.com.*
- *Y. Cui and Q. He are with the Department of Civil, Structural and Environmental Engineering, SUNY Buffalo, Buffalo, NY 14221 USA. E-mail: {ycui4, qinghe}@buffalo.edu.*
- *L. Su and J. Gao are with the Department of Computer Science and Engineering, SUNY Buffalo, NY 14221 USA. E-mail: {lusu, jing}@buffalo.edu.*

Manuscript received 15 May 2018; revised 18 Apr. 2019; accepted 5 June 2019. Date of publication 10 June 2019; date of current version 14 Jan. 2022. (Corresponding author: Chuishi Meng.)
Recommended for acceptance by H. Xiong.
Digital Object Identifier no. 10.1109/TBDATA.2019.2921823

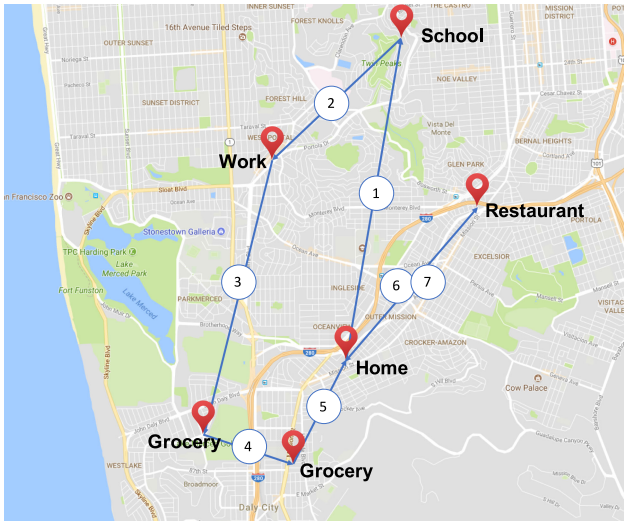


Fig. 1. A typical user's daily trips.

Because the trip purpose is a people-centric concept, the popularity of the POIs would be useful for the inference.

In this work, we propose to infer trip purposes from a large amount of heterogeneous data sets, i.e., users' trajectories, POIs and social media messages. In order to capture the sequential property of the trips, this paper proposes a dynamic Bayesian network model in which the trip purposes are hidden variables. By this means, we incorporate the knowledge of all other trips in the sequence to infer the trip purposes. Another advantage of the proposed method is that we can derive ranked results of purpose inference with corresponding confidence, e.g., 60 percent shopping, 35 percent eat out and 5 percent recreation. In fact, the results with higher ranks are very helpful, especially for trips with vague purposes, such as having lunch while shopping in a mall. In addition, this paper proposes to incorporate POI popularity in the trip purpose inference, and mine the popularity of POIs from geo-tagged social media data. This popularity information can provide a different perspective on the functionality of trip end locations, and it can help us infer the activities performed. However, the social media data is noisy and it is usually very hard to extract relevant information. In order to solve this problem, we propose an effective approach with local candidate pools to extract POI mentions from geo-tagged Twitter data. Then the popularity of a POI can be captured by the number of mentions in the social media. Moreover, trip data collected in the real-world applications usually have imbalanced distributions across different activities. For example, more trips would be annotated with "Transportation" because people spend a lot of trips on daily transit. In order to solve this problem, we propose to enhance the DBN method with sampling and ensemble techniques. As a result, the model can reduce the bias towards these "popular" labels. Furthermore, we also propose an incremental DBN method (*iDBN*) which can handle the large amount of streaming trip data effectively and efficiently.

In summary, this work makes the following contributions:

- We propose to infer individual's trip purposes by combining the knowledge from heterogeneous data sources including trajectories, POIs and social media data.
- The proposed dynamic Bayesian network model captures the sequential property of people's activities. By this means, a sequence of trips is considered as an integral part to infer the past or future trip purposes. In addition, the ranked results derived can help us handle trips with vague purposes.
- We propose an effective method to extract mentioned POIs from geo-tagged social media messages, and model POIs' popularities in trip end locations. This popularity knowledge can help improve the inference performance.
- We propose an ensemble DBN method (*eDBN*) to tackle the problem of imbalanced distributions in trip data, and also propose an incremental DBN method (*iDBN*) to handle the large amout of streaming trip data in real-world applications.
- We conduct extensive experiments on real-world data with trajectories of 8,361 residents and 6.9 million geo-tagged tweets in the Bay Area, CA. The results demonstrate the advantage of the proposed method on correctly inferring users' trip purposes.

The rest of the paper is organized as follows. We formally define the trip purpose inference problem in Section 2. The proposed methods are detailed in Section 3, 4, 5, and experiments are shown in Section 6. We review the related work in Section 7 then conclude the paper in Section 8.

2 Overview

In the following, we introduce several important concepts that will be used throughout the work, then formally define the trip purpose inference problem.

Definition 1. *A* Trajectory (Tr) *is a sequence of spatial points* l_i *with time stamps* t_i, $Tr : (l_1, t_1) \rightarrow (l_2, t_2) \rightarrow \cdots \rightarrow (l_n, t_n)$ *where each point* l *is represented by a pair of GPS coordinates, i.e., longitude and latitude.*

In this paper, we regard "trip" as the movement from one location to another, e.g., $l_1 \rightarrow l_2$, and we refer these GPS points l as "trip end locations". Take the trajectory in Fig. 1 as an example. The trajectory comprises eight GPS points which follows the sequence of $l_{Home} \rightarrow l_{School} \rightarrow l_{Work} \rightarrow l_{Grocery} \rightarrow l_{Grocery} \rightarrow l_{Home} \rightarrow l_{Restaurant} \rightarrow l_{Home}$. From the perspective of trips, it has seven trips labelled by blue circles. However, in most cases, we cannot know the activities performed in a location without users' input. In other words, we don't know whether a user is shopping in a grocery store or having meal at a restaurant, given only the geo-coordinates of trip end locations.

Definition 2. Trip Purpose *is the activity that a user performed at a trip end location.*

As the name refers, it denotes the purpose of a trip. In the following sections, we use "trip purpose" and "activity" interchangeably. As shown in Table 1, we categorize all trip purposes into eight categories, i.e., "Home", "Education", "Shopping", "EatOut", "Recreation", "Personal", "Work", and "Transportation". The example activities shown in Table 1 are defined by the California Household Travel Survey (CHTS) [5] which collected people's daily trajectories and

TABLE 1
Activity Categories

Category	Example Activities
Home	Any activities performed at home
Education	School, Class, Laboratory, Meal at college, etc.
Shopping	Groceries, Clothing, Gas, etc.
EatOut	Drive through meals, Meal at restaurant/dinner
Recreation	Indoor or outdoor exercise, Health care, etc.
Personal	Household errands, Religious activities, etc.
Work	All activities performed at the work place, etc.
Transportation	Change type of transportation, Pick up/drop off

activities. More details about this data set will be discussed in Section 6.

Definition 3. Point of Interest (POI) *is a specific location that someone may find useful. In this work, they represent venues in the physical world, e.g., banks and shopping malls. Each* POI *is associated with properties such as name, address, coordinates, category and etc.*

Definition 4. Geo-tagged Tweet *is a Twitter message associated with a pair of GPS coordinates where the message was generated.*

Problem Definition. *Given* the trajectories of users, the points of interest and the Twitter messages near trip end locations, our *Objective* is to infer the purposes of trips.

Note that some of the trips have labels with corresponding purposes. These labels can be manually recorded by users or mined from social media messages. However, acquiring these labels is very difficult, and it is usually assumed to be unavailable for a large portion of trips.

