Data & Knowledge Engineering 88 (2013) 248–266

Contents lists available at ScienceDirect

Data & Knowledge Engineering

journal homepage: www.elsevier.com/locate/datak

Top-k best probability queries and semantics ranking properties on probabilistic databases

Trieu Minh Nhut Le *, Jinli Cao, Zhen He

ARTICLE INFO

Available online 24 April 2013

Keywords:
Top-*k* query
Ranking query
Probabilistic data
Query processing
Uncertain data

ABSTRACT

There has been much interest in answering top-*k* queries on probabilistic data in various applications such as market analysis, personalized services, and decision making. In probabilistic relational databases, the most common problem in answering top-*k* queries (ranking queries) is selecting the top-*k* result based on scores and top-*k* probabilities. In this paper, we firstly propose novel answers to top-*k* best probability queries by selecting the probabilistic tuples which have not only the best top-*k* scores but also the best top-*k* probabilities. An efficient algorithm for top-*k* best probability queries is introduced without requiring users to define a threshold. The top-*k* best probability approach is more efficient and effective than the probability threshold approach (PT-*k*) [1,2]. Second, we add the "*k-best ranking score*" into the set of semantic properties for ranking queries on uncertain data proposed by [3,4]. Then, our proposed method is analyzed, which meets the semantic ranking properties on uncertain data. In addition, it proves that the answers to the top-*k* best probability queries overcome drawbacks of previous definitions of the top-*k* queries on probabilistic data in terms of semantic ranking properties. Lastly, we conduct an extensive experimental study verifying the effectiveness of answers to the top-*k* best probability queries compared to PT-*k* queries on uncertain data and the efficiency of our algorithm against the state-of-the-art execution of the PT-*k* algorithm using both real and synthetic data sets.

1. Introduction

Uncertain data has arisen in some important applications such as personalized services, market analysis and decision making, because data sources of these applications are collected from data integration, data analysis, data statistics, data classification, and results prediction. These data are usually inconsistent [5] or contain likelihood information [6]. Thus, selecting the best choice from various alternatives of uncertain data is an important challenge facing these applications. The top-*k* queries that return the *k* best answers according to a user's function score are essential for exploring uncertain data on these applications [6]. Uncertain data have been studied extensively by many researchers in areas such as modeling uncertain data [7,8], managing uncertain data [9], and mining uncertain data [10,11].

1.1. Motivation

In business, investors often make decisions about their products based on analysis, statistical data and mining data [11], which provide predictions relating to successful and unsuccessful projects. To analyze the market, investors firstly collect the historical

* Corresponding author.
E-mail addresses: trieule@sgu.edu.vn (T.M.N. Le), j.cao@latrobe.edu.au (J. Cao), z.he@latrobe.edu.au (Z. He).

0169-023X/$ – see front matter © 2013 Published by Elsevier B.V.
http://dx.doi.org/10.1016/j.datak.2013.04.005

Table 1
Predicted profits of USD $100 investment on products.

Tuple	Product ID	Profit of USD $100 investment	Probabilistic
t_1	A	25	0.29
t_2	B	18	0.3
t_3	E	17	0.8
t_4	B	13	0.4
t_5	C	12	1.0
t_6	E	11	0.2

statistical data, and then use the data to predict future market trends with probabilistic prediction. This is known as probabilistic data. For example, assume that the data in Table 1 have been collected and analyzed statistically, according to historical data resources [12]. Each tuple represents an investment project of USD $100 to produce a specific product (Product ID). Investing in products based on their probabilities (Probability) will result in an estimated amount of profit. In tuple t_1, a businessman invests USD $100 on product A, and it has a 0.29 chance of obtaining a profit of USD $25.

In the real world, when analyzing historical data, predictions on future market trends return two or more values per product with probabilities that the predictions are correct. Therefore, some tuples in Table 1 have the same product ID with different profit. In the probabilistic data model, these tuples are mutually exclusive, and controlled by a set of rules (generation rule) [1,2,6,13]. For example, tuples t_2 and t_4 as projects that invest in product B have a 0.3 probability of producing a USD $18 profit and 0.4 probability of producing a USD $13 profit. In this case, if the prediction for tuple t_2 is true, then the prediction for tuple t_4 will not be true. It is impossible for both profits to be true for the same product ID. They are mutually exclusive predictions. In Table 1, the probabilistic data are restricted by the exclusive rules $R_1 = t_2 \oplus t_4$ and $R_2 = t_3 \oplus t_6$.

Top-k queries can be used to help investors make business decisions such as choosing projects which have the top-2 highest profits. On probabilistic databases, top-k queries can be answered by using the probability space that enumerates the list of all possible worlds [1,2,7,14–16]. A possible world contains a number of tuples in the probabilistic data set. Each possible world has a non-zero probability for existence and can contain k tuples with highest profits. Different possible worlds can contain different sets of k tuple answers. Therefore, it is necessary to list all possible worlds of Table 1 to find the top-2 answers for the top-2 query of the probabilistic database. Thus, Table 2 lists three dimensions: the possible world, the probability of existence, and the top-2 tuples in each possible world.

According to Table 2, any tuple has a probability of being in the top-2. Therefore, Table 3 lists the tuples, profit, probability, and top-2 probability to analyze the top-2 answers of probabilistic databases. The top-2 probability of a tuple is aggregated by the sum of its probabilities of existence in the top-2 in Table 2.

In previous research [1,2], the top-k answers are found using the probability threshold approach called PT-k. The PT-k queries return a set of tuples with top-k probabilities greater than the users' threshold value. For example the answer to the PT-2 query with threshold 0.3 in the example listed in Table 3 is the set containing 4 tuples $\{t_2,t_3,t_4,t_5\}$. We have identified three drawbacks with PT-k queries. These are listed below:

- The PT-k queries may lose some important results. According to PT-2 query, tuple $t_1(25, 0.29)$ is eliminated by the PT-2 algorithm because its top-2 probability is less than threshold 0.3. In this case, we recommend that tuple t_1 should be in the result, the reason being that tuple $t_1(25, 0.29)$ is not worse than tuple $t_4(13, 0.3072)$, when comparing both attributes of profit and top-2 probability. That is, t_1.profit (25) is significantly greater than t_4.profit (13) and t_1.top-2 probability (0.29) is slightly

Table 2
List of all possible worlds and top-2 tuples.

Possible world	Probability of existence	Top-2 tuples in possible world
$W_1 = \{t_1,t_2,t_3,t_5\}$	0.0696	t_1,t_2
$W_2 = \{t_1,t_2,t_5,t_6\}$	0.0174	t_1,t_2
$W_3 = \{t_1,t_3,t_4,t_5\}$	0.0928	t_1,t_3
$W_4 = \{t_1,t_4,t_5,t_6\}$	0.0232	t_1,t_4
$W_5 = \{t_1,t_3,t_5\}$	0.0696	t_1,t_3
$W_6 = \{t_1,t_5,t_6\}$	0.0174	t_1,t_5
$W_7 = \{t_2,t_3,t_5\}$	0.1704	t_2,t_3
$W_8 = \{t_2,t_5,t_6\}$	0.0426	t_2,t_5
$W_9 = \{t_3,t_4,t_5\}$	0.2272	t_3,t_4
$W_{10} = \{t_4,t_5,t_6\}$	0.0568	t_4,t_5
$W_{11} = \{t_3,t_5\}$	0.01704	t_3,t_5
$W_{12} = \{t_5,t_6\}$	0.0426	t_5,t_6