Under consideration for publication in Knowledge and Information Systems

Efficient Location-based Search of Trajectories with Location Importance

Da Yan[§1], James Cheng[§2], Zhou Zhao[†3] and Wilfred Ng[†4]

[§]*Department of Computer Science and Engineering, the Chinese University of Hong Kong*
{[1]yanda, [2]jcheng}@cse.cuhk.edu.hk
[†]*Department of Computer Science and Engineering, the Hong Kong University of Science and Technology*
{[3]zhaozhou, [4]wilfred}@cse.ust.hk

Abstract. Given a database of trajectories and a set of query locations, location-based trajectory search finds trajectories in the database that are close to all the query locations. Location-based trajectory search has many applications such as providing reference routes for travelers who are planning a trip to multiple places of interest. However, previous studies only consider the spatial aspect of trajectories, which is inadequate for real applications. For example, one may obtain the reference route of a tourist who just passed by a place of interest without paying a visit. We propose the k *Important Connected Trajectories* (k-*ICT*) query by associating trajectories with *location importance*. For any query location, the result trajectories should contain an *important* point close to it. We describe an effective method to infer the importance of trajectory points from the temporal information. We also propose efficient R-tree based and grid-based algorithms to answer k-ICT queries, and verify the efficiency of our algorithms through extensive experiments on both real and synthetic datasets.

Keywords: Trajectory; location importance; threshold algorithm; Voronoi diagram

1. Introduction

With the popularity of location-acquisition technology, huge amounts of trajectory data are being generated at an unprecedented scale. We differentiate two types of trajectory data. The first type is simply a sequence of time-stamped locations, usually generated by mobile devices such as cell phones and GPS receivers at a relatively high sampling rate. The sample points in such trajectories have very little or no semantics, and many

Received Oct 23, 2013
Revised Jul 01, 2014
Accepted Aug 31, 2014

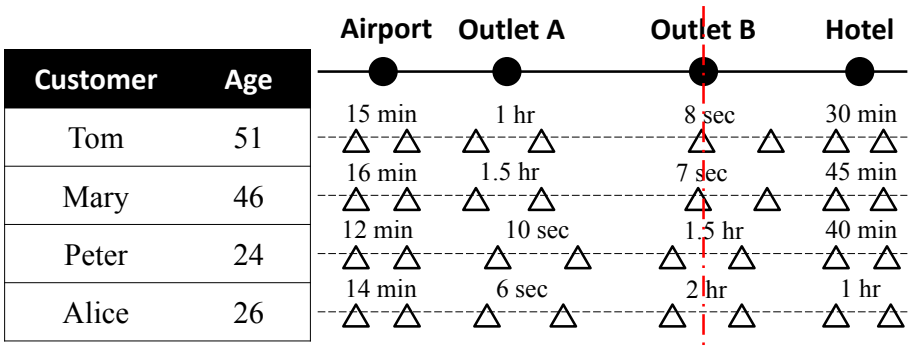


Figure 1. Illustration of the weakness of the k-BCT query

recorded locations are not important. The second type of trajectory is a sequence of locations with semantics, where each recorded location is usually important. One example of such a trajectory is a sequence of geo-tagged photos taken by a traveler in a trip. Numerous such trajectories can be obtained from photo-sharing websites such as Flickr (`www.flickr.com`), and people usually take photos at locations they like. Another example of such a trajectory is a sequence of check-in records of some traveler at the places he/she cares. Such trajectories are available from location-based social network services such as FourSquare (`foursquare.com`).

The proliferation of trajectory data has spawned many novel applications. One example is searching trajectories by locations (Chen et al, 2010; Shang et al, 2012; Zheng et al, 2013). Location-based trajectory search was first proposed in Chen et al (2010) as the k *Best-Connected Trajectories* (k-BCT) query. Given a few query locations, a k-BCT query finds k trajectories that are close to all query points from a trajectory database. Location-based trajectory search can benefit users in many real life applications. For example, it can help travelers who are planning a trip to multiple places of interest in an unfamiliar city, by providing similar routes traveled by other people for reference. Location-based trajectory search is also useful in human behavior analysis, where the query locations can be tourist attractions (specified by a travel agency) or the stops of a new metro line (specified by the transport department).

The k-BCT query, however, considers only the spatial aspect of trajectories, which is inadequate for many real applications. Consider a travel agency that queries a database of tourist trajectories for market analysis. Figure 1 shows a database with four trajectories, each belonging to a different tourist. For simplicity, we assume the data space to be 1D rather than 2D, and we only mark the relevant trajectory samples using $\triangle$. For example, Tom spent 15 minutes at the airport (for check-out), 1 hour at Outlet A (for shopping), 8 seconds at Outlet B (just passing by), and 30 minutes at the hotel (for check-in and taking a rest). From Figure 1, we can see that young people (e.g., Peter and Alice) may usually go shopping at Outlet B on their way from the airport to the hotel, while middle-aged people (e.g., Tom and Mary) would prefer to go shopping at Outlet A. Unfortunately, a 2-BCT query over the database with query locations, {Airport, Outlet B, Hotel}, would return the trajectories of Tom and Mary (who actually went shopping at Outlet A), since the 5-th sample in the trajectories of Tom and Mary is closer to Outlet B than any of the samples in the trajectories of Peter and Alice. As a result, the travel agency may make a wrong arrangement: when a tourist bus picks up

a group of middle-aged tourists at the airport and goes to the hotel, it would stop at Outlet B for the tourists to go shopping.

This example demonstrates that it is necessary to take location importance into consideration. Although Tom and Mary have a trajectory sample close to Outlet B, the importance of the sample with respect to the whole trajectory is low since Tom and Mary just passed by Outlet B. On the contrary, Peter and Alice went shopping at Outlet B, though their trajectory samples are farther away from Outlet B (the samples were probably recorded at a car park nearby).

In this paper, we propose a new type of location-based trajectory search called the k *Important Connected Trajectories* (k-ICT) query, over a database of trajectories associated with location importance. We discuss how to derive the importance of trajectory sample points from their timestamps, and develop efficient algorithms for answering k-ICT queries.

The main contributions of this paper are summarized as follows:

- We propose the k-ICT query over a database of trajectories with location importance, which returns trajectories of much higher utility compared with the k-BCT query (Chen et al, 2010).
- We design a practical method for deriving the importance of trajectory sample points from their timestamps.
- We propose two R-tree based algorithms for answering k-ICT queries, founded on two variants of *Threshold Algorithm* (TA) for top-k queries.
- We further develop two grid-based algorithms, which process k-ICT queries using our grid index built from the *Multiplicatively Weighted Voronoi Diagram* (MWVD) of trajectories. The grid-based algorithms address the drawbacks of the R-tree based algorithms. Experiments show that the grid-based algorithms are more efficient in terms of both time and space.

The rest of this paper is organized as follows. Section 2 reviews the related work. In Section 3, we formulate the k-ICT query. Section 4 discusses how to derive trajectory location importance from raw GPS data. We present our R-tree based algorithms in Section 5, and describe the grid-based algorithms in Section 6. We report experimental results in Section 7 and conclude the paper in Section 8.

2. Related Work

Conventional Trajectory Search. Given a query trajectory, conventional trajectory search finds k trajectories with the shortest distances to the query trajectory. Definitions of the distance function include (Yi et al, 1998; Vlachos et al, 1998; Chen and Ng, 2004; Chen et al, 2005). However, these definitions ignore the time dimension of the trajectory samples, and thus may overrate insignificant trajectory samples.

Trajectory Search by Locations. Location-based trajectory search was first proposed by Chen et al (2010), where the query input is a set of locations. Compared with searching trajectories by a complete query trajectory, it is more practical to search trajectories by locations of interest. Consider the example where a traveler is planning a trip to an unfamiliar city. He/she can easily specify the places he/she intends to visit as the query points, by clicking them on a digital map. On the other hand, it is difficult for a new comer to specify a preferred route as the query trajectory. However, as we shall discuss in Section 3, Chen et al (2010) adopts a distance measure that is undesirable. Instead,

Tang et al (2011) proposes to use the sum-of-distance measure which is more reasonable in real life applications. Recent research starts to enhance location-based trajectory search with keywords (Shang et al, 2012; Zheng et al, 2013).

The problem with these works is that they do not consider location importance and thus queries may easily overrate insignificant trajectory samples, as illustrated by the example described in Section 1.

Trajectory Search by Patterns. The location-based trajectory search mentioned above does not allow users to specify any constraints other than a set of query location. This kind of query is easy to specify and satisfies the requirement of many applications. For example, a tourist to an unfamiliar city may have some famous scenic spots in mind, but does not have a concrete plan yet. He/she may use the location-based trajectory search to find some related trajectories for reference.

However, there are also cases where users would like to add more constraints to the query. For example, a tourist in Seattle may want to visit the museums first as they only open in the daytime, and then goes to the Space Needle which stays open at night. In fact, the city view may be more beautiful at night. In this case, a user may formulate a spatial-temporal pattern and find the trajectories that match the pattern for reference, as is done in Vieira et al (2010) and in Hadjieleftheriou et al (2005).

Mining Important Locations from Trajectories. There are studies on how to mine important locations from trajectory data, such as raw GPS data (Cao et al, 2010) and Flickr data (Yang et al, 2011). These works measure location importance from all the trajectories. Another work (Spaccapietra et al, 2008) finds important locations from a single trajectory. The work models a trajectory by *stops* and *moves*, where a stop is a semantically important part of the trajectory. They proposed the IB-SMoT algorithm to generate stops: given a database of geographic objects, if a part of trajectory intersects with the object, and the time span of the sub-trajectory is above a minimum time threshold, then the sub-trajectory is identified as a stop. Later work uses density based clustering of the trajectory samples to find stops, such as CB-SMoT (Tietbohl et al, 2008) and DB-SMoT (Rocha et al, 2010). Conceptually, the samples of a stop are important, while the samples of a move are immaterial. However, these methods do not provide a concrete importance score for the samples (or stops), and thus it is impossible to compare the importance of different samples (or stops).

Other Topic about Trajectory Processing. Sometimes the trajectory data are sampled in a very low rate like every several minutes or even every several hours. In this case, the behavior between two consecutive trajectory point is missing. Zheng et al (2012) studies how to discover the top-k possible routes sequentially passing the queried locations from such uncertain trajectories, where the road network information is used to reduce the uncertainty caused by low sampling rate.

There are also works that use the trajectories to discover regions of different functions in a city. For example, Yuan et al (2012) uses topic modeling to learn the rich structure of different functional sections of Beijing, by using data sources such as points of interest (POIs) and GPS readings collected from taxis.

3. Problem Formulation

We now formally define the k-ICT query. Let D be a database of trajectories, where each trajectory $T \in D$ is a sequence of points $(p_1, p_2, \ldots, p_\ell)$. We assume that each point p_i is associated with a score $w(p_i) \geq 0$, which corresponds to the importance

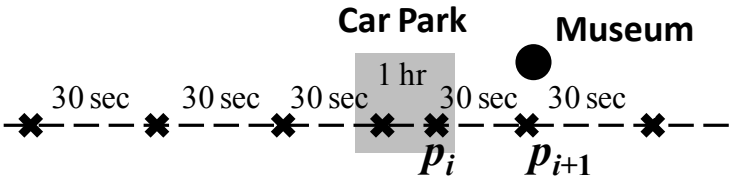


Figure 2. Intuition behind distance function in Eq. (2)

of p_i in trajectory T. For raw GPS data, we can derive the importance score using the time stamps of the trajectory samples, which we will further discuss in Section 4. For trajectories obtained from Flickr photos, the location importance of a photo can be derived using the number of page visits; the score can also be manually set by the photo owner.

A k-ICT query, Q, is represented by a set of m locations (or points): $Q = \{q_1, q_2, \ldots, q_m\}$. We first introduce the distance functions that define how a k-ICT query is to be evaluated.

Distance Related to One Query Point. Let us first focus on a specific query point $q_i \in Q$. We define the weighted distance between query point q_i and a trajectory point p_j as follows:

$$d(q_i, p_j) = \frac{\|q_i p_j\|}{w(p_j)}, \tag{1}$$

where we use $\|pq\|$ to denote the Euclidean distance between two points p and q. Note that a larger importance score $w(p_j)$ makes p_j closer to q_i (since $d(q_i, p_j)$ is smaller).