3 Trip Purpose Inference with Dynamic Bayesian Network

A trip's purpose is determined by many factors, such as other activities of the day, the category of the visited venue, the functionality and the popularity of the destination area. Before discussing the proposed methods, we first shed some light on how these factors associate with the trip purpose inference.

Sequential Activities of the Day. Common sense tells us that users' activities usually follow some patterns, and there are intrinsic relationships among the sequence of activities. For instance, parents may drop off their children at school before going to work, people would eat in a restaurant after shopping in a mall, and etc. Similar patterns among sequential activities widely exist, and it is very useful information for the purpose inference.

The category of the visited venue usually correlates with the trip purpose. For example, people arriving at a restaurant are very likely to have lunch or dinner; checking in at a mall tells us he will be shopping. There will be close relations between the category of venue people visited and their trip purposes. Unfortunately, the GPS devices are not accurate enough to pinpoint the venues visited. In addition, it is also not easy to acquire this knowledge from people because of the efforts it takes and potential privacy concerns.

The functionality of the trip end area reveals the general usage of the nearby area. When we are not aware of the specific venue that a user visited, the nearby POIs can give us a hint about what the trip purpose would be. For example, arriving at a place with many shops nearby means people will go shopping with a higher probability. Specifically, the distribution of POI categories is a good feature to denote the functionalities of a location.

The Popularity of the Trip End Area. Although the nearby POIs can help us understand the functionalities of a location, it cannot capture how people think about this area. Obviously, not all the venues attract equal attentions. In other words, some of them are more popular than the others. The popularity of the venues can be a useful feature for the purpose inference task, because the trip purpose is indeed a people-centric concept. Fortunately, social media can help us out here. Take Twitter as an example, people can send geo-tagged messages, and many of them contain the comments towards nearby POIs. By matching these geo-tagged tweets to real-world POIs, we can reveal venues' popularities accordingly.

A good trip purpose inference method needs to consider all these factors and the intrinsic relationships among them. In the following sections, we will first describe the proposed method for POI popularity modeling with social media data, then demonstrate the proposed Dynamic Bayesian Network approach.

3.1 POI Popularity Modeling

Based on the above reasoning, a POI's popularity can be captured if we can accurately identify them from tweet messages. The basic idea is that a POI is more popular if it has been mentioned by more tweets. However, this is a very challenging task. First, social media data are very short. Existing named entity extraction methods perform poorly on these messages with very limited contexts. For example, "apple" may refer to the IT company or the fruit. Second, social media data are also very noisy. People usually use informal languages and names in tweets, such that we cannot expect to match a POI's full name in tweets. For instance, a tweet "dinner 2 tacos from lacorneta" mentions a restaurant "La Corneta Taqueria" without the full name, but rather with an abbreviation.

In this section, we propose a method to learn POIs' popularities from geo-tagged tweets. It can be much easier to identify mentioned POIs from these tweets because their associated geo-coordinates give us a good hint. Specifically, we can narrow down the search space to the nearby POIs. Compared with traditional methods [6] that work with a city-wide or world-wide POI knowledge base, our method can restrict the number of POI candidates within several dozens in a local area.

In general, the proposed method works as follows. For each geo-tagged tweet, we first construct a local candidate pool with nearby POIs. Then a match index is calculated between the tweet content and each candidate POI names. Among the candidates, the POI with the largest match index above the threshold is marked as matched. Finally, the POI popularity of a trip end area can be derived by aggregating

TABLE 2
POI Categories

Category	Google Place Type
Money	accounting, atm, bank, post office, finance
Leisure	art gallery, gym, movie rental, movie theater, museum, etc.
Food	bakery, cafe, food, restaurant, meal delivery
Bar	bar, night club
Care	beauty salon, hair care, spa
Store	book store, clothing store, grocery, supermarket, etc.
Transportation	bus station, subway station, train station, taxi stand, etc.
Auto	car repair, car wash, gas station
Religion	cemetery, church, funeral home, mosque, etc.
Civic	courthouse, lawyer, police, fire station, city hall, etc.
Health	dentist, doctor, health, hospital, pharmacy, etc.
Improve	electrician, locksmith, painter, real estate agency, etc.
Education	library, school, university
Lodge	rv park, lodging, campground

all the identified POIs from geo-tagged tweets. By this means, we can identify the mentioned POIs from geo-tagged tweets effectively and efficiently. In the following, we describe each component in detail.

POI Local Candidate Pool Construction. In order to identify POI mentions for a geo-tagged tweet, we first construct a local candidate pool with all the POIs near the geo-coordinates of the tweet. The range limitation should be set considering the accuracy of GPS devices. In this work, we set the range as 200 meters which usually results in candidate pools with several dozen POIs. In addition, we consider the accuracy of the GPS signals, and a 200-meter threshold can tolerant most of the localization errors.

Calculate POI Match Index. After constructed a local candidate pool for each tweet, the next step is to find the best matched POI among all candidates. To this end, we design a match index to measure the similarity between a tweet and a POI name. This match index considers two essential factors:

- The number of matched terms. The match index should be larger if there are more terms matched between a tweet and a POI name. For example, a tweet "Was just told by a teenager working at this Jamba Juice, that I looked like a young Walter White" mentions the POI "Jamba Juice Redwood City", and two terms are matched between them. However, there is another nearby POI "Geoff White Photographers" which matches a term "white" to the tweet. In this case, "Jamba Juice" with 2 matched terms should be weighed higher than the other one with only 1 matched term.
- The rareness of the matched terms. Some terms may frequently appear in the candidate pool. For example, there is no surprise that many POI names contain "San Francisco" in the Bay area. Then these terms should have less impact on the matching index. On the contrary, terms such as "Corneta" is relatively rare. In fact, this term only appears in a restaurant named "La Corneta Taqueria". No doubt that these terms should have larger impact on the matching index. In other words, if terms like "Corneta" matched between a tweet and a POI name, we should have a high belief that the tweet mentioned the restaurant "La Corneta Taqueria".

In this work, we propose a POI Match Index which characterizes the aforementioned factors. Specifically, each Tweet T_i is represented by a set of terms, i.e., $T_i = \{u_1, u_2, \ldots, u_m\}$. Similarly, each candidate POI's name is represented by $P_j = \{v_1, v_2, \ldots, v_n\}$. The set of Matched Terms MT between T_i and P_j are

$$MT(T_i, P_j) = T_i \cap P_j. \tag{1}$$

Then we can calculate the Match Index as follows which borrowed the idea of the TF-IDF in the Information Retrieval field

$$MI(T_i, P_j) = |MT(T_i, P_j)| \times \log \frac{N_{pool}}{1 + \sum_{k=1}^{N} \mathbb{1}(MT(T_i, P_k) \in P_k)}, \tag{2}$$

where N_{pool} denotes the size of the candidate pool. $\mathbb{1}(\cdot)$ is the indicator function which returns 1 if and only if the condition holds. In addition, $\sum_{k=1}^{N} \mathbb{1}(MT(T_i, P_k) \in P_k)$ calculates the frequency of the matched terms MT in the POI candidate pool. As shown in the Equation (2), the first term considers the number of matched terms, and the second term considers the rareness of the matched terms. Note that $|MT(T_i, P_j)|$ could be zero, i.e., there are no matched terms between T_i and P_j. In this case, the Match Term Index will be 0 which is reasonable.

