IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 10, 2019, accepted June 28, 2019, date of publication July 5, 2019, date of current version August 12, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2926824

Search Model of the Region With the Maximum Coverage Value Based on Trajectory Data

ZHONGWEI YUE[1,2], JINGWEI ZHANG[1], RU CHEN[1], YA ZHOU[1], AND QING YANG[3]
[1]Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin 541004, China
[2]School of Data Science and Engineering, East China Normal University, Shanghai 200062, China
[3]Guangxi Key Laboratory of Automatic Measurement Technology and Instrument, Guilin University of Electronic Technology, Guilin 541004, China

Corresponding author: Qing Yang (gtyqing@hotmail.com)

ABSTRACT The wide application of mobile terminals has given rise to a large number of trajectory data. These data record spatio-temporal mobility of mobile objects and have important value for urban planning, traffic congestion detection, and other applications. In view of the important reference value of the trajectory data for commercial location selection, this paper proposes the search model of the region with the maximum coverage value based on trajectory data, which aims to maximize the sum of weights of sampling points covered by circular regions. The model considers the difference between the sampling points of different users and the interaction between the sampling points of the same user so as to adapt to different application scenarios such as location selection of signal stations and location selection of shopping malls. In order to further improve computing performance, this paper proposes two distributed schemes for this model. Finally, the extensive experiments on three real data sets demonstrated that the distributed schemes outperformed the centralized scheme and the application scenarios of the two schemes are summarized based on the experimental results.

INDEX TERMS Trajectory data, maximum coverage value, application scenarios, distributed schemes.

I. INTRODUCTION

The rapid development of mobile Internet and the wide application of mobile terminals have given rise to a large number of trajectory data. These data record spatio-temporal mobility of mobile objects and contain the behavior information and interaction information of the individual or group. It has important value for urban planning, commercial location selection, traffic congestion detection and other applications. For example, [1] optimizes traffic routes using taxi trajectory data; [2] uses shared bicycle trajectory data to plan bicycle lanes; and [3] uses human trajectory data to provide solutions for urban planning.

At present, the research on trajectory data can be roughly divided into the following four directions: preprocessing of trajectory data, indexing and retrieval of trajectory data, mining of trajectory data, and privacy protection of trajectory data [4], [5]. Among them, the mining of trajectory data is a hot direction of trajectory data research, which aims to find valuable knowledge and models from trajectory data. Trajectory clustering refers to the collection of trajectory data through a certain rule to obtain valuable information. People usually use the method of trajectory clustering to perform related trajectory data mining work, such as frequent visits to popular areas [6], frequent passing routes [7], and high-impact locations(shopping malls, restaurants, tourist attractions)[8]. In this paper, the region which have maximum coverage value is obtained by clustering.

This paper proposes a novel problem based on trajectory data—the region with the maximum coverage value. The region with maximum coverage value requires that the sum of the weights of the sample points covered by the circle whose center is in this region is the maximum. A special case is to make the circle cover the most sampling points for maximizing weight value. For the region search model proposed in this paper, there are multiple real-life application scenarios. For example, the location of the signal station should be chosen to cover more of the sampling points within its circular influence range to maximize the effect of the signal station; the location of the mall should be chosen to cover more objects of sampling points within its circular influence so that more users go shopping. There are many

The associate editor coordinating the review of this manuscript and approving it for publication was Xiao Liu.

regions that meet the maximum coverage value. In order to give the user more choices; the region with the maximum coverage value proposed in this paper refers to all regions that meet the maximum coverage value requirement.

The search model of the region with the maximum coverage value has the following challenges: (1) the shape of the region with the maximum coverage value is irregular, which increases the difficulty of this search model; (2) different application scenarios have different requirements for the region with the maximum coverage value; for example, signal stations need to cover more sampling points, and shopping malls need to cover more objects of sampling points; (3) in order to cope with the challenges of storage space and computing performance caused by the growing trajectory data, how to design a distributed solution of the model is important.

In summary, our contributions are as follows:

- We propose the search model of the region with the maximum coverage value based on trajectory data, which can find all regions that meet the designated requirements. The corresponding proofs are given to prove that the region searched by the model is the region with the maximum coverage value.
- In order to apply the search model to different application scenarios, the model sets two adjustable parameters: the weight value of the sampling point and the influence value between the sampling points of the same moving object. At the same time, the parameter values of some application scenarios are given.
- In order to cope with the challenges of storage space and computing performance caused by the growing trajectory data, we designed two distributed schemes of the search model. The efficiency of the two distributed schemes is verified by three real data sets, and the application scenarios of the two distributed schemes are summarized based on the experimental results.

The rest of this paper is organized as follows. Section 2 of this paper briefly describes the related work of trajectory clustering. Relevant definitions are introduced in section 3. Section 4 introduces in detail the search model of the region with the maximum coverage value proposed in this paper. In section 5, two distributed schemes of the search model are described in detail. The related experiments are presented in Section 6. The last section is the conclusion.

II. RELATED WORK

The search model proposed in this paper belongs to trajectory clustering, so the research related to trajectory clustering is introduced here. According to whether the trajectory data are directly used for clustering, the trajectory clustering is introduced from the following two aspects: trajectory direct clustering and trajectory indirect clustering.

Trajectory direct clustering refers to clustering directly based on the information of trajectory data. The location information is basic information of the trajectory data, and clustering the trajectories by the location information of the trajectory data is an effective strategy. For example, [9] finds interesting places by performing density clustering on trajectory data; [10] uses trajectory clustering to discover the convey which is a set of objects moving together in a continuous period of time; [11] proposed a framework called clustering and aggregating clues of trajectories to find out the routes that users frequently pass. The trajectory data not only contain the location information of the moving object but also the corresponding time information, so time information is also a frequently considered factor when employing trajectory clustering. [12] pointed out that many studies have used the time information of trajectory data to find linear patterns in the clustering process. [13] and [14] considered periodic patterns and time intervals respectively in trajectory clustering.

Trajectory indirect clustering is to convert the trajectory data into other information (such as road network information, semantic information) and then perform clustering. In the process of trajectory data, the semantic information (such as shopping malls, gas stations, tourist attractions) represented by the trajectory data is often the object of people's concern and has important applications in real life scenes. [15], [16] and [17] are to perform similar user mining, user's next destination prediction, and hot spot region identification by clustering the semantic information of the trajectory data, respectively. Moving objects such as cars and bicycles usually travel on roads, so converting the trajectory data generated by moving objects into corresponding road network coordinates is another research direction for trajectory data clustering. Reference [18] calculates the most popular route from the point of departure to the destination by mapping the trajectory data to the road network. Reference [19] conducts traffic anomaly detection through the road network information of the trajectory data.

