Journal of Computer and System Sciences ••• (••••) •••–•••

Contents lists available at ScienceDirect

Journal of Computer and System Sciences

www.elsevier.com/locate/jcss

Efficient monochromatic and bichromatic probabilistic reverse top-*k* query processing for uncertain big data

Guoqing Xiao [a], Kenli Li [a,b,*], Xu Zhou [a], 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, NY 12561, USA

ARTICLE INFO

Article history:
Received 7 March 2016
Received in revised form 20 May 2016
Accepted 31 May 2016
Available online xxxx

Keywords:
Big data
Data management
Probabilistic reverse top-*k* queries
Query processing
Uncertain data

ABSTRACT

There has been an increasing growth in numerous applications that naturally generate large volumes of uncertain data. By the advent of such applications, the support of advanced analysis query processing such as the top-*k* and reverse top-*k* for uncertain big data has become important. In this paper, we model firstly probabilistic reverse top-*k* queries over uncertain big data for the discrete situation, in both *monochromatic* and *bichromatic* cases, denoted by MPRT and BPRT queries, respectively. We determine the partitions of solution space of MPRT queries and provide in theory a mathematical model for solving arbitrary dimensional data space. Additionally, we propose effective pruning heuristics to reduce the search space of BPRT queries. Moreover, efficient query procedures are presented seamlessly with integration of the proposed pruning strategies. Extensive experiments demonstrate the efficiency and effectiveness of our proposed approaches with various experimental settings.

1. Introduction

Recently, the support of rank-aware query processing has played an increasingly important role in many real-world applications, such as market analysis, multi-criteria decision making, environmental surveillance, quantitative economics research, etc. This has created a need for data management algorithms and applications, among which a pivotal technique in this respect is the top-*k* query, which is a classic problem in the area of databases and information retrieval. Top-*k* query retrieves only the *k* objects that best match the user preferences based on a weighting function, thus avoiding huge and overwhelming answer sets [11]. 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 returns a set of users (or customers), named potential users, who regard the product as one of their top-*k* result sets [39]. A product in the top-*k* answer set of a user implies that the product meets 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 becomes a good estimate of the popularity of the product in the market.

Tables 1 and 2 illustrate an example of top-*k* queries and reverse top-*k* queries over data collected from a real estate deal database containing information about different houses as well as user preferences. For each of the three houses, the price and area are recorded, and smaller value on each attribute is more preferable. The database also stores the preferences of

* Corresponding author at: College of Information Science and Engineering, Hunan University, Changsha 410082, Hunan, China.
E-mail addresses: xiaoguoqing@hnu.edu.cn (G. Xiao), lkl@hnu.edu.cn (K. Li), zhouxu@126.com (X. Zhou), lik@newpaltz.edu (K. Li).

http://dx.doi.org/10.1016/j.jcss.2016.05.010
0022-0000/

Table 1
A product data set and its reverse top-2 results.

House	Price	Area	Prob.	Reverse top-2 results
a	0.8	0.1	0.6	{Alice, Chris}
b	0.2	0.7	0.7	{Bob, Chris}
c	0.5	0.5	0.5	{Alice, Bob}

Table 2
A user preference data set and its top-2 results.

User	Preference		
	μ [price]	μ [area]	Top-2 results
Alice	0.1	0.9	{*c*, *a*}
Bob	0.8	0.2	{*b*, *c*}
Chris	0.5	0.5	{*a*, *b*}

three users (Alice, Bob, and Chris) in terms of weights on the attributes. Different users can have various preferences about a potential house. For example, Alice prefers houses with spacious room, whereas Bob is interested in cheap houses. Chris is indifferent or values equally price and area. In order to recommend the most popular houses to users, the brokers need to know the potential user community of each house property. This can be implemented by executing a reverse top-k query. As illustrated in bold in Table 1, the reverse top-2 result set, i.e., the two potential users, for house property a, b, and c are {Alice, Chris}, {Bob, Chris}, and {Alice, Bob}, respectively. Furthermore, in Table 2, the top-2 houses are described in bold for each user along with their aggregate scores.

In addition, lots of other real-life applications, such as e-commence and online transactions etc., can be taken as the motivating examples to model the uncertain/probabilistic databases. For example,

(1) A motivating application can be the advertisement for e-commerce since there are a lot of identical items, the agency maintains an item database which is available for buyers (users) in the market. Additionally, the agency regularly collects the current preferences of the buyers (users) for favored characteristics of items, as determined by the market demand. In order to facilitate a beneficial strategy for the promotion of an item, the agency needs to identify the most influential items in the market such that the agents can prioritize showing such items to users. Reverse top-k queries can be applied to find which l items should be on advertisement.
(2) In a stock online exchange system, the bourse is interested in finding l stocks which have the largest investment potential based on historical stock market information and trader preferences, in order to perform direct marketing to its investor in a timely fashion. The l stocks form the most influential stocks of the bourse.

However, in reality, the items for advertising in the Internet and the stocks in a stock-trading system, due to data collecting, transmission delay, and inaccuracy or incompletion in conversion, etc., each item/stock can be regarded as an uncertain data record and can be given a definite probability value for characterizing its confidence or authenticity [2]; likewise, in a rent/sale system of house properties, each property is associated with a trustability value which is derived from customers feedback on the property quality, surrounding environment and traffic conditions, etc. This trustability value can also be regarded as existence probability of the property since it represents the probability that the house occurs exactly as described in the advertisement in terms of quality and surrounding environment [49]. Thus the trustability value is an important factor whether the users rent/purchase the house. In general, this kind of databases can be modeled as an uncertain/probabilistic database. For instance, in Table 1, the probability column (prob.) shows the trustability of each house property in the market.

As a matter of fact, due to data noise, transmission delay, environmental factors, and so on, the uncertainty of data is existent in essence. Similarly, surveys and imputation techniques create data which is uncertain in nature. There are numerous other examples which demonstrate the existences of uncertainty in the collected data, such as sensing data [12,1,27], moving object search [5,18,52], location-based services (LBS) [6,34,4], etc. These applications have created a need for uncertain big data management algorithms, among which a pivotal technique in this respect is the query processing over uncertain databases, such as top-k queries and reverse top-k queries. With the rapid development of data collection methods and the practical applications, the issue of uncertain data query has drawn wide attention in both academia and industry.

Although previous works [39,42,40,41,43] have studied the reverse top-k search problem over precise data, they cannot directly handle the uncertain databases. Jin et al. [24] proposed firstly reverse top-k queries upon uncertain databases. However, in this work, each object in the uncertain database is described as a probabilistic distribution function (i.e., the continuous situation) rather than a vector (i.e., the discrete situation) which is more practical for describing the attributes of an object and the preferences of a user. Thus, the present work cannot handle the discrete case directly. We have to redefine new query semantics for probabilistic reverse top-k queries.

Please cite this article in press as: G. Xiao et al., Efficient monochromatic and bichromatic probabilistic reverse top-k query processing for uncertain big data, J. Comput. Syst. Sci. (2016), http://dx.doi.org/10.1016/j.jcss.2016.05.010