Optimizing Quality for Probabilistic Skyline Computation and Probabilistic Similarity Search

(Extended Abstract)

Xiaoye Miao[†][*] Yunjun Gao[*] Linlin Zhou[*] Wei Wang[‡] Qing Li[§]

[†]*Center for Data Science, Zhejiang University, Hangzhou, China*
[*]*College of Computer Science, Zhejiang University, Hangzhou, China*
[‡]*School of Computer Science and Engineering, University of New South Wales, Sydney, Australia*
[§]*Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China*
[†][*]`{miaoxy, gaoyj, zlinlin}@zju.edu.cn` [‡]`weiw@cse.unsw.edu.au` [§]`csqli@comp.polyu.edu.hk`

Abstract—**Probabilistic queries usually suffer from the noisy query result sets, due to data uncertainty. In this paper, we propose an efficient optimization framework, termed as QueryClean, for both probabilistic skyline computation and probabilistic similarity search. Its goal is to optimize query quality by selecting a group of uncertain objects to clean under limited resource available, where an entropy based quality function is leveraged. We develop an efficient index to organize the possible result sets of probabilistic queries, which is able to help avoid multiple probabilistic query evaluations over a large number of possible worlds for quality computation. Moreover, using two newly presented heuristics, we present *exact* and *approximate* algorithms for the optimization problem. Extensive experiments on both real and synthetic data sets demonstrate the efficiency and scalability of QueryClean.**

I. Introduction

Uncertain data exist in many real-life applications due to a variety of reasons, e.g., the noise in sensor inputs or errors in wireless transmission, missing or incorrect values in data integration, etc. As a result, probabilistic queries have received much attention, including probabilistic skyline computation [1], probabilistic nearest neighbor search [2], and so forth. A probabilistic query returns, from an uncertain database, the objects with non-zero probabilities to be query results. Hence, the data uncertainty propagates to the query results, even though users usually expect to obtain correct and accurate results. It is thereby difficult for users to identify desirable objects and make correct decisions from the result sets with much noise, especially for the high uncertainty. Furthermore, critical decisions based on poor-quality data have very serious consequences. As reported by Gartner [3], poor data quality is a primary reason for 40% of all business initiatives failing to achieve their targeted benefits, and data quality affects overall labor productivity by as much as a 20%.

It is well-known that data cleaning is an effective way to improve data quality. Nevertheless, in most cases, data cleaning is a labor-intensive, time-consuming, and expensive process, and thus, it is infeasible to clean all data objects due to limited resources available. As a consequence, complementary to the fruitful work upon probabilistic models and queries, in this paper, we aim to improve the quality of probabilistic query results via making full use of limited budget to find the beneficial objects (to clean) for quality improvement. Since optimization methods are *query-dependent*, existing strategies for max/region query [4] and PT-k query [5] cannot efficiently support the quality optimization on the probabilistic skyline query and probabilistic similarity search. In brief, the key contributions of the paper are summarized as follows.

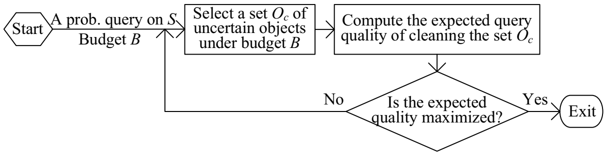

Start
A prob. query on $\mathcal{S}$, Budget B
Select a set O_c of uncertain objects under budget B
Compute the expected query quality of cleaning the set O_c
No
Is the expected quality maximized?
Yes
Exit

Fig. 1: The flowchart of QueryClean

- We propose an efficient framework, termed as QueryClean, to optimize the quality of probabilistic skyline computation and probabilistic similarity search.
- Based on a novel index structure, we develop an effective strategy for quality computation. Using two newly introduced heuristics, we present three efficient algorithms for object selection.
- Extensive experiments with both real and synthetic data sets demonstrate the performance of QueryClean.

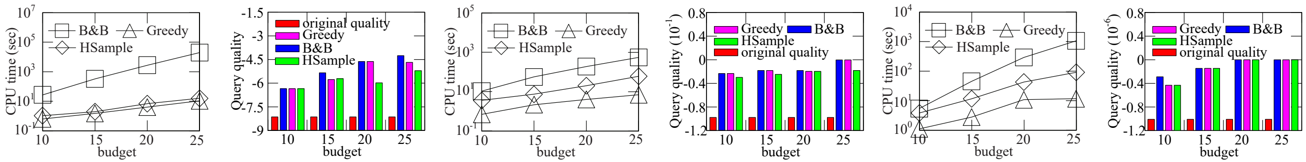
II. Preliminaries

Definition 1: **(Query quality)**. Given an uncertain dataset $\mathcal{S}$ and a query object q (if it exists), the quality of a probabilistic query ϕ w.r.t. q and $\mathcal{S}$, denoted by $\kappa(\phi(q;\mathcal{S}))$, is defined in Eq. 1. It is assumed all possible query result sets, denoted as R, are collected in a set Ω, and each R contains a group of objects from $\mathcal{S}$. $\Pr(R)$ denotes the probability of R being an answer set.

$$\kappa(\phi(q;\mathcal{S})) = \sum_{R\in\Omega} \Pr(R)\log_2 \Pr(R) \tag{1}$$

Definition 2: **(Expected query quality)**. The expected quality of a probabilistic query ϕ when a set O_c of uncertain objects is chosen to clean, denoted by $\mathbb{E}[\kappa(\phi(q;\mathcal{S})|O_c)]$ (abbrev. as $\mathbb{E}[\kappa(\phi|O_c)]$), is defined in Eq. 2. $\kappa(\phi|T_c)$ is the quality of query ϕ after the objects in O_c are cleaned as the

