Generalized Top-k Joins

Mr. QI Shuyao

Abstract:

Consider two collections of objects R and S, where each object is assigned a score (e.g., ranking). Given a join predicate and an integer k, a top-k join query returns the k pairs of objects which have the highest combined score (based on an aggregate scoring function) among all object pairs in R × S that qualify the join predicate. Despite the fact that databases are becoming increasingly more complex over the years and objects are commonly assigned several types of auxiliary information such as textual descriptions and spatial locations, the top-k join has only been studied in a relational database context, where the join predicate is equality. In addition, the main goal in processing top-k joins has been the minimization of accessed tuples instead of the computational cost. We address these issues by proposing a novel evaluation paradigm for computing top-k joins which operates on any type of join attributes and predicates. By examining blocks of data from input collections R and S ordered by the object scores our paradigm manages to access only a limited number of objects and minimize the computational cost. Extensive experiments using numerical, spatial and string join attributes demonstrate the advantage of our methodology compared to baseline approaches that primarily focus either on the top-k or the join component of the query.

About the Speaker:

QI Shuyao is a full-time PhD student at the Department of Computer Science, the University of Hong Kong. His supervisor is Prof. Nikos Mamoulis. His research interest is query processing, data management and information retrieval.