We define the weighted distance between query point q_i and a trajectory $T = (p_1, p_2, \ldots, p_\ell)$ as the weighted distance between q_i and its closest trajectory point in T:

$$d(q_i, T) = \min_{p_j \in T} \{d(q_i, p_j)\}. \tag{2}$$

We now illustrate the intuition behind the distance function in Equation (2). Consider the vehicle GPS trajectory fragment shown in Figure 2, which is generated as follows. A traveler rented a GPS-equipped car to travel around a city. He drove to a car park near a museum, parked his car, stayed in the museum for an hour, and then drove to the next destination. Since the car was turned off when it was parked, the on-board GPS device was also off. As a result, no sample was generated during that one hour when the car was parked, and $p_i \in T$ is the first sample after the traveler drove the car away from the car park.

In this example, the query point q in question is the museum. Now assume that $w(p)$ is proportional to the time the car stopped at point p. Although $\|qp_{i+1}\| < \|qp_i\|$, it is obvious that $w(p_{i+1}) \ll w(p_i)$ and thus $d(q, p_{i+1}) > d(q, p_i)$ according to Equation (1), i.e. p_i is closer to q than p_{i+1}. Therefore, $d(q_i, T) = d(q, p_i)$ by Equation (2). Note that $d(q, p_i)$ correctly estimates the confidence that the traveler of T visited the museum, since he may instead visit an aquarium nearby after parking his car. In the latter case, $d(q, p_{i+1})$ overestimates the confidence that the traveler visited the museum since he actually visited the aquarium nearby the point p_{i+1}. Thus, $d(q, p_i)$ presents an accurate estimate in this case.

Overall Distance Function. We now define the weighted distance between a query

Q and a trajectory T, by aggregating $d(q_i, T)$ for all query points $q_i \in Q$. Since we want to find trajectories close to all query points in Q, we define the overall weighted distance as:

$$d(Q, T) = \sum_{i=1}^{m} d(q_i, T). \tag{3}$$

Intuitively, $d(Q, T)$ is the total distance of traveling from the closest position of T to q_i for all $q_i \in Q$.

In addition to the physical meaning described above, Equation (3) is also meaningful from the probabilistic point of view, which we discuss next. Let us denote p_{n_i} to be the trajectory point of T closest to q_i (in terms of weighted distance), then $d(q_i, T) = d(q_i, p_{n_i})$. We also denote $p(q_i, T)$ to be the probability that the owner of the trajectory T visited q_i, and a reasonable assumption is that $p(q_i, T)$ decays exponentially as $d(q_i, p_{n_i})$ increases. Using the PDF (Probability Density Function) of the exponential distribution, we have $p(q_i, T) = \lambda e^{-\lambda \cdot d(q_i, T)}$. Since we have no preference of one query point over another, we use the same λ for all $q_i \in Q$. Since we want a result trajectory to be close to all query points, the probability that the owner of trajectory T visited all $q_i \in Q$ is:

$$\prod_{i=1}^{m} p(q_i, T) \quad \propto \quad e^{-\lambda \sum_{i=1}^{m} d(q_i, T)}, \tag{4}$$

where we assume that "whether the owner visited one query location" is independent of "whether he visited another query location".

Since we want to maximize the probability value of Equation (4), it is equivalent to minimize $d(Q, T) = \sum_{i=1}^{m} d(q_i, T)$.

The k-BCT query (Chen et al, 2010) adopts a similarity function $sim(Q, T) = \sum_{i=1}^{m} e^{-d(q_i, T)}$. If we fix $\lambda = 1$, then $sim(Q, T) = \sum_{i=1}^{m} p(q_i, T)$. Compared with Equation (4), this similarity function is undesirable, since the similarity value is high as long as one query point is close to T, even if all other query points are far from T. Similar observation is mentioned in Tang et al (2011), which proposes to use a sum-of-Euclidean-distance measure. In this paper, we use the sum-of-weighted-distance measure to incorporate object importance.

We define k-ICT querying as follows.

Definition 1 (k-ICT Querying). Given a database of trajectories $D = \{T_1, \ldots, T_n\}$ ($n \geq k$), a set of query locations Q, a k-ICT query is to find a set of k trajectories, $R \subseteq D$, such that

$$d(Q, T) \leq d(Q, T'),\ \forall T \in R,\ \forall T' \in D - R.$$

Complexity Analysis. Let us denote the size of a trajectory T (i.e., the number of trajectory points in T) by $|T|$, and the the database size (i.e. the total number of trajectory points in D) by $\|D\| = \sum_{T \in D} |T|$. Note that $\|D\|$ is different from the number of trajectories in D, which is given by $n = |D|$. Given a k-ICT with m query locations, for each trajectory T, we may compute the closest point to each query location in $O(|T|)$ time, and thus compute the closest points to all query locations in $O(m|T|)$ time. We may then compute the sum-of-weighted-distance score for each trajectory in $O(|T|)$ time. Therefore, we may compute the scores for all trajectories in $\sum_{T \in D} O(m|T|) = O(m \sum_{T \in D} |T|) = O(m\|D\|)$ time. Given the trajectories scores,

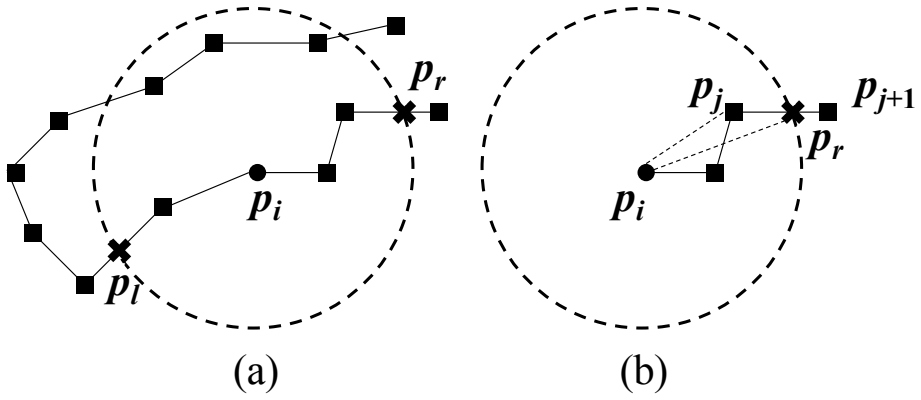


Figure 3. Illustration of the evaluation of location importance

we may then find the top-k trajectories in $O(n \log k)$ time using a priority queue of size at most k. Therefore, the time complexity of a k-ICT is bounded by $O(m\|D\|+n \log k)$. However, this brute force approach requires scanning the whole database once, as the time complexity is linear to $\|D\|$ and n. In the rest of this paper, we consider how to access only a small fraction of the whole trajectory database to find the top-k trajectories.

4. Location Importance

In this section, we discuss how to compute the importance of trajectory samples from raw GPS data.

Formulation. A GPS reading can be represented by a triplet $(latitude, longitude, timestamp)$. In order to manipulate the data in Euclidean space, we map the coordinates of all sample points from the GPS coordinates $(latitude, longitude)$ to Universal Transverse Mercator (UTM) coordinates $(easting, northing)$, or simply (x, y).

Given a trajectory $T = (p_1, p_2, \ldots, p_\ell)$, where each sample point $p_i = (p_i.x, p_i.y, t(p_i))$, we want to compute the importance $w(p_i)$ for all the sample points $p_i \in T$.

We define the neighborhood of a sample point $p_i \in T$, denoted by $Cir(p_i)$, to be a circle centered at p_i with radius r, where r is a user-specified parameter. Figure 3(a) shows the circle $Cir(p_i)$ and the trajectory T. Let p_l (or respectively, p_r) be the first location on T reaching the boundary of $Cir(p_i)$ when going backward (or respectively, forward) from p_i along T. Note that p_l and p_r may not be an existing trajectory sample, but rather the intersection point between $Cir(p_i)$ and a segment $p_j p_{j+1}$ as shown in Figure 3(b). In this case, we use linear interpolation to compute the location and time stamp of p_r (or p_l). Another extreme case is that p_l (or respectively, p_r) may be the first (or respectively, last) trajectory sample that is inside $Cir(p_i)$, since in this case we cannot go backward (or respectively, forward) from p_i along T.

We define the following measure using p_l and p_r:

$$\Delta t(p_i) = \max\{t(p_r) - t(p_i), t(p_i) - t(p_l)\}. \quad (5)$$

Here, $(t(p_r) - t(p_i))$ is the time spent before the traveler left $Cir(p_i)$ from p_i in the forward direction, while $(t(p_i) - t(p_l))$ is the time spent before the traveler left $Cir(p_i)$ from p_i in the backward direction (or more intuitively, the time spent from

when the traveler stepped in $Cir(p_i)$ until he reached p_i). Intuitively, $\Delta t(p_i)$ is defined such that as long as the traveler stopped near p_i (no matter in the forward or backward direction), the importance of p_i is promoted. The greater $\Delta t(p_i)$ is, the more important the trajectory sample p_i is.

This definition of $\Delta t(p_i)$ has two benefits. First, even when the traveler is in an important location (e.g., a marketplace), he may still be walking around and the accumulated distance can be large. Using a neighborhood circle to cover the marketplace, we can correctly identify that the locations in the marketplace are important. Second, when a GPS-equipped car is turned off, so is the GPS device. Thus, we can only consider the last few samples before the car stops, or the first few samples after the car starts as important (these locations may still be inside the car park), which are better covered using the neighborhood circle.

The next issue is how to compute $w(p_i)$ using $\Delta t(p_i)$. Obviously, $w(p_i)$ should increase fast with $\Delta t(p_i)$ when $\Delta t(p_i)$ is small, but increase slowly when $\Delta t(p_i)$ is large. For example, a 1-minute stop may not be important since it may be due to a red traffic light, while a 10-minute stop is more likely to be important. On the other hand, a 1-hour stop quite certainly implies an important location, and the score should not increase too much even if $\Delta t(p_i)$ becomes 2 hours.

When $\Delta t(p_i) = 0$, we want the importance $w(p_i) = 0$. Furthermore, we want $w(p_i)$ to be within $[0, 1]$ so as to carry a probability meaning: the confidence that the traveler of trajectory T stops at p_i. As a result, we define our importance score as follows:

$$w(p_i) = 1 - e^{-\alpha \cdot \Delta t(p_i)}, \tag{6}$$

where α controls how fast $w(p_i)$ increases with $\Delta t(p_i)$.

Computation Details. Next, we discuss the details of computing $\Delta t(p_i)$. We first describe how we compute p_r (the computation of p_l is similar) for the scenario in Figure 3(b), where $p_j \in T$ is inside $Cir(p_i)$ and the next sample p_{j+1} is outside of $Cir(p_i)$.

Instead of directly computing p_r, we first compute segment length $\|p_j p_r\|$. According to the Cosine Law, we can compute $\cos \angle p_i p_j p_r$ as follows:

$$\cos \angle p_i p_j p_r = \frac{\|p_i p_j\|^2 + \|p_j p_{j+1}\|^2 - \|p_i p_{j+1}\|^2}{2 \cdot \|p_i p_j\| \cdot \|p_j p_{j+1}\|},$$

where $\|p_i p_j\|$, $\|p_j p_{j+1}\|$ and $\|p_i p_{j+1}\|$ can be easily computed from the coordinates of p_i, p_j and p_{j+1}.

Then, according to the Cosine Law, $\|p_j p_r\|$ can be obtained by solving the following quadratic equation:

$$\begin{aligned} \|p_i p_r\|^2 = \|p_i p_j\|^2 + \|p_j p_r\|^2 - \\ 2 \cdot \|p_i p_j\| \cdot \|p_j p_r\| \cdot \cos \angle p_i p_j p_r, \end{aligned} \tag{7}$$

where $\|p_i p_j\|$ is computed from the coordinates of p_i and p_j, and $\|p_i p_r\| = r$. The roots

of Equation (7) are:

$$\|p_j p_r\|^{(1)} = \|p_i p_j\| \cdot \cos \angle p_i p_j p_r + \sqrt{\|p_i p_j\|^2 \cdot \cos^2 \angle p_i p_j p_r - \|p_i p_j\|^2 + r^2}, \tag{8}$$

$$\|p_j p_r\|^{(2)} = \|p_i p_j\| \cdot \cos \angle p_i p_j p_r - \sqrt{\|p_i p_j\|^2 \cdot \cos^2 \angle p_i p_j p_r - \|p_i p_j\|^2 + r^2}. \tag{9}$$

However, the second root $\|p_j p_r\|^{(2)}$ can be discarded since its value is negative. To see this, recall that p_j is inside $Cir(p_i)$, and hence $\|p_i p_j\| < r$. Thus, we have

$$\begin{aligned} &\sqrt{\|p_i p_j\|^2 \cdot \cos^2 \angle p_i p_j p_r - \|p_i p_j\|^2 + r^2} \\ &> \sqrt{\|p_i p_j\|^2 \cdot \cos^2 \angle p_i p_j p_r - r^2 + r^2} \\ &> \|p_i p_j\| \cdot \cos \angle p_i p_j p_r. \end{aligned}$$

Therefore, $\|p_j p_r\|^{(2)} < 0$ according to Equation (9), and we conclude that the value of $\|p_j p_r\|$ is given by Equation (8).