2375-026X/19/$31.00 ©2019 IEEE
DOI 10.1109/ICDE.2019.00259



(a) Probabilistic skyline queries on *CarDB* (b) Probabilistic kNN queries on *Forest* (c) Probabilistic range queries on *Forest*

Fig. 2: Performance evaluation on QueryClean

tuple set T_c, and $\mathcal{T}$ is supposed to contain all the possible tuple sets that O_c could be cleaned as.

$$\mathbb{E}[\kappa(\phi|O_c)] = \sum_{T_c \in \mathcal{T}} \Pr(T_c)\kappa(\phi|T_c) \qquad (2)$$

In this paper, we consider data cleaning as a user-provided module, where a variety of existing cleaning techniques can be employed, e.g., crowdsourcing. Without loss of generality, for an uncertain object o, our cleaning module simply *determines* o as one of its tuples o^t at the probability of $\Pr(o^t)$, which incurs cleaning cost, denoted by $c(o)$.

Definition 3: **(Our problem)**. Given an uncertain dataset $\mathcal{S}$ and a cleaning budget B, the goal of our problem studied in this paper is to find out a set $O^\star \subseteq \mathcal{S}$ of uncertain objects to clean, such that the expected quality of a probabilistic query ϕ is maximized. Let $c(o)$ denote the cost of cleaning an object o. Formally,

$$O^\star = \arg_{O_c} \max\{\mathbb{E}[\kappa(\phi|O_c)] \mid O_c \subseteq \mathcal{S}, \sum_{o \in O_c} c(o) \leq B\}$$

III. Optimization Framework

Figure 1 illustrates the general flowchart of our proposed optimization framework QueryClean. Given the cleaning budget B, for a probabilistic query over an uncertain dataset $\mathcal{S}$, QueryClean iteratively selects a group of uncertain objects (stored in a set O_c) for cleaning *with limited resource B available*. QueryClean terminates until the maximum expected query quality of cleaning the set O_c is achieved. Specifically, we minimize the processing costs of the *quality computation* and the *object selection*, respectively.

We first propose a novel *answer-set based indexing structure*, called ASI, which indexes all the possible answer sets for a probabilistic query ϕ. ASI supports to *directly* derive the probability of every answer set for any chosen object set O_c to clean, where an effective *probability update strategy* is leveraged for efficiency enhancement. Using ASI, we present an efficient algorithm RrB for quality computation.

On the other hand, for optimizing the object selection, we present two effective heuristics stated below. The first heuristic offers a small candidate set for the object selection, resulting in fewer possible chosen object sets to clean. The second heuristic reveals the monotonic property of the expected quality function, meaning that the query quality does not drop with an enlarging cleaning object set.

Using the two newly proposed heuristics, we present a *Branch and Bound* (B&B) algorithm, an improved Greedy algorithm, and HSample algorithm to support QueryClean. The idea of B&B is to partition the feasible set, i.e., $\mathcal{O}$, into smaller subsets, and then to eliminate its subsets from further consideration if the cleaning cost of the current chosen object set to clean is within the budget. Greedy selects one cleaning object from $\mathcal{O}$ with the maximum unit-cost expected quality every time until the budget is used up. In contrast, HSample is to capture a group of chosen object sets to clean from the power set of $\mathcal{O}$, and to return the *sampled* object set having the maximum expected quality with a guaranteed accuracy.

IV. Experimental Evaluation

In this section, we evaluate the performance of our proposed framework QueryClean including object selection algorithms B&B, Greedy, and HSample, using both real and synthetic data sets. We explore the influence of budget B on the performance of QueryClean. The experimental results are shown in Figure 2 for probabilistic skyline queries, probabilistic kNN search, and probabilistic range retrieval. It is observed that, B&B is inferior to Greedy and HSample in terms of time cost in most cases. For execution time and quality, Greedy is slightly better than HSample (where sample size is 100) in some cases. With the growth of budget B, the query quality improves gradually, and the time cost becomes larger accordingly. The reason is that, as budget B turns larger, there is enough resource to choose more uncertain objects to clean, resulting in the higher query quality yet more time consumption. In particular, for every dataset, with the support of a constant ASI index, it needs at most 10 seconds for the two approximate algorithms to select the objects for cleaning over the dataset of 100,000 uncertain objects.

Acknowledgments

This work was supported in part by the 973 Program of China under Grant No. 2015CB352502, NSFC Grants under No. 61522208 and 61379033, the NSFC-Zhejiang Joint Fund under Grant No. U1609217, and the Fundamental Research Funds for the Central Universities. Wei Wang was supported by ARC DP 170103710, D2D CRC Grants DC25002 and DC25003. Yunjun Gao is the correspondin author of the work.

References

[1] J. Pei, B. Jiang, X. Lin, and Y. Yuan, "Probabilistic skylines on uncertain data," in *VLDB*, pp. 15–26, 2007.

[2] R. Cheng, J. Chen, M. Mokbel, and C.-Y. Chow, "Probabilistic verifiers: Evaluating constrained nearest-neighbor queries over uncertain data," in *ICDE*, pp. 973–982, 2008.

[3] T. Friedman and M. Smith, *Measuring the business value of data quality*. Gartner, 2011.

[4] R. Cheng, J. Chen, and X. Xie, "Cleaning uncertain data with quality guarantees," in *VLDB*, pp. 722–735, 2008.

[5] L. Mo, R. Cheng, X. Li, D. W. Cheung, and X. S. Yang, "Cleaning uncertain data for top-k queries," in *ICDE*, pp. 134–145, 2013.