A Probabilistic Approach for GNN Queries in LBS

Peng Chen, Junzhong Gu*, Xin Lin and Rong Tan

Department of Computer Science and Technology, East China Normal University, 200241 Shanghai, China
*{pchen, xlin, rtan}@ica.stc.sh.cn; *jzgu@cs.ecnu.edu.cn;*

Abstract

Range-based Probabilistic Group Nearest Neighbor (in short RP-GNN) query has recently gain much attention, due to its wide usage in many Location Based Services (LBSs). Previous works mainly focus on the uncertainty of data objects (P). While the uncertainty of query objects (Q) is prevailing in reality. In this paper, a comprehensive discussion on uncertain query objects is presented. Meanwhile two novel pruning methods are proposed to improve the performance of RP-GNN: one is Query points pruning (Q_pruning) and the other is Geometric pruning (G_pruning). Q_pruning reduces the number of query objects needed to be considered. And G_pruning method exploits the geometric properties of the RP-GNN problem to narrow down the search space. Extensive experiments show the effectiveness, efficiency and scalability of proposed methods.

Keywords: *RP-GNN, Q_pruning, G_pruning, Location Based Services*

1. Introduction

The combination of mobile technologies, Internet and GIS has flourished Location Based Services (LBSs) in recent years [1]. The main usage of LBSs is to provide mobile users with timely information at the right place for their decision making [2]. As an important decision support query, GNN query has received considerable attention from LBS research community.

A typical scenario of GNN query is to find a facility which minimizes the maximum (minimum or total) travel distance for a group of users. This, in turn, leads to the latest (earliest or total) time that a user (users) will arrive at the facility [3]. However, LBS scenarios are not that ideal. When privacy concerns, sampling precisions, and network transmission delays are taken into consideration, locations become uncertain in LBS applications. In those new emerging scenarios, the location of each object can be modeled as a so-called uncertain region $UR(p_i)$ with center C_i and radius r_i [4, 5, 6]. The RP-GNN query can be illustrated as Figure 1, where each uncertain object (data object, p_i, or query object q_i) can locate within a circle with arbitrary distribution. Similarly, the uncertain region of q_i, $UR(q_i)$, is centered at Cq_i and has radius qr_i.

Previous works [3, 4, 7] mainly focus on the scenarios when data objects (P) are uncertain regions. While, very little work has done to the scenario when query objects (Q) are also uncertain regions. In this paper, a comprehensive discussion is presented for this extended scenario. And two novel pruning methods are proposed to improve the performance of RP-GNN. Extensive experiments are conducted to evaluate the effectiveness, efficiency and scalability of proposed methods under various experiment settings.

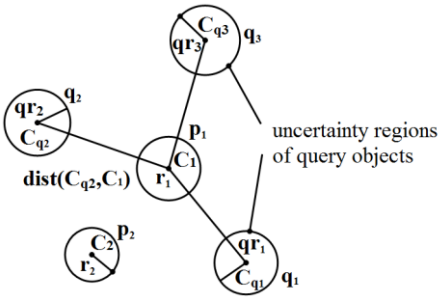

Figure 1. A Range-based Probabilistic Group Nearest Neighbor (RP-GNN) Query

2. Related Work

Range-based query in LBS. The traditional queries in LBS are based on exact location measurement, such as Skyline [2], GNN [7]. While in many real-world applications, locations inherently contain uncertainties. Privacy consideration, sampling precision, and network transmission delay are the three major causes of location uncertainty [2]. For a better approximation of the real world and more concerns about privacy protection, range-based queries are promising methods to solve uncertain location problems. When the exact location is taken place by an uncertain region, GNN results are represented by interested locations or items, with probabilities showing the reliabilities [8].

GNN query in LBS. GNN query is first introduced in [7] and has a wide usage in many LBS scenarios, typhoon monitoring, forest fire suppression and etc. Various variants are proposed subsequently, including ANNs [3], PGNN [4]. The range-based probabilistic GNN query is first studied in [4]. However, [4] mainly focuses on the situation when the data objects are uncertain regions. Only a very brief discussion on the effect of uncertain query objects is presented. In this paper, a more comprehensive research for uncertain query objects is conducted under various experiment settings.