III. RELATED DEFINITION

In order to clearly introduce the search model of the region with the maximum coverage value, the related definitions and application scenarios are given in detail below. The frequently used notations are listed in Table 1.

TABLE 1. Notations.

Notation	Description
P	sampling information of a moving object at a certain time, where the P is called the sampling point
T	a sequence of sampling points of a moving object in chronological order
R_t	a set of T
R_p	a set of sample points
$Number(R_p)$	the number of sampling points in the R_p
R_pK	R_p consisting of sampling points of moving object K
f	the influence parameters between sampling points belonging to the same moving object, $f > 0$
$Circle(c, r)$	a circle with c as the center and r as the radius
$Weight(Q)$	the weight of Q, where Q represents the set of sampling points
$Rect((x, y) \rightarrow (xx, yy))$	A rectangle, where (x, y) is the coordinate point of the lower left corner of the rectangle, and (xx, yy) is the coordinate point of the upper right corner of the rectangle

Definition 1 (Sampling point): The sampling point is the sampling information of the moving object at a certain time, which generally includes the location information, sampling time, the identifier of the moving object, the speed of the moving object, etc. The sampling point is denoted as P, where $P.t$ represents the sampling time of the sampling point and $P.ID$ represents the identifier of the moving object of the sampling point.

Definition 2 (Trajectory): A trajectory is a sequence of sampling points of a moving object in the order of sampling time. It is recorded as $T = \{P_1, P_2, P_3, P_4, \ldots\}$, where if $i > j$ then $P_i.t < P_j.t$ and $P_i.ID = P_j.ID$ for any i and j. $T.P$ is recorded as the sampling point of the trajectory.

Definition 3 (Trajectory set): The trajectory set is a set of trajectories, which is denoted as $R_t = \{T_1, T_2, T_3, \ldots\}$.

Definition 4 (Sample point set): The sample point set is a set of sample points, which is denoted as $R_p = \{P_1, P_2, P_3, P_4, \ldots\}$. If the sample points of the R_p belong to the same moving object whose identifier is K, the sample point set is recorded as R_pK.

Definition 5 (Influence range): The influence range refers to a circular region with a certain coordinate point c as the center and a length r as the radius, which is denoted as $Circle(c, r)$. The reason for defining the influence range is to correspond to the specific scene in life. For example, the influence range of the signal station is represented by the circle centered on the location of the signal tower, and the influence range of the rescue station is represented by the circle centered on the location of the rescue station.

Definition 6 (Sample point weight): The sampling point weight is used to indicate the importance of the sampling point in a specific scene and is recorded as $Weight(p)$. Since the weight of the sampling point is determined by the moving object, the weights of the sampling points belonging to the same moving object are equal, that is, if $P_i.ID = P_j.ID$, then $Weight(Pi) = Weight(Pj)$.

Definition 7 (Influence parameter): The influence parameter indicates the interaction between the sampling points of the same moving object, denoted as f; the weight value of the sample point set R_pK is recorded as $Weight(R_pK)$. If $R_pK = \{P_1, P_2, \ldots, P_n\}$, then

$$Weight(R_pK) = Weight(P_1) * f^0 + Weight(P_2) * f^1 + \ldots + Weight(P_n) * f^{n-1} \quad (1)$$

From (1), it can be concluded that $f = 1$ means that there is no influence between the sampling points; $f < 1$ means that the influences between the sampling points are weakened by each other; and $f > 1$ means that the influences between the sampling points are mutually enhanced. Suppose the sample point set R_p contains sampling points of n moving objects, which is denoted as $Rp = R_p1 \cup R_p2 \cup \ldots \cup R_pn$($R_pi$ represents the trajectory generated by the moving object whose identifier is i), and the weight of the sampling point set R_p is:

$$Weight(R_p) = Weight(R_p1) + Weight(R_p2) + \ldots + Weight(R_pn) \quad (2)$$

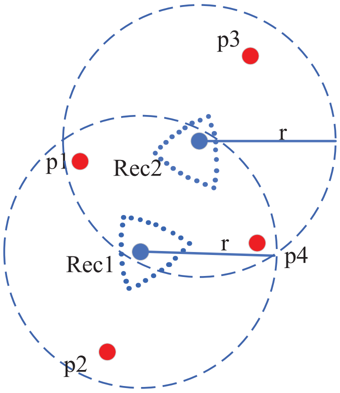


FIGURE 1. The region with the maximum coverage value.

Definition 8 (Point with maximum coverage value): In the trajectory set R_t, a coordinate point c is selected. If the set of sampling points composed of sampling points in $Circle(c, r)$ gets the maximum $Weight(Rp)$; the coordinate point c is called the point with the maximum coverage value.

Definition 9 (Region with the maximum coverage value): The region with the maximum coverage value is the region consisting of all the points with the maximum coverage value and is recorded as Rec.

In order to more specifically describe the search model of region with the maximum coverage value, the application scenario is illustrated based on Figure 1, in which $Rec1$ and $Rec2$ represent the region, and P_1, P_2, P_3, and P_4 represent sampling points.

Application Scenario 1. In Figure 1, a signal station is established, which requires more sampling points to be covered in the influence range of radius r. The position of the signal station is selected at any point of $Rec1$ and $Rec2$ to satisfy requirement of the most covered sampling points. The model in this paper only needs to set the radius $r, f = 1$ and the weight of all sampling points is 1; the $Rec1$ and $Rec2$ which are the region with the maximum coverage value can be obtained.

Application Scenario 2. In Figure 1, a shopping mall is established, and the radius of its influence range is r. P_1, P_2, P_3, and $P4$ belong to 4 different moving objects respectively. The moving objects of P_1 and P_2 often go shopping at the mall, while the moving objects of P3 and P4 rarely go to the mall to shop. Then any point in the $Rec1$ is the best location for the mall. The search model in this paper only needs to set the radius $r, f = 1$, $Weight(P1) = Weight(P2) = 2$ and $Weight(P3) = Weight(P4) = 1$; the $Rec1$ which is the region with the maximum coverage value can be obtained.