After calculated the Match Index between T_i and every P_j in its candidate pool, we can return the one with the highest index as the identified POI. However, we still need to set a threshold to the MI, and return non-identified if none of MIs exceed the threshold. In sum, With the constructed local POI candidate pool and the proposed Match Term Index, we can accurately identify the nearby POIs that mentioned in the Tweets.

Note that, in real-world applications, we are always facing the GPS inaccuracy issues. Fortunately, researchers have developed effective and efficient algorithms for the map matching problem, e.g., Lou et al. [7]. In addition, the proposed framework can also tolerant the GPS inaccuracy issue by constructing the POI local candidate pool with a range limitation.

POI Popularity Modeling. After extracted the mentioned POIs from social media data, we can further represent the POI popularity across different categories by counting the corresponding mentions from social media. For example, if restaurants are mentioned by 10 different tweets, we will count 10 towards the popularity of the POI category "Food". Then the counts can be normalized into a distribution across all the POI categories in Table 2.

3.2 Dynamic Bayesian Network Construction

In this work, we propose a Dynamic Bayesian Network (*DBN*) to model people's sequential activities. As shown in the Fig. 2, $a \in A$ denotes the activity performed (or trip

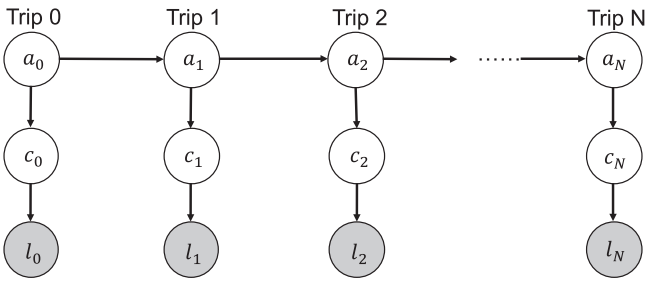


Fig. 2. The dynamic Bayesian network.

purpose), $c \in C$ denotes the category of POI a user visited, and l is the trip end Location. All activities (A) and POI categories (C) defined in this paper are shown in Tables 1 and 2.

The ***DBN*** model can be interpreted in a generative process. For each trip i, a user first decides an activity a_i (or purpose) based on his previous one a_{i-1}. Then he chooses a venue category c_i based on this choice of activity. At last, he chooses a geo-location l_i to finally perform the activity a_i in venue c_i. This process continues until the last trip.

The likelihood function of the proposed ***DBN*** model is as follows:

$$P(a,c,l) = P(a_0)P(c_0|a_0)P(l_0|c_0) \cdot \left(\prod_{i=1}^{N} P(a_i|a_{i-1})P(c_i|a_i)P(l_i|c_i)\right), \tag{3}$$

where $P(a_i|a_{i-1})$ is the probability of the activity a_i given previous activity a_{i-1}, $P(c_i|a_i)$ is the probability of the visited POI category given current activity a_i, and $P(l_i|c_i)$ is the probability of chosen location l_i given currently chosen POI category c_i. In the proposed model, the visited location l_i is always observed from the trajectory data, such that we can use Bayes's rule to approximate $P(l_i|c_i)$ as follows:

$$P(l_i|c_i) \propto \frac{P(c_i|l_i)}{P(c_i)} \propto \frac{P_{POI}(c_i|l_i)P_{tweet}(c_i|l_i)}{\sum_j P_{POI}(c_j|l_i)P_{tweet}(c_i|l_i)} \times \frac{1}{P(c_i)}. \tag{4}$$

In the Equation (4), $P(c_i|l_i)$ denotes the POI category distribution given a geo-location l_i, and it is defined as the distribution of nearby POI categories near the location l. Specifically, it is determined by two aforementioned factors: the functionality distribution $P_{POI}(c_i|l_i)$, and the popularity distribution $P_{tweet}(c_i|l_i)$. The first distribution is obtained from populating the nearby POIs, and the second distribution is obtained by extracting POI mentions from nearby tweets.

3.3 DBN Parameter Learning

There are two sets of parameters in the ***DBN*** model: the transition probabilities $P(a_i|a_{i-1})$ and the emission probabilities $P(c_i|a_i)$. Note that in our problem, the activities a_i and visited venues c_i are not fully observed. In other words, many activities and corresponding venues are not labelled in the data. In order to learn the parameters from such incomplete data, we adopt the EM algorithm [8]. The process is summarized in Algorithm 1. It starts with an initial set of parameters. In each Expectation step (E-step), we compute the expected sufficient statistics for the parameter variables. Then in each Maximization step (M-step), we treat the expected sufficient statistics as observed, and perform Maximum Likelihood Estimation to estimate a new set of parameters. The algorithm continues between these two steps until converges.

Algorithm 1. DBN Parameter Learning

Input: DBN structure $\mathcal{G}$,
Initial set of parameters $\theta_0 = \{P(a_i|a_{i-1}), P(c_i|a_i)\}$
Partially observed trip data set $\mathcal{D}$,
Output: DBN parameters θ

for $t \leftarrow 1, 2, \ldots,$ until convergence **do**
  for each $a \in A$ and each $c \in C$ **do** ▷ **Initialize**
    $M_t[a_i, a_{i-1}] = 0$
    $M_t[c_i, a_i] = 0$
  end for
  for each $d \in D$ **do** ▷ **E-Step**
    Inference on graph $\langle \mathcal{G}, \theta_{t-1} \rangle$ using evidence d
    for each $a \in A$ and each $c \in C$ **do**
      $M_t[a_i, a_{i-1}] \leftarrow M_t[a_i, a_{i-1}] + P(a_i, a_{i-1}|d)$
      $M_t[c_i, a_i] \leftarrow M_t[c_i, a_i] + P(c_i, a_i|d)$
    end for
  end for
  for each $a \in A$ and each $c \in C$ **do** ▷ **M-Step**
    $P_t(a_i|a_{i-1}) \leftarrow \frac{M_t[a_i, a_{i-1}]}{M_t[a_{i-1}]}$
    $P_t(c_i|a_i) \leftarrow \frac{M_t[c_i, a_i]}{M_t[a_i]}$
    $\theta_t \leftarrow \{P_t(a_i|a_{i-1}), P_t(c_i|a_i)\}$
  end for
end for
return θ_t.

3.4 DBN Prediction

With the learned parameters of the ***DBN*** mode, we can infer possible activities and their corresponding probabilities for any given trip. Specifically, we can calculate the posterior probability of the jth activities given a user's trajectory $tr \in Tr$, where Tr denotes all users' trajectories. This estimates the probability of activity a_j out of all possible activities A, as shown in Equation (5)

$$P(a_j|tr) = \frac{P(a_j, tr)}{P(tr)}, \forall a_j \in A, \tag{5}$$

The returned results are possible activities ranked by their probabilities. Note that, generating a ranked result is a great advantage by adopting the Bayesian method, especially compared with traditional methods which can only provide a best guess. In fact, the top ranked inference results are very useful in real-world applications. Many classification tasks may have very vague decision boundaries, and usually the best guess results in poor performance. However, a ranked list with corresponding confidence can help us identify several meaningful results and improve the inference accuracy. The experiment results shown in Section 6.3 provide a good demonstration.

