Finding nearest neighbor with uncertain data using uncertain query

Abhay Mone, 810963969, amone@kent.edu

Introduction

Due to intrinsic uncertainty in the data, it is essential to provide meaningful answers with some probability. For example, uncertainty models have been proposed for moving object environments in order to help the address the problem. Suppose we can provide a guarantee that, at the time the query is evaluated, o_1 and o_2 could be no further than some distances d_1 and d_2 from their locations stored in the database, respectively [1]. With this information, we can state with confidence that o_1 is the nearest neighbor of q. In general, the uncertainty of the objects may not allow us to determine a single object as the nearest neighbor. Instead, several objects could have the possibility(probability) of being the nearest neighbor.

Project Description

The aim of this project was to find nearest neighbor of submitted uncertain query for uncertain data. Providing probabilistic answers to nearest-neighbor queries is difficult because, for nearest-neighbor queries with uncertain query point, the interplay between objects is critical and the probability that an uncertain object is the closest to the uncertain query is greatly influenced by the position and uncertainty of the other objects and query point itself. In this paper, I presented the technique for finding the neighbor queries with uncertain submitted query point. As an overview, proposed algorithm first eliminates all the objects that have no chance of being the nearest neighbor. Then, for every object that may be the nearest neighbor, its probability is evaluated by summing up the probability of being the nearest neighbor for all its possible locations.

Related Work

In the paper [1], author has discussed the method to calculate the nearest neighbor with uncertain objects where submitted query is a point query. However, in this paper for the first time, I have proposed a method where submitted query is uncertain with my best knowledge.

The problems of indexing and efficient access of spatiotemporal objects have been addressed in [2], [3], [4], [5], [6]. The issues of dynamic attributes indexing were discussed in [7]. The processing of nearest-neighbor queries in a moving-object environment is discussed in [8]. Song and Roussopoulos [8] investigate how to execute k-nearest neighbor queries for moving query point efficiently.

Motivation

The motivation for doing this project was primarily an interest in implementing the solution for a situation where a moving sheep goes out to rescue the other moving sheeps who are out of fuels. In such situation, with uncertain object, submitted query is also uncertain. Considering our motivation example, to rescue the sheeps, it is important or sensible to find nearest neighbor.

Problem Definition

In this section, we describe a model of uncertainty for moving objects. Based on this uncertainty model, we introduce the concept of probabilistic nearest-neighbor queries.

Definition 1. *An uncertainty region of an object O_i at time t, denoted by $U_i(t)$, is a closed region such that O_i can be found only inside this region.*

Definition 2. *The uncertainty probability density function of an object O_i , denoted by is a probability density function of O_i's location (x,y) at time t, that has a value of 0 outside $U_i(t)$.*

In the above definitions, we assume each object is a point, i.e., its spatial extents are not considered. Also, since $f_i(x, y, t)$ is a probability density function, it has the property that $\int_{U_i(t)} f_i(x, y, t) = 1$. We do not limit how the uncertainty region evolves over time or what the probability density function of an object is inside the uncertainty region. The only requirement for the probability density function is that its value is 0 outside the uncertainty region. A trivial probability density function is the uniform density function, which depicts the worst-case or "most uncertain" scenario. Usually, the scope of uncertainty is determined by the recorded location of the moving object, the time elapsed since its last update, and other application- specific assumptions. For example, one may decide that the uncertainty region of an object contains all the points within distance $(t - t_u) * v$ from its last reported position, where t_u is the time that the reported position was sent and v is the maximum speed of the object. One can also specify that the object location follows the Gaussian distribution inside the uncertainty region.

Definition 3. *PDF over circle : Probability that a randomly chosen point will fall between r and r + dr on a circle of radius R, we assume the all points on the circle are equally likely and is given by pr(r).*

