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

Contents lists available at ScienceDirect

Journal of Computer and System Sciences

www.elsevier.com/locate/jcss

Range-max queries on uncertain data

Pankaj K. Agarwal [a], Nirman Kumar [b], Stavros Sintos [a,*], Subhash Suri [c]

[a] Department of Computer Science, Duke University, Durham, USA
[b] Department of Computer Science, University of Memphis, Memphis, USA
[c] Department of Computer Science, UC Santa Barbara, Santa Barbara, USA

ARTICLE INFO

Article history:
Received 8 April 2017
Accepted 16 September 2017
Available online xxxx

Keywords:
Data structures
Algorithms
Data uncertainty
Range-max queries
Orthogonal query ranges
Lower bounds
Skylines

ABSTRACT

Let P be a set of n uncertain points in $\mathbb{R}^d$, where each point $p_i \in P$ is associated with a real value v_i and exists with probability $\alpha_i \in (0, 1]$ independently of the other points. We present algorithms for building an index on P so that for a d-dimensional query rectangle ρ, the expected maximum value or the most-likely maximum value in ρ can be computed quickly. Our main contributions include the following: (i) The first index of sub-quadratic size to achieve a sub-linear query time in any dimension. (ii) A conditional lower bound for most-likely range-max queries, based on the conjectured hardness of the set-intersection problem. (iii) A near-linear-size index for estimating the expected range-max value within approximation factor $1/2$ in $O(\text{polylog}(n))$ time. (iv) Extensions of our algorithm to more general uncertainty models and for computing the top-k values of the range-max.

1. Introduction

Query-driven data management is an important function in most database systems, and range query is a common tool for summarizing information about the objects lying in a query range. In recent years, motivated by applications in sensor networks, data cleaning, data integration, pervasive computing and scientific data analysis, there has been a growing interest in managing *uncertain data* [6,14]. In these settings, uncertainty is typically captured using stochastic data models, and summarizing or querying data requires computing *statistics* about the probabilistic behavior of the underlying data.

In this paper we study the *range-max* query on uncertain data, which asks for statistics on the maximum value of the data inside a query range. Formally, we consider database records of the form (p, v, α), where p is a *point* in $\mathbb{R}^d$ (attribute values of the record) with $d \geq 1$ a constant, along with a positive *scalar value* $v \in \mathbb{R}^+$ and a probability $\alpha \in (0, 1]$. We refer to this simple model as the *existence-uncertainty* model. The value v is the metric of interest in our range queries and the probability α reflects our confidence in this record. Given a collection of n such records, our goal is to construct an index to answer queries of the following form efficiently: report the *expected maximum value* (the EM problem) or the *most likely maximum value* (MLM problem) of the records whose attribute values (points) lie in a d-dimensional orthogonal query rectangle ρ. We use P to denote the set of points corresponding to the input records. Throughout the paper, we assume the real-RAM model of computation, which counts word-level operations instead of bits: specifically, each memory cell can store an arbitrary real number and any arithmetic or relational operation between two operands takes $O(1)$ time.

* Corresponding author.
E-mail addresses: pankaj@cs.duke.edu (P.K. Agarwal), nkumar8@memphis.edu (N. Kumar), ssintos@cs.duke.edu (S. Sintos), suri@cs.ucsb.edu (S. Suri).

https://doi.org/10.1016/j.jcss.2017.09.006
0022-0000/

Table 1
Summary of main results. Lower bounds are conditional to the set-intersection conjecture and hold for $d \geq 2$; RPM is the random permutation model; and t_q is the query time.

Model	Algorithm	EM		MLM	
		Space	Query	Space	Query
Existence	Exact	$O\left(n^{(2d-1)t+1}\right)$	$O\left(n^{1-t}+\log n\right)$	$O\left(n^{(2d-1)t+1}\right)$	$O\left(n^{1-t}+\log n\right)$
			$\Omega\left(\left(\frac{n}{t_q \operatorname{polylog}(n)}\right)^2\right)$	t_q	
	Approx.	$O\left(n\log^d n\right)$	$O\left(\log^{d+1} n\right)$		
	RPM		$O\left(n\log^{d+1} n\right)$	$O\left(\log^{d+3} n\right)$	
Location	Exact	$O\left(n^{(2d-1)t+1} f^{2d}\right)$	$O\left(n^{1-t}+\log(nf)\right)$	$O\left(n^{(2d-1)t+1} f^{2d}\right)$	$O\left(n^{1-t}+\log(nf)\right)$
		$\Omega\left(\left(\frac{n}{t_q \operatorname{polylog}(n)}\right)^2\right)$	t_q	$\Omega\left(\left(\frac{n}{t_q \operatorname{polylog}(n)}\right)^2\right)$	t_q
	Approx.	$O\left(nf\log^d(nf)\right)$	$O\left(\log^{d+1}(nf)\right)$		
	RPM		$O\left(nf^{2d}\log^{2d+1}(nf)\right)$	$O\left(\log^{2d+3}(nf)\right)$	