4 *eDBN*: Ensemble DBN on Imbalanced Data

In Section 3.3, we present how to learn the parameters of the ***DBN*** model with all the trip data collected, and there is an

TABLE 3
Activity Distribution

Activity	Number of Occurrences	Probability
Education	1285	5.6%
Shopping	4478	19.6%
EatOut	2425	10.6%
Recreation	2695	11.8%
Personal	3013	13.2%
Transportation	8945	39.2%

underlying assumption that the class labels are balanced, i.e. the number of trip purposes are evenly distributed across different classes. However, this "balanced data" assumption usually does not hold in real-world applications. Table 3 populates the different purposes in our collected trip data, and it clearly shows the imbalanced distribution among different activities. The most popular activity recorded is "Transportation" which appears almost 7 times more often than the "Education". This is not a surprise because people spend a lot of trips on daily transits. This data imbalance problem can cause severe problems because the model can be biased towards those "popular" classes.

In light of these challenges, we propose to improve the ***DBN*** method by adopting both sampling and ensemble techniques.

Sampling. We can split the trip data according to their trip purposes. As we discussed, these trips of different activities are usually imbalanced. To alleviate this issue, we form a new training dataset by randomly under-sample the data. The sample rate $R(a)$ is determined with respect to the probability of the corresponding activity $P(a)$ (e.g., the third column in Table 3)

$$R(a_i) = \frac{\min_{a_j \in A}(P(a_j))}{P(a_i)}, a_i \in A. \quad (6)$$

This ensures the trips of rare activities are sampled with higher probability, and vice versa.

Algorithm 2. Ensemble DBN Parameter Learning

Input: DBN structure $\mathcal{G}$,
Initial set of parameters $\theta_0 = \{P(a_i|a_{i-1}), P(c_i|a_i)\}$
Partially observed trip data set $\mathcal{D}$,
number of ensembles K,
Output: DBN parameters θ
for $k \leftarrow 1, 2, \ldots, K$ **do**
  Populate the class distribution $\Phi(\mathcal{D})$
  Under-sample the training data $\mathcal{D}_s$ with Equation (6)
  θ_k = DBN_learning($\mathcal{D}_s$, θ_0, $\mathcal{G}$) ▷ **Algorithm 1**
end for
return $\theta = \{\theta_1, \ldots, \theta_K\}$.

Ensemble. Although the sampling method may alleviate the bias towards the "popular" activities, the ***DBN*** method may miss important information because the sampling technique under-samples the records from the majority activities and high variance is thus introduced. Therefore, we propose to adopt ensemble method to reduce the variance by training several ***DBN*** models on different sampled data. As demonstrated in [9], the variance could be reduced by training multiple models as long as the samples are uncorrelated. Suppose we have learned a series of ***DBN*** models on different data samples. Given any trip, we can infer its purpose using all the models, and then aggregate the results by majority voting.

Algorithm 2 summarizes the parameter learning process for the ***eDBN*** method.

5 *iDBN*: Incremental DBN Learning

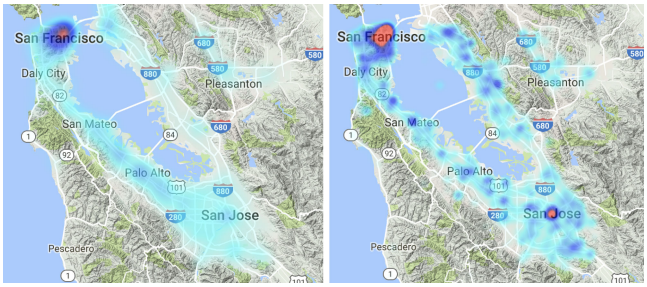
In real-world applications, trip data are generated on a daily basis. Such data are continuously collected in a "streaming" manner. Taking more data as input would help the ***DBN*** model to refine the model and make better inference. However, unnecessary computation may be introduced if historical data needs to be re-visited each time. Eventually, the ***DBN*** method would suffer from the ever-increasing training time. To reduce the computational cost, we propose an incremental learning method for ***DBN*** which guarantees constant running time at each time epoch by processing the trip data only once.

By closely examining the DBN learning method (Algorithm 1), we realize that the most computation-intensive operation lies in the EM algorithm, and more specifically, in the expectation step. At every iteration of the E-step, we need to first perform inference on the DBN to estimate purposes for the trips without labels. This inference takes up lots of computational cost since it evaluates all the possibilities.

Algorithm 3. Incremental DBN Parameter Learning

Input: DBN structure $\mathcal{G}$,
Parameters θ_{t-1} on time epoch $t-1$,
Trip data $\mathcal{D}_t$ on time epoch t,
Output: DBN parameters θ_t
while not converge **do**
  for each $a \in A$ and each $c \in C$ **do** ▷ **Initialize**
    $M_t[a_i, a_{i-1}] = 0$
    $M_t[c_i, a_i] = 0$
  end for
  for each $d \in \mathcal{D}_t$ **do** ▷ **Incremental E-Step**
    Inference on graph $\langle \mathcal{G}, \theta_{t-1} \rangle$ using evidence d
    for each $a \in A$ and each $c \in C$ **do**
      $M_t[a_i, a_{i-1}] \leftarrow M_t[a_i, a_{i-1}] + P(a_i, a_{i-1}|d)$
      $M_t[c_i, a_i] \leftarrow M_t[c_i, a_i] + P(c_i, a_i|d)$
    end for
  end for
  $M_t[a_i, a_{i-1}] = \gamma M_t[a_i, a_{i-1}] + M_{t-1}[a_i, a_{i-1}]$
  $M_t[c_i, a_i] = \gamma M_t[c_i, a_i] + M_{t-1}[c_i, a_i]$
  for each $a \in A$ and each $c \in C$ **do** ▷ **M-Step**
    $P_t(a_i|a_{i-1}) \leftarrow \frac{M_t[a_i, a_{i-1}]}{M_t[a_{i-1}]}$
    $P_t(c_i|a_i) \leftarrow \frac{M_t[c_i, a_i]}{M_t[a_i]}$
    $\theta_t \leftarrow \{P_t(a_i|a_{i-1}), P_t(c_i|a_i)\}$
  end for
end while
return θ_t.

In the following, we assume there are T time epochs in total, and the trip data $\mathcal{D}_t$ are collected at each time epoch $t \in 1, 2, \ldots, T$. Let's first examine how ***DBN*** performs on this streaming data. At epoch 1, ***DBN*** takes in the data $\mathcal{D}_1$; At epoch 2, it takes in data $\{\mathcal{D}_1, \mathcal{D}_2\}$; similarly, at epoch t, it takes in the data collected from current and all previous



(a) Trip ends (b) Geo-tagged Tweets

Fig. 3. Trip ends and Geo-tagged tweets.

time epochs, i.e., $\{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_t\}$. As a result, historical data is re-evaluated too many times, and this results in huge computational cost.

In order to tackle this challenge, we propose ***iDBN***, an incremental DBN learning method. The algorithm is detailed in Algorithm 3. At each time epoch t, ***iDBN*** takes in the parameters θ_{t-1} (transition probabilities and emission probabilities) learned from the previous epoch $t-1$, and updates the parameters with the trip data recently collected $\mathcal{D}_t$. This schema, known as the incremental EM algorithm, was proposed in [10]. It has many advantages comparing to the traditional EM algorithm including fast convergence rate and comparable estimation accuracy. In addition to this "incremental E-step", the ***iDBN*** method shares the same M-step as the ***DBN*** method.

6 Experiments

The proposed method is evaluated with real-world data sets including human trajectories, point of interests, and tweets in the Bay area, CA. In the following sections, we discuss the data sets, the baselines, and the evaluation results.

