



Article

Utilizing MapReduce to Improve Probe-Car Track Data Mining

Li Zheng ¹ , Meng Sun ¹, Yuejun Luo ^{2,3,*}, Xiangbo Song ³, Chaowei Yang ⁴ , Fei Hu ⁴ and Manzhu Yu ⁴

¹ School of Geodesy and Geomatics, Wuhan University, Wuhan 430072, China; lzhang@sgg.whu.edu.cn (L.Z.); mengsun_2016@163.com (M.S.)

² School of Resource and Environmental Sciences, Wuhan University, Wuhan 430072, China

³ Wuhan Kotei Infomatics Co., Ltd., Wuhan 430072, China; xiangbos@kotei-info.com

⁴ Department of Geography and GeoInformation Sciences, College of Science, George Mason University, Fairfax, VA 22030, USA; cyang3@gmu.edu (C.Y.); fhu@masonlive.gmu.edu (F.H.); myu7@gmu.edu (M.Y.)

* Correspondence: yuejunl@kotei-info.com; Tel.: +86-27-8785-5598

Received: 29 May 2018; Accepted: 19 July 2018; Published: 23 July 2018



Abstract: With the rapidly increasing popularization of the automobile, challenges and greater demands have come to the fore, including traffic congestion, energy crises, traffic safety, and environmental pollution. To address these challenges and demands, enhanced data support and advanced data collection methods are crucial and highly in need. A probe-car serves as an important and effective way to obtain real-time urban road traffic status in the international Intelligent Transportation System (ITS), and probe-car technology provides the corresponding solution through advanced navigation data, offering more possibilities to address the above problems. In addition, massive spatial data-mining technologies associated with probe-car tracking data have emerged. This paper discusses the major problems of spatial data-mining technologies for probe-car tracking data, such as true path restoration and the close correlation of spatial data. To address the road-matching issue in massive probe-car tracking data caused by the strong correlation combining road topology with map matching, this paper presents a MapReduce-based technology in the second spatial data model. The experimental results demonstrate that by implementing the proposed spatial data-mining system on distributed parallel computing, the computational performance was effectively improved by five times and the hardware requirements were significantly reduced.

Keywords: Probe-car track; spatial data-mining; big data; MapReduce

1. Introduction

With the rapid development of urbanization and the economy, the automobile has become essential in everyday life. For example, China's national motor vehicle ownership reached 310 million as of the end of 2017, which includes 217 million cars [1]. As a consequence, various problems have emerged, including traffic congestion, energy crises, traffic safety, and environmental pollution. The increasing traffic demands and tightened automotive emission standards urge infrastructure operators and the automotive industry to act. Faced with these problems, urgent demands have been set on green driving, safe driving, congestion relief, and so on. To meet these demands, more advanced navigational data are needed, including dynamic traffic status information, passage cost information, road accident information, and road fuel consumption information. Probe-cars provide an opportunity to obtain these data by participating in the traffic flow and determining self-experienced traffic conditions, and transmitting these to a traffic center.

The probe-car, known as the Global Positioning System (GPS) rover, is an important and effective way to obtain urban road traffic status in the international Intelligent Transportation System (ITS) [2]. It is an advanced road traffic information collection technology in the field of international ITS [3]. By mounting a GPS device on a vehicle, the vehicle's position information, behavioral information, and event information are transferred to a data center in real time or offline. By combining massive probe-car tracks and behavioral data obtained from the data center with the existing map data, a variety of advanced navigation data can be mined to address certain problems, including the discovery of new roads, the actual cost of road traffic, the peak periods of roads, and accident-prone areas. Probe-car data can range throughout the region and be collected 24/7. This technology greatly improves the efficiency of information collection by wireless real-time transmission and center-through processing. Meanwhile, it lowers the cost of maintenance of acquisition equipment by the installation of GPS and other communication network resources [4].

A probe-car system consists of a wireless communication network that includes GPS and wireless communication capabilities and an information processing center. Probe-car data (PCD) systems are composed of three parts: the data acquisition system of the probe-car, the traffic information processing system, and the real-time traffic information distribution system. The probe-car is driven on city roads and uploads the collected real-time raw data to the probe-car data acquisition system. The traffic information processing system is responsible for the raw data preprocessing, coordinate conversion, geographic information system (GIS) electronic map matching, as well as travel time calculation. The information distribution system releases the traffic information by the processing system to provide the public with real-time road traffic reference information by General Packet Radio Service (GPRS), the Internet, and other means. The architecture of a probe-car data system is shown in Figure 1.

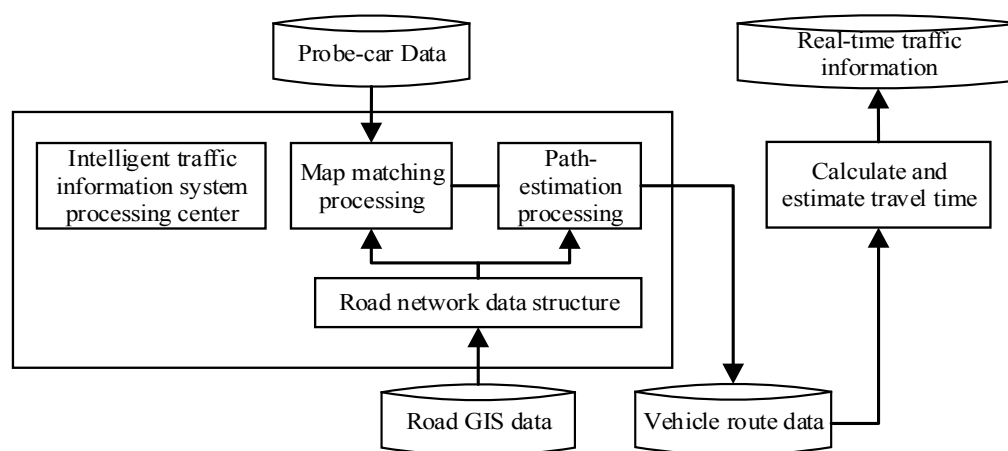


Figure 1. Technical architecture of a probe-car system.

As an important component of ITS, the urban traffic dynamic route guidance system obtains vehicle locations under real-time road traffic conditions and provides the best route guidance information. It helps direct travelers in order to improve traffic conditions, reduce traffic congestion, and achieve a reasonable distribution of traffic flow on the roads. ITS experts and enterprises have conducted theoretical research and developed applications based on vehicle locations. Countries in Europe have developed various new technologies and made intelligent road transport systems available in various cities. The ADVANCE real-time traffic releasing system for probe-cars was initiated by the State of Illinois with the US Federal Highway Administration as its partner [5]. The purpose of this system is to determine whether drivers need real-time information to avoid congestion, in order to increase capacity. The UK traffic master offers a series of traffic information services, where the data are mainly provided from fixed sensors and supplemented by PCD [6].