Algorithm 1 Computing $(t(p_r) - t(p_i))$

Input: Trajectory $T = (p_1, p_2, \ldots, p_\ell)$
Output: $\Delta t = t(p_r) - t(p_i)$

```
1: Δt ← 0;
2: for p_j := p_i to p_{ℓ−1} do
3:     if p_{j+1} is inside Cir(p_i) then
4:         Δt ← Δt + (t(p_{j+1}) − t(p_j));
5:     else
6:         Compute ‖p_j p_r‖ by Equation (8);
7:         Δt ← Δt + (‖p_j p_r‖ / ‖p_j p_{j+1}‖)(t(p_{j+1}) − t(p_j));
8:         return Δt;
9: return Δt;
```

Now we discuss how to compute $(t(p_r) - t(p_i))$. The value of $(t(p_i) - t(p_l))$ can be computed similarly, and both of them are then used to compute $\Delta t(p_i)$ according to Equation (5). The algorithm is described in Algorithm 1. We check samples forward along T starting from p_i (Line 2). If the next sample p_{j+1} is inside $Cir(p_i)$, then the whole segment $p_j p_{j+1}$ is inside $Cir(p_i)$ (due to the convexity of circles), and we accumulate the time spent on $p_j p_{j+1}$ to the result (Lines 3-4). Otherwise, we compute p_r and accumulate the time spent on $p_j p_r$ to the result (Lines 5-7). Note that in the latter case, we already reach the boundary of $Cir(p_i)$ and thus the accumulated time is directly returned (Line 8).

Parameter Setting. We have two parameters: (1) radius r of $Cir(p_i)$, and (2) the decay rate α in Equation (6). Typically, r is set as the diagonal length of a market place, or the distance between a car park and the intended destination. While the parameter

choice is application-dependent, our experiments on several vehicle GPS datasets show that our method always provides reasonable importance score (judged by human) when $r = 50$m and $\alpha = 0.002$. The details are omitted due to the space limitation.

5. R-Tree Based Algorithms

In this section, we introduce two R-tree based algorithms for answering k-ICT queries. Before we present our algorithms, we first describe the *Threshold Algorithm* (TA) (Fagin et al, 2001), since our algorithms adopt the TA framework for top-k query processing.

5.1. Threshold Algorithm and Its Variants

TA (Fagin et al, 2001) has been widely adopted for processing top-k queries, including the k-BCT querying algorithm (Chen et al, 2010) and the keyword-aware variants (Shang et al, 2012; Zheng et al, 2013). In the setting of Fagin et al (2001), we are given a database table D of n tuples, where the schema of the table is $(A_1, A_2, \ldots, A_m)$. For each attribute A_i, a list L_i is built by sorting all the tuples in non-decreasing order of the values of attribute A_i, and stored on disk. Each entry in L_i is a pair (id, val), where id is the id of the corresponding tuple, and val is the value of attribute A_i for the tuple. We describe two algorithms that use the lists to find the top-k tuples, where the ranking score of a tuple equals the summation of the values of all its m attributes (a smaller score is preferred).

Fagin's Algorithm (FA). FA finds the top-k tuples in three steps. *Step 1*: read a (id, val) pair from each list in a round-robin manner, until there are k tuples whose id's have been seen from all the m lists. *Step 2*: for each tuple id seen (from any list), retrieve the tuple from the table D if any of its attribute values are missing (random access to D is needed). *Step 3*: compute the ranking score by summing the attribute values for each tuple whose id has been seen, and return the k tuples with the smallest summation values.

Threshold Algorithm (TA). Unlike the *filter-and-refine* framework of FA where random access to D is only used in the refinement step, TA adopts a more aggressive approach. TA also reads a (id, val) pair from each list in a round-robin manner, but for each tuple id seen, TA immediately retrieves the tuple from D by random access, computes the ranking score, and updates the current top-k tuples. Meanwhile, for each list L_i, TA maintains a variable τ_i equal to val of the last (id, val) pair read from L_i. The round-robin operation stops when the ranking score of the current top k-th tuple is equal to or smaller than $\sum_{i=1}^{m} \tau_i$.

In general, TA reads less pairs from the lists than FA, but performs more random accesses to D than FA. While we focus on FA and TA when introducing our algorithms, other variants of TA may also be adopted by our algorithm.

5.2. R-Tree Based Algorithms

We now present our R-tree based algorithms. We first describe a key operator used by our algorithms: the *incremental weighted nearest-neighbor* (*NN*) algorithm.

Incremental Weighted NN Algorithm. Unlike Chen et al (2010), in our problem,

each trajectory point p_i is associated with an importance score $w(p_i)$, and its distance to a query point q is evaluated as $\frac{\|qp_i\|}{w(p_i)}$ using Equation (1). Thus, R-tree is no longer sufficient for solving our problem. Instead, we index the trajectory points by an *aggregate R-tree* (*aR-tree*) (Lazaridis and Mehrotra, 2001) with aggregate function MAX, called *MAX R-tree*.

Compared with a traditional R-tree, each node entry e of a MAX R-tree maintains the maximum importance score among all points indexed under e (i.e., indexed in the subtree rooted at the node pointed to by e). Given an R-tree node entry e, we denote its MAX aggregate value by $w(e)$. We also denote the *Minimum Bounding Rectangle* (*MBR*) of e by $mbr(e)$. Then, for any trajectory point p indexed under e, its weighted distance to query point q is given by:

$$d(q,p) = \frac{\|qp\|}{w(p)} \geq \frac{mindist(q, mbr(e))}{w(e)}, \tag{10}$$

where $mindist(q, mbr(e))$ is the distance from q to its closest point in $mbr(e)$. The inequality holds as $mindist(q, mbr(e)) \leq \|qp\|$ and $w(e) \geq w(p)$.

For simplicity, given an R-tree node entry e and a query point q, we define:

$$LB(q,e) = \frac{mindist(q, mbr(e))}{w(e)}. \tag{11}$$

According to Equation (10), $LB(q,e)$ lower bounds the weighted distance from any point indexed under e to q.

Algorithm 2 Computing the Next Weighted NN of q_i

Input: query location q_i, priority queue *min-heap*, Max R-tree $tree$
Output: $(d(q_i,p),p)$ where p is the next NN of q_i

```
1:  while min-heap is not empty do
2:    (LB(q_i, e), e) ← min-heap.dequeue();
3:    if e is a leaf node entry then
4:      p ← the trajectory point pointed to by e;
5:      return (LB(q_i, e), p);
6:    else
7:      node ← the R-tree node pointed to by e;
8:      for each entry e' of node do
9:        Compute LB(q_i, e');
10:       min-heap.enqueue(LB(q_i, e'), e');
```

Algorithm 2 describes our incremental weighted NN algorithm. When processing a k-ICT query, we maintain a priority queue of R-tree node entries for each query point q_i, so that the next NN of q_i can be incrementally obtained using Algorithm 2. Initially, the priority queue *min-heap* contains only the root node of the Max R-tree $tree$, and each call of Algorithm 2 updates *min-heap* and retrieves the next NN of q_i.

We now explain Algorithm 2 in details. In each round, the entry e with the smallest $LB(q_i,e)$ is dequeued from *min-heap* (Line 2). If e is an entry of a leaf node, then it points to a trajectory point p and $LB(q_i,e) = d(q_i,p)$. In this case, we can conclude that p is the next NN (Line 5), since any unseen trajectory point p' is indexed under some node entry en in *min-heap*, and $d(q_i,p') \geq LB(q_i,en) \geq LB(q_i,e)$. Otherwise, e is

an entry of a non-leaf node $node$, and we enqueue all the entries of $node$ into *min-heap* (Lines 7-10).

R-Tree based FA and TA. We now introduce our two R-tree based algorithms for answering k-ICT queries, one based on FA and the other based on TA. Both algorithms use Algorithm 2 for sequentially accessing the next NN of each query point q_i.

Algorithm 3 presents the R-tree based FA for answering k-ICT queries. Similar to FA, Algorithm 3 has two phases: the filtering phase (Lines 3-15) and the refinement phase (Lines 16-20).

Algorithm 3 R-tree based FA for Answering k-ICT Queries

Input: k, query set Q, trajectory database D, Max R-tree $tree$
Output: k-ICT (the top-k trajectories)

1: $table \leftarrow \emptyset, d \leftarrow 1$;
2: $N \leftarrow$ number of trajectory points in D;
3: **while** $d \leq N$ **do**
4: **for each** $q_i \in Q$ **do**
5: Retrieve the d-th NN p of q_i, together with the weighted distance $d(q_i, p)$, using Algorithm 2;
6: $T \leftarrow$ the trajectory that p belongs to;
7: **if** $T \notin table$ **then**
8: Insert T into $table$;
9: $T[i] \leftarrow d(q_i, p)$; $/* \ T[i]$ represents $d(q_i, T) \ */$
10: **else**
11: **if** $T[i]$ is not yet assigned **then**
12: $T[i] \leftarrow d(q_i, p)$;
13: **if** there are k trajectories in $table$ whose attribute values are all assigned **then**
14: **goto** Line 16;
15: $d \leftarrow d + 1$;
16: **for each** $T \in table$ **do**
17: Read T;
18: For any $T[i]$ not yet assigned: $T[i] \leftarrow d(q_i, T)$;
19: Compute $d(Q, T) = \sum_{i=1}^{m} T[i]$;
20: Update the top-k trajectories;
21: **return** the top-k trajectories;

In each round of the filtering phase, Algorithm 3 obtains the next NN of each q_i for processing (Line 4), i.e., the NNs of the query points are processed in a round-robin manner. Since there are N trajectory points indexed by $tree$, there are at most N rounds (Line 3). All the seen trajectories are maintained using a hash table $table$, where the hash key is the trajectory id. If the trajectory of the obtained point p has not been seen yet, we know that $d(q_i, T) = d(q_i, p)$, and thus we insert T into $table$ and record $d(q_i, p)$ as the value of the i-th attribute, denoted by $T[i]$ (Lines 7-9). Otherwise, T is already in $table$, and we check whether $T[i]$ has been assigned a value (Line 11). If $T[i]$ has already been assigned a value, we ignore the obtained point p since the point in T that is closest to q_i has already been processed before. Otherwise, p is the point in T closest to q_i, and we set $T[i]$ to be $d(q_i, p)$. The filtering phase terminates once k tuples are seen with the value of $T[i]$ assigned for all $q_i \in Q$, which is similar to the traditional FA.

In the refinement phase, we first compute $d(q_i, T)$ for any $T[i]$ whose value has not

Algorithm 4 R-tree based TA for Answering k-ICT Queries

Input: k, query set Q, trajectory database D, Max R-tree $tree$
Output: k-ICT (the top-k trajectories)

1: $table \leftarrow \emptyset, d \leftarrow 1$;
2: $N \leftarrow$ number of trajectory points in D;
3: *max-heap*$\leftarrow \emptyset$;
4: **while** $d \leq N$ **do**
5: $\tau \leftarrow 0$;
6: **for each** $q_i \in Q$ **do**
7: Retrieve the d-th NN p of q_i, together with the weighted distance $d(q_i, p)$, using Algorithm 2;
8: $T \leftarrow$ the trajectory that p belongs to;
9: **if** $T \notin table$ **then**
10: $T[i] \leftarrow d(q_i, p)$;
11: $\forall j \neq i$, compute $T[j] = d(q_j, T)$ by accessing T;
12: Insert T into $table$;
13: Compute $d(Q, T) = \sum_{i=1}^{m} T[i]$;
14: *max-heap*.$enqueue(d(Q, T), T)$;
15: If *max-heap*.$size() > k$: *max-heap*.$dequeue()$;
16: $\tau \leftarrow \tau + d(q_i, p)$;
17: **if** *max-heap*.$size() = k$ **and** *max-heap.top*$() \leq \tau$ **then**
18: **goto** Line 20;
19: $d \leftarrow d + 1$;
20: **return** the k trajectories in *max-heap*;

yet been assigned (a more efficient method of obtaining $T[i]$ is actually used, which we will discuss in Section 6.2). Then, for all the seen trajectories $T \in table$, $d(Q, T)$ is computed and the k tuples with the smallest values of $d(Q, T)$ are returned.