Application Scenario 3. In Figure 1, a rescue station covering more moving objects is created, and the radius of its influence range is r. P_1, P_2, and P_4 belong to the same moving object, and P_3 belongs to another moving object. Then any point in the $Rec2$ is the best location for the rescue station to be located. The model in this paper only needs to set the radius $r, f < 1/4$ and $Weight(P1) = Weight(P2) = Weight(P3) = Weight(P4) = 1$; the $Rec2$ which is the region with the maximum coverage value can be obtained.

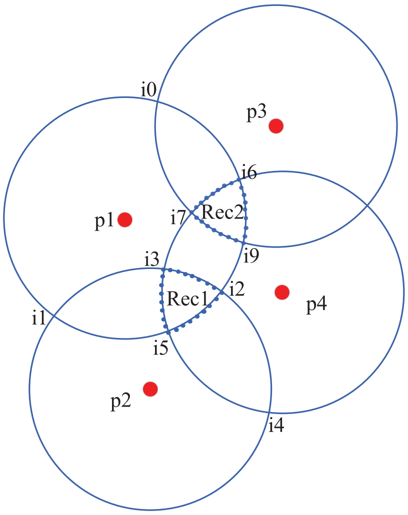


FIGURE 2. The search process of the model.

IV. THE SEARCH MODEL OF THE REGION WITH THE MAXIMUM COVERAGE VALUE AND ITS ANALYSIS

In order to maximize the sum of weights of sampling points covered by circular regions, the search model of the region with the maximum coverage value based on trajectory data is proposed. The main idea of this model is to find the region with the largest number of overlaps of circles centered on sampling points. The model is described as the following steps:

(1) With each sampling point as the center and r as the radius, the intersecting arc is obtained.

(2) Calculate the weight value of the intersecting arc on each circle. The weight value of the intersecting arc is the weight value of the set of sampling points that meet the condition that the circle whose center is the sample point and radius equals r can completely cover the intersecting arc. Therefore, the intersecting arc with maximum weight value on the circle associated with each sample point can be obtained. If there is no intersecting arc, the whole circle is the arc with maximum weight value.

(3) The arc with the maximum weight value on each circle is compared for weight values, and the resulting maximum weight value arc is the edge of the region with the maximum coverage value.

In order to express the search model more clearly, we will carry out the relevant description of the above steps based on Fig. 2. In Figure 2, P_1, P_2, P_3, and P_4 represent 4 sampling points. P_1 and P_2 belong to the same moving object and their weight values are all 1; P_3 belongs to a moving object and its weight value is 3; P_4 belongs to a moving object and its weight value is 2. The influence parameter $f = 0.6$ is set. The calculation process of the search model is as follows:

(1) Draw four circles with the radius of r as the center of P_1, P_2, P_3, and P_4, and obtain the intersection points $I_0, I_1, I_2, I_3, I_4, I_5, I_6, I_7, I_8$, and I_9.

(2) The intersection arc of the circle centered at P_1 and other circles is $I_1 \rightarrow I_5, I_5 \rightarrow I_2, I_2 \rightarrow I_9, I_9 \rightarrow I_6, I_6 \rightarrow I_0$. The set of sampling points whose circle can completely cover the intersecting arc $I1 \rightarrow I5$ is $\{P_1, P_2\}$. P_1 and P_2 belong to the same moving object, so the weight of $I1 \rightarrow I5$ is $Weight(P_1) + 0.6 * Weight(P_2) = 1.6$. For the intersecting arc $I_5 \rightarrow I_2$, the set of sampling points whose circle can completely cover this arc is $\{P_1, P_2, P_4\}$. P_1 and P_2 belong to the same moving object and P_4 belongs to another moving object, so the weight of $I_5 \rightarrow I_2$ is $Weight(P1) + 0.6 * Weight(P_2) + Weight(P_4) = 3.6$. For the intersecting arc $I_2 \rightarrow I_9$, the set of sampling points whose circle can completely cover this arc is $\{P_1, P_4\}$, so the weight of $I_2 \rightarrow I_9$ is $Weight(P_1) + Weight(P_4) = 3$. The set of sampling points whose circle can completely cover the intersecting arc $I_9 \rightarrow I_6$ is $\{P_1, P_4, P_3\}$; then the weight of $I_9 \rightarrow I_6$ is $Weight(P_1) + Weight(P_4) + Weight(P_3) = 6$. The set of sampling points whose circle can completely cover the intersecting arc $I_6 \rightarrow I_0$ is $\{P_1, P_3\}$, so the weight of $I_6 \rightarrow I_0$ is $Weight(P_1) + Weight(P_3) = 4$. Then, the arc with maximum weight value of the circle centered on P_1 is $I_9 \rightarrow I_6$ and its weight value is 6. Similarly, the arc with maximum weight value of the circle centered on P_2 is $I_2 \rightarrow I_3$ and its weight value is 3.6; the arc with maximum weight value of the circle which is centered on P_3 is $I7 \rightarrow I9$ and its weight value is 6; The arc with the maximum weight value of the circle which is centered on P_4 is $I6 \rightarrow I7$, and its weight value is 6.

(3) By comparing the four arcs with the maximum weight values obtained above, we can get that $I_6 \rightarrow I_7, I_7 \rightarrow I_9, I_9 \rightarrow I_6$ are maximum weight value arcs on all circle, so the $Rec2$ surrounded by these maximum weight value arcs is the region with maximum coverage value.

To prove that the region obtained by the search model is the region with the maximum coverage value. There will always be a boundary for the region with the maximum coverage value, and if the circle with radius r is centered at any point on this boundary, the coverage value will reach the maximum value. However, if a circle with radius r is centered at any point outside the boundary, the coverage value will not reach its maximum due to less coverage of some sampling points. Therefore the distance between any point on the boundary of the region with maximum coverage value and some sampling point P is radius r. That is to say, the boundary of all regions with the maximum coverage value can be obtained by searching the intersecting arcs of the circle whose center is the sample point and radium equals r, so the region with the maximum coverage value can be obtained.