6.1 Data Sets

Trajectories. The California Household Travel Survey (CHTS) [5] collected travel information from residents across California's 58 counties. The survey was designed to obtain detailed information about the household socioeconomic characteristics and their travel behaviors. Among its various achievements, the survey collected 8,631 participants' GPS trajectories for a week. In addition, each trajectory is accompanied with a detailed trip diary which records visited POIs and trip purposes. The trip purposes are labelled by users with pre-defined categories, and some of them are shown on the second column in Table 1. However, the diaries are far from exhaustive and many trips recorded by GPS devices are not logged in the diary. This is a common issue for traditional surveys because it requires too much efforts from the participants. The heat map of trip end locations are shown in Fig. 3a.

Twitter Data. We collected 6.9 million geo-tagged tweets in the Bay area from Jan 31 2013 to Feb 16 2017. They are queried through Twitter APIs and filtered by geo-coordinates. The heat map of geo-tagged tweets are shown in Fig. 3b.

Point of Interest. The POIs are queried through Google Places API [11]. In this work, we use its Nearby Search request to get the nearby POIs of any given geo-coordinates, and the returned POI names and types are utilized by the proposed method. Some Google Place types are shown in Table 2.

6.2 Baselines

In the following experiment, we compare the proposed method with the following baseline methods,

- Random Forest (***RF***) is an ensemble learning method for classification. It is constructed by a large amount of decision trees with sub-samples. The prediction result is decided by taking the majority vote across all trees.
- Support Vector Machine (***SVM***) tries to find the optimal decision boundary with the largest margins to classify data from different classes.
- Artificial Neural Network (***ANN***) can be trained to perform classification tasks. In the experiment, we adopt Multi-layer Perceptron, one typical kind of ANN, to predict the trip purpose.
- K-nearest Neighbor (***KNN***) finds a predefined number of training samples closest in distance to the new point, and predict the label from these samples.

In the following, we perform extensive experiments on the proposed ***DBN*** method, and demonstrate its advantages compared with baselines. Specifically, the performance of trip purpose inference is discussed in Section 6.3, the proposed POI mention identification method with social media is evaluated in Section 6.4, and we discuss how POI popularity features can affect the performance of trip purpose inference in Section 6.5.

6.3 Trip Purpose Inference

To compare the performance of all the methods on the trip purpose inference, we conduct experiments on the collected real-world data set. The features used for training include travel mode, previous activity category, activity time and duration, nearby POI category distribution, and nearby POI popularity distribution. We randomly select 80 percent trips as training data, and leave the rest for testing. All the methods are evaluated by 10 times, and the average results are reported. In addition, there are about one third trip end locations in the dataset are either home or work, because the survey collected trips from people's daily lives. Using these trips to train the model would greatly bias the performance. As a result, we only test the inference results on other activities, and assume the home and work trips are known. In fact, it is also easy to infer users' home and work locations from their trajectories, because normal people stay in these locations for the most time. In other words, we are interested in non-trivial tasks of inferring non-home and non-work trip purposes.

Note that the proposed ***DBN*** model can output a ranked list of activities with corresponding confidence. This is a great advantage by adopting the Bayesian method. As we have discussed in Section 3.4 the top ranked inference results are very useful in real-world applications, especially when the classification boundaries among classes are vague. As a result, we also evaluate the ***DBN*** performance with top-2 and top-3 inference results in the experiment, denoted as DBN-top-2, and DBN-top-3 respectively. In these cases, we regard the inference as correct if the ground truth activity is among the top-2 or top-3 results.

The inference accuracy and F1 scores are shown in Table 4, and the average results are also compared in Fig. 4. We can observe that the proposed ***DBN*** model, including DBN-top-1, DBN-top-2 and DBN-top-3, outperform other

TABLE 4
Performance of Trip Purpose Inference

Accuracy	SVM	ANN	KNN	RF-top-1	DBN-top-1	RF-top-2	DBN-top-2	RF-top-3	DBN-top-3
Education	27.0%	41.9%	34.4%	40.2%	**52.3%**	71.0%	**72.9%**	78.5%	**78.4%**
Shopping	59.6%	65.3%	52.5%	78.6%	**80.1%**	82.3%	**94.2%**	87.9%	**98.4%**
EatOut	4.8%	30.0%	32.1%	60.6%	**79.0%**	66.9%	**83.2%**	76.2%	**85.5%**
Recreation	18.6%	39.0%	30.6%	55.5%	**62.7%**	71.9%	**77.4%**	83.3%	**84.7%**
Personal	7.7%	21.7%	22.5%	**50.4%**	42.2%	**68.5%**	65.0%	85.5%	**90.0%**
Transportation	**84.6%**	74.2%	59.1%	75.4%	68.3%	**93.9%**	84.3%	**97.9%**	89.6%
Average	33.7%	45.4%	38.5%	60.1%	**64.1%**	75.8%	**79.5%**	84.9%	**87.8%**
F1 Score	SVM	ANN	KNN	RF-top-1	DBN-top-1	RF-top-2	DBN-top-2	RF-top-3	DBN-top-3
Education	0.362	0.461	0.358	0.419	**0.484**	**0.777**	0.757	**0.854**	0.840
Shopping	0.517	0.600	0.452	0.756	**0.754**	0.842	**0.853**	0.899	**0.909**
EatOut	0.087	0.349	0.320	0.635	**0.712**	0.742	**0.835**	0.825	**0.879**
Recreation	0.281	0.430	0.342	0.585	**0.592**	**0.746**	0.721	**0.856**	0.826
Personal	0.134	0.299	0.266	**0.550**	0.476	0.729	**0.737**	0.871	**0.920**
Transportation	0.625	0.655	0.588	0.722	**0.735**	0.859	**0.869**	0.910	**0.919**
Average	0.334	0.465	0.387	0.611	**0.626**	0.783	**0.795**	0.871	**0.882**

baselines on almost every activity category with higher accuracy and F1 score. This is because ***DBN*** model captures the intrinsic relationships among sequential activities, trip end locations' POI distributions and the popularities identified from the Twitter data. On average, the ***DBN*** model can reach 64 percent accuracy with the top-1 inference result. This demonstrates the top-ranked results generated by the ***DBN*** model are very useful in the trip purpose inference. Comparing with the random forest method, the proposed ***DBN*** also demonstrates better performance with respect to the average accuracy and F1.

In the Section 3.1, we construct a POI local candidate pool in order to identify the matched POIs with the social media data. A 200 meter range was set as the threshold when constructing the candidate pool by considering both the GPS signal accuracy and the POI density. In the Fig. 5, we examine the robustness of this threshold, and it turns out that the performance of the proposed method is not greatly affected by the threshold, we believe this is because the GPS signal is quite accurate in the collected trip data set, such that narrowing down the threshold does not introduce much errors. However, the performance does diminish to some extent with larger thresholds, for example, increased from 200 meters to 500 meters, because the POI candidate pool is larger and this makes it harder to correctly match the tweets with the nearby POIs. As for other real-world data sets with less GPS signal accuracy, we expect much worse performance with respect to larger or smaller thresholds.

