Neurocomputing 000 (2017) 1–7

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Probabilistic routing using multimodal data

Shunzhi Zhu[a], Yan Wang[a], Shuo Shang[b,c], Guang Zhao[d], Jiye Wang[e,*]

[a] School of Computer and Information Engineering, Xiamen University of Technology, China
[b] Department of Computer Science, China University of Petroleum, China
[c] Beijing Key Laboratory of Big Data Management and Analysis Methods, China
[d] Xiamen Great Power GeoInformation Technology Co., Ltd, State Grid Information & Telecommunication Group, China
[e] State Grid Corporation of China, China

ARTICLE INFO

Article history:
Received 13 May 2016
Revised 19 August 2016
Accepted 21 August 2016
Available online xxx

Keywords:
Multimodal data
Probabilistic routing
Transfer cost
Query processing
Public transportation networks

ABSTRACT

Human tracking and prediction are pervasive nowadays. It is possible to integrate multi-source human tracking data and location based social media data, which includes spatial data, temporal data, and textual data, to make human-mobility prediction and over-crowded station detection, and then to plan convenient bus/subway routes for passengers. This study is useful in many real applications, including convenient travel route recommendation and location based services in general. We face two challenges in this study: (1) how to use multi-source human tracking data to model probabilistic transfer cost between different bus/subway lines practically, and (2) how to compute convenient bus/subway routes efficiently. To overcome these challenges, we define a set of probabilistic spatial metrics and propose a travel-time threshold and a transfer-cost threshold convenient route planning queries. A series of optimization techniques are developed to enhance the query efficiency. We also conduct extensive experiments to verify the performance of the proposed algorithms.

1. Introduction

With the rapid development of GPS-enabled mobile devices, intelligent transportation systems, and online mapping services, human-mobility tracking and prediction are pervasive [35–37]. As an example, in many large cities like Hong Kong, the number of commuters is usually very huge at peak hours (e.g., millions of commuters travel around 7:00am∼ 8:30am). In a public transportation network, once a subway station is over-crowded, the transfer cost of commuters may increase. In this work, we integrate multi-source human tracking data and location based social media data, such as spatial data (locations), temporal data (time stamps), and textual data (geo-tagged tweets and geo-tagged microblogs) according to the approaches introduced in [33,34], to make human-mobility prediction and over-crowded station detection. Then, we define a set of probabilistic spatial metrics to describe transfer cost between two different bus/subway lines, and then we try to compute the convenient bus/subway routes (transfer-cost threshold query) efficiently in such public transportation networks. This type of query is useful in many real applications, including convenient travel route recommendation and location based services in general.

We give an example in Fig. 1, where exist 3 subway lines (blue, green, and red lines) and 9 stations p_1, p_2, ..., and p_9. Assume p_1 and p_4 are a source station and a destination station for a passenger, and $r_1 = \langle p_1, p_2, p_3, p_5, p_4 \rangle$ is a travel route. A passenger may take the blue line and at p_4 he may transfer to the red line. The local travel time of r_1 is 20 min, excluding transfer cost. At peak hours, p_5 has 35% over-crowded probability, and its transfer cost is $p_5.tc = 1 + 30 \times 35\% = 11.5$ min (if a station is over-crowded, it may be closed for 30 min, and the transfer time at normal time is 1 minute). Thus, the global travel time of r_1 is 31.5 min. Route $r_2 = \langle p_1, p_2, p_6, p_5, p_4 \rangle$ is an alternative travel route from p_1 to p_4, and it has p_2 and p_6 two transfer stations. The local travel time of route r_2 is 30 min, excluding transfer cost. At peak house, p_2 has 5% over-crowded probability, and its transfer cost is $p_2.tc = 1 + 30 \times 5\% = 2.5$ min. Thus, the global travel time of r_2 is 32.5 min. For a travel-time threshold routing query, if the travel-time threshold $\tau.t = 35$ in., r_2 is returned because it has a lower transfer cost (compared to r_1) and its travel time does not exceeds the travel-time threshold $\tau.t$. For a transfer-cost threshold routing query, if the transfer-cost threshold $\tau.tc = 12$ min, r_1 is returned because it has a lower travel time and its transfer cost does not exceeds the transfer-cost threshold.

* Corresponding author.
E-mail addresses: zhusz66@163.com (S. Zhu), wangyan@xmut.edu.cn (Y. Wang), jedi.shang@gmail.com (S. Shang), zhguang@163.com (G. Zhao), Jiyewang@126.com (J. Wang).

http://dx.doi.org/10.1016/j.neucom.2016.08.138
0925-2312/

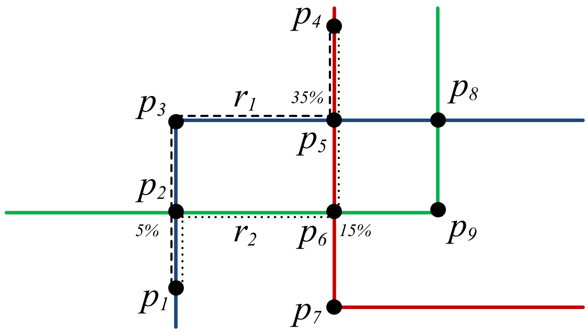

Fig. 1. An example of probabilistic routing.