To the authors' knowledge, research on large-scale PCD processes is still in the preliminary stages. The existing research mainly focuses on cost, probe-car size, system architecture, and precision [7,8]. With the development of PCD technology and the popularity of GPS devices, the amount of data will grow dramatically. Due to the specific mobility of probe-car data and the limitation of the size of the cars, real-time probe-car data is unable to cover all of the road network. It is essential to address the missing data issues in the road network analysis and improve the application's efficiency.

We describe the basic principles and key issues of mining probe-car data in Section 2. Section 3 describes a MapReduce approach to accelerate the map-matching of massive probe-car tracking data. Section 4 demonstrates the experimental results of the MapReduce approach. Section 5 concludes the research and introduces future work.

2. Literature Review and Key Issues of Mining Probe-Car Data

2.1. Literature Review

Probe-car technology research can be traced back decades. Considering the complex construction of intelligent transportation systems, the relevant agencies have internationally conducted a great amount of research and application, as shown in Table 1. For example, the UK's probe-car data system was developed to collect and analyze traffic information and was invested in by ITIS Holdings Plc, a typical successful probe-car system [6]. The data sources include both real-time and historical information from the Automobile Association Traffic Control Centre (AATCC). The probe-car data system uses the GPS/wireless data transmission mode. After the acquired data are processed, the system predicts the traveling time for users in real time and continues to update the information. Another example is the American ADVANCE system, which was an experimental project of dynamic road induction in an Illinois suburban area conducted by the Federal Highway Administration (FHA), Illinois Transportation Authority (ITA), Motorola, Transportation Research Institute at Illinois State University, and other agencies in 1991. Its goal was to determine whether traffic guidance information is helpful to avoid traffic congestion and improve driving quality [6]. The Vehicle Information and Communication System (VICS) in Japan is one of the successful applications in the field of intelligent transportation. VICS acquires traffic data via GPS navigation devices and releases accurate traffic guidance information and real-time traffic information to travelers by FM radio and wireless data transmission [9]. The Korea Road Traffic Information Center (KORTIC) system of Korea, developed by the Korean Road Safety Association (KRSA), combines toroidal coil, GPS probe-car, and Closed-Circuit Television (CCTV) surveillance equipment for traffic information collection. Then it extracts traffic information after data fusion, analysis, and processing, and determines the traffic status to reduce the estimated error probability for obtaining road travel time to 10% or less [10]. In 2001, the German Space Center Transportation Institute (GSCTI) integrated probe-cars with 300 floating taxis to collect and analyze their location, speed, and other information in Berlin [11]. Jan Fabian Ehmke predicted time-dependent travel time and assessed the resulting road information using data-mining methods through different levels of aggregation for the large amount of probe-car data [12]. With its acquisition of Waze in 2013, Google added a human element to its traffic calculations. Drivers can use the Waze app to report traffic incidents including accidents, disabled vehicles, slowdowns, and even speed traps [13].

Table 1. Research and applications of intelligent transport systems. AATCC, Automobile Association Traffic Control Centre; FHA, Federal Highway Administration; ITA, Illinois Transportation Authority; GPS, Global Positioning System; VICS, Vehicle Information and Communication System.

System	Country	Developer	Main Characteristics/Data Source
Probe-car data	UK	ITIS Holdings Plc	From AATCC and real-time and historical data
ADVANCE system	USA	FHA, ITA, and other agencies	Avoid traffic congestion, improve driving quality
VICS	Japan	\	From GPS navigation devices
KORTIC	Korea	Korean Road Safety Association	From the toroidal coil, GPS probe-cars, and so on
Probe-car Data System	Germany	Space Center Transportation Institute	From floating taxis
Google Maps	USA	Google	A human element is added to its traffic calculations, report traffic incidents from drivers

Probe-car technology research in China started relatively late, but the progress is rapid. Various research institutes and scholars have made great achievements in theory and practice. Based on GPS navigation and the positioning of vehicles, Li [14] used velocity and time information to obtain the average speed, traffic, travel time, and other traffic information through a mathematical model to achieve real-time road detection. With the urban road network data acquired by probe-cars equipped with GPS devices, Dong [15] analyzed the road network level and obtained the travel conditions and functioning of the road network at different levels. Zhang [16] used pattern recognition, statistical forecasting, time series, and intelligent algorithms through traffic parameters of probe-car data acquisition to detect traffic incidents. Xin [17] analyzed the space and time distribution characteristics of urban road networks based on probe-car data, adopting the coverage and intensity in a certain coverage as indicators. Xin pointed out that the coverage and intensity of probe-car data have similar peak hours on weekdays. The higher the level of the road, the higher the coverage and intensity of probe-car data. Li [18] presented a mathematical model of probe-car coverage in a single section and the whole road network on the basis of the minimum requirements for probe-car samples in a single section and verified it by simulation. This model considered various factors such as computing interval, average traffic flow density, average travel speed, mistake matching scores of probe-cars, and so on. The simulated results showed that the coverage rate calculated by this model can ensure a 93.7% link of the road network through collecting data. Zhang [19] described the composition of probe-car systems and the optimization theory for probe-car sampling. By considering velocity and analyzing the random signal and the spectrum of the Fourier transform, the optimal sampling frequency is determined by the Shannon sampling theory. The results showed that the more optimal the sampling frequency obtained, the higher the data accuracy, which is suitable for practical applications. Weng [20] categorized probe-car data into three stages, historical data applications, historical traffic state data applications, and dynamic traffic state data applications, on the basis of summarizing the research of probe-car traffic information applications. He analyzed the urban transport operating characteristics of probe-car data in Beijing, such as the distribution characteristics of road traffic and the utilization of different levels of road mileage. Feng [21] proposed a probe-car map-matching algorithm based on the search for local paths by analyzing the characteristics of the collection of raw data. Making use of GPS points matched previously, it greatly reduces the search space to achieve better positional accuracy of probe-car data, so as to determine the vehicle track.

As the probe-car tracking data is massive, the above traditional data-processing methods cannot meet the current data-processing needs. This paper proposes a parallel algorithm combined with cloud computing technology to process the data.

2.2. Probe-Car Collection Interval

The timeliness and reliability of traffic status identification can be largely impacted by the configuration of essential parameters based on the data collection of GPS probe-cars, including the data-sampling interval and the sampling ratio. The collection interval refers to the time interval of uploading the collected data of the probe-car to the information processing center; in other words, the time period of the probe-car data collection. Generally speaking, the higher the frequency of data collection, the more accurate the real-time road traffic information.