6.4 POI Mention Extraction from Geo-Tagged Tweets

In Section 3.1, we propose to extract mentioned POIs from geo-tagged tweets. For each geo-tagged tweet, we can accurately identify the mentioned POI with a local candidate pool and the match index. In practice, it is very hard to evaluate the performance of POI mention extraction from tweets, because we have so many nearby tweets and trips in the data set. Fig. 6 shows the histogram of tweets near trip ends. On average, there are 2,607 geo-tagged tweets near each trip end location (within 200 meters). Actually, it is impossible to be evaluated without a standard data set labelled by human workers. To this end, we recruited volunteers to label the POI mentioned tweets near 50 random trip end locations. Given a list of tweets, the volunteers are asked to judge whether those tweets mentioned any nearby POI, and whether the identified POIs are correct. In Table 5, we present the results on several example trip end locations and the average performance. For each location, we populate the total number of nearby tweets, the number of POI-mentioned tweets which are identified by human workers. Some of the tweets are formatted by third party Apps, such as Foursquare and Instagram. We can easily parse POIs from these well formatted tweets, for example, "Im at Applewood Pizza in San Carlos Ca" and "Bagel time! @ Bagel Street Cafe Town Center Alameda". However, there are still many POI-mentioned tweets without these formats. For instance, "Catch us at amc Mercado 20 12 am insurgent!!!!". The better performance shown in Table 5 indicates

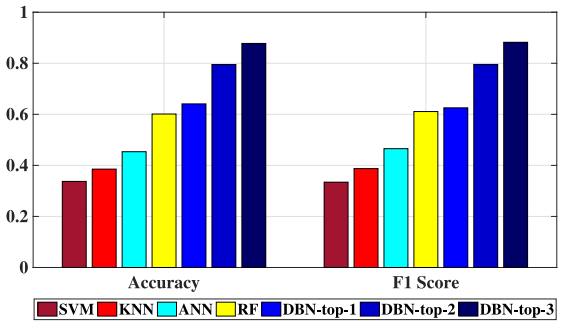


Fig. 4. Average performance of trip purpose inference.

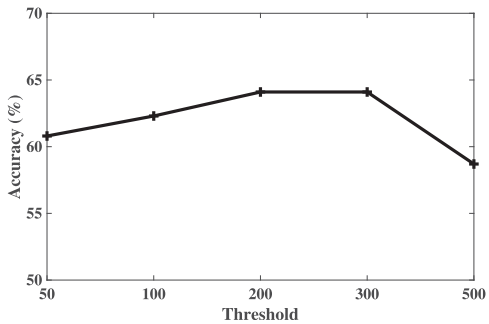


Fig. 5. Robustness of the range threshold.

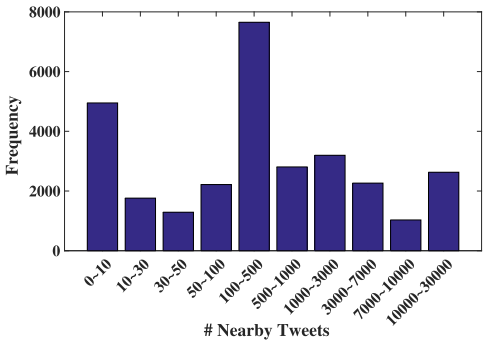


Fig. 6. Histogram of tweets near trip ends.

the proposed method can extract mentioned POIs from tweets, no matter they are well formatted or not.

6.5 Impact of the POI Popularity Features

For each trip end, we can model its nearby POIs' popularity based on the social media data discussed in Section 3.1. In this section, we perform experiments to evaluate how this popularity distribution can help us infer people's trip purposes. We first evaluate its impact based on the Random Forest model. In this experiment, we train two Random Forest models, one with the POI popularity features (RF-Tweet) and the other without them (Rf-noTweet). As shown in Table 6. the POI popularity features mined from the Twitter data can improve the inference accuracy on every class.

In addition, we further evaluate the POI popularity features on the ***DBN*** model. Two ***DBN*** models are trained and compared. One without the popularity features (DBN-noTweet), and the other with the popularity features (DBN-Tweet). The results are shown in Table 6. Interestingly, this popularity feature has different impacts on different activity categories. Specifically, incorporating the popularity features leads to higher accuracy on the inference of "Education", "EatOut", and "Personal". But it results in slightly lower accuracy on the inference of "Shopping", "Recreation", and "Transportation". The reason lies in the different properties between the cyber and physical worlds. Take the activity "Education" as an example. One educational institute, e.g., high school, would be surrounded by many other POIs, e.g., restaurants. In this case, the weight of high school will be under-estimated in the POI distribution. Fortunately, we can identify more people's tweets in the school than in the restaurant, i.e., the school is more popular than other POIs in the social media. By this means, the POI's popularity feature can help us recognize the importance of the school. On the other hand, the POI's popularity may fail to capture the importance of certain POIs if there are not enough people discussing them in the social media. This may result in slightly worse inference performance, such as in "Recreation" and "Transportation", because people tweet about these activities less frequently.

TABLE 5
Performance of POI Extraction from Geo-Tagged Tweets

Example Locations	Total Nearby Tweets	POI-mentioned Tweets	App-formatted Tweets	POI-identified Tweets
37.4028125, -121.881519784	575	184	116 (63.0%)	159 (86.4%)
37.3896875, -122.03080036	294	103	80 (77.7%)	93 (90.3%)
37.3278125, -121.8842176	1188	368	247 (67.1%)	305 (82.9%)
37.6715625, -122.472796763	1232	596	446 (74.8%)	495 (83.1%)
Average Recall	-	-	68.8%	83.5%

TABLE 6
Accuracy Comparison with POI Popularity Features

Accuracy	RF-noTweet	RF-Tweet	DBN-noTweet	DBN-Tweet
Education	40.3%	**40.5%**	50.0%	**52.3%**
Shopping	77.0%	**78.6%**	**82.9%**	80.1%
EatOut	60.0%	**60.6%**	78.9%	**79.0%**
Recreation	53.2%	**55.5%**	**65.7%**	62.7%
Personal	47.7%	**50.4%**	38.9%	**42.2%**
Transportation	75.1%	**75.4%**	**69.5%**	68.3%

6.6 Evaluation on the *eDBN* Method

In this section, we conduct experiments to illustrate the effectiveness of the ensemble DBN method. As discussed in Section 4, the trip purposes are not balanced in the training data (Table 3) such that the model could be biased by "popular" activities, e.g., Transportation. In this experiment, we first sample the training data with Equation (6), then learn an ensemble of multiple models according to Algorithm 2.

We compare the ***eDBN*** with two baselines—the DBN method and the DBN method with sampling. The results are summarized in the Table 7. As expected, the ***eDBN*** method improves the inference performance on "less-popular" activities, such as Education, EatOut, Recreation and Personal. Overall, the ***eDBN*** method also shows better performance than the baselines.

6.7 Evaluation on the *iDBN* Method

We also evaluate the performance of the proposed incremental DBN method. The trip data is first split into 10 parts, and each part contains trips collected at one time epoch. We then feed the trip data to ***DBN*** and ***iDBN*** respectively. Table 8 shows the inference accuracy of both methods. Obviously, the incremental DBN has the same accuracy as the non-incremental method ***DBN***. However, the running time of ***iDBN*** is consistently low comparing to that of ***DBN*** which increases dramatically over time (Fig. 7). This experiment clearly demonstrates the effectiveness and efficiency of the ***iDBN*** method.

7 Related Work

There are several research fields related to this work, and we summarize them in this section.

