Electronic Commerce Research and Applications 14 (2015) 431–443

Contents lists available at ScienceDirect

Electronic Commerce Research and Applications

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

Cloud computing-based map-matching for transportation data center

Jian Huang [a,b], Jinhui Qie [c], Chunwei Liu [d], Siyang Li [c], Jingnong Weng [a,e,*], Weifeng Lv [a,*]

[a] School of Computer Science and Engineering, Beihang University, Beijing, PR China
[b] Jiangsu Province Collaborative Innovation Center of Modern Urban Traffic Technologies, Nanjing, PR China
[c] School of Software, Beihang University, Beijing, PR China
[d] National Space Science Center, Chinese Academy of Sciences, Beijing, PR China
[e] International School, Beihang University, Beijing, PR China

ARTICLE INFO

Article history:
Received 17 November 2014
Received in revised form 16 February 2015
Accepted 6 March 2015
Available online 14 May 2015

Keywords:
Map-matching
MapReduce
Hadoop
Vehicle tracking data
GPS privacy protection

ABSTRACT

Transportation data center has recently become a common practice of modern integrated transportation management in major cities of China. Being the convergence center of large-scale multi-source vehicle tracking data, it caused great challenge on GPS map-matching efficiency and privacy protection. In this paper, we propose a secure parallel map-matching system based on Cloud Computing technology to meet the demand of transportation data center. The main contributions are as follows: (1) we propose a leap-frog method to improve the efficiency of traditional serial map-matching algorithm on the increasingly common high sampling rate GPS data; (2) we adapt the serial leapfrog map-matching algorithm for cloud computing environment by reforming it in the MapReduce paradigm; (3) we propose a privacy-aware map-matching model over hybrid clouds to realize the sensitive GPS data protection. We implemented the proposed map-matching system in the hadoop platform and tested its performance with a large-scale vehicle tracking dataset, which exceeds 100 billion records. The experimental results show that our approach is highly efficient and effective on massive vehicle tracking data processing.

1. Introduction and motivation

The new economy of E-commerce heavily relies on just-in-time performance and integrated transportation logistics systems. To satisfy the challenges and demands that are placed on transportation systems and trucking services, the emerging and evolving technology known as Intelligent Transportation System (ITS) has become the key solution to sustain the new economy's growth and competitiveness.

Vehicle location awareness technology, particularly the wide used Global Positioning System (GPS), is notably essential to ITS. Many important ITS technologies, such as vehicle monitoring, route suggestion, and traffic flow prediction, etc., rely on the analysis of vehicle tracking data that are collected using GPS-instrumented probe vehicles.

Unfortunately, because of the measurement error caused by the limited GPS accuracy, and the sampling error caused by the sampling rate (Brakatsoulas et al. 2005), the reported locations in GPS points are not precise for the post hoc analysis. Thus, **M**ap **M**atching (MM), which can integrate positioning data with the road network to identify the correct path on which a vehicle is traveling, is an indispensable step of GPS applications.

Large amounts of research works on MM algorithm have been conducted over the past decades (Quddus et al. 2007). In earlier studies, the proposed MM algorithms were usually designed for the sole use of one application (such as the on-board navigation system). In these one-to-one scenarios, vehicle tracking data are collected from one or few GPS devices; thus, the data scale is relatively small and can be processed promptly even with hand held devices. Therefore, the researchers generally focused most of their efforts on improving the matching accuracy. The matching speed, which is another important measure of the algorithm performance, has not been pushed to the extreme.

However, the once sufficient matching speed has been facing severe challenge in recent years. Consider the scenario of map-matching process in transportation data center (Zhou 2012). As an important infrastructure of the emerging Smart City development strategy, a transportation data center is build by the municipal government to collect and pre-process huge volumes of multi-source traffic data all together, and share them with various modern integrated transportation management systems. In this one-many scenario, the MM program that runs in the data center must process a huge amount of vehicle tracking data that are

* Corresponding authors at: School of Computer Science and Engineering, Beihang University, Beijing, PR China.
E-mail addresses: wengjn@buaa.edu.cn (J. Weng), lwf@nlsde.buaa.edu.cn (W. Lv).

http://dx.doi.org/10.1016/j.elerap.2015.03.006
1567-4223/

required by various upper-layer ITS applications in notably tight time schedule; thus, the matching speed becomes a vital measure for the MM algorithm.

For example, Beijing Transportation Data Center collects vehicle tracking data from public transit vehicles and commercial fleet of more than 60,000 taxis, 30,000 buses and many other probe vehicles. The tracking data upload rate reaches 50,000 GPS records per second, and the accumulated historical raw GPS data set that waits for MM processing has already exceeded 100 billion records. Mean while, the reported processing speeds of the current MM algorithms are at most 10,000 GPS records per second (when the average GPS sampling period is 15 s), according to the result of ACM SIGSPATIAL GIS Cup 2012 (Ali et al. 2012), a well-known competition on MM algorithms. The processing efficiency is apparently beyond the demand of the transportation data center.