Proposed Architecture. The proposed methods are implemented as a part of Spatial Decision Support Server of GaCAM [1]. GaCAM is a middleware system to support the construction and running of LBS applications. In our architecture, users periodically report their locations sampled by cell phones. Combining RP-GNN results with user profiles and GIS information, personalized information are sent back to users.

3. Formal Definition of the Problem

Given two uncertain object sets P and Q, a RP-GNN query retrieves object $o \in P$ with α probability to minimize the maximum distance from o to Q. Similar to [4], the reliability of RP-GNN query is defined as:

$$\alpha = \int_{r_{\min}}^{r_{\max}} (\Pr\{adist(o,Q) = r\} \cdot \prod_{\forall p \in P \setminus \{o\}} \Pr\{adist(p,Q) \geq r\}) dr \quad (1)$$

$$\text{where} \quad adist(o,Q) = \max_{i=1}^{n} dist(o,q_i)$$

It can be seen from equation (1) that, a reduction on either Q or P will improve the performance of RP-GNN algorithm. In this paper, two novel pruning methods are proposed to reduce Q and P, respectively.

4. RP-GNN with Uncertain Query Objects

4.1 Q_pruning

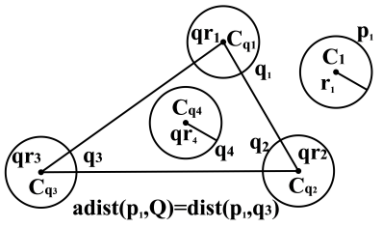

Figure 2. Example of Q_pruning Method

For any point in the plane, its farthest point in the set *Q* must be a point that lies on the convex hull of *Q* [9]. In Figure 2, q_4 can be pruned, because it has no effect on neither decision of the convex nor the GNN result. Instead of concerning all *n* uncertain query objects, only the uncertain query objects those intersect with the convex are considered. Theoretically, |*Q*| can be reduced to |*Q*'|, where |*Q*'|>=2.

4.2 G_pruning

G_pruning is based on the observation that GNN results are tending to appear at (or nearby) the center of query set. This is similar with the idea of traditional SPM and MBM. After a deeper research of the geometric properties of the problem, G_pruning method is proposed to narrow down the search space of RP-GNN algorithm. The main idea of G_pruning is to calculate an idea area, Idea-GNN (in short I-GNN, idea means without data objects considerations), within which RP-GNN results should appear.

As illuminated in Figure 3, two circles, Cir_1 and Cir_2, are made. The smallest closing circle of Cq_i (the center of q_i) is centered at Op and has radius R. Cir_1 and Cir_2 are both centered at Op. And the radii of Cir_1 and Cir_2 are $r_1 = R - \max\{radius(q_i)\}$ and $r_2 = R + \max\{radius(q_i)\}$, respectively. Moreover, Cir_1 is the smallest circle where, for each q_i, $Cir_1 \cap UR(q_i) \neq \varnothing$ holds. And Cir2 covers all $UR(q_i)$.

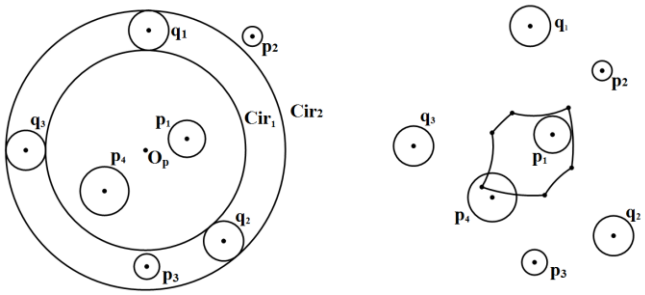

Figure 3. Example of G_pruning Method and its I-GNN Area

For each q_i, $r_1 <= dist(I\text{-}GNN, q_i) <= r_2$ holds. When this is considered reversely, since $dist(I\text{-}GNN, q_i)$ equals to $dist(q_i, I\text{-}GNN)$, I-GNN should be the intersection of annuluses centered at q_i and has radius r_1 and r_2. A detailed proof of the correctness of G_pruning method is reported in [10].