The implementation of the search model of the region with the maximum coverage value is shown in Algorithm 1. The time complexity of the Algorithm 1 is mainly determined by the Line 4 and Line 19, and the time complexity of the Line 4 is $O(n)$, and the time complexity of the Line 19 is $O(n \log n)$. Since the Line 19 is nested in the Line 4, the time complexity of the search model of the region with the maximum coverage value is $O(n^2 \log n)$. In this algorithm, Line 6–18 is to calculate the intersection arc of a circle with other circles; Line 4–18 is to calculate the intersection arc between all circles corresponding to the first step of the model. Line 19–40 is to calculate the intersection arcs that is the maximum weight value arcs of the current sample point set, which is equivalent to the combination of the second step and the third step of the model. For each iteration, Line 19–40 will get the region with maximum coverage value of the current traversed sample point and the trajectory set composed of the previously traversed sample points, so the end of the iteration will get the region with the maximum coverage value of all sample point sets.

It is assumed that the sampling point of the s moving objects constitute the sampling point set R_p, and the weight of each sampling point is W. There are two sampling point sets R_p1 and R_p2 in R_p ($R_p1 \subset R_p, R_p2 \subset R_p$). R_p1 contains the trajectory set of n moving objects, which are recorded as $R_p1 = \{R_{p1}1, R_{p1}2, \ldots, R_{p1}n\}$; R_p2 contains the trajectory set of m moving objects, which are recorded as $R_p2 = \{R_{p2}1, R_{p2}2, \ldots, R_{p2}m\}$. Then the influence parameter f is discussed as follows:

(1)$f = 1$:

$$\begin{aligned} Weight(R_{p1}) &= Weight(R_{p1}1) + Weight(R_{p1}2) \\ &\quad + \ldots + Weight(R_{p1}n) \\ &= (Number(R_{p1}1) + Number(R_{p1}2) \\ &\quad + \ldots + Number(R_{p1}n)) * W \\ &= Number(R_{p1}) * W \qquad (3) \\ Weight(R_{p2}) &= Weight(R_{p2}1) + Weight(R_{p2}2) \\ &\quad + \ldots + Weight(R_{p2}m) \\ &= (Number(R_{p2}1) + Number(R_{p2}2) \\ &\quad + \ldots + Number(R_{p2}m)) * W \\ &= Number(R_{p2}) * W \qquad (4) \end{aligned}$$

The weight values of R_{p1} and R_{p2} in (3) and (4) are determined by the number of sample points trajectory set contain. That is to say, the more sampling points are included, the greater the weight value is. Therefore, when $f = 1$, the search model of the region with the maximum coverage value can be used to mine the region covering the most sampling points.

(2)$f \neq 1$:

In order to make the search model of region with maximum coverage value mine the region covering the most users; suppose $s \geq n \geq m \geq 0$ and $v = Max((Number(R_{p2}1), Number(R_{p2}2), \ldots, Number(R_{p2}m))$. Then the maximum weight value of $Weight(R_{p2})$ is that the number of sampling points of m moving objects is v, that is,

$$\begin{aligned} Weight(R_{p2}) \leq m * W + f * m * W + f^2 * m * W \\ + \ldots + f^{(v-1)} * m * W \qquad (5) \end{aligned}$$

The minimum weight value of R_{p1} is that there is only one sampling point for each moving object, that is,

$$Weight(R_{p1}) = n * W \qquad (6)$$

Algorithm 1 The Search Model of the Region With the Maximum Coverage Value

```
Input: influence range radius: $r$, influence parameter: $f$,
       sampling point weight value: $weight$
Output: the set of the starting point of the arc with
        maximum weight value: $M1$, the set of the end
        points of the arc with maximum weight value:
        $M2$, the set of the center of intersecting arc:
        $M3$, the maximum coverage value: $Best_{max}$
1  $Best_{max} = 0$;
2  $N = \{\}, M1 = \{\}, M2 = \{\}, M3 = \{\}$;
3  //The trajectory data set is represented by $R_p$;
4  for each cell $i \in R_p$ do
5      $N = \{\}$;
6      for each cell $j \in R_p$ do
7          if $Distance(i, j) \leq 2 * r$ then
8              //The $s$ and $e$ represent the start point and
               end point of the intersecting arc centered on
               $i$, respectively;
9              $s.value = the\ polar\ angle\ of\ s$;
10             $e.value = the\ polar\ angle\ of\ e$;
11             $s.flag = true$;
12             $e.flag = flase$;
13             $count = 0$;
14             $Map = (j.user \rightarrow (j.weight, count))$;
15             $s.user = j.user$;
16             $e.user = j.user$;
17             $N \leftarrow s$;
18             $N \leftarrow e$;
19     $Sort(N)$ //Sort by polar angle;
20     $Cnt = 0$;
21     for each cell $k \in N$ do
22         $Mid = Map.get(k.user)$;
23         if $k.flag == true$ then
24             $Cnt = Cnt + Mid.weight * f^{Mid.count}$;
25             $Mid.count = Mid.count + 1$;
26             if $Cnt > Best_{max}$ then
27                 $M1 = \{\}$;
28                 $M2 = \{\}$;
29                 $M3 = \{\}$;
30                 $M1 \leftarrow k$;
31                 $M3 \leftarrow i$;
32             if $Cnt == Best_{max}$ then
33                 $M1 \leftarrow k$;
34                 $M3 \leftarrow i$;
35         if $k.flag == false$ then
36             if $Cnt == Best_{max}$ then
37                 $M2 \leftarrow k$;
38             $Cnt = Cnt - mid.weight * f^{Mid.count}$;
39             $Mid.count = Mid.count - 1$;
40         $Best_{max} = Max(Cnt, Best_{max})$;
41 return $(M1, M2, M3, Best_{max})$;
```

If the search model of the region with the maximum coverage value is required to cover more users, only the following conditions need to be satisfied:

$$\begin{aligned} Weight(R_{p1}) &> Weight(R_{p2}) \\ &\Leftrightarrow n > m*(1-f^{v})/(1-f) \end{aligned} \tag{7}$$

$f < 1:$

$$\begin{aligned} n &> m*(1-f^{v})/(1-f) \\ &\Leftrightarrow (n-m) > nf - mf^{v} \\ &\Leftrightarrow 1 > nf - mf^{v} \end{aligned} \tag{8}$$

Since the value of mf^{v} is greater than or equal to 0, the (8) is equivalent to $nf < 1$, then $f < 1/n$. Since $n \leq s$, it is only necessary to set $f < 1/s$ to find the region whose circle covers the most users in the sampling point set R_p. $f > 1:$

$$\begin{aligned} n &> m*(1-f^{v})/(1-f) \\ &\Leftrightarrow (n-m) < nf - mf^{v} \\ &\Leftrightarrow 1 < nf - mf^{v} \end{aligned} \tag{9}$$

Since the value of v may be infinite, the value of mf^{v} may be infinite, and $f > 1$ can't mine the region whose circle covers the most users. In summary, when $f < 1/s$, the search model of the region with the maximum coverage can be used to mine the region covering the most users.