The max is a special case of the widely studied top-k summary of data (i.e. $k = 1$) but is also a popular aggregation operator in its own right. For example, the suprema of random processes (a collection of random variables arising from given distributions) are widely studied in random matrix theory, control of empirical processes in statistics and machine learning, random optimization problems, and probability in Banach spaces. The reason for this keen interest can be understood from the following hypothetical but representative application. The points might represent a set of geographical locations (cities) $\{p_1, p_2, \ldots, p_n\}$, each associated with a probability α_i of being struck by a natural disaster (flood, earthquake, fire) during next year, and v_i a measure of the *cost* (damage) incurred at that location due to that disaster.[1] In this case, a range query asks for the expected value of the maximum damage suffered within the range. Similarly, an insurance company may associate probabilities of financial claims with various entities, and need to analyze the profile of its maximum loss portfolio. Indeed, in natural disasters such as earthquakes or flooding, the impact is highly *non-linear*—even hundreds of small quakes hardly cause serious financial or social harm, but a single large one can be catastrophic. Thus, it is far more important to be able to carry out analysis on the profile of the *maximum values*, and not on simpler aggregates such as sum or average.

Range-max queries on uncertain data are also relevant when dealing with spatially distributed noisy data sources (e.g. sensors) where unusually high measurements might be cause for concern, but only if they deviate from the norm. A probabilistic profile of the expected max in a range can serve as the benchmark for deciding when a sensor measurement is abnormal.

We point out that, given a query rectangle ρ, the expected value of *range-SUM*, namely, $\sum_{p_i \in \mathsf{P} \cap \rho} \alpha_i v_i$ is easy to compute: we simply assign each point p_i a *weight* of $\alpha_i v_i$ and compute the weighted range sum using traditional (non-uncertain) techniques. In contrast, computing the expected value of *range-max*, our EM problem, seems much harder and it is not clear how to use any existing range-aggregation index to answer this query. The difficulty arises in part because the EM problem is not *range decomposable*, i.e., if P_1, P_2 is a partition of P then for a query rectangle ρ, the EM value of $\mathsf{P} \cap \rho$ cannot be quickly computed from the EM values of $\mathsf{P}_1 \cap \rho$ and $\mathsf{P}_2 \cap \rho$. The interaction of probabilities rules out traditional tree-style range-searching indexes where the value at each node is inferred simply from values at its children. On the positive side, while the probability distribution of range-SUM can take exponentially many possible values, the probability distribution of range-max only has linear size. Therefore, computing statistics on the distribution of range-max (e.g. most likely max) might be easier than a similar statistics on range-SUM; the latter is known to be $\#P$-hard [19].

For both the EM and MLM queries in 1-dimension, an index structure with $O(n^2)$ size and $O(\log n)$ query time is straightforward: we can precompute answers for each of the $O(n^2)$ *combinatorially distinct* intervals defined by pairs of input points. The interesting question is whether these queries can be answered in sublinear time using subquadratic space. In this paper, we answer this question affirmatively.

Our results. Our main results are shown in Table 1, and can be summarized as follows:

(A) We design the first sub-quadratic size index that achieves a sub-linear query time in any dimension $d \geq 1$ for both EM and MLM problems. For any $t \in [0, 1]$, the index answers a query in $O(n^{1-t} + \log n)$ time, has size $O(n^{(2d-1)t+1})$, and takes $O(n^{(2d-1)t+1} \log n)$ time to build. The tunable parameter t gives the index a continuum of trade-offs between query time and its size. In particular, for $d = 1$ the index can achieve a query time of $O(\sqrt{n})$ with $O(n^{3/2})$ size and $O(n^{3/2} \log n)$ preprocessing.

[1] More generally, each point can be associated with not just a single pair (value, probability) but an entire distribution. For the ease of exposition, we initially focus on the single value case, but remark on how to generalize our results to distributional settings later.

Please cite this article in press as: P.K. Agarwal et al., Range-max queries on uncertain data, J. Comput. Syst. Sci. (2018), https://doi.org/10.1016/j.jcss.2017.09.006