Algorithm 4 presents the R-tree based TA for answering k-ICT queries. Recall that the conventional TA maintains a variable τ_i for each list L_i, whose value equals the attribute value of the last accessed entry. TA stops when the ranking score of the current top k-th tuple is equal to or smaller than $\sum_{i=1}^{m} \tau_i$. In our problem, $\tau_i = d(q_i, p)$ where p is last accessed NN of q_i. We set τ to 0 at the beginning of a round-robin processing round (Line 5), and add $\tau_i = d(q_i, p)$ to τ for each query point q_i (Line 16). Therefore, at the end of the round-robin processing round, $\tau = \sum_{i=1}^{m} \tau_i$ is exactly the pruning threshold, which is then compared with the top k-th trajectory in Line 17 to determine the stopping condition.

In Lines 9-15, we only process the trajectory T of the current point p if T is not in $table$, by accessing T to assign $T[i]$ (a more efficient method discussed in Section 6.2 is actually used here), computing $d(Q, T)$ and updating the top-k results. Note that if T is in $table$, then $T[i]$ must have been assigned for all $i = 1, \ldots, m$ (Lines 10-12), and hence we can ignore T.

Finally, we note that the correctness of both Algorithms 3 and 4 is easy to see by following the correctness of FA and TA (Fagin et al, 2001). We thus omit the details here.

Limitations of R-Tree Based Algorithms. We identify the following limitations of using an R-tree index built over all the trajectory points in the database, which motivates our grid-based algorithm to be introduced in Section 6.

Firstly, the incremental NN search for each query q_i is done over an R-tree that contains all the trajectory samples. However, if we know q_i beforehand, then only one sample per trajectory requires examining (i.e., the sample with the shortest weighted distance to q_i), and there are totally $n = |D|$ such samples, much less than the number of all samples in D. Therefore, there is huge room for improvement in terms of sample candidate pruning.

Secondly, much of the computation done by the R-tree based algorithms could be wasteful. This is because consecutive samples of a trajectory are close in space, and are very likely indexed under the same R-tree node. As a result, in consecutive calls of Algorithm 2 for retrieving the NNs of a query point q_i, many returned NNs may come from the same trajectory.

Finally, we use the maximum importance $w(e)$ of an R-tree node entry to compute the lower bound in Equation (11), which is not tight. As long as there is one point indexed under e with a large weight, the whole entry e has to be accessed early even if all the other points have very low weight, resulting in the addition of all its child nodes into the priority queue.

6. Grid-Based Algorithms

In this section, we present the grid-based algorithms.

Overview. We first give an overview of how our grid-based algorithms address all the three drawbacks of the R-tree based algorithms mentioned in Section 5.2.

Firstly, to avoid doing NN search over all trajectory points, we divide the data space by a grid, so that each grid cell covers a small region. We observe that only a small fraction of samples per trajectory have the chance to be the NN of some location in a cell. Thus, if a query point locates in a grid cell, we only need to check the samples relevant to the cell.

Secondly, to avoid checking a lot of samples of a trajectory that do not contribute to the top-k answers, we propose to pre-compute the *Multiplicatively Weighted Voronoi Diagram* (*MWVD*) of the points of each trajectory. Note that a sample p_i is the weighted NN of q if and only if q locates inside the Voronoi cell of p_i.

Finally, to avoid the interference of samples from different trajectories, we treat trajectories as the first-class citizen (while the R-tree index treats the trajectory points as the first-class citizen). Given a grid cell, we group all its relevant samples by trajectories, and the NN search is done in the unit of trajectories rather than trajectory points.

We discuss these ideas in details in the following subsections.

6.1. Trajectory Preprocessing by MWVD

For each trajectory $T = (p_1, p_2, \ldots, p_\ell)$, we pre-compute the MWVD (OKabe et al, 2009) of its points, which is then used to build our grid index. We first briefly review the MWVD and then show how we use it in our solution.

Let U be the data domain. Given two samples p and p', the *dominant region* of p over p' is defined as:

$$R_{p|p'} = \{q \in U \mid d(q,p) \leq d(q,p')\}.$$

We now consider the shape of $R_{p|p'}$. Let us first assume that $w(p) < w(p')$, then $R_{p|p'}$ is characterized by the region within circle $C_{p|p'}$, whose center c and radius r are

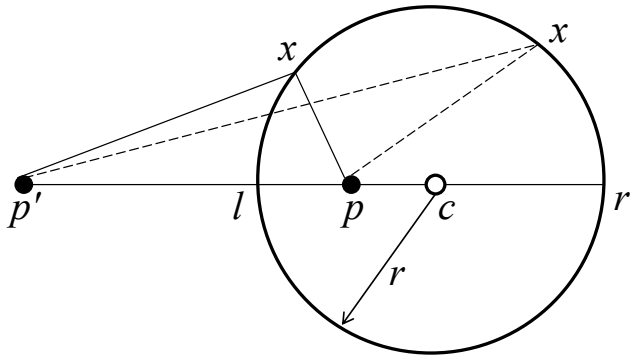


Figure 4. Illustration of dominant regions

given as follows:

$$c = \left(\frac{w^2(p') \cdot p.x - w^2(p) \cdot p'.x}{w^2(p') - w^2(p)}, \frac{w^2(p') \cdot p.y - w^2(p) \cdot p'.y}{w^2(p') - w^2(p)} \right)$$
$$r = \frac{w(p) \cdot w(p') \cdot \|pp'\|}{w^2(p') - w^2(p)}.$$

Figure 4 illustrates the concept of dominant region with circle $C_{p|p'}$. In fact, $C_{p|p'}$ is an Apollonius circle, since for any point x on its boundary, $\frac{\|px\|}{\|p'x\|} = \frac{w(p)}{w(p')}$.

When $w(p) > w(p')$, $R_{p|p'}$ is characterized by the region outside of circle $C_{p|p'}$. For example, in Figure 4 where $w(p') > w(p)$, $R_{p'|p} = U - C_{p|p'}$. Finally, when $w(p) = w(p')$, the perpendicular bisector of pp' divides the space into two half planes, and $R_{p|p'}$ corresponds to the half plane that contains p, denoted by $H_{p|p'}$.

The Voronoi cell of a trajectory point $p \in T$ is given by:

$$VC(p) = \bigcap_{p' \in T - \{p\}} R_{p|p'}, \tag{12}$$

since any point in $VC(p)$ should be in $R_{p|p'}$ for any $p' \in T - \{p\}$. Given a sample $p \in T$, we divide the other samples in T into three sets: T^+ contains all samples p' with $w(p') > w(p)$, T^- contains all samples p' with $w(p') < w(p)$, and T^0 contains all samples p' with $w(p') = w(p)$.

Equation (12) implies that $VC(p)$ may be represented by $\ell - 1$ circles or lines in the worst case. In fact, not all circles/lines contribute to the final shape of $VC(p)$ and many of them can be pruned by the six pruning rules presented in Wu et al (2011). We adopt the best-first search algorithm of Wu et al (2011) for MWVD computation, but the computation is done in memory since the number of points in each trajectory is usually not large.

6.2. Grid Index

Next, we describe two indices used in our grid-based algorithms. In our problem, we assume that there exists a rectangular data space U, such that all trajectory points and

query points locate inside U. For example, U can be the bounding box of a city region. Our grid-based approach divides U by an $N \times N$ grid, denoted by G.

For each trajectory T, we build a *random access* index, denoted by $RAI[T]$, which returns $d(q, T)$ given a query point q; while for each grid cell $G[i, j]$, we build a *sequential access* index, denoted by $SAI[i, j]$, which returns trajectories in non-decreasing order of $d(q, T)$ for a query point q falling in $G[i, j]$.

Random Access Index (RAI). We now describe how we build $RAI[T]$. First, we compute $VC(p)$ for all $p \in T$, where $VC(p)$ is represented by a set of pairs $(p', R_{p|p'})$. We say that p' is *related to* $VP(p)$ if $(p', R_{p|p'}) \in VC(p)$. Then, for each grid cell $G[i, j]$, we compute the set of Voronoi cells overlapping with the rectangular region R that $G[i, j]$ represents. We denote the set by $S(R) = \{VC_R(p_{i_1}), VC_R(p_{i_2}), \ldots, VC_R(p_{i_s}))\}$. Note that a Voronoi cell $VC_R(p)$ may contain less pairs of $(p', R_{p|p'})$ than the original $VC(p)$, since we only need to characterize its shape within R. If $VC(p)$ does not overlap with R, p cannot be the weighted NN of any location in R, and is thus pruned.

We now consider how to compute $VC_R(p)$ from the original $VC(p)$. We divide the trajectory points p' related to $VC(p)$ into three sets S^+, S^- and S^0, according to whether p' belongs to T^+, T^- and T^0, respectively. We check each $(p', R_{p|p'}) \in VC(p)$ in turn for the following:

- *Cell Pruning:* if $R_{p|p'}$ does not overlap with R, we prune $VC_R(p)$ immediately since $VC(p) \cap R_{p|p'} = \emptyset$;
- *Pair Pruning:* if $R_{p|p'}$ contains R, then p' has no contribution to the shape of $VC_R(p)$, and thus $(p', R_{p|p'})$ is not included in $VC_R(p)$;
- Otherwise, $(p', R_{p|p'})$ is added to $VC_R(p)$.

Figure 5 lists the conditions for Cell Pruning and Pair Pruning when $p' \in S^+$, S^- and S^0.

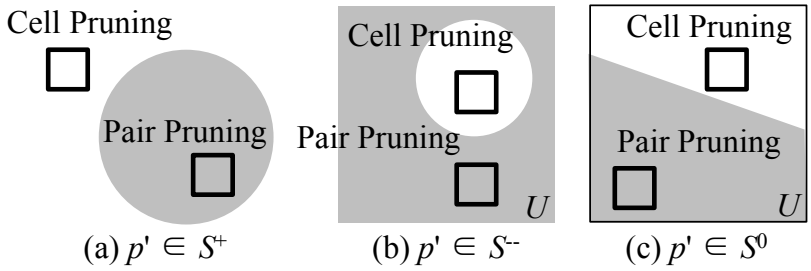
In our implementation, we do not compute $S(R)$ for each grid $G[i, j]$ with region R directly from the original Voronoi set. Instead, we perform the computation by building a quadtree $qtree$ whose leaf nodes correspond to the grid cells. By specifying the height of the quadtree as h, we obtain a $2^h \times 2^h$ grid (i.e., $N = 2^h$).

Algorithm 5 Computing Quadtree Node $node$

Input: Current node $node$, Parent node par, current level $level$

1: $S(node.R) \leftarrow \emptyset$;
2: **for each** $VC_{par.R}(p) \in S(par.R)$ **do**
3: Compute $VC_{node.R}(p)$ by checking the pairs in $VC_{par.R}(p)$, and do the pruning listed in Figure 5;
4: If $VC_{node.R}(p)$ is not pruned, add it to $S(node.R)$;
5: **if** $level < h$ and $|S(node.R)| > 1$ **then**
6: Split $node.R$ into four equal quadrants, $R_i,\ i = 1, 2, 3, 4$;
7: Create child nodes $ch_i,\ i = 1, 2, 3, 4$ with $ch_i.R = R_i$;
8: Recurse over each child node;

Each quadtree node, $node$, is associated with a region $node.R$ and a set of the Voronoi cells overlapping with $node.R$, i.e., $S(node.R)$. Algorithm 5 shows how we compute $S(node.R)$ for each quadtree node $node$ in a recursive manner. Let the quadtree



Condition 1	Condition 2	Action
$p' \in S^+$	$G[i,j]$ is outside of $C_{p \mid p'}$	Prune $VC(p)$
$p' \in S^+$	$C_{p \mid p'}$ contains $G[i,j]$	Prune $(p', R_{p \mid p'})$
$p' \in S^{--}$	$C_{p' \mid p}$ contains $G[i,j]$	Prune $VC(p)$
$p' \in S^{--}$	$G[i,j]$ is outside of $C_{p' \mid p}$	Prune $(p', R_{p \mid p'})$
$p' \in S^0$	$H_{p' \mid p}$ contains $G[i,j]$	Prune $VC(p)$
$p' \in S^0$	$H_{p \mid p'}$ contains $G[i,j]$	Prune $(p', R_{p \mid p'})$

Figure 5. Cell Pruning & Pair Pruning

root be $root$ with $root.R = U$ and $S(root.R) = \{VC(p_1), \cdots, VC(p_\ell)\}$, the recursion is initiated over each child node of $root$ with $level = 1$. For each node, we compute its Voronoi cell set only from that of its parent (Line 3). If the set contains only one Voronoi cell $V_{node.R}(p)$, then for any location in $node.R$, p is its weighted NN. We stop recursion in that case (Line 5). Otherwise, if the current level is not the leaf level, we continue to split $node$ and construct its four children (Lines 6-8).