V. THE DISTRIBUTED SCHEMES OF THE SEARCH MODEL

With the continuous growth of trajectory data, centralized processing has been difficult to meet the requirements of storage space and computing performance. For this reason, two different distributed schemes are proposed for this search model. The architecture of a distributed system is shown in Figure 3. The architecture of a centralized system is shown in Figure 4. From the two figures, it can be concluded that the centralized system processes all the data by one single machine; while the distributed system divides the data into slave nodes through the master node, and then the data is processed by the slave node. The main idea of a distributed solution is:

(1) Divide all the data into several parts according to certain rules, and then send these parts to the slave node.
(2) Calculate the arc with maximum weight of the sampling point set in each slave node.
(3) Send the arc with the maximum weight value on all slave nodes to the master node.
(4) Finally, the master node obtains the arc with the maximum weight value of all the sampling point sets according to the arc with the maximum weight value on all the slave nodes, that is, the region of the maximum coverage value is obtained.

Steps 3 and 4 above can be easily implemented. Therefore, the distributed scheme here mainly describes how to divide the trajectory data for the first step and how to process trajectory data for the slave node in the second step.

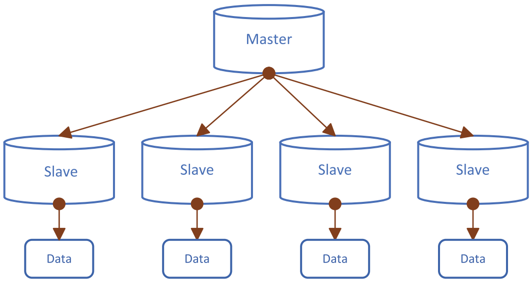


FIGURE 3. The architecture of a distributed system.

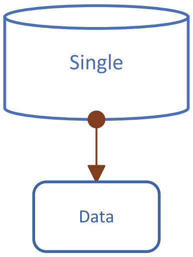


FIGURE 4. The architecture of a centralized system.

A. DISTRIBUTED SCHEME I

Assume that there are n slave nodes, numbered 0, 1, 2, 3, …, $n-1$. The division scheme of the trajectory data is as follows:

(1) Traverse all sampling points.
(2) Assign the j-th traversed sample point to the slave node numbered $j\%n$.
(3) Assign all the sampling points to each slave node again.

The step 2 above causes each slave node to obtain a set of sample points, which is represented by R_{p1}. The step 3 above causes each slave node to obtain a set of sample points, which is represented by R_{p2}. The processing algorithm of the slave node for its data set is through the modification of the Algorithm 1: replace R_p in the Line 4 of the Algorithm 1 with R_{p1}, and replace R_p in the Line 6 of the Algorithm 1 with R_{p2}.

In distributed scheme I, the sampling point of the data set in the Line 4 and the sampling point of the data set in the Line 6 are rarely intersected. However, the distance between the sampling points in Line 4 and all the sampling points in Line 6 is calculated. Therefore, the distributed scheme II is proposed to reduce the number of sampling points in Line 6.

B. DISTRIBUTED SCHEME II

Suppose there are n slave nodes whose numbers are 0, 1, 2, 3, …, $n-1$. The rectangle is denoted as $Rect((x, y \rightarrow (xx, yy))$; then the division scheme of the trajectory data is as follows:

(1) The minimum outer rectangle composed of sampling points is equally divided into n small rectangles, which are numbered 0, 1, 2, …, $n-1$, respectively.
(2) Send the sampling points in the small rectangle to the corresponding numbered slave nodes.
(3) The condition that the two circles of radius r generate intersecting arcs is that the distance between the two centers is not more than $2r$, so the center of the circle that intersects the circle centered on the sampling point

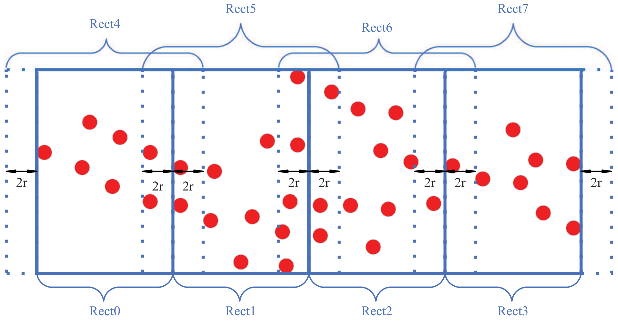


FIGURE 5. **The division scheme of trajectory data set.**

in the rectangle $Rect((x, y) \to (xx, yy))$ must be in the rectangle $Rect((x-2r, y) \to (xx+2r, yy))$. Therefore, the sampling point in the rectangle $Rect((x-2r, y) \to (xx+2r, yy))$ is sent to the salve node where the sampling point in $Rect((x, y) \to (xx, yy))$ is located.

In order to show the division scheme more vividly, this paper uses Figure 5 as an example to explain the division process ($n = 4$):

(1) The minimum outer rectangle of the sampling point in Fig. 5 is equally divided into four small rectangles, which are represented as $Rect0$, $Rect1$, $Rect2$, and $Rect3$, respectively.

(2) Send the sampling points in $Rect0$, $Rect1$, $Rect2$, and $Rect3$ to $Slave0$, $Slave1$, $Slave2$, and $Slave3$ respectively.

(3) Send the sampling points in $Rect4$, $Rect5$, $Rect6$, and $Rect7$ to $Slave0$, $Slave1$, $Slave2$, and $Slave3$ respectively.

The step 2 above causes each slave node to obtain a set of sample points, which is represented by R_{p1}. In the step 3 above, each slave node obtains a set of sample points, and this set is represented by R_{p2}. The processing Algorithm of the slave node for its data set is through the modification of Algorithm 1: replace R_p in the Line 4 of Algorithm 1 with R_{p1}, and replace R_p in the Line 6 of Algorithm 1 with R_{p2}.