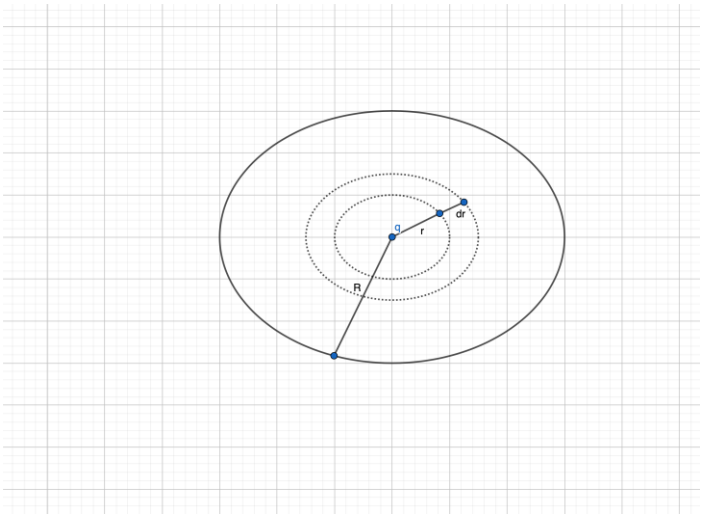

Figure 1

As shown in fig.1 above, assume a circle of radius R has divided into rings of length "dr". And we have given the probability density function of a point lies on one of the circular rings "dr".

Definition 4: *Probabilistic Nearest-Neighbor Query (PNNQ). For a set of n objects with uncertainty regions and probability density functions given at time t_0, a PNNQ for uncertain region Uq(r) with radius "r" is a query that returns a set of tuples of the form (O_i , p_i), where p_i is the nonzero probability that O_i is the nearest neighbor of uncertain region Uq(r) at time t_0.*

PNNQ = Prob(a point lies inside uncertain query region Uq(r)).

$\int_{ni}^{f}$ *Prob(Oi lies on the boundry of Cq(r')).Prob(other objects lie outside Cq(r'))dr'*

$$= pr(r). \int_{ni}^{f} pr_i(r). \prod_{k=1 \Lambda k \neq i}^{|S|} (1 - p_k(r))dr \quad \text{- (1)}$$

f = minimum of largest distance
ni= minimum distance from query point to each object
pri(r) = probability density function of r such that Oi is located on the boundry of Cq(r)
pi(r) = probability that Oi is located inside the circle Cq(r)

|S| = object set after pruning phase

Proposed Solution

Processing a PNNQ involves evaluating the probability of each object being closest to a uncertainty region query. We will explain parameters and theirs inference used in equation (1) step by step.

Pruning Phase:

The first objective is to find f, the minimum of the longest distances of the uncertainty regions from uncertainty region Uq(r), and eliminate any object with shortest distance to q larger than f. This algorithm is inspired from paper [1], where author eliminates the objects in the refining step.

1. **for** $i \leftarrow 1$ **to** $|S|$ **do**
   (a) Let n_i be the shortest distance of $U_i(t_0)$ from q
   (b) Let f_i be the longest distance of $U_i(t_0)$ from q
2. $f \leftarrow min_{i=1,\ldots,|S|} f_i$
3. $m \leftarrow |S|$
4. **for** $i \leftarrow 1$ **to** m **do**
   (a) **if** $(n_i > f)$ **then** $S \leftarrow S - O_i$
5. **return** S

Figure 2.

After applying pruning phase, we need to work on only those objects which are under the Cq(r) as show in the fig 3. For each element in |S|, there is no need to examine all portions in the uncertainty region. We only have to look at the regions that are located no farther than f from Ur(q). We do this conceptually by drawing a bounding circle C of radius f, centered at q. Any portion of the uncertainty region outside C can be ignored.