After the quadtree $qtree$ is constructed, for all its nodes $node$, $S(node.R)$ is already computed. Then, for each grid cell $G[i,j]$ with region R, we compute the set of trajectory points whose Voronoi cells overlap with R, denoted by $C_T[i,j]$. We compute $C_T[i,j]$ by finding the leaf node, $leaf$, that contains the center of R using $qtree$; and for each $VC_{leaf.R}(p) \in S(leaf.R)$, we add the corresponding trajectory point p into $C_T[i,j]$.

It is easy to see that, for any query location in R, its weighted NN must be some trajectory point in $C_T[i,j]$. We call $C_T[i,j]$ as the candidate set of $G[i,j]$ from now on. For each trajectory T, we store C_T, which is an $N \times N$ array of trajectory point lists, on disk as the random access index $RAI[T]$.

Given a query point q, we identify the grid cell $G[i,j]$ that q locates in, load the list $C_T[i,j]$ into memory, and compute $d(q,T)$ as follows:

$$d(q,T) = \min_{p \in C_T[i,j]} \{d(q,p)\}. \tag{13}$$

Compared with loading the whole trajectory T in memory, it is more efficient to obtain $d(q,T)$ using this random access index, since $|C_T[i,j]|$ is much smaller than the trajectory length ℓ. Therefore, in our implementation, we use this index to compute

$d(q,T)$ instead of accessing T directly (recall Lines 17-18 of Algorithm 3 and Line 11 of Algorithm 4).

Sequential Access Index (SAI). For each grid cell $G[i,j]$, we also build a list $L[i,j]$ for retrieving trajectories T in non-decreasing order of $d(q,T)$, where query point q locates in $G[i,j]$. Since q can be any location in $G[i,j]$, the value of $d(q,T)$ is not fixed beforehand. We compute the lower bound of $d(q,T)$ instead, denoted by $LB(q,T)$, which is given by:

$$LB(q,T) = \min_{p \in C_T[i,j]} \left\{ \frac{mindist(p,R)}{w(p)} \right\}, \tag{14}$$

where R is the region of $G[i,j]$.

Each trajectory T has an entry in $L[i,j]$, represented by $en(T) = (T, C_T[i,j], LB(q,T))$. The list $L[i,j]$ is constructed by sorting the entries in non-decreasing order of $LB(q,T)$. We store the $N \times N$ list array L on disk.

Given a query point q that falls in $G[i,j]$, in order to retrieve trajectories in non-decreasing order of $d(q,T)$ using $L[i,j]$, we maintain a priority queue *min-heap* in main memory. We get the next trajectory T with the smallest value of $d(q,T)$ in two steps:

- We read the next entry $en(T)$ from $L[i,j]$, evaluate $d(q,T)$ using Equation (13), and add $(T, d(q,T))$ into *min-heap*. The process is repeated until the value $d(q,T')$, where $T' =$ *min-heap*$.top()$, is smaller than the $LB(q,T)$ of the last accessed entry $en(T)$. Note that all subsequent entries have lower bound values larger than $d(q,T')$.
- We return $T' =$ *min-heap*$.top()$ as the next NN, and remove it from *min-heap*.

The priority queue *min-heap* is a memory buffer that reorders the trajectories in $L[i,j]$ by $d(q,T)$, and we call it as the sequential access index of $G[i,j]$, denoted by $SAI[i,j]$.

Grid-based Algorithms. Our two grid-based algorithms also follow the FA and TA frameworks, respectively, but use the grid index (i.e., the RAI and SAI) in replace of the R-tree index.

The grid-based FA differs from Algorithm 3 in the following aspects:

- Line 2 now becomes "$N \leftarrow n$", where $n = |D|$;
- Line 5 is now replaced by "retrieve the d-th NN of q_i using $SAI[j,k]$, where q_i falls in $G[j,k]$";
- Line 6 is no longer necessary since $SAI[j,k]$ directly returns the trajectory T along with $d(q_i,T)$;
- Lines 9 and 12 now become "$T[i] \leftarrow d(q_i,T)$";
- We no longer need to do the checking in Line 11, since each T will be accessed only once for each query point q_i.

The grid-based TA differs from Algorithm 4 in the following aspects:

- Line 2 now becomes "$N \leftarrow n$";
- Line 7 is now replaced by "retrieve the d-th NN of q_i using $SAI[j,k]$, where q_i falls in $G[j,k]$";
- Line 8 is no longer necessary;
- Line 10 now becomes "$T[i] \leftarrow d(q_i,T)$";

Extension to Skewed Trajectory Distribution. Our current algorithm uses a uniform grid to partition the rectangular data space U. Our experiments show that our algorithm works quite well on the datasets with trajectories relatively uniformly distributed over U. However, it is not the best choice when the trajectory distribution is skewed.

Although the road network of most regions occupies the majority of the region's bounding box U (e.g., Colorado), it is not always true. For example, in the bounding box of Florida, most regions correspond to the ocean where no trajectory can exist, and it is meaningless to divide such regions into grid cells. Furthermore, there are usually much more trajectory points in city centers than in outskirts, and thus dense regions should be divided into finer granularity.

We proposes a heuristics to handle data skewness. Specifically, we first build a linear quadtree index over all the trajectory points. Then, we build our RAI and SAI indices over the leaf nodes of the linear quadtree. We have conducted experiments to compare the performance of using uniform grid with that of using linear quadtree over skewed trajectory data, and found that the latter is an order of magnitude faster than the former, and achieves similar performance compared with using uniform grid over relatively uniform trajectory data.

7. Experimental Results

In this section, we evaluate the performance of our algorithms: *RTree-TA*, *RTree-FA*, *Grid-TA*, and *Grid-FA*. We implemented our algorithms in JAVA. All the experiments were run on a public Linux server with eight 3GHz Intel CPU and 32GB memory.

7.1. Datasets and Query-sets

We first describe the datasets and query-sets used in our experiments.

Datasets. We evaluate the efficiency of our algorithms and the quality of the result trajectories using four real trajectory datasets that are publicly available:

- *Trucks*[1]: This dataset consists of 276 trajectories of 50 trucks delivering concrete to several construction places around Athens metropolitan area in Greece for 33 distinct days.
- *SchoolBuses*[2]: This dataset consists of 145 trajectories of 2 school buses collecting (and delivering) students around Athens metropolitan area in Greece for 108 distinct days.
- *Geolife*[3]: This GPS trajectory dataset was collected in MSRA Geolife project by 182 users in a period of over five years (from April 2007 to August 2012).
- *T-Drive*[4]: This dataset contains the GPS trajectories of 10,357 taxis during the period of Feb. 2 to Feb. 8, 2008 within Beijing.

For the two small datasets *Trucks* and *SchoolBuses*, the length of the trajectories is in the order of hundreds. We choose these datasets for empirical evaluation since there

[1] http://www.chorochronos.org/?q=node/5
[2] http://www.chorochronos.org/?q=node/6
[3] http://research.microsoft.com/en-us/downloads/b16d359d-d164-469e-9fd4-daa38f2b2e13/default.aspx
[4] http://research.microsoft.com/apps/pubs/?id=152883

exists some important locations in their underlying applications, such as construction places and schools.

Unlike the *Trucks* and *SchoolBuses* datasets where most trajectories have a sampling rate of every 30 seconds, the sampling rate of the large datasets (i.e., *Geolife* and *T-Drive*) varies a lot. Specifically, the *Geolife* dataset contains 10,373 trajectories recorded by different GPS loggers and GPS phones with a variety of sampling rates, where 91.5% of the trajecotries are logged in a dense representation, like every 1–5 seconds or every 5–10 meters per trajectory point. As a result, the length of the trajectories is in the order of thousands to tens of thousands. For the *T-Drive* dataset, the sampling rate is much lower, with the average sampling interval being about 177 seconds with a distance of about 623 meters. The length of the trajectories in *T-Drive* is in the order of thousands.

The dense sampling rate of *Geolife* is not useful, since two consecutive trajectory points that are 1 second or 1 meter apart usually refer to the same Point of Interest (POI). Therefore, we re-sample the trajectories as follows:

- We always sample the first trajectory point;
- If the next trajectory point is less than 30 seconds or 1 meter apart from the last sampled trajectory point, we skip the trajectory point.

We also do the trajectory re-sampling over *T-Drive* despite its low sampling rate, because we find that *T-Drive* records a lot of samples for a taxi even when it stops, and we want to remove the redundant samples that refer to the same stop locations. After the re-sampling process, the length of the trajectories in both datasets becomes shorter, in the order of thousands.

Another issue with the large datasets (i.e., *Geolife* and *T-Drive*) is that, the spatial distribution of the trajectory points is highly skewed. For *Geolife*, the majority of the data was created in Beijing, China. However, some trajectory points may locate in other cities like those in the USA and Europe. If we use our grid-based indexing approach, the majority of trajectory points in Beijing are clustered in only a small number of grids while many grids are empty. Moreover, a traveler usually use the k-ICT query to find some reference trajectories for planning a trip to multiple POIs in an unfamiliar city such as Beijing, and it is unlikely that the trajectory points abroad are of any interest. Therefore, for *Geolife*, we only use the 8,726 trajectories that are totally in the 5-th ring road of Beijing, which account for 85.75% of the trajectories in the dataset. As for *T-Drive*, we eliminate the skewed spatial distribution of trajectory points similarly, by using only the 7,450 trajectories that are totally inside Beijing, which account for 71.93% of the trajectories in the dataset.

Query-sets. We do not generate query locations randomly, since trajectories usually follow the underlying road network. Moreover, a query location in a sparse region not covered by the road network is meaningless in real applications.

We generate a meaningful query-set containing m query locations in the following way: (1) randomly pick a trajectory from the trajectory dataset to query over; (2) pick the top-10% points of the trajectory in terms of importance; (3) randomly select m locations from these points without replacement; (4) shift these locations in a random direction by a small randomly generated distance (within 200m), and add them to the query-set.

In this way, we are generating meaningful query locations which are important and correlated for at least one trajectory in the dataset.

7.2. Evaluation Measures

The k-ICT query has two query parameters: (1) the number of query points, m; and (2) the number of trajectories, k, that the user wants the query to return. These parameters are usually small in real applications. We also have a parameter for the dataset, which is the number of trajectories, n.

We measure the following four costs of our algorithms when the above parameters change: (1) CPU time; (2) number of blocks accessed by sequential index (the Max R-tree, or the grid index SAI); (3) number of blocks accessed by random index (the grid index RAI); (4) number of priority queue entries in main memory.

Since our algorithms are I/O bound, the number of blocks accessed by sequential/random indices are the most important performance criteria. When using the grid index $SAI[i, j]$ for a query point locating in $G[i, j]$, we maintain a main memory buffer of one block which is refilled from $L[i, j]$ whenever it is used up. Therefore, we can use the number of blocks accessed to evaluate the I/O cost. As for R-tree, the nodes are loaded in blocks, and thus the number of blocks accessed can be measured.

The smaller memory a query requires, the more queries a server can handle simultaneously. Therefore, we also measure the memory cost of our algorithms. For the R-tree based algorithms, the memory cost is dominated by the priority queue *min-heap* used for NN search (see Algorithm 2), while for grid-based ones, the memory cost is dominated by the priority queue of $SAI[i, j]$ for reordering $L[i, j]$ (see Section 6.2). The total number of memory entries equals the sum of the entries in the priority queue for each query point q_i, and we report the maximum number among all the round-robin iterations.

For the *Trucks* and *SchoolBuses* datasets, we manually mark the points of some trajectories with human-specified importance values, and then find the values of r and α that best fits these marked data using Equation (6). We find that the choice of $r = 50$ m and $\alpha = 0.002$ effectively distinguishes between the important and unimportant trajectories points, and we use these parameter values when generating sample importance by the method discussed in Section 4.

The trajectories in the *Geolife* dataset have various speed, as they records different outdoor movements such as walking, biking, or driving. Therefore, we fix $r = 50$ m but use various values of α for different trajectories. Note that if we fix $\alpha = 0.002$ which is appropriate for a vehicle trajectory, the importance scores of most trajectory points in the trajectory of a walking person are close to 1, although many of them are not important. We use the following approach to set α of a trajectory, which is observed to well characterize the location importance of the trajectories: in Equation (6), we set α such that when $\Delta t(p_i)$ (when $r = 50$ m) equals the longest one among all points p_i in the trajectory, the score $w(p_i) = 0.99$.

As for the *T-Drive* dataset, the sampling rate is very low and thus a radius of $r = 50$ m is not able to cover a sufficient number samples around the current trajectory point. We set the parameters as follows, which is observed to well characterize the location importance of the trajectories: r is fixed to 500 m and we set α such that when $\Delta t(p_i) = 500$ s (when $r = 500$ m), the score $w(p_i) = 0.99$.