VI. EXPERIMENT

A. EXPERIMENTAL ENVIRONMENT AND DATA

All experiments were run on a distributed computing platform Spark cluster with one master node and four slave nodes. Each node is equipped with a 2-core Intel Xeon CPU W3505@2.53GHz, and each node of the Spark cluster has a running memory of 2.6G. Each node in the cluster is connected to a TP-link(TL-SF1024D) switch, and each node is configured with Red Hat 4.4.7-3, Hadoop 2.6.4 and Spark 1.6.0[20]. The centralized experiment is to send all the data of the cluster to the master node, and then the master node processes the data separately. This is equivalent to a centralized system running the search model.The experiment uses three real data sets. The first data sets is the trajectory data set of the Beijing taxi, which contains 21,658,278 sample points. The second data set is the trajectory data set of the Suzhou taxi, which contains 11,741,688 sampling points. The last data set is a trajectory data set provided by Microsoft [21]–[23]. This experiment uses part of the data set, which includes 23,668,828 sample points. In addition, the trajectory data of the three data sets all include the identifier of the moving object, longitude, latitude, and sampling time.

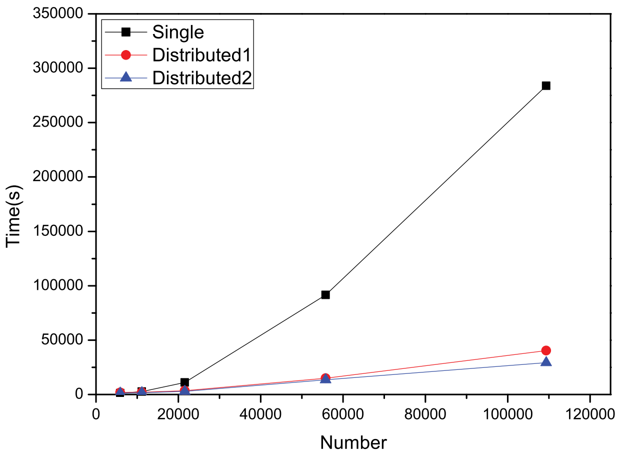


FIGURE 6. **Running time for suzhou trajectory data set.**

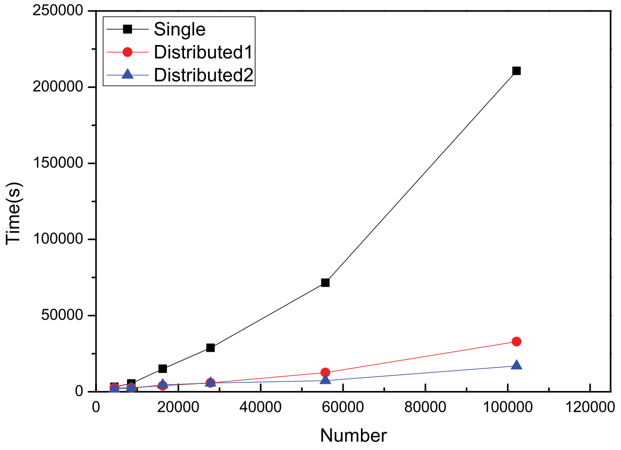


FIGURE 7. **Running time for beijing trajectory data set.**

The experiment first puts these three data sets on Hadoop's HDFS system, then reads the data set into the memory of the Spark cluster, using the time-based trajectory data partitioning technology(the number of partitions is 16 here) [24]. Then the R-tree [25] index is established on the slave nodes. Finally, the trajectory data set is obtained by range query. The location information of sampling points in these trajectory data sets is identified by latitude and longitude. For the convenience of calculation, the longitude and latitude coordinates are transformed into plane rectangular coordinates by Gauss projection. This paper conducts related experiments based on the transformed data sets.

B. EXPERIMENTAL RESULTS AND ANALYSIS

In order to compare the computational efficiency of the centralized scheme and distributed schemes of the search model, three experiments were performed, as shown in Fig. 6, Fig. 7, and Fig. 8. In these figures, *Number* represents the

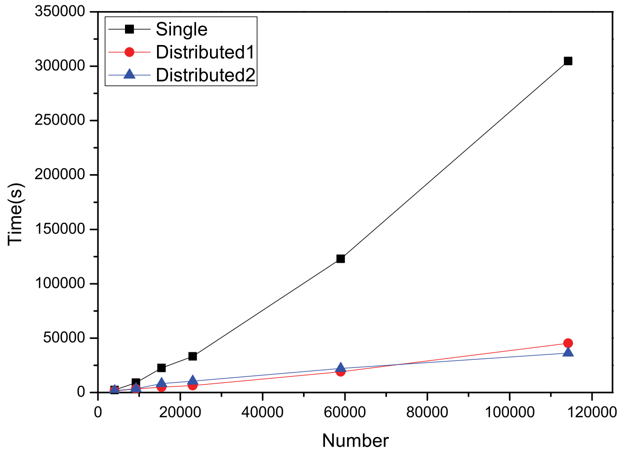


FIGURE 8. Running time for microsoft trajectory data set.

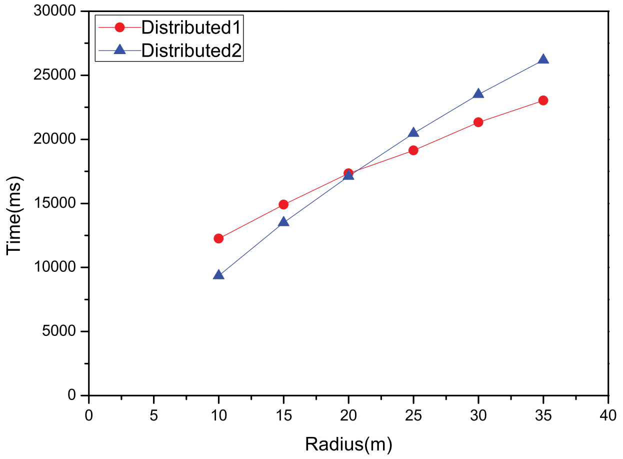


FIGURE 9. Running time comparison of two distributed schemes for suzhou trajectory data set.