However, if the acquisition time interval is too short, it will not only increase the cost of acquisition, but also result in higher data redundancy, and the road traffic conditions will be very similar. On the contrary, if the time interval is too long, it will miss important data, which would lead to not reflecting the dynamic traffic conditions precisely. Therefore, it is very important to set an appropriate collection interval for probe-car data.

Yamane conducted research on urban probe-cars in Osaka, Japan [22], and found that different data-collection intervals would result in different road map matching results. Researchers also found that the longer the time interval of probe-car data acquisition, the worse the accuracy of the reflection of real-time road traffic conditions, in spite of the relatively low cost of the acquisition. If the probe-car collection interval is short, the reflected effect of real-time road traffic conditions is better. Though the time cost of the acquisition will be relatively higher, the effect of GIS map-matching is better. From different tests, we found that when the probe-car collection interval is 30 s, we can achieve the best balance among the acquisition cost, the reflection accuracy of real-time traffic conditions, and the results of matching. Therefore, the optimal sampling interval of the probe-car is 30 s in this research.

2.3. The Key Issues in Probe-Car Data Mining

Probe-car track data consists of sampling points with a series of latitude and longitude information and other vehicle behavior information. The characteristics of probe-car data are as follows [23,24]:

- (1) Position information (latitude and longitude).
- (2) Position information with noise, and the noise is affected by a variety of factors (GPS noise, clouds, the status of buildings nearby, indoor and outdoor conditions, and so forth).
- (3) Loss of spatial information: the sampling interval of probe-car tracks is usually long (tens of seconds or minutes), which will result in the loss of shape information.
- (4) Spatial information redundancy caused by intensive sampling and low speed.
- (5) The temporal correlation between some series of track points.
- (6) Additional property information: incidental event information, driving behavior information, sensor parameters while sampling points.

For the above features of probe-car data tracks, the first five characteristics are the inherent basic properties of time and space, which will be fully considered during the data matching. The last characteristic is the additional information provided by the probe-car, which may vary greatly (not available for all vehicles) according to the vehicle's own situation, but it can be used to mine more behavioral factors.

In the spatial data mining of probe-car tracks, there are several key issues for the above data features [25,26]:

- (1) The close correlation of spatial data. Since the electronic map of the road network topology has a close correlation, it often needs to load all of the road network data into memory to handle all the tracks. Since the road network data is massive, it demands high performance for the hardware. Moreover, it results in low performance for searching the matching data in the entire road network to process each specific track, which is a fatal flaw for the massive spatial data mining of the probe-car track.

- (2) True path restoration for the probe-car track. It is crucial to combine the electronic map to restore the true path of probe-car tracks as accurately as possible. Since the sampling interval is sparse, the distance between two adjacent track points might be far. Due to data noise, the position of each track point may greatly deviate from the position of the real road, so large errors result from the conventional method, which matches the best roads according to each track point location on an electronic map. As shown in Figure 2a, if it is a partial match, a part marked I in the whole track will be matched to road 1. However, the whole track should be matched to road 2, as shown in Figure 2b. Therefore, combining the spatial characteristics of the entire track optimally and globally restores the path, while the local single track point cannot be matched to the best road.

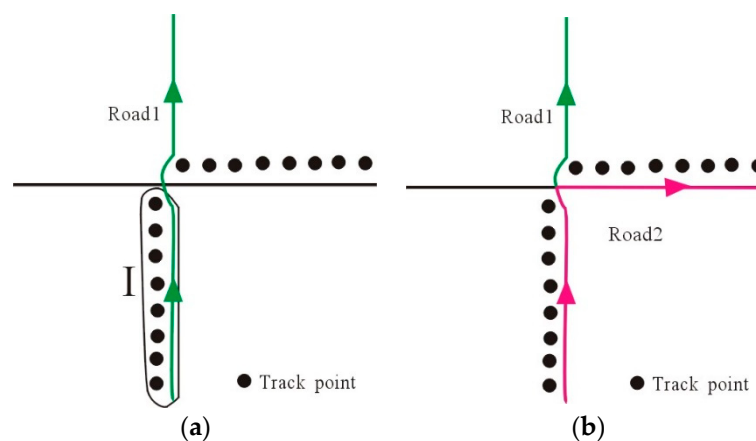


Figure 2. (a) Global Positioning System (GPS) track; (b) track point matching.

To obtain city-wide real-time traffic information, large-scale data are collected in real time and analyzed to ensure the timely release of traffic information. Using a general process to deal with these data, PCD will encounter great bottlenecks, thus the real-time data cannot be obtained, and there is no way to use a more sophisticated processing algorithm to improve the accuracy of the processing results. MapReduce adopts a distributed parallel computing model and makes it easy to implement parallel computing, load balancing, and other excellent properties [27–30], thus it is very suitable for big data mining. However, due to the close spatial correlation characteristic of the probe-car track data, the entire road network of the electronic map data needs to be read for each computing node to handle the massive track, which increases the hardware requirements for each computing node and impairs computing performance.

To address the above-mentioned problems of spatial data mining on probe-car tracks, massive probe-car track data and map data need to be loaded. By combining road topology with map matching, strong correlations can be made between track and map data. When matching and analyzing the data, the data are not locally matched, but globally matched, so as to avoid the case of Figure 2. For high data throughput and limited computing performance, we propose a parallel processing algorithm to reduce the probe-car tracks, namely, two-step MapReduce technology in the spatial data model.

3. Methodology

Two types of data were processed by different MapReduce algorithms to enable parallel computing of spatial data. A two-step MapReduce was conducted: (1) for space partitioning to analyze the relatively independent data, that is, the full tracks in one spatial division range that can be matched with the real road net, which is global matching; (2) for the cross-regional track processing to match with the road network of the electronic map in the designated cross-region.

3.1. MapReduce Parallel Distributed Computing Model

Dean and Ghemawat proposed a MapReduce distributed computing model for the analysis of web log files [27]. The Hadoop project implemented this computational model, which used a cluster consisting of thousands of computers to analyze the massive server files. The MapReduce model is realized mainly through two functions: mapping (Map) and reduction (Reduce). The main process [31] is shown in Figure 3.