Trip Purpose Inference. The ubiquitous adoption of GPS-integrated devices has enabled extensive studies on the trip purpose inference [12], [13], [14]. In [15], they proposed to annotate geo-tagged tweets with POI categories. Detailed feature extraction methods are introduced and traditional classification methods are utilized for the annotation task. [3] proposed a method to predict the POIs that users visited,

TABLE 7
Performance of the Ensemble DBN

Accuracy	DBN	DBN sample	DBN Ensemble	F1 Score	DBN	DBN sample	DBN Ensemble
Education	51.3%	72.3%	**73.9%**	Education	0.472	0.486	**0.489**
Shopping	**79.8%**	78.8%	79.2%	Shopping	0.752	0.765	**0.771**
EatOut	77.5%	**80.3%**	**80.3%**	EatOut	**0.703**	0.683	0.681
Recreation	62.3%	62.4%	**65.0%**	Recreation	**0.574**	0.562	0.562
Personal	42.1%	45.8%	**51.4%**	Personal	0.421	0.476	**0.508**
Transportation	**68.5%**	53.9%	57.6%	Transportation	**0.734**	0.666	0.701
Average	63.6%	65.6%	**67.9%**	Average	**0.618**	0.606	**0.618**

then infer the activities with a pre-defined mapping from POIs. In [4], [16], [17], they use taxi trajectories and POIs to capture the human mobility patterns in an urban area. It is worth mentioning that taxi trajectories are quite different from individual's daily trajectories, and it results in different strategies between [4], [16], [17] and this work. In [18], they propose a relational travel topic model to infer the personal travel preferences of aviation customers. In addition, traditional classification methods were widely utilized to infer the trip purposes, such as random forest [2], artificial neural network [19], decision tree [20], SVM [1] and etc. Compared with the aforementioned work, the model proposed in this work incorporates heterogeneous data sets, including trajectories, POIs and social media messages. In addition, the proposed method captures the sequential properties of individual's trip activities which results in better inference results.

Trajectory Mining. Human trajectory data mining [21], [22], [23] has attracted lots of studies recently. Wu et al. [24] proposed a Markov Random Field model to infer the visited POIs given users' trajectories. It aims to answer the question "if a person is observed at certain location and time, which venue is the true destination of this person". In [25], Zhang et al. adopted Hidden Markov model to capture group-level human mobilities with social media data, and latent activity states are represented by topical words. In order to capture the topical information of trajectories, Kim et al. [26] proposed a probabilistic model to cluster different trajectories. With this method, significant movement patterns that appear frequently in data can be recognized. In [27], [28], they propose methods to predict future locations of users. In addition, there have been studies [29], [30] that extract the patterns underlying people's activities. Our task is different from these studies, as none of them attempts to capture the high-level trip purpose with both the knowledge from POI and social media.

POI Identification from Social Media Data. As we have mentioned, social media messages are quite noisy. This makes it very difficult to perform Named Entity Recognition from such short and noisy texts. In order to extract fine-grained location information from tweets, Li et al. [31] proposed a Conditional Random Field model to identify POI mentions from social media messages, and they further proposed a method [32] to link the POI name with Foursquare inventory. In [33], Flatow et al. proposed a data-driven approach to identify phrases associated with regions. Then it is used to label non-geotagged tweets with a regional area. In addition, a supervised Bayesian Model [34] is proposed to annotate POIs with tweet information. The above research aim at identifying POIs from general social media messages, especially for tweets without geo-tags. However, in this work, we propose to extract POI mentions from geo-tagged tweets. The proposed method with local POI candidate pools can solve this problem effectively and efficiently.

TABLE 8
Accuracy of ***iDBN*** and ***DBN***

	Accuracy	F1 score
DBN	62.63%	0.614
iDBN	62.57	0.613

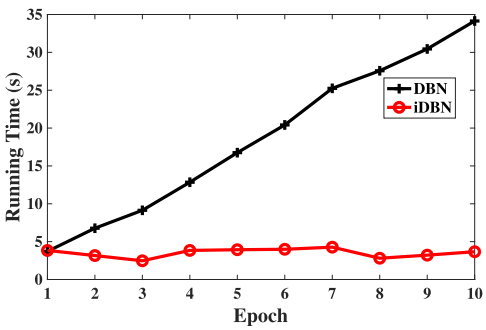


Fig. 7. Running time of ***iDBN***.

8 Conclusions

The knowledge of people's daily activities is very useful which can benefit both the government and residents. In this paper, we propose to infer people's trip purposes with heterogeneous data sets including trajectories, POIs, and social media messages. The proposed dynamic Bayesian network model can capture the intrinsic relationships among sequential activities. In addition, we also propose to incorporate POIs' popularity information near trip end locations. This information can give us good hints about people's activities. Moreover, in order to deal with the noisy social media data, we propose an effective method with local POI candidate pool to identify POIs from geo-tagged tweets. Furthermore, we also propose ensemble DBN and incremental DBN methods which can handle real-world challenges such as data unbalance and large data volume. Extensive experiments were conducted on real-world data sets, and the results demonstrate the advantages of the proposed method on accurately inferring the trip purposes.

Acknowledgments

The work was supported by US National Science Foundation (IIS-1553411, and CNS-1737590), Region II University

Transportation Center, and Transportation Informatics (TransInfo) University Transportation Center at University at Buffalo.

REFERENCES

[1] Z. Zhu, U. Blanke, and G. Tröster, "Inferring travel purpose from crowd-augmented human mobility data," in *Proc. 1st Int. Conf. IoT Urban Space*, 2014, pp. 44–49.

[2] A. Ermagun, Y. Fan, J. Wolfson, G. Adomavicius, and K. Das, "Real-time trip purpose prediction using online location-based search and discovery services," *Transp. Res. Part C: Emerging Technol.*, vol. 77, pp. 96–112, 2017.

[3] B. Furletti, P. Cintia, C. Renso, and L. Spinsanti, "Inferring human activities from GPS tracks," in *Proc. 2nd ACM SIGKDD Int. Workshop Urban Comput.*, 2013, Art. no. 5.

[4] P. Wang, Y. Fu, G. Liu, W. Hu, and C. Aggarwal, "Human mobility synchronization and trip purpose detection with mixture of hawkes processes," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2017, pp. 495–503.

[5] California household travel survey. 2019. [Online]. Available: http://www.dot.ca.gov/hq/tpp/offices/omsp/statewide_travel_analysis/chts.html

[6] W. Hua, K. Zheng, and X. Zhou, "Microblog entity linking with social temporal context," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2015, pp. 1761–1775.

[7] Y. Lou, C. Zhang, Y. Zheng, X. Xie, W. Wang, and Y. Huang, "Map-matching for low-sampling-rate GPS trajectories," in *Proc. 17th ACM SIGSPATIAL Int. Conf. Advances Geograph. Inf. Syst.*, 2009, pp. 352–361.

[8] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA, USA: MIT Press, 2009.

[9] J. Gao, W. Fan, J. Han, and P. S. Yu, "A general framework for mining concept-drifting data streams with skewed distributions," in *Proc. SIAM Int. Conf. Data Mining*, 2007, pp. 3–14.

[10] R. M. Neal and G. E. Hinton, "A view of the Em algorithm that justifies incremental, sparse, and other variants," in *Learning in Graphical Models*. Berlin, Germany: Springer, 1998, pp. 355–368.

[11] Google places API. 2019. [Online]. Available: https://developers.google.com/places

[12] J. Wolf, R. Guensler, and W. Bachman, "Elimination of the travel diary: Experiment to derive trip purpose from global positioning system travel data," *Transp. Res. Rec.: J. Transp. Res. Board*, vol. 1768, pp. 125–134, 2001.

[13] L. Shen and P. R. Stopher, "A process for trip purpose imputation from global positioning system data," *Transp. Res. Part C: Emerging Technol.*, vol. 36, pp. 261–267, 2013.