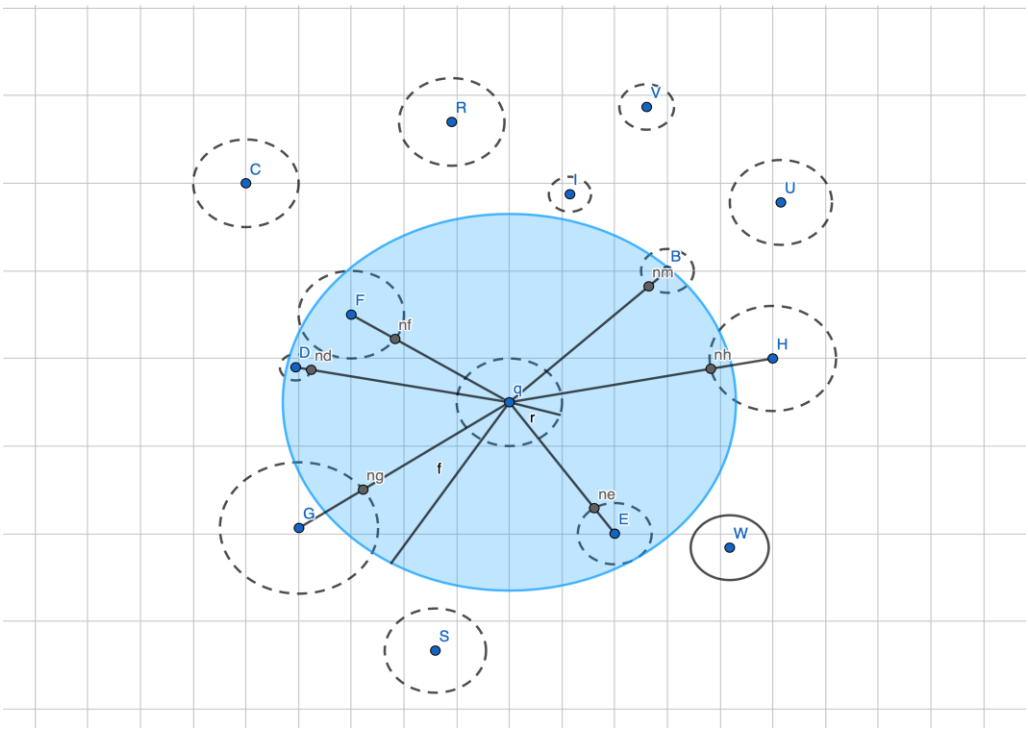

Figure 3 Pruning phase

Execution Phase:

Based on S and the bounding circle C, our aim is to calculate, for each object in S, the probability that it is the nearest neighbor of q. The solution is based on the fact that the probability of an object o being the nearest neighbor with distance r to q is given by the probability of o being at distance r to q times the probability that every other object is at a distance of r or larger from q. As introduced before, Pi(r) is the probability that O_i is located inside the uncertainty region, and pri(r) is probability density function of r such that Oi is located on the boundary of Cq(f) . Note that, in the fig 4, the uncertain query object lies on the boundary of uncertain region Uq(r). It can lie anywhere inside the uncertain region Uq(r). In the execution phase, we assume that the outer ring of length r' + dr [9], is divided into the ring of length dr' where r' = f – R . And we will find the probability for the uncertain object being nearest neighbor of uncertain region Uq(r) lies on the ring of length dr' which is given by,

$$= \sum_{j=i}^{|S|} pr(r). \int_{n_j}^{n_{j+1}} pr_i(r'). \prod_{k=1 \Lambda k \neq i}^{j} (1 - p_k(r'))dr' \quad \text{- (2)}$$

