Reverse Nearest Neighbor Search in Peer-to-Peer Systems

Dehua Chen, Jingjing Zhou, and Jiajin Le

Col. of Computer Science, University of Donghua
P. O. Box 324, 200051, Shanghai, PRC
lydehua@mail.dhu.edu.cn
lejiajin@dhu.edu.cn

Abstract. Given a query point *Q*, a *Reverse Nearest Neighbor* (**RNN**) Query returns all points in the database having *Q* as their nearest neighbor. The problem of **RNN** query has received much attention in a centralized database. However, not so much work has been done on this topic in the context of Peer-to-Peer (P2P) systems. In this paper, we shall do pioneering work on supporting distributed **RNN** query in large distributed and dynamic P2P networks. Our proposed **RNN** query algorithms are based on a distributed multi-dimensional index structure, called *P2PRdNN*-tree, which is relying on a super-peer-based P2P overlay. The results of our performance evaluation with real spatial data sets show that our proposed algorithms are indeed practically feasible for answering distributed **RNN** query in P2P systems.

1 Introduction

The problem of *Reverse Nearest Neighbor* (RNN) Query [1,2,3,4,5,6,7,8,9,10] is to retrieve all data points in given multi-dimensional data sets whose Nearest Neighbor (NN) is a given query point. Although RNN is a complement of NN problem it is more complex than NN problem. The solutions from NN query cannot be directly applied to RNN query. This is because of the asymmetric relationship between NN/RNN: if a data point p is an RNN(q) (q is the nearest neighbor of p), it does not imply that p is the nearest neighbor NN(q) of q. The RNN problem has recently received considerable attention in the context of centralized database system due to its importance in a wide range of applications such as decision support system, profile-based marketing, document databases etc.

Nowadays, Peer-to-Peer (P2P) systems have become popular for sharing resources, information and services across a large number of autonomous peers in Internet. Especially, the applications of sharing multi-dimensional data (e.g. spatial data, documents, image files) in P2P systems are now being widely studied in the literatures [11,12,13,14,15,16,17]. However, most of these applications focus mainly on two types of queries: Range query and Nearest Neighbor (NN) query on the distributed data sets. And not so much effort is taken to support RNN search in such large distributed and ad-hoc environment. However, we believe that like its importance in the centralized database system, RNN query will become a practical

H. Legind Larsen et al. (Eds.): FQAS 2006, LNAI 4027, pp. 87 – 96, 2006.

and important class of queries in P2P systems. Let us first consider an example in the P2P Geographic Information System (GIS) application. Suppose a large-scale chain supermarket is to open up a new supermarket at a location, the RNN query can be used to find the subset of existing supermarkets will be affected by the new supermarket, assuming people choose the nearest supermarket to consume. Another example is that when a new document is inserted into a P2P digital library, the RNN query can be used to ask the subset of authors of other documents who will find the new document interesting based on similarity to their documents. Therefore, this paper will investigate RNN search in distributed and dynamic P2P systems.

Like most of previous researches for RNN query in centralized database system, our proposed methods also build on tree-based multi-dimensional index structures (e.g. the R-tree family [18,19,20]). However, instead of maintaining a centralized multi-dimensional index in one centralized server, we propose a distributed multi-dimensional index, called P2PRdNN-tree, supported by a super-peer-based P2P overlay network. The P2PRdNN-tree structure enables efficient RNN search in large distributed environment. Like Rdnn-tree [2] proposed for centralized database context, our proposed distributed P2PRdNN-tree index structure stores extra information about nearest neighbor of data points in tree nodes. The extra information can efficiently reduce the search space and network communication

The remainder of this paper is organized as follows: Section 2 overviews the previous work. Section 3 presents our proposed super-peer-based P2P overlay and P2PRdNN-tree structure. Section 4 presents our proposed distributed RNN search algorithms. Section 5 provides experimental results and Section 6 summarizes our work.

2 Related Work

In Section 2.1, we shall briefly describe previous work on RNN query in centralized database systems. Section 2.2 overviews multi-dimensional data sharing in P2P systems.

2.1 RNN Search in Centralized Database Systems

Algorithms for processing RNN query in centralized databases can be classified into two categories depending on whether they require pre-computation or not.

The problem of RNN was first studied in [1]. The idea of the authors is to pre-compute, for each data point d, the distance dnn to its nearest neighbor NN(d). Thus, each data point is represented as a circle, whose center is the data point and whose radius is its dnn. Besides the R-tree that indexes the original data point, a separate R-tree is maintained which indexes the sets of such circles. The problem of finding RNN of a query point Q is then reduced to finding the circles that contain Q.

In order to avoid maintaining two separate R-trees, [2] combines the two indexes in the Rdnn-tree (R-tree containing Distance of Nearest Neighbors) index structure. Rdnn-tree differs from standard R-tree by storing extra information about NN of the data points for each tree node: for every leaf node, its record stores dnn, and for every non-leaf node, it record stores max_dnn (the maximum distance from every point in