Throughout the experiments, we fix the size of a block as 512 bytes. We generate 1000 queries in each experiment, and all results are averaged over the 1000 runs.

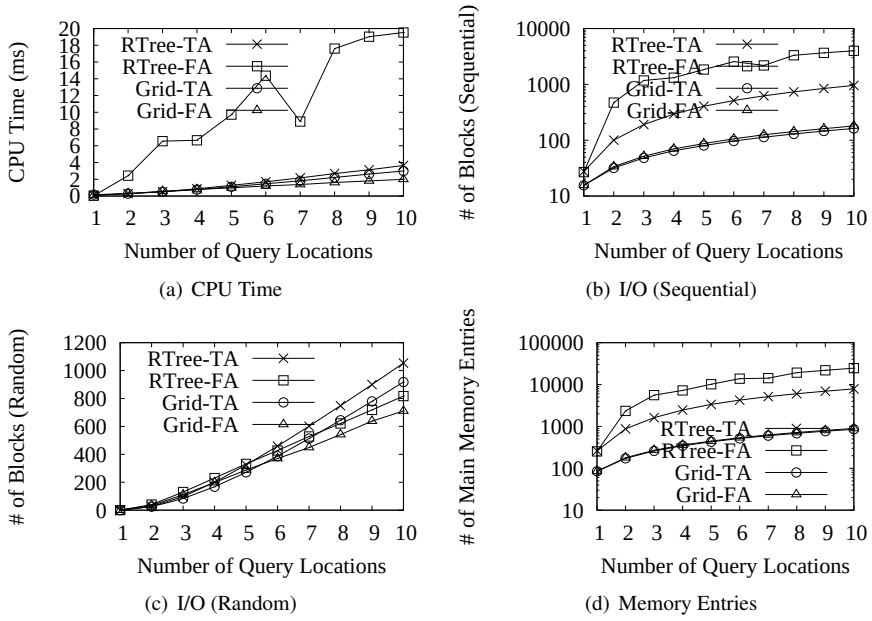


Figure 6. Effect of m using the Trucks dataset

7.3. Effect of Query Parameters

We first study the performance of our algorithms with respect to the query parameters m and k. For the relatively small *Trucks* and *SchoolBuses* datasets, we build grid indices by constructing a quadtree of height $h = 5$. Accordingly, the grid we use is of size 32×32. For the large datasets *Geolife* and *T-Drive*, we build a grid index by constructing a quadtree of height $h = 7$. Accordingly, the grid we use is of size 128×128.

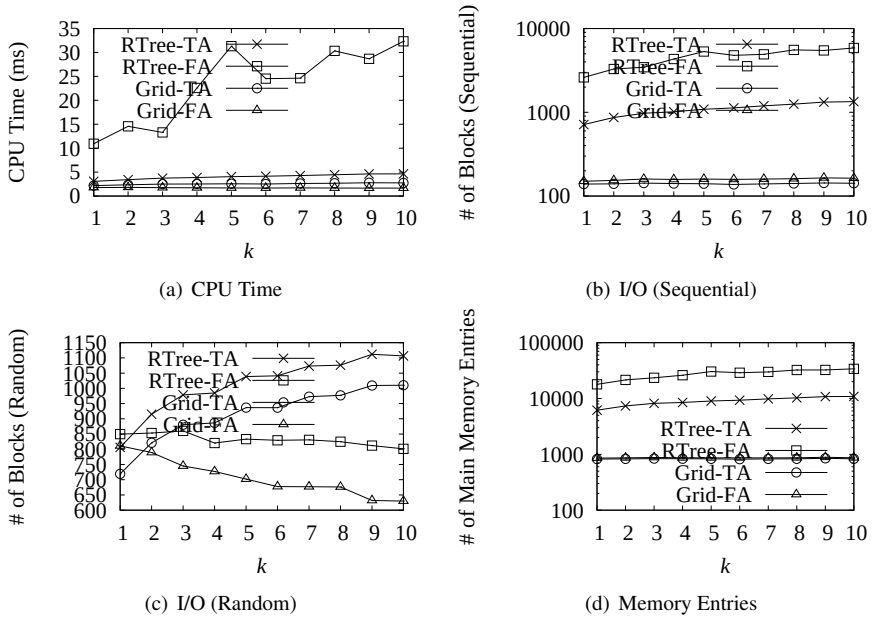
To study the effect of m, we fix k as 5 and process queries with $m = 1, 2, \ldots, 10$. On the other hand, to study the effect of k, we fix m as 3 and process queries with $k = 1, 2, \ldots, 10$.

Figure 6 reports the performance of our algorithms for processing k-ICT queries over the *Trucks* dataset when $k = 5$ and the number of query points m increases from 1 to 10.

Figure 6(a) shows that the CPU time of *RTree-FA* is much larger than the other three algorithms, while the grid-based algorithms record the shortest CPU time.

Since all our algorithms are I/O bound, the results reported in Figure 6(b) and (c) dominate the overall performance of query processing. According to Figure 6(b), *RTree-FA* requires reading a lot of blocks (or R-tree nodes) for the incremental NN search, and both of the R-tree based algorithms read significantly more blocks for sequential access than the grid-based algorithms. For example, when $m = 5$, *RTree-FA* reads over 1844 blocks while *Grid-FA* reads only 87 blocks. For random access, Figure 6(c) shows that *Grid-FA* (or respectively, *Grid-TA*) also reads fewer blocks than *RTree-FA* (or respectively, *RTree-TA*), though the difference is not as big as in the case of sequential access.

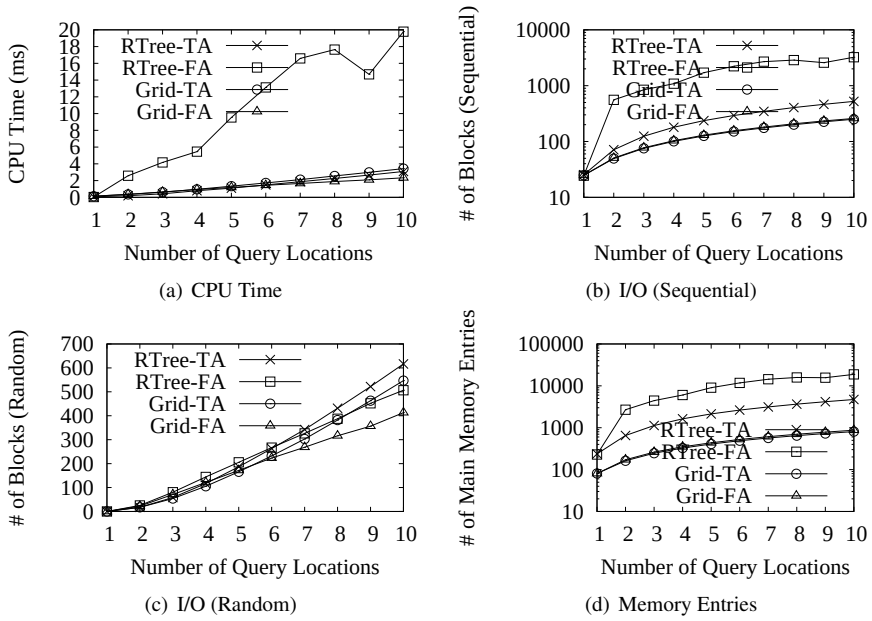
Overall, *Grid-TA* is around 1.3 times faster than *Grid-FA*, several times faster than *RTree-TA*, and an order of magnitude faster than *RTree-FA*.



(a) CPU Time (b) I/O (Sequential)

(c) I/O (Random) (d) Memory Entries

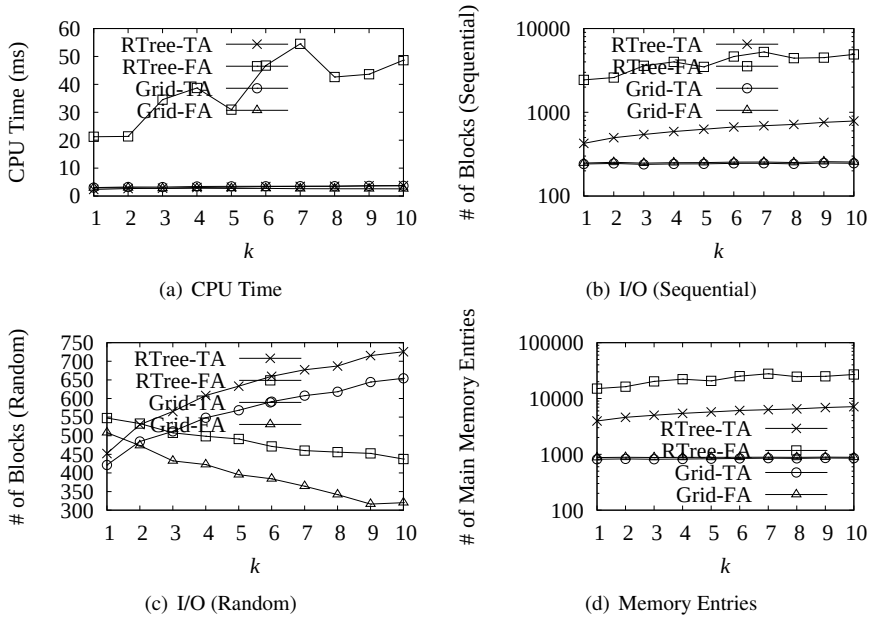
Figure 7. Effect of k using the Trucks dataset



(a) CPU Time (b) I/O (Sequential)

(c) I/O (Random) (d) Memory Entries

Figure 8. Effect of m using the SchoolBuses dataset



(a) CPU Time (b) I/O (Sequential)

(c) I/O (Random) (d) Memory Entries

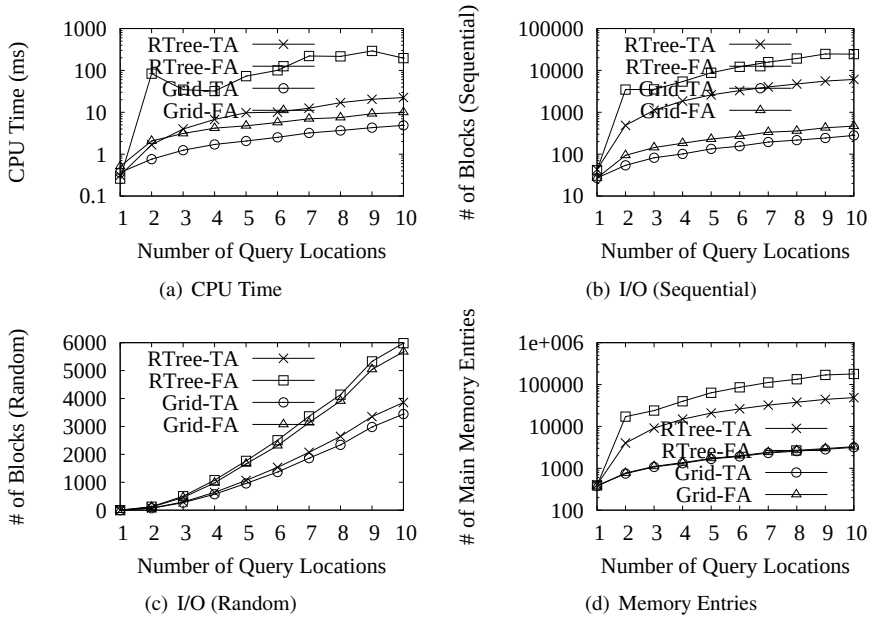
Figure 9. Effect of k using the SchoolBuses dataset

Figure 6(d) shows that the number of data entries maintained in memory by *RTree-TA* and by *RTree-FA* is from several times to tens of times larger than that by both of the grid-based algorithms. Given the fact that the size of an entry maintained by the grid index is much smaller than an R-tree node entry (which contains MBR and weight besides the node pointer), the grid-based algorithms are much more memory-efficient than the R-tree based ones.

Figure 7 reports the performance of our algorithms over the *Trucks* dataset when $m = 3$ and k increases from 1 to 10. The results are similar to that of increasing m we just discussed, except for the I/O cost of random access. As shown in Figure 7(c), the two FA-based algorithms, *RTree-FA* and *Grid-FA*, read fewer blocks when k increases, while the two TA-based algorithms, *RTree-TA* and *Grid-TA*, read more blocks when k increases. This is because FA adopts a filter(sequential access)-and-refine(random access) approach. A larger k requires that FA do more sequential accesses, and since more trajectories are accessed, the need for random access is reduced.