4.3 Procedure of RP-GNN

After the construction of Rtree over uncertain data objects, the constructed Rtree and uncertain query objects are pruned by G_pruning and Q_pruning respectively. Then all the candidates of RP-GNN are refined by sampling points in uncertain objects and calculating α according to equation (1). The same as [4, 7], the nearest neighbor algorithm used in our implementation is incremental. The RP-GNN algorithm is output sensitive, and its time complexity is O($ks|Q|$), where k is the number of RP-GNN results, s is the number of samples per object taken in refining phase (s=100, in the following experiments), and $|Q|$ is the number of query objects needed to be considered.

5. Performance Evaluation

In this section, Q_pruning and G_pruning are evaluated empirically to show their effectiveness, efficiency and scalability. Since the real data sets are not available, the proposed methods are evaluated over synthetic data sets in a d-dimensional data space [0, 1000]^d, similar to [4, 5, 6]. Four data sets (IUrU, IUrG, ISrU, and ISrG) are synthetized. IUrU (IUrG) denotes the data set with centers of Uniform distribution and radii of Uniform distribution (Gaussian distribution, with mean = (r_{max} + r_{min})=2 and variance = (r_{max} − r_{min})=5). Similarly, ISrU (ISrG) represents the data set with centers of Skew distribution (skewness = 0.8) and radii of Uniform distribution (Gaussian distribution, with the same setting)

PSPM and PMBM both introduced by [4] are taken as the benchmark algorithms. Another benchmark algorithm is linear scan, which sequentially scans all the objects in data set to check the condition in equation (1). All experiments are conducted on a Pentium IV 2-GHz PC with 2-Gbyte memory, and the reported results are the average of 100 queries.

5.1 Experiment Results

The Effectiveness of Proposed Pruning Methods. In the first set of experiments, the effectiveness of proposed pruning methods is evaluated. PSPMQ (PSPMG) and PMBMQ (PMBMG) denote the algorithms with only Q_pruning (G_pruning) method, respectively. Two proposed pruning methods are implemented in both PSPMB and PMBMB. The numbers over columns are speed-up ratios of proposed methods, compared with the linear scan.

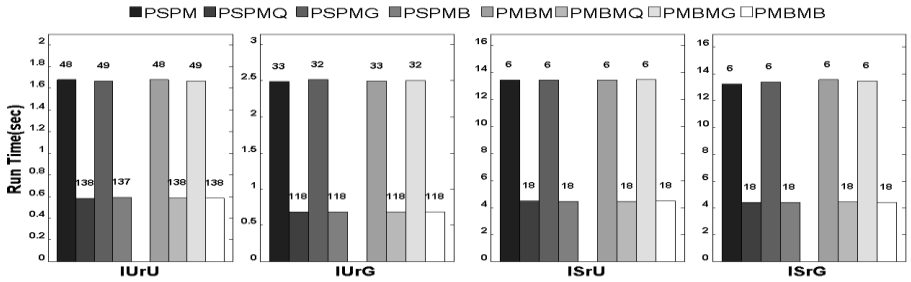

Figure 4. The Effectiveness of Proposed Pruning Methods

It can be seen from Figure 4 that, both pruning methods are effective, and Q_pruning has a better effect than G_pruing. Since a reduction of Q will sharply decrease the refining time. Runtimes over ISrU and ISrG are much higher than those over IUrU and IUrG. This is because most of data objects concentrate in a small region in Skew distributions. And more time is spent on pruning and refining. Both PSPMB and PMBMB outperform the benchmark

algorithms. And high speed-up ratios are achieved over different data sets (by 1-2 orders of magnitude).The data set size is set to 30K.

Performance versus |*P*|. Figure 5 tests the scalabilities of pruning methods. The runtime increases with the increasing data set size (from 30K to 1000K). Speed-up ratios are still high (by 1-4 orders of magnitude), which indicate good scalabilities with respect to the data size over various data sets.