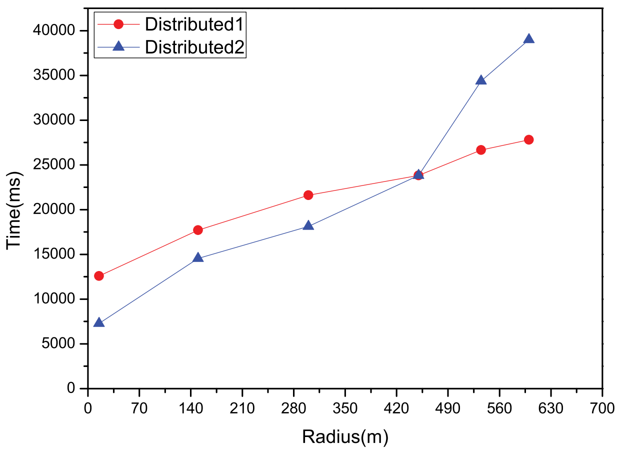


FIGURE 10. Running time comparison of two distributed schemes for beijing trajectory data set.

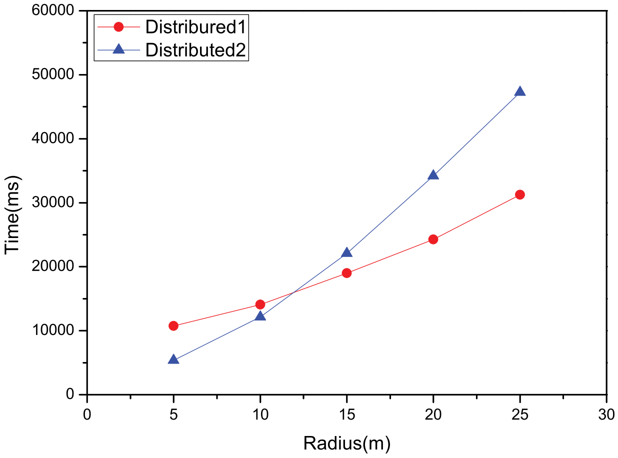


FIGURE 11. Running time comparison of two distributed schemes for microsoft trajectory data set.

number of sampling points calculated by the search model; *Single* represents centralized calculation. *Distributed*1 represents the distributed scheme I; *Distributed*2 represents the distributed scheme II, where $r = 15, f = 1$, and the weight of sampling points is 1.

The running time of centralized search model is the time spent in processing trajectory data. The running time of the distributed search model includes the trajectory data transmission time between the master node and the slave node and the time taken by the cluster to process the trajectory data. Distributed scheme is superior to centralized scheme in processing trajectory data. Therefore, if the number of sampling points is small, the data transmission time between nodes is not negligible relative to the time taken by the cluster to process data, and the distributed processing performance advantages are not apparent. With the increase of the number of sampling points, the data transmission time between nodes can be neglected relative to the time taken by the cluster to process data, so that the distributed processing performance advantages appear. As can be seen from Fig. 6, Fig. 7 and Fig. 8, with the increase in the number of sampling points, the running speed of the distributed scheme is faster and faster than that of the centralized scheme.

The first distributed scheme does not change as the radius changes, while the second distributed scheme changes as the radius changes. Therefore, in order to detect the influence of radius on the second distributed scheme, three experiments were performed, as shown in Figure 9, Figure 10, and Figure 11. In these figures, *Radius* represents the radius of the influence range; *Distributed*1 represents the distributed scheme I; *Distributed*2 represents the distributed scheme II. The number of sampling points in Figures 9, Figures 10 and Figures 11 are 55782, 55763, 58989, respectively.

It can be observed that the distributed efficiency of *Distributed*2 is high when the value of *Radius* is small, and the efficiency of *Distributed*1 is high when the value of *Radius* is large. The reason for this phenomenon is that when the *Radius* is small, the step 3 of the division scheme of *Distributed*2 only transmits a small part of the data set; so the *Distributed*2's slave node processes the trajectory data much faster than the *Distributed*1's slave node. When the *Radius* is large, the step 3 of the division scheme of *Distributed*2 transmits most of the data sets, so the slave node of *Distributed*2 has no time advantage in processing data; at the same time, the division scheme of *Distributed*2 takes more time than that of *Distributed*1. In summary, the distributed scheme I is suitable for the case that the radius is larger than the minimum outer rectangle of the trajectory data set, while the distributed scheme II is suitable for the case that the radius is smaller than the minimum outer rectangle of the trajectory data set.

VII. CONCLUSION

This paper proposes the search model of the region with the maximum coverage values, which has important reference value for commercial location. When the user uses the model, the weight value of sampling point, the radius, and the influence range can be set according to different application scenarios; the flexibility of the search model for different application scenarios is fully demonstrated. Considering the challenges of storage space and computing performance due to the continuous growth of trajectory data, this paper proposes two distributed solutions of the model. Users can choose one of them according to the specific conditions of their application to achieve desired results. In addition, the processing object of the search model proposed in this paper is a coordinate point, so the model can not only adapt to the trajectory data, but also can process other data with coordinates, such as spatial building coordinate data and road network coordinate data.

REFERENCES

[1] J. Yuan, Y. Zheng, X. Xie, and G. Sun, ''T-drive: Enhancing driving directions with taxi drivers' intelligence,'' *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 1, pp. 220–232, Jan. 2013. doi: 10.1109/TKDE.2011.200.

[2] J. Bao, T. He, S. Ruan, Y. Li, and Y. Zheng, ''Planning bike lanes based on sharing-bikes' trajectories,'' in *Proc. KDD*, Halifax, NS, Canada, 2017, pp. 1377–1386.

[3] J. Yuan, Y. Zheng, and X. Xie, ''Discovering regions of different functions in a city using human mobility and POIs,'' in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Beijing, China, 2012, pp. 186–194.

[4] Q. Gao, F. L. Zhang, R. J. Wang, and F. Zhou, ''Trajectory big data: A review of key technologies in data processing,'' *RuanJian Xue Bao/J. Softw.*, vol. 28, no. 4, pp. 959–992, Apr. 2017. doi: 10.13328/j.cnki.jos.005143.

[5] Y. Zheng, ''Trajectory data mining: An overview,'' *ACM Trans. Intell. Syst. Technol.*, vol. 6, no. 3, 2015, Art. no. 29. doi: 10.1145/2743025.

[6] K. E. Liu, J. C. Xiao, Z. M. Ding, and M. S. Li, ''Discovery of hot region in trajectory databases,'' *Ruan Jian Xue Bao/J. Softw.*, vol. 24, no. 8, pp. 1816–1835, 2013. doi: 10.3724/SP.J.1001.2013.04340.

[7] Z. Chen, H. T. Shen, and X. Zhou, ''Discovering popular routes from trajectories,'' in *Proc. 27th Int. Conf. Data Eng.*, Hannover, Germany, 2011, pp. 900–911.