As for the *SchoolBuses* dataset, Figure 8 reports the performance of our algorithms when m changes, and Figure 9 reports the performance of our algorithms when k changes. It can be observed that the performance trend of the algorithms is similar to that of the *Trucks* dataset discussed above (we thus omit the details).

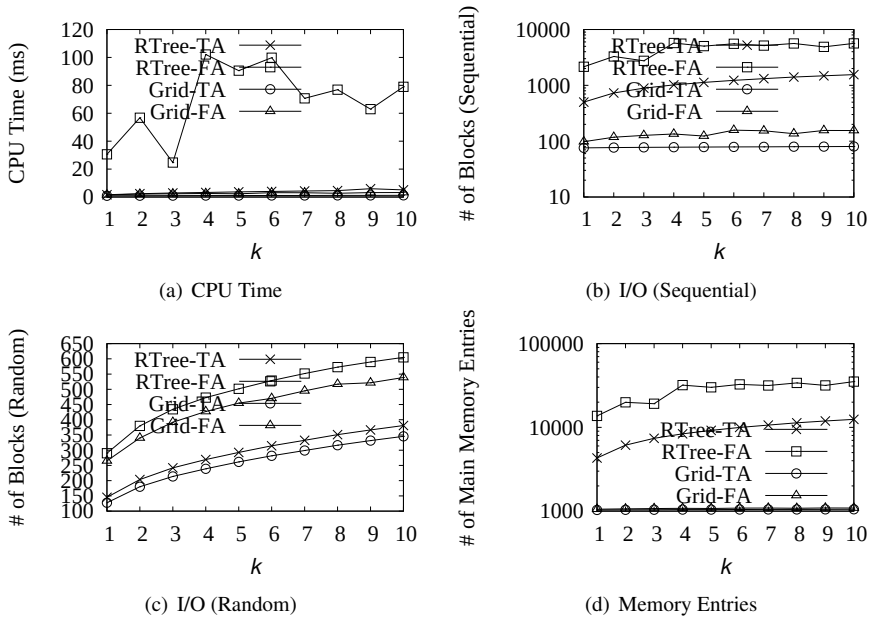
For the *Geolife* dataset, Figure 10 reports the performance of our algorithms when m changes, and Figure 11 reports the performance of our algorithms when k changes. The performance trend of the algorithms is mostly similar to that of the *Trucks* and *SchoolBuses* datasets discussed above, except for the random IO cost shown in Figure 10(c) and Figure 11(c), where we can see that the FA-based algorithms incur much more random IO cost than the TA-based algorithms. This shows that the effectiveness of TA over FA is more prominent for a larger trajectory dataset. Also, Figure 11(c) shows that the random IO cost of the FA-based algorithms read more blocks when k increases,



(a) CPU Time (b) I/O (Sequential)

(c) I/O (Random) (d) Memory Entries

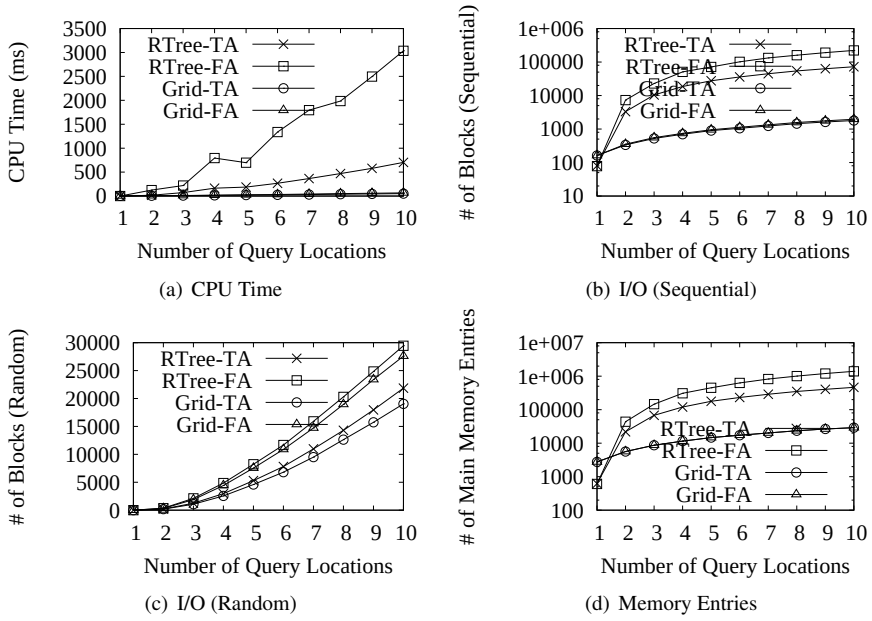
Figure 10. Effect of m using the Geolife dataset



(a) CPU Time (b) I/O (Sequential)

(c) I/O (Random) (d) Memory Entries

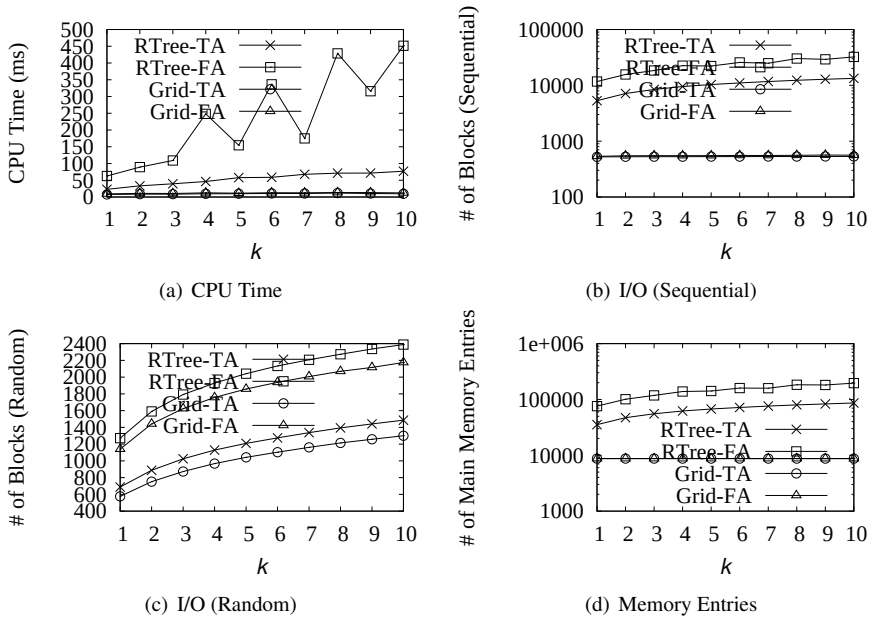
Figure 11. Effect of k using the Geolife dataset



(a) CPU Time (b) I/O (Sequential)

(c) I/O (Random) (d) Memory Entries

Figure 12. Effect of m using the T-Drive dataset



(a) CPU Time (b) I/O (Sequential)

(c) I/O (Random) (d) Memory Entries

Figure 13. Effect of k using the T-Drive dataset

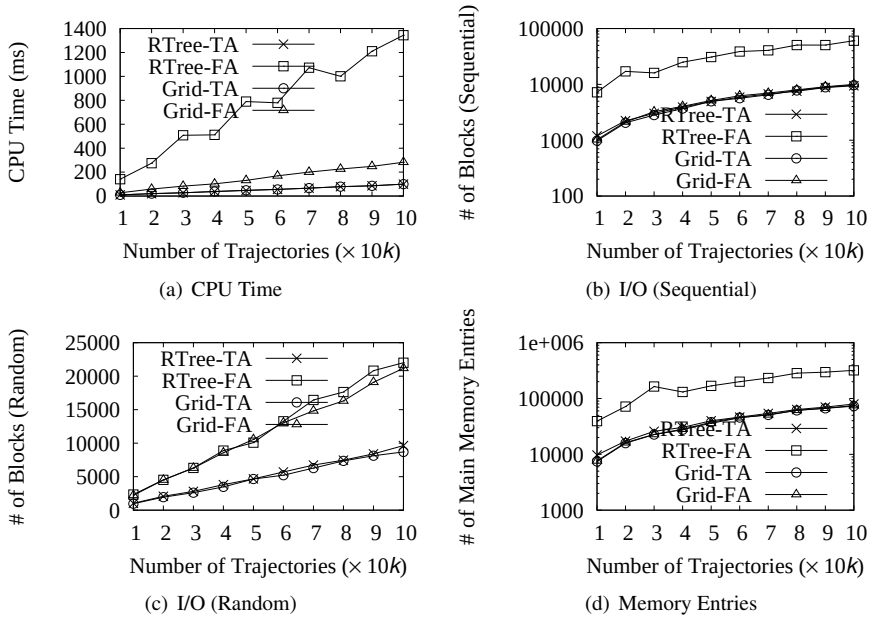


Figure 14. Scalability results

which is different from that observed from Figure 7(c) and Figure 9(c). This is because the filtering phase of FA is not as effective on *Geolife* as on the other two datasets, and thus, the refinement phase still requires more random accesses as k increases. Overall, *Grid-TA* is around twice faster than *Grid-FA*, 3–19 times faster than *RTree-TA* and 20–70 times faster than *RTree-FA*.

As for the *T-Drive* dataset, Figure 12 reports the performance of our algorithms when m changes, and Figure 13 reports the performance of our algorithms when k changes. It can be observed that the performance trend of the algorithms is similar to that of the *Geolife* dataset discussed above. Overall, *Grid-TA* is around 1.3 times faster than *Grid-FA*, 3–13 times faster than *RTree-TA* and 10–50 times faster than *RTree-FA*.

7.4. Results of Scalability Test

To study the scalability of our algorithms when the number of trajectories increases, we generate synthetic datasets based on the *Trucks* dataset. Specifically, to generate a dataset with n trajectories, we repeat the following operations n times: (1) randomly pick a trajectory from the *Trucks* dataset; (2) shift it in a random direction by a small randomly generated distance (within 200m); (3) insert the new trajectory into the synthetic dataset. We generate synthetic datasets from a real dataset since we want the generated trajectories to exhibit the properties of real trajectories.

We generate synthetic datasets D with $|D| = 10k, 20k, \cdots, 100k$, and process queries with $m = 3$ and $k = 5$. The grid indices are built by constructing a quadtree of height $h = 6$, and accordingly, the grid is of size 64×64.

Figure 14 shows the scalability of our algorithms when the number of trajectories increases. We can see from Figure 14(c) that when the data size is large, FA-based algorithms require reading many more blocks using random access than TA-based al-

	Traj ID	Weighted Distance Sum	Distance Sum	$\Delta t(p_i)$		
				q_1's Match	q_2's Match	q_3's Match
Top-1	231	1155.21	107.90 m	15709.34 s	3097.60 s	3761.37 s
Top-2	24	1737.65	208.38 m	695.38 s	2086.14 s	1469.69 s
Top-3	214	2612.54	270.07 m	8913.50 s	1643.62 s	57.59 s
Top-4	89	3031.17	202.23 m	2226.35 s	790.20 s	2275.79 s
Top-5	274	3189.64	2086.82 m	3382.07 s	916.64 s	63.22 s

(a) Top-5 Trajectories Found by Sum-of-Weighted-Distance

	Traj ID	Weighted Distance Sum	Distance Sum	$\Delta t(p_i)$		
				q_1's Match	q_2's Match	q_3's Match
Top-1	231	1155.21	107.90 m	4.46 s	3097.60 s	3761.37 s
Top-2	89	3031.17	202.23 m	3.28 s	30.53 s	27.53 s
Top-3	24	1737.65	208.38 m	4.47 s	63.99 s	12.94 s
Top-4	83	3355.39	236.52 m	3.49 s	63.22 s	5.36 s
Top-5	269	3680.68	242.84 m	3.78 s	31.61 s	7.36 s

(b) Top-5 Trajectories Found by Sum-of-Distance

Figure 15. Top-5 Result Trajectories

gorithms do. Otherwise, the performance trend is quite consistent with the results on the *Trucks* dataset reported in Section 7.3.

Overall, our algorithms scale well with the data size. *Grid-TA* is slightly (less than 10%) faster than *Grid-FA*, 2.7–3 times faster than *RTree-TA* and 13–18.5 times faster than *RTree-FA*.

7.5. Quality of Trajectory Answers

So far, we have only studied the performance of our algorithms. In this subsection, we compare the quality of the trajectories found by our sum-of-weighted-distance measure with that of the traditional sum-of-Euclidean-distance (or simply, sum-of-distance) measure. We use the *Trucks* dataset for quality evaluation, since it only contains the trajectories of trucks, and thus, the result trajectories exhibit similar characterizations.