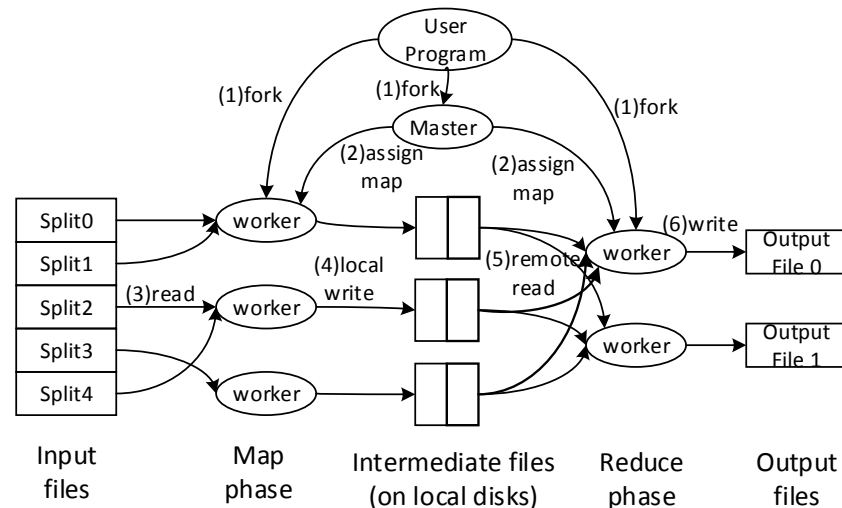


Figure 3. Main processes of MapReduce.

Google's MapReduce programming model serves to process large datasets in a massively parallel manner (subject to a MapReduce implementation). The programming model is based on the following simple concepts: (i) iteration over the input; (ii) computation of key/value pairs from each piece of input; (iii) grouping of all intermediate values by key; (iv) iteration over the resulting groups; and (v) reduction of each group.

The model is stunningly simple and effectively supports parallelism. The programmer may abstract from the issues of distributed and parallel programming, because MapReduce implementation takes care of load balancing, network performance, fault tolerance, and so forth. The seminal MapReduce paper [28,32,33] described one possible implementation model based on large networked clusters of commodity machines with local storage. The programming model may appear restrictive, but it provides a good fit for many problems encountered in the practice of processing large datasets. Additionally, expressiveness limitations may be alleviated by decomposing problems into multiple MapReduce computations or by escaping to other (less restrictive, but more demanding) programming models for subproblems.

3.2. MapReduce Method for Data Mining in Probe-Car Tracks

The size division is based on the distribution and density of the trajectory. After dividing the total area into the target areas as described above, a nested MapReduce approach can be used to achieve spatial data mining for the probe-car tracks. The implementation is described as follows:

(1) Level-1 Map function design

The Level-1 Map function is mainly responsible for dealing with the track in a small designated area of the probe-car, and the algorithm flow is shown in Figure 4.

The pseudocode is as follows:

```

Enter the range A1 to be processed;
A2 = A1 + 0.1 ° // Extend the range A1 as A2
M = LoadData (A2); // Load the map data of the range A2 into memory
Enter the track collection D
While(D != NULL)
{
    P = Read(D); // Read one track in M in order
    if (P ∈ A1) // Track P belongs to the range A1
    {
        Match track P with the electronic map M;
        If (matched)
        {
            log(P and M match relations);
        }
        else
        {
            log (non-matched trajectory P);
        }
    }
    else if (Part of P belongs to A1) // Track P is cross-regional
    {
        log(Track P, the latitude, and the longitude of track P);
    }
    D = D – P;
}

```

The algorithm is explained as follows:

Step 1: A small area is to be assigned and processed, such as the example of the spatial division range ($8^{\circ} \times 5^{\circ}20'$), extended to a range (for example, 0.1°) on the basis of a small area range. The algorithm will read the road network data in the electronic map of the extended region into memory.

Step 2: The algorithm will sequentially read all the probe-car tracks and determine whether the probe-car track is in the scope of the small area being processed currently.

Step 3: If the tracks of the probe-car are entirely in the small area, then it will match the track data with the road network data and record the matching relationship and a portion of the nonmatching to the road network in the electronic map.

Step 4: If the track of the probe-car is partly in the small area or out of the small area to be processed, it will record the track and range of latitude and longitude.

Step 5: If the tracks of the probe-car are entirely outside the small area to be processed, then it will do nothing.

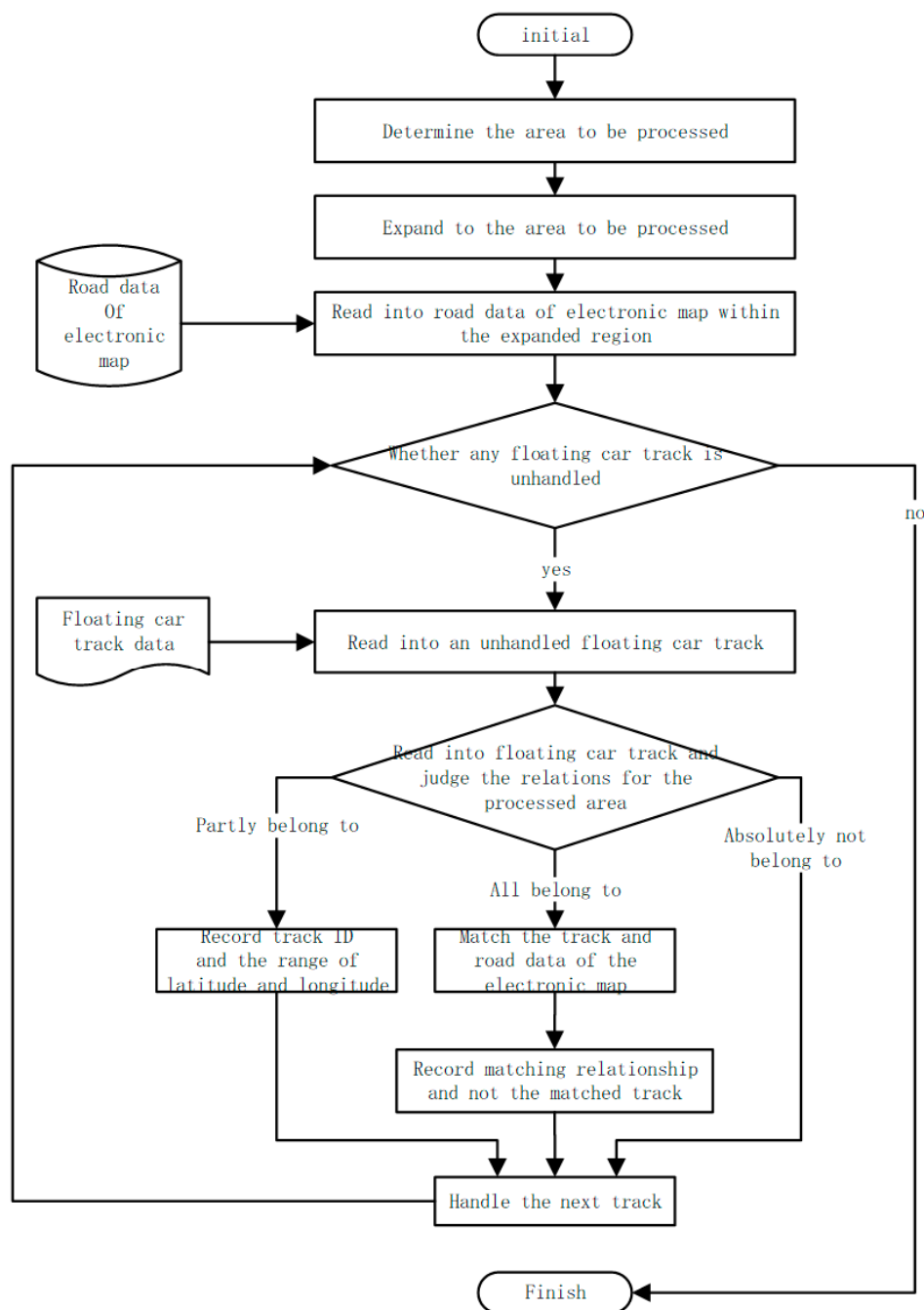


Figure 4. Workflow for the first map algorithm.