[8] X. Cao, G. Cong, and C. S. Jensen, ''Mining significant semantic locations from GPS data,'' *Proc. VLDB Endowment*, vol. 3, no. 1, pp. 1009–1020, 2010. doi: 10.14778/1920841.1920968.

[9] A. T. Palma, V. Bogorny, B. Kuijpers, and L. O. Alvares, ''A clustering-based approach for discovering interesting places in trajectories,'' in *Proc. ACM Symp. Appl. Comput.*, Fortaleza, Brazil, 2008, pp. 863–868.

[10] H. Jeung, M. L. Yiu, X. Zhou, C. S. Jensen, and H. T. Shen, ''Discovery of convoys in trajectory databases,'' *Proc. PVLDB*, vol. 1, no. 1, pp. 1068–1080, 2008. doi: 10.14778/1453856.1453971.

[11] C.-C. Hung, W.-C. Peng, W.-C. Lee, C. S. Jensen, and H. T. Shen, ''Clustering and aggregating clues of trajectories for mining trajectory patterns and routes,'' *VLDB J.*, vol. 24, no. 2, 2015, pp. 169–192. doi: 10.1007/s00778-011-0262-6.

[12] S. Mitsch, A. Müller, W. Retschitzegger, A. Salfinger, and W. Schwinger, ''A survey on clustering techniques for situation awareness,'' in *Proc. 15th Asia–Pacific Web Conf.*, Sydney, NSW, Australia, 2013, pp. 815–826.

[13] D. Birant and A. Kut, ''ST-DBSCAN: An algorithm for clustering spatial-temporal data,'' *Data Knowl. Eng.*, vol. 60, no. 1, pp. 208–221, 2007. doi: 10.1016/j.datak.2006.01.013.

[14] M. Nanni and D. Pedreschi, ''Time-focused clustering of trajectories of moving objects,'' *J. Intell. Inf. Syst.*, vol. 27, no. 3, pp. 267–289, 2006. doi: 10.1007/s10844-006-9953-7.

[15] X. Xiao, Y. Zheng, Q. Luo, X. Xie, and W. Schwinger, ''Finding similar users using category-based location history,'' in *Proc. 18th ACM SIGSPATIAL Int. Symp. Adv. Geograph. Inf. Syst.*, San Jose, CA, USA, 2010, pp. 442–445.

[16] J.-C. Ying, W.-C. Lee, T.-C. Weng, and V. S. Tseng, ''Semantic trajectory mining for location prediction,'' in *Proc. 19th ACM SIGSPATIAL Int. Symp. Adv. Geograph. Inf. Syst.*, Chicago, IL, USA, 2011, pp. 34–43.

[17] S. Liu, Y. Liu, M. L. Ni, J. Fan, and A. M. Li, ''Towards mobility-based clustering,'' in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Washington, DC, USA, 2010, pp. 919–928.

[18] W. Luo, H. Tan, L. Chen, and M. Lionel Ni, ''Finding time period-based most frequent path in big trajectory data,'' in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, New York, NY, USA, 2013, pp. 713–724.

[19] B. Pan, Y. Zheng, D. Wilkie, and A. C. Shahabi, ''Crowd sensing of traffic anomalies based on human mobility and social media,'' in *Proc. 21st SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst.*, Orlando, FL, USA, 2013, pp. 334–343.

[20] *Apache Zookeeper*. Accessed: Oct. 4, 2018. [Online]. Available: https://zookeeper.apache.org/

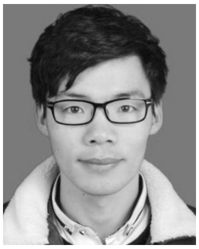
[21] Y. Zheng, L. Zhang, X. Xie, and W. Y. Ma, ''Mining interesting locations and travel sequences from GPS trajectories,'' in *Proc. 18th Int. Conf. World Wide Web*, Madrid, Spain, 2009, pp. 791–800.

[22] Y. Zheng, Q. Li, Y. Chen, X. Xie, and W. Y. Ma, ''Understanding mobility based on GPS data,'' in *Proc. ACM Conf. Ubiquitous Comput. (UbiComp)*, Seoul, South Korea, 2008, pp. 312–321.

[23] Y. Zheng, X. Xie, and W.-Y. Ma, ''GeoLife: A collaborative social networking service among user, location and trajectory,'' *IEEE Data Eng. Bull.*, vol. 33, no. 2, pp. 32–39, Jun. 2010.

[24] Z. Yue, J. Zhang, H. Zhang, and Q. Yang, ''Time-based trajectory data partitioning for efficient range query,'' in *Proc. Int. Workshops Database Syst. Adv. Appl.*, Gold Coast, QLD, Australia, 2018, pp. 24–35.

[25] A. Guttman, ''R-trees: A dynamic index structure for spatial searching,'' in *Proc. Annu. Meeting SIGMOD*, Boston, MA, USA, 1984, pp. 47–57.

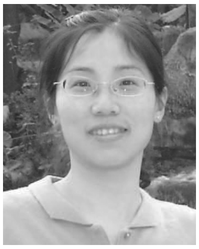


ZHONGWEI YUE received the bachelor's degree from the Zhengzhou Shengda University of Economics, Business and Management, in 2016, and the master's degree from the Guilin University of Electronic Technology, in 2019. He is currently pursuing the Ph.D. degree with the East China Normal University, China. His research interests include spatial data management, distributed systems, database transaction processing, Remote Direct Memory Access (RDMA).

JINGWEI ZHANG received the Ph.D. degree from East China Normal University, China, in 2012. He is currently a Professor with the School of Computer and Information Security, Guilin University of Electronic Technology, China. His research interests include massive data management, distributed computing frameworks, web data analysis, and big data analytics for emerging applications.

RU CHEN received the bachelor's degree from the Hebei Normal University of Science and Technology, China, in 2016. She is currently pursuing the master's degree with the Guilin University of Electronic Technology, China. Her research interests include database management, big data similarity join, and optimization.

YA ZHOU received the master's degree from Fudan University, China. She is currently a Professor with the School of Computer and Information Security, Guilin University of Electronic Technology, China. Her research interests include massive data management, web services, distributed systems, and big data analytics for emerging applications.



QING YANG is currently an Associate Professor with the School of Electronics Engineering and Automation, Guilin University of Electronic Technology, China. Her research interests include intelligent information processing, social network analysis, and large-scale data processing optimization.