We now report the result when $m = 3$ and $k = 5$. Figure 15(a) shows the top-5 trajectories found by our sum-of-weighted-distance measure (i.e., a 5-ICT query), for a randomly generated query-set with three query locations. For each result trajectory in Figure 15(a), we show its trajectory ID, the values of sum-of-weighted-distance and sum-of-distance. Let us define the match of a query point q_j in trajectory T as the point $p_i \in T$ closest to q_j in terms of weighted distance. Then, for each trajectory in Figure 15(a), we also show $\Delta t(p_i)$ of the match p_i of each of the three query locations q_j. Obviously, the top-1 trajectory is of high quality, since (1)the sum-of-distance is only 107.90 meters, which means that the trajectory is physically close to each query location, and (2)$\Delta t(p_i)$ is long for the match p_i of each query location q_j, which means that the truck spent a while in the 50-meter-radius neighborhood $Cir(p_i)$ and thus p_i is important. Similarly, the other four trajectories shown in Figure 15(a) have relatively high quality and is likely to be helpful to the user.

For the same query locations, we also find the top-5 trajectories found by the tra-

	Top-1	**Top-2**	**Top-3**	**Top-4**	**Top-5**
Weighted Distance Sum	6552.83 s	6609.30 s	6963.46 s	7034.18 s	7276.55 s
Distance Sum	1064.17 s	905.08 s	1002.74 s	759.88 s	802.34 s

Figure 16. Avg Sum-of-$\Delta t(p_i)$ of Top-k Trajectories on *Truck*

m	**1**	**2**	**3**	**4**	**5**
Weighted Distance Sum	1675.23 s	6907.76 s	6416.46 s	5468.21 s	10525.52 s
Distance Sum	3.30 s	3460.82 s	24.82 s	3398.88 s	4239.07 s

Figure 17. Avg Sum-of-$\Delta t(p_i)$ of Top-1 Trajectories on *Truck* with Different Query Parameter m

ditional sum-of-distance measure, which are shown in Figure 15(b). In Figure 15(b), we define the match of a query point q_j in trajectory T as the point $p_i \in T$ closest to q_j in terms of Euclidean distance. We can see that most matching trajectory points p_i has a small value of $\Delta t(p_i)$, which means that these locations are not very important. The sum-of-distance measure fails to find important trajectories like Trajectory 214, and even though it finds some important trajectories like Trajectory 89, the matches of the query locations are of low quality.

We now show that sum-of-weighted-distance is superior to sum-of-distance in general, by randomly generating 1000 queries and report the quality measures averaged over the 1000 query results. We define the quality of a trajectory as the sum of $\Delta t(p_i)$ for all matches p_i of the query locations q_j, and a larger value implies a higher quality. Intuitively, sum-of-$\Delta t(p_i)$ represents the total time spent by the trajectory at the locations of interest. Figure 16 shows the average sum-of-$\Delta t(p_i)$ of the top-ith trajectory found by sum-of-weighted-distance and by sum-of-distance, where we set the query parameter $m = 3$. The figure clearly shows that trajectories found by sum-of-weighted-distance have higher quality. We also test the average sum-of-$\Delta t(p_i)$ of the top-1 trajectory found by sum-of-weighted-distance and by sum-of-distance, by varying the query parameter m. The result is shown in Figure 17, which also confirms that trajectories found by sum-of-weighted-distance have higher quality.

We also compare the quality of the trajectories found by our sum-of-weighted-distance measure with that of the sum-of-distance measure on the *T-Drive* dataset. Figure 18 shows the average sum-of-$\Delta t(p_i)$ of the top-ith trajectory found by sum-of-weighted-distance and by sum-of-distance, where the query parameter $m = 3$. Figure 19 shows the average sum-of-$\Delta t(p_i)$ of the top-1 trajectory found by sum-of-weighted-

	Top-1	**Top-2**	**Top-3**	**Top-4**	**Top-5**
Weighted Distance Sum	1734.40 s	1936.08 s	1782.64 s	1776.27 s	1750.20 s
Distance Sum	209.22 s	40.92 s	92.89 s	43.21 s	62.21s

Figure 18. Avg Sum-of-$\Delta t(p_i)$ of Top-k Trajectories on *T-Drive*

m	1	2	3	4	5
Weighted Distance Sum	551.11 s	973.96 s	1634.67 s	2582.91 s	4131.28 s
Distance Sum	63.14 s	181.73 s	217.65 s	283.85 s	463.04 s

Figure 19. Avg Sum-of-$\Delta t(p_i)$ of Top-1 Trajectories on *T-Drive* with Different Query Parameter m

distance and by sum-of-distance, by varying the query parameter m. Both figures confirm that trajectories found by sum-of-weighted-distance have higher quality.

7.6. Summary of Experimental Results

To sum up, we have the following observations: (1) the grid-based algorithms are significantly more efficient than the R-tree based algorithms; (2) the TA-based algorithms are more efficient than the FA-based algorithms; and (3) *Grid-TA* is much faster than the other three algorithms on large datasets.

8. Conclusions

We proposed the new problem of k *Important Connected Trajectories* (k-ICT) query processing over trajectories with location importance. We designed effective methods to infer the importance of trajectory locations from the temporal information, and developed four algorithms to answer the queries: two based on the R-tree index, and the other two based on an efficient grid index. The R-tree index based algorithms are adaptations of the algorithms in Chen et al (2010) to querying trajectories with location importance. However, the R-tree index only captures the spatial aspects of the trajectory points, and location weights are only considered during R-tree querying. On the other hand, our grid index includes the location weights as first-class citizen, and is thus more suitable for querying trajectories with location importance.

We showed by experiments on both real and synthetic datasets that our algorithms are efficient for answering k-ICT queries. The grid index based algorithms are especially efficient in terms of both time and space: they incur one to two orders of magnitude less sequential IO cost and computational overhead compared with R-tree index based algorithms, due to the more effective pruning power of the grid index. As for trajectory traversal, TA is more effective than FA since the aggressive strategy of TA tightens the pruning threshold much faster. Overall, the combination of TA with grid index offers the best performance.

Acknowledgements. We thank the reviewers for giving us many constructive comments, with which we have significantly improved our paper. This research is supported in part by GRF grant HKUST 617610, SHIAE Grant No. 8115048 and MSRA Grant No. 6903555.

References

Chen Z, Shen HT, Zhou X, Zheng Y and Xie X (2010) Searching trajectories by locations - an efficiency study. Proceedings of the 2010 ACM SIGMOD international conference on management of data (SIGMOD), June 2010, pp 255–266.

Yi BK, Jagadish H and Faloutsos C (1998) Efficient retrieval of similar time sequences under time warping. Proceedings of the 14th IEEE International Conference on Data Engineering (ICDE), February 1998, pp 201–208.

Vlachos M, Kollios G and Gunopulos D (2002) Discovering similar multidimensional trajectories. Proceedings of the 18th IEEE International Conference on Data Engineering (ICDE), February 2002, pp 673–684.

Chen L and Ng R (2004) On the marriage of lp-norms and edit distance. Proceedings of the 30th International Conference on Very Large Data Bases (VLDB), August 2004, pp 792–803.

Chen L, Özsu MT and Oria V (2005) Robust and fast similarity search for moving object trajectories. Proceedings of the 2005 ACM SIGMOD international conference on management of data (SIGMOD), June 2005, pp 491–502.

Shang S, Ding R, Yuan B, Xie K, Zheng K and Kalnis P (2012) User oriented trajectory search for trip recommendation. Proceedings of the 15th International Conference on Extending Database Technology (EDBT), March 2012, pp 156–167.

Zheng K, Shang S, Yuan NJ and Yang Y (2013) Towards efficient search for activity trajectories. Proceedings of the 29th IEEE International Conference on Data Engineering (ICDE), April 2013, pp 230–241.

Cao X, Cong G and Jensen CS (2010) Mining significant semantic locations from GPS data. Proceedings of the 36th International Conference on Very Large Data Bases (VLDB), September 2010, pp 1009–1020.

Cao X, Cong G and Jensen CS (2010) Mining significant semantic locations from GPS data. Proceedings of the 36th International Conference on Very Large Data Bases (VLDB), September 2010, pp 1009–1020.

Yang Y, Gong Z and U LH (2011) Identifying points of interest by self-tuning clustering. Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), July 2011, pp 1009–1020.

Spaccapietra S, Parent C, Damiani ML, de Macêdo JA, Porto F and Vangenot C (2008) A conceptual view on trajectories. Data & Knowledge Engineering (DKE), 65(1): 126–146.

Tietbohl A, Bogorny V, Kuijpers B and Alvares LO (2008) A clustering-based approach for discovering interesting places in trajectories. Proceedings of the 2008 ACM Symposium on Applied Computing (SAC), March 2008, pp 863–868.

Rocha JAMR, Oliveira G and Bogorny V (2010) DB-SMoT: a direction-based spatio-temporal clustering method. 5th IEEE International Conference on Intelligent Systems (IS), July 2010, pp 114–119.

Fagin R, Lotem A and Naor M (2001) Optimal aggregation algorithms for middleware. Proceedings of the 20th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS), May 2001.

Lazaridis I and Mehrotra S (2001) Progressive approximate aggregate queries with a multi-resolution tree structure. Proceedings of the 2001 ACM SIGMOD international conference on management of data (SIGMOD), May 2001, pp 401–412.

OKabe A, Boots B, Sugihara K and Chiu SN (2009) Spatial tessellations, concepts and applications of Voronoi diagrams. Vol. 501. John Wiley & Sons, 2009.

Wu D, Yiu ML, Jensen CS and Cong G (2011) Efficient continuously moving top-k spatial keyword query processing. Proceedings of the 27th IEEE International Conference on Data Engineering (ICDE), April 2011, pp 541–552.

Tang LA, Zheng Y, Xie X, Yuan J, Yu X and Han J (2011) Retrieving k-nearest neighboring trajectories by a set of point locations. Advances in Spatial and Temporal Databases - 12th International Symposium (SSTD), August 2011, pp 223–241.

Vieira MR, Bakalov P and Tsotras VJ (2011) Querying trajectories using flexible patterns. Proceedings of the 13th International Conference on Extending Database Technology (EDBT), March 2010, pp 406–417.

Hadjieleftheriou M, Kollios G and Bakalov P (2005) Complex spatio-temporal pattern queries. Proceedings of the 31th International Conference on Very Large Data Bases (VLDB), September 2005, pp 877–888.

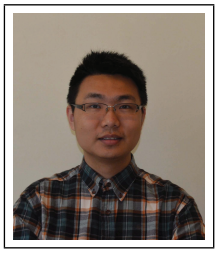
Zheng K, Zheng Y, Xie X and Zhou X (2012) Reducing uncertainty of low-sampling-rate trajectories. Proceedings of the 28th IEEE International Conference on Data Engineering (ICDE), April 2012, pp 1144–1155.

Yuan J, Zheng Y and Xie X (2012) Discovering regions of different functions in a city using human mobility and POIs. The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), August 2012, pp 186–194.

Author Biographies

Da Yan received his Ph.D. degree in Computer Science from the Hong Kong University of Science and Technology, Kowloon, Hong Kong, in 2014; and received his B.S. degree in Computer Science from Fudan University, Shanghai, in 2009. He is currently a postdoctoral fellow in the Department of Computer Science and Engineering, the Chinese University of Hong Kong. His research interests include distributed graph computing systems, cloud computing and big data, spatial data management, uncertain data management and data mining.

James Cheng is currently an Assistant Professor with the Department of Computer Science and Engineering at the Chinese University of Hong Kong (CUHK). Dr. Cheng received his Ph.D., M.Phil., and B.Eng. (First Class Honors) degrees in Computer Science from the Hong Kong University of Science and Technology in August 2008, 2004, and 2003, respectively. Before he joined CUHK, he was an Assistant Professor with the School of Computer Engineering at Nanyang Technological University, Singapore, from May 2009 to Dec 2012. His research focuses on large scale data analytics and distributed computing systems.



Zhou Zhao received his B.S. degree in Computer Science from the Hong Kong University of Science and Technology (HKUST), in 2010. He is currently a Ph.D. student in the Department of Computer Science and Engineering, HKUST. His research interests include data cleansing and data mining.

Wilfred Ng received his MS.c. (Distinction) and Ph.D. in Computer Science from the University of London. Currently he is an Associate Professor of Computer Science and Engineering at the Hong Kong University of Science and Technology, where he is a member of the database research group. His research interests are in the areas of databases, data mining and information Systems, which include Web data management and XML searching. Further Information can be found at the following URL: `http://www.cs.ust.hk/faculty/wilfred/index.html`.

Correspondence and offprint requests to: Da Yan, Department of Computer Science and Engineering, the Chinese University of Hong Kong, Shatin, N.T., Hong Kong. Email: yanda@cse.cuhk.edu.hk