(2) Level-1 Reduce function design

The Level-1 Reduce function is responsible for writing the results processed for the probe-car tracks within the designated area to the master server. The algorithm is as follows:

Step 1: Write the correspondence relation for matching the local probe-car track and the electronic road map to the master server.

Step 2: Write the local records that the probe-car track does not match with the road in the electronic map into the master server.

(3) Level-2 Map function design

The Level-2 Map function is mainly responsible for processing the existence of probe-car tracks across the region of the designated small area currently. The algorithm is shown in Figure 5.

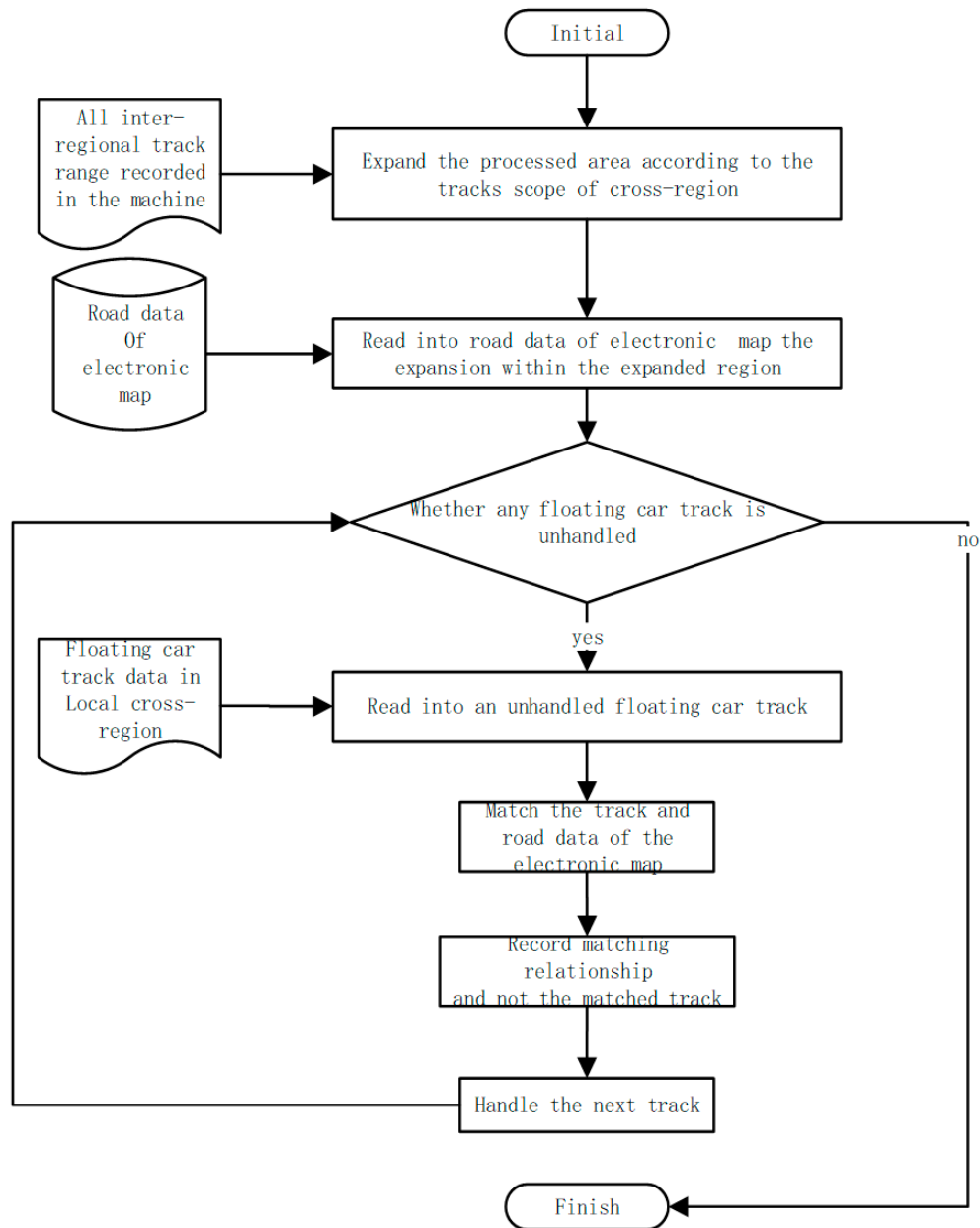


Figure 5. Second map algorithm flow.

The pseudocode is as follows:

Enter the range A1 to be processed;

Enter the set D of the cross-regional tracks to be processed

Read the latitude and longitude range A2 in D;

$A3 = A1 + A2$;

$M = LoadData(A3)$; // Load the map data of A3 into memory

While(D != NULL)

{

```

P = Read(D); // Read one of the cross-region tracks in M in order
Match track P with the electronic map M;
If(matched)
{
    log(P and M match relations);
}
else
{
    log (non-matched trajectory P);
}
D = D - P;
}
while (log != NULL)
{
    l = Read(log); // read a log in the log, matched relationship or non-matched track P

    if (l not in ControlServe)
    {
        l in the ControlServe;
    }
    log = log - l;
}

```

The algorithm is explained as follows:

Step 1: According to the latitude and longitude range A2 across the whole region for probe-car tracks, extended on the current designated area A1, the algorithm will read them into the memory for all the road network data A3(A3 = A1 + A2) in the electronic map of the extended region.

Step 2: Sequentially read all the probe-car tracks across the region.

Step 3: Match the track data with the road network data, then record the matching relation and nonmatching relation to the road network of the electronic map to the machine.

(4) Level-2 Reduce function design

The Level-2 Reduce function is mainly responsible for writing the matching relation and nonmatching relation to the road network of the electronic map in the designated cross-area to process into the master server and remove the duplicate part. The algorithm is explained as follows:

Step 1: Read the results of the probe-car track in the cross-region.