The centralized GPS data processing and sharing mode of transportation data center also brings up another serious issue – vehicle tracking data privacy protection. Despite all the merits (such as cost-effective computing and storage power, interoperability between systems, etc.), some organizations are still reluctant to handle their sensitive vehicle tracking data to transportation data center for map-matching process, due to the concern that data exposure may leak their operational information or customers' privacy, and subsequently made them suffer from financial loss or legal judgment. For example, the users' moving trajectories recorded by navigation/LBS service providers can easily be used by malicious attackers to infer user's detailed activities, or to track and predict the user's daily movements (Krumm 2009). Accordingly, measures must be taken to protect the sensitive GPS data during map-matching process in transportation data center.

To address the unprecedented challenges of large-scale vehicle tracking data processing in urban transportation data centers, this paper proposes a secure parallel map-matching system that aims to achieve high efficiency and user privacy protection. The main ideas of this work include the following:

1. Topological MM algorithm improvement with the *leapfrog* method
   The traditional MM algorithm tends to exploit all available GPS points in the tracking dataset to increase the matching accuracy. That excessive effort causes extra needless computation in certain circumstances.
   We observed that in the case of a high GPS sampling rate or a long road segment, a large part of the intermediate GPS points between two consecutive road intersections do not contribute to the vehicle travel path determination. Based on this observation, we implanted a new *leapfrog* method into the popular topological MM algorithm to find these superfluous points and save considerable computing time by fully skipping them in the candidate path determination.
2. Parallelization of the MM program
   Most existing MM algorithms are designed to run serially. Their capabilities are constrained by the computation capacity of a single computer, which is far below the MM requirement in a transportation data center. Hence, it is natural to split the huge GPS dataset to a number of compute nodes to gain a multiplied computing power.
   Instead of distributing serial MM programs to standalone computers, we turn to cloud computing techniques to gain more power from connected computers (Li et al. 2011). In other words, we use the MapReduce paradigm to parallelize our improved topological MM algorithm and implement it in the hadoop platform. The well-designed parallelization solved former problems in serial MM, such as large search space, slow data sorting, and long map-loading time. The inherent merits of the hadoop platform, such as load balance and fault tolerance, also make our system more adaptive to big data processing in the transportation data center environment.
3. Privacy-Aware MM model on hybrid cloud

Current researches on GPS data privacy protection have mainly focused on introducing uncertainty or error into location data, e.g., location anonymity and obfuscation (Seidl 2014). These approaches fall short because they all degrade the quality of positioning data, thus caused trouble in subsequent use of ITS applications.

In this paper we propose a novel solution aimed at preserving the location privacy without sacrificing the accuracy of MM system. We first build a hybrid cloud in transportation data center. For the sensitive GPS data map-matching tasks, most stages of MM are accomplished in public cloud, while the final stage of merging intermediate results into discernible travel paths is carried out within the organization's private cloud. A control program is implemented to automatically arrange data isolation and placement among private cloud and public cloud according to the data privacy characteristic.

The remainder of this paper is organized as follows. In Section 2, we present some preliminary information of our work. The proposed leapfrog method is discussed in Section 3, the parallel map-matching algorithm is presented in Section 4, and the privacy-aware map-matching model is described in Section 5. Section 6 reports the empirical experiment results and describes the evaluation in terms of efficiency, accuracy and scalability. Finally in Section 7, we conclude the paper.

2. Background and motivation

2.1. Imprecise nature of vehicle tracking data

Vehicle tracking data are a stream of spatio-temporal points that represent the trajectory of moving vehicle. For a probe car, a trajectory P consists of n observed positioning points $(p_i)_{i=t_0}^{t_n}$ during the time interval $[t_0, t_n]$. Each point p_t (referred to as a GPS record) is a 4-tuple (CarID, X, Y, t), where X is the longitude, Y is the latitude, and t is a represents timestamp.

The values of X and Y in the GPS record is actually imprecise because of the *measurement error* caused by the limited GPS accuracy. The typical measurement error is in the range of 2–10 m. In certain situations (shadowed and reflected signals), the position recorded in P can differ from the actual location with an error of up to hundreds of meters (Brakatsoulas et al. 2005).

Because ITS applications usually must know the vehicle's travel path, the original tracking data P must be aligned with the road network on a given digital map using a pre-processing procedure known as map-matching. Taking $P = (p_i)_{i=t_0}^{t_n}$ as the input, the map matching procedure should choose from the road network a corresponding sequence of consecutively connected road links $L = (l_i)_{i=0}^{m}$ that represent the vehicle's travel path.

In addition to the measurement error, the correctness of the matching result is also deeply affected by the second error of the tracking data, which is known as the *measurement error* and is directly related to the frequency with which the position samples are taken (sampling rate) from the GPS device.

Considering a typical sampling rate of 30 s and a travel speed of 50 km/h, the moving vehicle may cover a distance of 417 m between two consecutive GPS sample points, with several possible routes for the vehicle to travel from the first point to the second one. Wenk et al. (2006) used the following figure (see Fig. 1) to show these two errors and their effect on the travel path determination. The green road links within the active region are identified as candidate paths according to their proximity to the sampling