[14] Z. Deng and M. Ji, "Deriving rules for trip purpose identification from GPS travel survey data and land use data: A machine learning approach," in *Proc. 7th Int. Conf. Traffic Transp. Stud.*, 2010, pp. 768–777.

[15] D. Falcone, C. Mascolo, C. Comito, D. Talia, and J. Crowcroft, "What is this place? Inferring place categories through user patterns identification in geo-tagged tweets," in *Proc. 6th Int. Conf. Mobile Comput. Appl. Serv.*, 2014, pp. 10–19.

[16] P. Wang, G. Liu, Y. Fu, Y. Zhou, and J. Li, "Spotting trip purposes from taxi trajectories: A general probabilistic model," *ACM Trans. Intell. Syst. Technol.*, vol. 9, no. 3, 2018, Art. no. 29.

[17] Y. Liu, C. Liu, X. Lu, M. Teng, H. Zhu, and H. Xiong, "Point-of-interest demand modeling with human mobility patterns," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2017, pp. 947–955. [Online]. Available: http://doi.acm.org/10.1145/3097983.3098168

[18] J. Liu, B. Liu, Y. Liu, H. Chen, L. Feng, H. Xiong, and Y. Huang, "Personalized air travel prediction: A multi-factor perspective," *ACM Trans. Intell. Syst. Technol.*, vol. 9, no. 3, 2018, Art. no. 30.

[19] G. Xiao, Z. Juan, and C. Zhang, "Detecting trip purposes from smartphone-based travel surveys with artificial neural networks and particle swarm optimization," *Transp. Res. Part C: Emerging Technol.*, vol. 71, pp. 447–463, 2016.

[20] M. Oliveira, P. Vovsha, J. Wolf, and M. Mitchell, "Evaluation of two methods for identifying trip purpose in GPS-based household travel surveys," *Transp. Res. Rec.: J. Transp. Res. Board*, vol. 2405, pp. 33–41, 2014.

[21] Y. Zheng, "Trajectory data mining: An overview," *ACM Trans. Intell. Syst. Technol.*, vol. 6, no. 3, 2015, Art. no. 29.

[22] L. Chen, M. Lv, Q. Ye, G. Chen, and J. Woodward, "A personal route prediction system based on trajectory data mining," *Inf. Sci.*, vol. 181, no. 7, pp. 1264–1284, 2011.

[23] J.-Y. Kang and H.-S. Yong, "Mining spatio-temporal patterns in trajectory data," *J. Inf. Process. Syst.*, vol. 6, no. 4, pp. 521–536, 2010.

[24] F. Wu and Z. Li, "Where did you go: Personalized annotation of mobility records," in *Proc. 25th ACM Int. Conf. Inf. Knowl. Manage.*, 2016, pp. 589–598.

[25] C. Zhang, K. Zhang, Q. Yuan, L. Zhang, T. Hanratty, and J. Han, "GMove: Group-level mobility modeling using geo-tagged social media," in *Proc. Int. Conf. Knowl. Discovery Data Mining*, 2016, Art. no. 1305.

[26] Y. Kim, J. Han, and C. Yuan, "TOPTRAC: Topical trajectory pattern mining," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 587–596.

[27] Y. Wang, N. J. Yuan, D. Lian, L. Xu, X. Xie, E. Chen, and Y. Rui, "Regularity and conformity: Location prediction using heterogeneous mobility data," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 1275–1284.

[28] J. Zhou, A. K. Tung, W. Wu, and W. S. Ng, "A semi-lazy approach to probabilistic path prediction in dynamic environments," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2013, pp. 748–756.

[29] Q. Yuan, W. Zhang, C. Zhang, X. Geng, G. Cong, and J. Han, "PRED: Periodic region detection for mobility modeling of social media users," in *Proc. 10th ACM Int. Conf. Web Search Data Mining*, 2017, pp. 263–272.

[30] C. Zhang, J. Han, L. Shou, J. Lu, and T. La Porta , "Splitter: Mining fine-grained sequential patterns in semantic trajectories," *Proc. VLDB Endowment*, vol. 7, no. 9, pp. 769–780, 2014.

[31] C. Li and A. Sun, "Fine-grained location extraction from tweets with temporal awareness," in *Proc. 37th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2014, pp. 43–52.

[32] Z. Ji, A. Sun, G. Cong, and J. Han, "Joint recognition and linking of fine-grained locations from tweets," in *Proc. 25th Int. Conf. World Wide Web*, 2016, pp. 1271–1281.

[33] D. Flatow, M. Naaman, K. E. Xie, Y. Volkovich, and Y. Kanza, "On the accuracy of hyper-local geotagging of social media content," in *Proc. 8th ACM Int. Conf. Web Search Data Mining*, 2015, pp. 127–136.

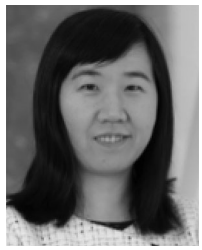
[34] K. Zhao, G. Cong, and A. Sun, "Annotating points of interest with geo-tagged tweets," in *Proc. 25th ACM Int. Conf. Inf. Knowl. Manage.*, 2016, pp. 417–426.

Chuishi Meng received the BE degree in electronic information engineering from Tianjin University, China, in 2010, the MSc degree in information science from the University of Science and Technology of China, in 2013, and the PhD degree from the Computer Science Department University at Buffalo, State University of New York, in 2018. Now, he is a research scientist of the JD Intelligent Cities Business Unit and JD Intelligent Cities Research. His research interest include data and information analysis with a focus on data mining and machine learning. In particular, he is interested in information integration and data mining applications in various domains, such as smart transportation, environmental monitoring, urban computing, and social sensing. He is a member of the IEEE.

Yu Cui received the BE degree in transportation management from Dalian Maritime University, China, in 2013. She is working toward the PhD degree specializing in transportation engineering at the University at Buffalo (UB), SUNY. She is currently a research assistant and a teaching assistant with the Department of Civil, Structural and Environmental Engineering, UB. Her research interests include connected vehicle big data analysis, transportation big data analysis, and railroad data analysis. She is a member of the IEEE.

Qing He received the BS and MS degrees in electrical engineering from Southwest Jiaotong University, and the PhD degree in systems and industrial engineering from the University of Arizona, in 2010. From 2010 to 2012, he worked as a postdoctoral researcher with IBM T. J. Watson Research Center. He is currently the Stephen Still assistant professor in Transportation Engineering and Logistics, affiliated with both civil engineering and industrial engineering with the University at Buffalo, State University of New York since 2012. His research interests include traffic signal control and freeway operations, social media and transportation, railway predictive maintenance, transportation data analytics, and supply chain management and logistics. He is a member of the IEEE.

Lu Su received the MS degree in statistics and the PhD degree in computer science both from the University of Illinois at Urbana-Champaign, in 2012 and 2013, respectively. He is an assistant professor with the Department of Computer Science and Engineering, SUNY Buffalo. His research focuses on the general areas of cyber-physical systems, wireless and sensor networks, and mobile computing. He has also worked with IBM T. J. Watson Research Center and National Center for Supercomputing Applications. He is a member of the ACM and IEEE.



Jing Gao received the PhD degree from Computer Science Department, University of Illinois at Urbana Champaign, in 2011, and subsequently joined UB, in 2012. She is currently an associate professor with the Department of Computer Science, University at Buffalo (UB), State University of New York. She is broadly interested in data and information analysis with a focus on information integration, crowdsourcing, ensemble methods, mining data streams, transfer learning, and anomaly detection. She is a member of the IEEE.