Step 2: Check whether the processing results of the probe-car track are on the master server.

Step 3: If the results of the track handling do not exist in the master server, write the results to the master server. Otherwise, handle the next track.

4. Experimental Results and Discussion

4.1. Experiment Scenarios

In the experiment, the test environment is the Hadoop platform. The computer is configured with the following specifications: CPU Core Duo 2.7 GHz, memory 8 GB DDR3, Windows operating system; one is the master node server, the other three are the computing node servers. First, all the nodes on the master node are directly calculated for the execution time. Second, the dataset is divided into $N \times M$ intervals by the method described. The two-time MapReduce algorithm is used to compute the nodes in different operations. It is obtained through the experiment that the use of this algorithm can improve the operation time of each node in the computing process and cannot lose effective association rules.

The PCD parallel processing system consists of the collection server software, data preprocessor program, server-side program, node calculation program, database storage, and other components. The PCD acquisition server receives real-time transmissions of a large number of PCDs, establishes a database, and saves all the original PCD collected by the system. The collection server has a high-speed connection to the master server and is convenient for data to be transported to the processing center quickly.

The preprocessing program includes two parts. First, the map data are preprocessed, generating map grid information. This part belongs to the offline calculations, which are only done once. Second, each of the real-time PCD records obtained is pretreated, which removes some obviously wrong invalid records caused by equipment failure.

The server program runs on the master server KD50, which is responsible for the PCD tasks of decomposition, scheduling, and consolidating the results. The node program on each computing node will allocate computing resources in accordance with the instructions on the server side of the program. In addition, the server program needs to initialize the system and load the configuration file, map file, and historic mining information, initiate the shutdown node, and so on. The node program distributes each computing unit over KD50, which is used to complete the map-matching data-processing tasks distributed by the master server. KD50 has many computing units. After task scheduling by the master server, there will be multiple node programs running processing tasks for the computing units. Thus, it can achieve basic PCD parallel processing.

4.2. Results

Based on the methods described in the paper, the experimental verification is conducted based on the probe-car tracks during one month in Japan. The experimental data are the long-distance freight data shown in Table 2. Table 3 shows the PCD sample data format. The probe-car tracks are composed of the PCD data in Table 3, and are shown on the map in Figure 6.

Table 2. Source data.

Number of Probe-Car Tracks	Number of Vehicles	Number of Roads in the Electronic Map	Total Road Length (km)
2,056,489	51,973	12,538,343	21,453,863

Table 3. Probe-car sample data format.

RecordID	GpsDateTime	GpsLatitude	GpsLongitude	GpsAzimuth	GpsSpeed
204451900001	2017/8/31 4:24	38.45707628	141.2965907	49°	39.6 km/h
204451900002	2017/8/31 4:25	38.45992784	141.3001752	45°	37.8 km/h
204451900003	2017/8/31 4:33	38.46099555	141.3012858	347°	9.9 km/h
204451900004	2017/8/31 4:34	38.4589209	141.2973367	164°	45 km/h
204451900005	2017/8/31 4:35	38.46034831	141.2877371	166°	57.6 km/h
204451900006	2017/8/31 4:36	38.46274794	141.2803071	155°	10 km/h
204451900007	2017/8/31 4:37	38.46221842	141.2794217	255°	36 km/h
204451900008	2017/8/31 4:38	38.45871636	141.279031	276°	44.1 km/h
204451900009	2017/8/31 4:39	38.45225857	141.2803076	276°	4.5 km/h
204451900010	2017/8/31 4:40	38.44693848	141.2823128	272°	42.3 km/h

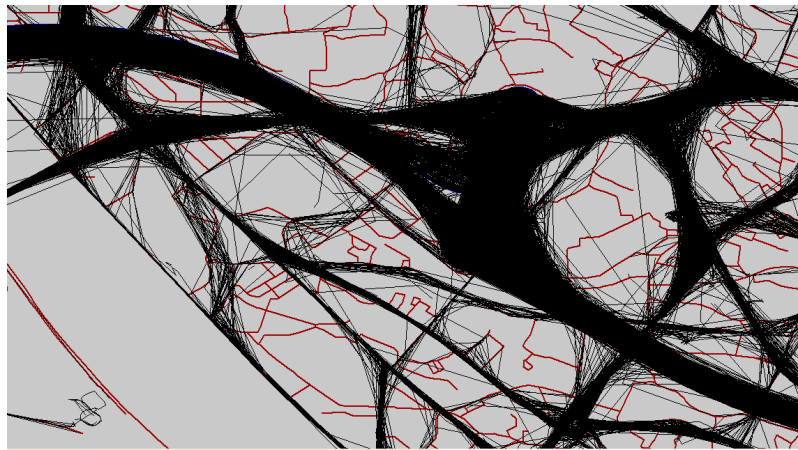


Figure 6. Vehicle tracks.

The different distributions and densities of the trajectories lead to different division sizes. The experiment for the different sizes of the divided region ($m \times n$) is shown in Figure 7. There are 4 machines to process the data by the proposed distributed parallel processing mode based on MapReduce. As shown in Figure 7, it uses the least time for $2^\circ \times 1^\circ 20'$ and the percentage of the cross-regional trajectories is 23.1%. It uses the most time for $12^\circ \times 8^\circ$ and the percentage of the cross-regional trajectories is 5.3%. The size division for m and n is based on the actual situation, such as the length, distribution, and density of the trajectory. Generally, the region should not be too small. It does not reflect the effect of MapReduce while the proportion of the cross-regional track is significant. Then the region should not be too large for the limitation of the consumption memory and computing resources.

Then the target area is divided into $n^\circ \times m^\circ$ small regions (n and m are positive numbers) according to the following principles, ensuring that most tracks (70–80%) are divided into no more than one small lattice:

- (1) n and m should not be too small, or the proportion of the tracks across the regions will be very significant, which cannot reflect the effect of MapReduce. n and m should be chosen so that the cross-regional trajectory ratio is kept around 25%. $m \times n$ is limited by memory limitations.
- (2) n and m should not be too large. If $m \times n$ is larger, consumption memory and computing resources will be greater. The maximum of $m \times n$ is limited by the resources a single computing unit can provide.
- (3) Within a single computing unit resource, it is not necessarily good to have larger $m \times n$ values, but it would be better to have a greater range for $m \times n$ and a lower marginal effect of the inter-regional track proportion. To meet the conditions in which the cross-regional track ratio is below a certain range, the smaller $m \times n$, the better.

In general, n and m are preferably larger than 1. The size of the divided region can be appropriately adjusted according to the number of working machines.

