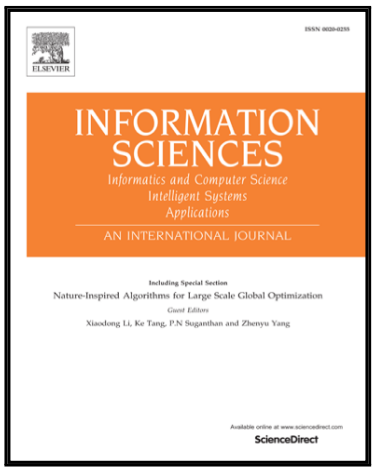

Reporting l most influential objects in uncertain databases based on probabilistic reverse top-k queries

Guoqing Xiao, Kenli Li, Keqin Li

PII: S0020-0255(17)30674-6
DOI: 10.1016/j.ins.2017.04.028
Reference: INS 12849

To appear in: *Information Sciences*

Received date: 20 January 2016
Revised date: 15 January 2017
Accepted date: 12 April 2017

Please cite this article as: Guoqing Xiao, Kenli Li, Keqin Li, Reporting l most influential objects in uncertain databases based on probabilistic reverse top-k queries, *Information Sciences* (2017), doi: 10.1016/j.ins.2017.04.028

Reporting *l* Most Influential Objects in Uncertain Databases Based on Probabilistic Reverse Top-*k* Queries

Guoqing Xiao[a,b], Kenli Li[a,b,*], Keqin Li[a,b,c]

[a] *College of Information Science and Engineering, Hunan University, Changsha 410082, Hunan, China.*
[b] *National Supercomputing Center in Changsha, Changsha 410082, Hunan, China.*
[c] *Department of Computer Science, State University of New York, New Paltz, New York 12561, USA.*

Abstract

Reverse top-*k* queries are proposed from the perspective of a product manufacturer, which are essential for manufacturers to assess the potential market. However, the existing approaches for reverse top-*k* queries are all based on the assumption that the underlying data are exact (or certain). Due to the intrinsic differences between uncertain and certain data, these methods cannot be applied to process uncertain data sets directly. Motivated by this, in this paper, we firstly model the probabilistic reverse top-*k* queries over uncertain data. Moreover, we formulate a probabilistic top-*l* influential query, that reports the *l* most influential objects having the largest impact factors, where the impact factor of an object is defined as the cardinality of its probabilistic reverse top-*k* query result set. We present effective pruning heuristics for speeding up the queries. Particularly, we exploit several properties of probabilistic threshold top-*k* queries and probabilistic skyline queries to reduce the search space of this problem. In addition, an upper bound of the potential users is estimated to reduce the cost of computing the probabilistic reverse top-*k* queries for the candidate objects. Finally, efficient query algorithms are presented seamlessly with integration of the proposed pruning strategies. Extensive experiments using both real-world and synthetic data sets demonstrate the efficiency and effectiveness of our proposed algorithms.

Keywords: Data management, Probabilistic reverse top-*k* queries, Probabilistic skyline queries, Probabilistic top-*l* influential queries, Uncertain databases

1. Introduction

Microeconomics shows that customer preference is an important factor in making decisions of product sales, which thus becomes one major concern in microeconomics [26]. Given a set of objects and an individual user, a top-*k* query helps the user (customer) find his/her favorite objects (products) that best match the user preference, thus avoiding huge and overwhelming answer sets. It is very important for a manufacturer that its products are returned in the highest ranked positions for as many different user preferences as possible.

On the other hand, a reverse top-*k* query for a product, from the perspective of producers, is proposed to estimate the impact of the product in the market. It returns a set of users, named potential users, who regard the product as one of their top-*k* result sets [32]. A product in the top-*k* answer set of a user implies that the product matches the preference of the user more than the products not in the top-*k* result set. Accordingly, the number of potential users in the reverse top-*k* answer set for a product is a good estimate of the influence of the product in the market. Reporting *l* most influential objects with reverse top-*k* queries from a given database of products is important for market analysis and decision-making of manufacturers and beneficial for many real-world applications.

Application 1. A motivating application can be the advertisement for e-commerce since there are a lot of identical items and we can use reverse top-*k* queries to find which *l* items should be on advertisement.

*Corresponding author
Email addresses: xiaoguoqing@hnu.edu.cn (Guoqing Xiao), lkl@hnu.edu.cn (Kenli Li), lik@newpaltz.edu (Keqin Li)