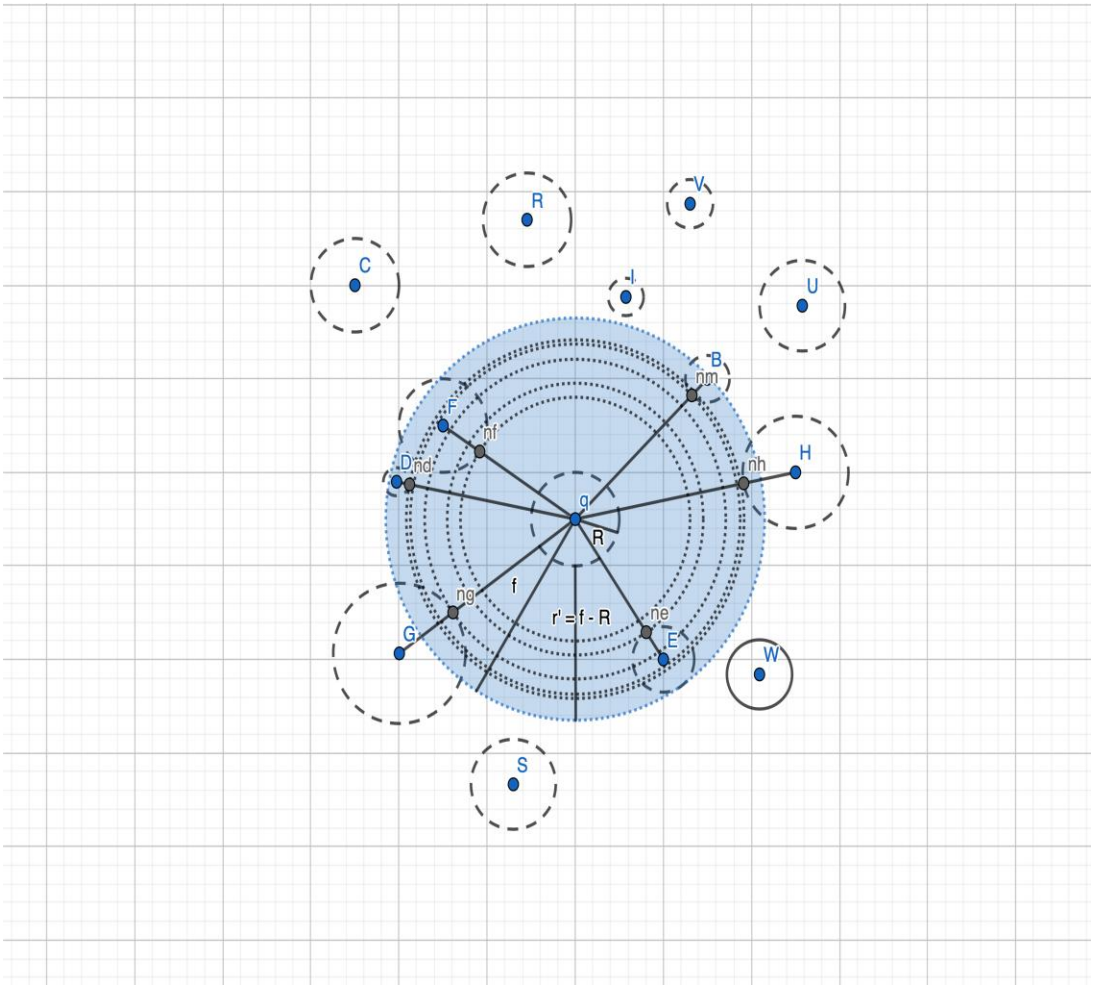

Figure 4 Execution Phase

Results and Future scope

Theoretically, I have proposed the method to calculate the nearest neighbor in uncertain region with submitted uncertain region. The solution is proposed considering we know the probability density function of each object. Although, in the pruning phase, we don't need to know the probability distribution(PDF) of objects. However, in the execution phase, given a query point in

the uncertain region, calculating nearest neighbor without knowing the probability distribution(PDF) of the query point will be in the future scope.

References :

1. Cheng, R., Kalashnikov, D. V., & Prabhakar, S. (2004). Querying imprecise data in moving object environments. *IEEE Transactions on Knowledge and Data Engineering*, *16*(9), 1112-1127.
2. Benetis, R., Jensen, C. S., Karĉiauskas, G., & Ŝaltenis, S. (2006). Nearest and reverse nearest neighbor queries for moving objects. *The VLDB Journal*, *15*(3), 229-249.
3. Lian, X., & Chen, L. (2009). Efficient processing of probabilistic reverse nearest neighbor queries over uncertain data. *The VLDB Journal—The International Journal on Very Large Data Bases*, *18*(3), 787-808.
4. Aggarwal, C. C., & Philip, S. Y. (2009). A survey of uncertain data algorithms and applications. *IEEE Transactions on Knowledge and Data Engineering*, *21*(5), 609-623
5. Qin, B., Xia, Y., Prabhakar, S., & Tu, Y. (2009, March). A rule-based classification algorithm for uncertain data. In *Data Engineering, 2009. ICDE'09. IEEE 25th International Conference on* (pp. 1633-1640). IEEE.
6. Cheng, R., Kalashnikov, D. V., & Prabhakar, S. (2003, June). Evaluating probabilistic queries over imprecise data. In *Proceedings of the 2003 ACM SIGMOD international conference on Management of data* (pp. 551-562). ACM.
7. Pei, J., Hua, M., Tao, Y., & Lin, X. (2008, June). Query answering techniques on uncertain and probabilistic data: tutorial summary. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data* (pp. 1357-1364). ACM.
8. Z. Song and N. Roussopoulos, "k-Nearest Neighbor Search for Moving Query Point," Proc. Symp. Spatial and Temporal Databases, pp. 79-96, 2001
9. G. Trajcevski, O. Wolfson, F. Zhang, and S. Chamberlain, "The Geometry of Uncertainty in Moving Object Databases," Proc. Eighth Int'l Conf. Extending Database Technology, Mar. 2002.