Information Sciences 318 (2015) 73–90

Contents lists available at ScienceDirect

Information Sciences

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

Trustworthy answers for top-k queries on uncertain Big Data in decision making

H.T.H. Nguyen, J. Cao *

Department of Computer Science and Computer Engineering, La Trobe University, Australia

ARTICLE INFO

Article history:
Received 22 January 2014
Received in revised form 22 August 2014
Accepted 29 August 2014
Available online 6 September 2014

Keywords:
Uncertain Big Data
Top-k query
Reliable and trustable answer
Useful knowledge

ABSTRACT

Effectively extracting reliable and trustworthy information from Big Data has become crucial for large business enterprises. Obtaining useful knowledge to enable better decisions to be made in order to improve business performance is not a trivial task. The most fundamental challenge for Big Data extraction is to handle the uncertainty of data to meet emerging business needs, such as marketing analysis, future prediction and decision making. In this paper, we firstly propose a novel approach called Dominating Top-k Aggregate Query (DA-Topk) to provide trustworthy and reliable informative knowledge from uncertain Big Data by combining the techniques of skyline and top-k queries. Then, we design a number of pruning rules to reduce the search space and terminate the ranking process as early as possible. Next, we provide a deeper analysis regarding the satisfaction of the six ranking properties (i.e. *exact-k*, *containment*, *unique-rank*, *value-invariance*, *stability* and *faithfulness*) between our approach and existing approaches to demonstrate that our method is the only one which satisfied all of these properties. Extensive experiments with both real and synthetic data sets have been conducted to verify the efficiency and effectiveness of our proposed approach compared to the state-of-the-art approaches. Our approach can help managers make strategic decisions quickly and accurately in competitive market places.

1. Introduction

Big Data has attracted huge attention not only from researchers in the fields of information sciences, but also decision makers in enterprises. The hidden informative knowledge excavated from massive volumes of data is extremely valuable since it can bring huge benefits to enterprises by enhancing business performance for market analysis, future prediction, decision making processes, etc. [2,10,20]. Taking an example of Wal-Mart in the retail industry, the effective use of valuable knowledge exploited from its data warehouse has significantly benefitted their pricing strategies and advertising campaigns [20]. However, since extracting valuable knowledge from Big Data is much more challenging compared with analyzing tasks in traditional data warehouses, it is impossible to perform effective analysis on Big Data by using existing traditional analytic techniques [5–7]. The most fundamental challenge for Big Data extraction is to handle the uncertainty of data. It is clear that the answers to analytical queries performed on imprecise data repositories are naturally associated with a degree of uncertainty [19]. As a result, decision makers must take the reliability of the exploited information into greater consideration due

* Corresponding author.
E-mail addresses: ht34nguyen@students.latrobe.edu.au (H.T.H. Nguyen), j.cao@latrobe.edu.au (J. Cao).

http://dx.doi.org/10.1016/j.ins.2014.08.065
0020-0255/

to the fact that correct data has a direct effect on the final decisions. Therefore, the need to develop and create new techniques to extract reliable and useful knowledge from such massive, distributed and large-scale data repositories has become critical.

One of the most important challenges when dealing with uncertain Big Data is to cope with increasing volumes of data [20]. As a result, top-k (or ranking) queries which return the most relevant/interesting data have been proven to be the most important technique for exploring uncertain Big Data. There have been many attempts to propose some algorithms and semantics for ranking queries on uncertain data [4,8,11,12,16,21,26]. Papers [4,21] have defined two important semantics of top-k queries, called Expected Rank and U-Topk, respectively. The Expected Rank approach [4] ranks uncertain tuples based on their expected values in all possible worlds and the U-Topk approach [21] returns the top-k list which is valid in some possible worlds and has the maximum aggregated probability of being top-k. However, these proposed algorithms have their own deficiencies and may not return reliable answers. In this paper, we propose a novel approach called Dominating Top-k Aggregate Query (DA-Topk) which combines the benefits of skyline and top-k queries to generate trustworthy and reliable answers for ranking queries over uncertain Big Data. The outcome of the research can provide trustworthy and useful knowledge to support data analysis and decision making by overcoming the weaknesses of the Expected Rank and U-Topk method. Our approach can help managers make strategic decisions quickly and accurately in competitive market places. We also guarantee the effectiveness of the proposed algorithms using a number of pruning rules to reduce the search space. We analyse in detail the advantages of our proposed approach with an explanation of the motivation scenario in the following sections.

2. Motivation scenario

To illustrate the advantages of our proposed approach, we consider a simplified business scenario related to decision making tasks for effective resource allocation in an organization.

Suppose that company X, which has a number of branches distributed around the world, needs to analyse a huge amount of data to extract hidden useful knowledge which will assist it to achieve its business goal. Data on historical sales over previous years is used to predict the profit each branch is likely to earn next year. As a result of this prediction, the estimated profit of each branch is associated with a successful probability. Table 1 shows the simplified data of company X.

The executives of the company have to make a decision regarding resource allocation so that the profit is maximized with the lowest risk. In particular, the limited resources can only be allocated to the top three branches with the highest profit and maximum successful probability. The problem is how to rank all the branches to select the three most reliable and trustworthy ones to help the executive make the best decision. For example, it is difficult to conclude which is the better branch: branch C which can earn a profit of approximately USD $80 million with a probability of 0.3 or branch D which can earn a lower amount of profit (USD $60 million but with higher probability (0.4). Moreover, the ranking processes will become more complicated if the organization has some data constraints on the final result. For instance, if the company wants to allocate product Z to three branches in three different locations, then the branches located in the same location suffer from mutually exclusive constraints, meaning that only one of them can appear in the top 3 results. Such kind of data constraints occur very frequently in many real-world applications [6]. In the example given in Table 1, branches A, C and D are in location L1, thus, they are mutually exclusive and only one of them can appear in the final top-k result. Similarly, branches B and E are located in location L2, so they suffer the same data constraints. Dealing with the two ranking criteria (profit and probability) and the complicated constraints of data at the same time make the ranking processes on uncertain Big Data more complex and the existing ranking techniques seem to be inapplicable [21].

In the literature, *possible world semantics* has been defined and extensively used to handle the uncertain nature of data as well as complicated data correlations. This means each uncertain database is viewed as a set of possible worlds [3,5,7,11,14,15,18,19]. Each possible world is a subset of tuples from the uncertain database and is viewed as a single deterministic relation which can perform any traditional query. The validity of these possible worlds is governed by a set of *generation rules* which constrain the occurrences of uncertain tuples (i.e. data constraints or data correlations). The probability of each possible world is computed based on both the existence of all tuples in the world and the absence of all other tuples.

Table 1
Example of uncertain data relation of company X.

	Branches	Location	Estimated profit (millions of USD)	Probability of success
t_1	A	L1	125	0.3
t_2	B	L2	110	0.4
t_3	C	L1	80	0.3
t_4	D	L1	60	0.4
t_5	E	L2	58	0.5
t_6	F	L3	56	1.0
t_7	G	L4	49	0.6