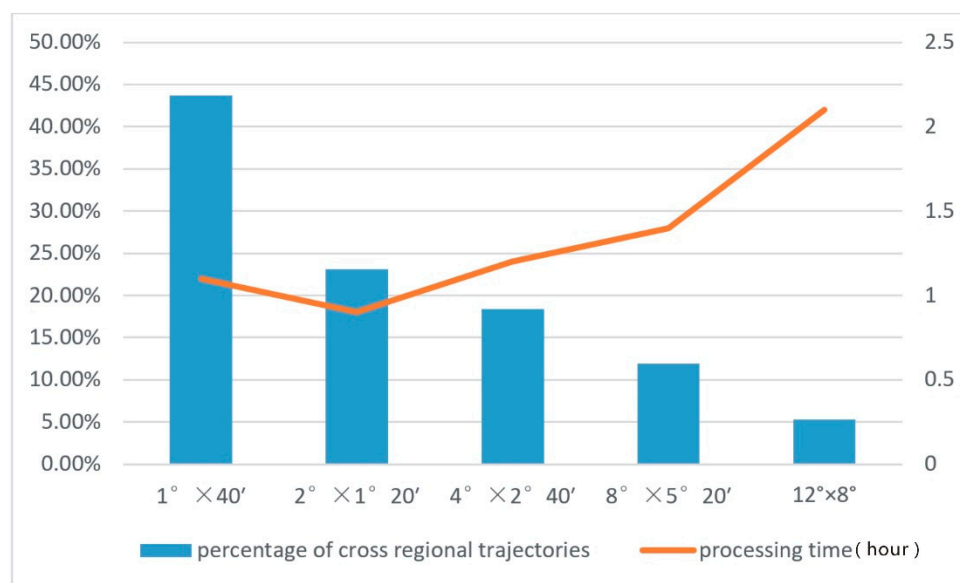


Figure 7. Different sizes of the divided region (m × n).

The traditional single server and the proposed approach in this paper, respectively, are used to process the road data matching in the electronic map for the probe-car track by the track-matching algorithm based on the spatial semantic features, where it is divided into the range of $2^{\circ} \times 1^{\circ}20'$ in the MapReduce model. The test results are shown in Table 4. The traditional single server takes about 10.5 h. By the proposed distributed parallel processing mode based on MapReduce, when the number of machines was respectively 4 units \rightarrow 8 units, the processing time required was less, from 2.5 h \rightarrow 2.1 h. Thus, the processing efficiency of the distributed parallel processing based on MapReduce shows great improvement.

Table 4. Experimental results ($2^{\circ} \times 1^{\circ}20'$).

Processing Approach	Single Processing	Distributed Processing	Distributed Processing
Machine type	IBM server	Ordinary desktop	Ordinary desktop
Machine physical memory	16 GB	2 GB	2 GB
Machine CPU	2.13 GB, 4-core	3.3 GB, 2 nuclear	3.3 GB, 2 nuclear
Number of machines	1	4	8
Processing time	10.5 h	2.5 h	2.1 h

5. Conclusions

This paper discusses the main issues of spatial data mining of probe-car tracks and presents the two-time MapReduce technology in a spatial data model to solve map-matching by using the massive trajectory data of floating vehicles and a special strong correlation of the data. Two types of data are processed by the MapReduce algorithms to enable parallel spatial data computing. The first MapReduce is conducted for space region partitioning to analyze relatively independent data, that is, full tracks as a global matching step. The second MapReduce is conducted for cross-regional track processing, as the cross-regional track. Experiments confirmed that MapReduce technology can be successfully used in data mining. Using distributed parallel computing for spatial data mining, the computing performance is significantly improved from the traditional approach of a single server and the hardware requirements are reduced. Using the proposed MapReduce model, the master server can better balance the tasks of a working host and achieve better load balancing and better stability. Based on the two-time MapReduce technology, the probe-car track is divided into spatial

region processes and track processes to solve the strong spatial correlation characteristics of the map data and achieve parallel processing of the probe car track reduction.

Based on the application of the parallel MapReduce algorithms, it is crucial to improve the probe-car track matching and the reduction algorithms for further research, so as to make better use of parallel computing to enhance the massive probe-car registration track performance. Based on the probe-car data and spatial data mining general model, we will conduct floating space vehicle trajectory data matching with reference to the specific electronic maps. According to the matching information, the track is divided into matched and not matched track information. From the not matched track information, new roads, accident-prone areas, and parking Point of Interests (POIs) can be discovered. From the matched track information, abandoned roads and shape-changing information for the roads and other map elements can be implied. Utilizing the parallel MapReduce algorithms, the lane path extracted by the vehicle tracks is the future research issue that emerges from our approach, shown in Figures 8 and 9. Based on the probe-car data, we will conduct floating space vehicle trajectory data matching with reference to the specific electronic map. According to the matching information, combining the spatial data mining model and artificial intelligence, the lane path will be extracted by the vehicle tracks, which will lay the foundation for the acquisition of high-precision maps.

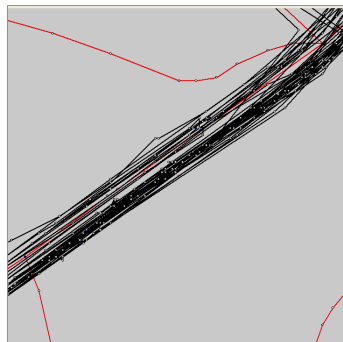


Figure 8. Vehicle track.

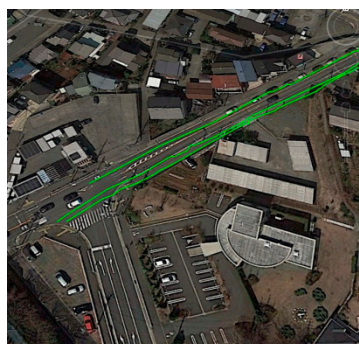


Figure 9. Lane path.

Author Contributions: L.Z. originated the idea, conducted the experiments, and wrote the manuscript. M.S. performed the experiments. Y.L. and X.S. designed and performed the experiments. C.Y., F.H., and M.Y. designed the presentation, commented on the content, and revised the manuscript.

Acknowledgments: The authors thank the anonymous reviewers and members of the editorial team for the comments and contributions. This work is supported by the National Natural Science Foundation of China (NSFC) under project number 41771486.

Conflicts of Interest: The authors declare no conflict of interest.