Human-mobility tracking data, such as IC-card data of public transforation, are pervasive, and its data form is (line, source, departure time), which includes spatial, temporal, and textual attributes. As an example in Fig. 1, the stored passenger data is (blue line, p_1, 9:00am). The drop-off data (station and time) are not recorded because in many large cities, buses and subways are using equal-price ticket hence there is no need to record these data for ticket-fare calculation. Moreover, location based social media data (e.g., geo-tagged microblogs) are also useful to make human-mobility prediction and over-crowded detection. The main challenge here is how to integrate these data to establish an effective human-mobility prediction model and then to detect over-crowded stations effectively.

According to the approaches introduced in [33,34], we use multi-source data (human-tracking data and location based social media data) to make human-mobility prediction and over-crowded detection effectively. Stations can be classified into several levels based on their attractiveness. A higher-level station means more attractiveness for passengers. If a station is over-crowded, its transfer cost may increase. We define a set of probabilistic spatial metrics to describe the transfer cost between different lines practically. Then, we propose a travel-time threshold and a transfer-cost threshold convenient route planning queries. A series of optimization techniques are developed to enhance the query efficiency.

Notice that we inherit the prediction model and over-crowded detection methods from existing studies [33,34]. We improve the transfer cost model, and define two novel routing queries namely travel-time threshold and transfer-cost threshold queries. Existing studies focus on the computation of congestion probability, and their techniques cannot be used in the computation of transfer costs. To sum up, we make the following contributions in this work.

- We propose and investigate a novel problem of planning convenient routes in public transportation networks. This study is useful in many real applications, such as travel planning and recommendation and location based services in general.
- We define two novel travel-time threshold and transfer-cost threshold routing queries and the corresponding probabilistic spatial metrics practically.
- We develop two efficient algorithms to compute the travel-time threshold and transfer-cost threshold routing queries in public transportation networks.
- We conduct an experimental study to verify the performance of the developed algorithms.

2. Preliminaries

2.1. Multimodal data

We model a public transportation network by a connected and undirected graph $G(V, E)$, where V is a set of vertices (stations) and E is a set of edges (paths between two adjacent stations). We assign a weight to each edge to represent its travel time. A public transportation network includes several bus lines and subway lines.

IC card data are stored in the form of (line, source, departure time). Location based social media data (e.g., geo-tagged tweets, geo-tagged microblogs) are also useful in station classification. These data are in the form of (location, timestamp, short text). During a time period, we use the density of geo-tagged microblogs within a region defined by (p, r), where p is a station and r is a radius, to define the attractiveness of p. Then, we use the attractiveness level to detect over-crowded. For example, during the peak hours 7:00am$\sim$ 8:30am, we use the number of geo-tagged microblogs spatially close to a station to define the station's attractiveness. According to the number of geo-tagged microblogs, the attractiveness is classified into several levels. A lower level means a higher attractiveness.

Notice that we share the definitions of public transportation networks, IC card data, and location based social media data with existing studies [33,34].

2.2. Human-mobility prediction and over-crowded Station Detection

As introduced in [33,34], the human-mobility prediction model is based on priority ranking. Given a subway line L, a start station p, a moving direction, and a following station p_i, the drop-off probability of p_i is computed by

$$p_i.prob = e^{-p_i.level}, \tag{1}$$

where, $p_i.prob$ is the drop-off probability of station p_i, $p_i.level$ is the priority level of p_i (the stations are classified into several levels based on their attractiveness, and a station with a higher level means more attractive to passengers), and p_i is a following station of p. For example, in Fig. 1, in route r_1, p_1 is a start station, and p_2, p_3, p_5, and p_4 are following stations of p_1. Then, we normalize the original probabilities as

$$p_i.prob_N = \frac{p_i.prob}{\sum_{p_j \in p.f} p_j.prob} \tag{2}$$

where $p_i.prob_N$ is the normalized probability of station p_i, and $p.f$ is the following station set of p.

The over-crowded station detection method is also detailed in [33,34]. Given a station $p \in G(V, E)$ and a time period (t_s, t_e), we count the drop-off probability of all passengers at p during (t_s, t_e) as $count = \sum_{t_s}^{t_e} p.prob_N$. If the value of $count$ exceeds the passenger capacity τ, p is set to "over-crowded" station and its over-crowded probability is set to 100% during (t_s, t_e). Otherwise, the over-crowded probability of p during (t_s, t_e) is computed by $p.prob = \frac{count}{\tau}$.

2.3. Transfer Cost

For a station p in $G(V, E)$, its transfer cost $cost(p)$ is defined by Eq. (3).

$$cost(p) = \begin{cases} 0 & \text{if } C_1 \\ p.tc + M \cdot p.prob & \text{if } C_2 \end{cases} \tag{3}$$

C_1: $p.pre$ and $p.next$ are two stations belonging to the same line, such as $p_2 \in r_1$ in Fig. 1, where $p_2.pre = p_1$ and $p_2.next = p_3$ in route r_1.

C_2: $p.pre$ and $p.next$ are two stations belonging to different lines, and p is a transfer station, such as $p_6 \in r_2$ in Fig. 1, where $p_6.pre = p_2$ and $p_6.next = p_5$ in route r_2. $p.tc$ is the normal transfer cost between two different lines, M is the delay time when the over-crowded occurs, and $p.prob$ is the over-crowded probability.

Please cite this article as: S. Zhu et al., Probabilistic routing using multimodal data, Neurocomputing (2017), http://dx.doi.org/10.1016/j.neucom.2016.08.138