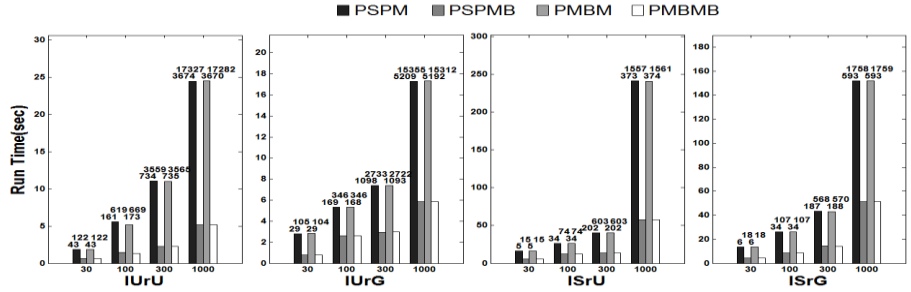

Figure 5. Performance versus |*P*|

Performance versus [qr_{min}, qr_{max}]. The performance over different *qr* is shown in Figure 6. Higher uncertainties of query objects lead to more data objects identified as RP-GNN candidates. The proposed methods outperform benchmarks in most cases. Moreover, both proposed methods gain 1-2 orders of magnitude against linear scan.

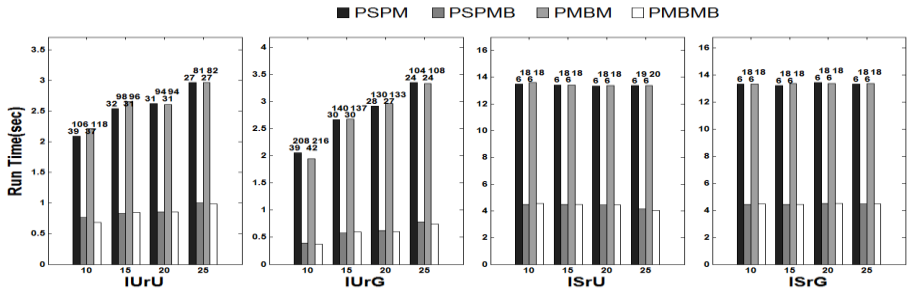

Figure 6. Performance versus [qr_{min}, qr_{max}].

Experiments are also conducted over data sets with other parameter values. Due to space limit, the results are not presented. They have the similar trends as the reported results.

6. Conclusions

In this paper, we focus on the RP-GNN query with uncertain query objects. Two novel pruning methods are proposed to improve the performance of RP-GNN query. Q_pruning aims at reducing the number of query objects needed to be considered. And the purpose of G_pruning is to narrow down the search space of RP-GNN. Extensive experiments demonstrate the effectiveness, efficiency and scalability of proposed pruning methods under various experiment settings.

References

[1] G. Junzhong, "Middleware for Physical and Logical Context Awareness", Communications in Computer and Information Science. 260, pp. 366--378 **(2011)**.

[2] L. Xin, X. JianLiang and H. Haibo, " Range-based Skyline Queries in Mobile Environments", TKDE. To be published **(2012)**.

[3] P. Dimitris, T. Yufei, M. Kyriakos and K. H. Chun, "Aggregate Nearest Neighbor Queries in Spatial Databases", TODS. 30, 2, pp.529-576 **(2005)**.

[4] L. Xiang and C. Lei, "Probabilistic Group Nearest Neighbor Queries in Uncertain Databases", TKDE. 20, 6, pp. 809-824 **(2008)**.

[5] C. Reynold, V. K. Dmitri and P. Sunil, "Querying Imprecise Data in Moving Object Environments", TKDE. 16, 9, pp. 1112-1127 **(2004)**.

[6] T. Yufei, C. Reynold, X. Xiaokui, K. N. Wang, K. Ben and P. Sunil, "Indexing Multi-Dimensional Uncertain Data with Arbitrary Probability Density Functions", In: VLDB, pp. 922-933 **(2005)**.

[7] P. Dimitris, S. Qiongmao, T. Yufei and M. Kyriakos, "Group Nearest Neighbor Queries. In: ICDE, pp. 301-312 **(2004)**.

[8] A. Lyublena, K. Christoph and O. Dan, "10^(10^6) Worlds and Beyond: Efficient Representation and Processing of Incomplete Information", VLDB, pp. 1021-1040 **(2009)**.

[9] D. B. Mark, C. Otfried,V. K. Marc and O. Mark, "Computational Geometry: Algorithms and Applications", 3rd. Springer-Verlag, Berlin Heidelberg **(2008)**.

[10] C. Peng, G. Junzhong, L. Xin and T. Rong, "A probabilistic approach for rendezvous decisions with uncertain data", Journal of Computational Information Systems. 7, 13, pp. 4668-4677 **(2011)**.