Nomenclature

ITS	Intelligent Transportation System
GPS	Global Positioning System
OEM	Original Equipment Manufacturer
GPRS	General Packet Radio Service
VICS	Vehicle Information and Communication System
AATCC	Automobile Association Traffic Control Centre
FHA	Federal Highway Administration
ITA	Illinois Transportation Authority
KRSA	Korean Road Safety Association
GSCTI	German Space Center Transportation Institute
KORTIC	Korea Road Traffic Information Center
CCTV	Closed-Circuit Television
POIs	Point of Interests

References

- China Car Survey and Market Prospect Forecast Report. Available online: <http://www.askci.com> (accessed on 31 December 2017).
- Geng, X.; Wang, S.; Ji, J. Fast road-matching algorithm of probe-car. *J. Water Resour. Archit. Eng.* **2013**, *11*, 122–125.
- Hao, Y.; Wu, G.; Zhou, S. A multi-vehicle speed fusion algorithm based on probe vehicle data. *J. Transp. Inf. Saf.* **2012**, *30*, 56–61.
- Zhu, T.; Guo, S. A study on floating car based information processing technology. *J. Image Graph.* **2009**, *14*, 1230–1237.
- Boyce, D.E.; Kirson, A.; Schofer, J.L. Design and implementation of advance: The Illinois dynamic navigation and route guidance demonstration program. In Proceedings of the Vehicle Navigation and Information Systems Conference, Dearborn, MI, USA, 20–23 October 1991; pp. 415–426.
- Cowan, K.; Gates, G. Floating vehicle data system—Realization of a commercial system. In Proceedings of the 11th International Conference on Road Transport Information and Control, London, UK, 19–21 March 2002; pp. 187–189.
- Pang, H. Research on Key Technologies of Urban Dynamic Traffic Guidance Systems Based on FCD. Master's Thesis, University of Science and Technology of China, Hefei, China, 2009.
- Wang, Y. Research on the Key Technology of Large-Scale Strategic Traffic Coordination & Control System. Ph.D. Thesis, Jilin University, Changchun, China, 2009.
- Vehicle Information and Communication System Center. *Introduction of VICS Ver. 2010*; Vehicle Information and Communication System Center: Tokyo, Japan, 2010.
- Han, W.; Choi, K.K. An implementation of a Korea traffic information center over metropolitan Seoul region. In Proceedings of the Mobility for Everyone World Congress on Intelligent Transport Systems, Berlin, Germany, 21–24 October 1997; pp. 21–24.
- Schafer, R.; Thiessenhusen, K.; Wagner, P. A traffic information system by means of real-time floating-car data. In Proceedings of the ITS World Congress, Chicago, IL, USA, 14–17 October 2002; pp. 1–8.
- Jan, F.E.; Stephan, M.; Stefan, E.; Dirk, C.M. Data chain management for planning in city logistics. *Int. J. Data Min. Model. Manag.* **2009**, *1*, 335–356.
- System and Method for Realtime Community Information Exchange. Available online: <https://www.cbinsights.com/company/waze-patents> (accessed on 1 March 2016).
- Li, X.; Meng, Q. The applications of GPS technology in the Real-Time detection of city traffic condition to GPS. *J. Ocean Univ. Qingdao* **2002**, *32*, 475–481.
- Dong, J.; Wu, J.; Guo, J. Assessment of road network based on GPS/GIS data: a practice in Beijing. *City Plan. Rev.* **2005**, *29*, 70–74.
- Zhang, Z.; Lin, X.; Lin, S. Traffic parameter features in traffic incidents based on probe-car data. *J. Transp. Inf. Saf.* **2011**, *29*, 94–98.

17. Xin, F.; Chen, X.; Lin, H. Research on time space distribution characteristics of probe-car data in road network. *China J. Highw. Transp.* **2008**, *4*, 105–110.
18. Li, Q.; Yin, J.; He, F. A coverage rate model of GPS probe-car for road networks. *Geomat. Inf. Sci. Wuhan Univ.* **2009**, *34*, 715–718.
19. Zhang, C. Research on the Traffic Data Collection and Data Processing Theory and Method Based on Probe-Car. Ph.D. Thesis, Tongji University, Shanghai, China, 2007.
20. Weng, J.; Zhou, X.; Zhai, Y. Applications of probe-car data in urban macroscopic traffic character study. *J. Wuhan Univ. Technol. (Transp. Sci. Eng.)* **2008**, *32*, 806–809.
21. Guo, J.; Wen, H.; Chen, F. Function analysis and application design of floating car system. *J. Transp. Syst. Eng. Inf. Technol.* **2007**, *7*. [[CrossRef](#)]
22. Yamane, K.; Endo, Y.; Fujiwara, J.; Kumagai, M. Estimation of statistical traffic data for navigation system. *Int. J. ITS Res.* **2004**, *2*, 1–9.
23. Gou, X.; Zuo, X.; Zhang, Y. Application of digital velocity model in urban traffic dynamics analysis based on probe-car. *Sci. Technol. Eng.* **2013**, *13*, 3172–3177.
24. Xie, J. Design of Vehicle Multimedia Navigation System Based on AU1200. Master's Thesis, Zhejiang University, Hangzhou, China, 2008.
25. Schroedl, S.; Wagstaff, K.; Rogers, S.; Langley, P.; Wilson, C. Mining GPS Traces for Map Refinement. *Data Min. Knowl. Discov.* **2004**, *9*, 59–87. [[CrossRef](#)]
26. Zheng, K.; Song, X.; Zhu, D. A New Method of Trajectory Restoration at Intersection. *TELKOMNIKA* **2015**, *13*, 563–570. [[CrossRef](#)]
27. Dean, J.; Ghemawat, S. Map/Reduce advantages over parallel databases include storage-system independence and fine-grain fault tolerance for large jobs. *Commun. ACM* **2010**, *53*, 72–77. [[CrossRef](#)]
28. Dean, J.; Ghemawat, S. Map/Reduce: Simplified data processing on large clusters. *Commun. ACM* **2008**, *51*, 107–113. [[CrossRef](#)]
29. Afrati, F.N.; Ullman, J.D. Optimizing joins in a map-reduce environment. *IEEE Trans. Knowl. Data Eng.* **2011**, *23*, 1282–1298. [[CrossRef](#)]
30. Biswapesh, C.; Liang, L.T. A SQL implementation on the MapReduce framework. In Proceedings of the VLDB Endowment, Athens, Greece, 12–16 June 2011; Volume 4, pp. 1318–1327.
31. Fang, S.; Zhou, J.; Zhang, M. Research of improvement selection algorithm in cloud-computing web data mining based on Map/Reduce. *Appl. Res. Comput.* **2013**, *30*, 377–379.
32. Dean, J. Experiences with MapReduce: An abstraction for Large-scale computation. In Proceedings of the IEEE 15th International Conference on Parallel Architectures and Compilation Techniques, Seattle, WA, USA, 16–20 September 2006.
33. Anand, R.; Jeffrey, D.U.; Wang, B. *Key Technologies and Application Research of Cloud Computing*; People's Posts and Telecommunications Press: Beijing, China, 2012.

