Theoretical Computer Science 609 (2016) 32–48

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Range queries on uncertain data [☆]

Jian Li [a], Haitao Wang [b,*]

[a] Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing 100084, China
[b] Department of Computer Science, Utah State University, Logan, UT 84322, USA

ARTICLE INFO

Article history:
Received 7 January 2015
Received in revised form 7 July 2015
Accepted 6 September 2015
Available online 10 September 2015
Communicated by G.F. Italiano

Keywords:
Range queries
Uncertain data
Top-k queries
Threshold queries
Data structures
Algorithms

ABSTRACT

Given a set P of n uncertain points on the real line, each represented by its one-dimensional probability density function, we consider the problem of building data structures on P to answer range queries of the following three types for any query interval I: (1) top-1 query: find a point in P that lies in I with the highest probability, (2) top-k query: given any integer $k \le n$ as part of the query, return the k points in P that lie in I with the highest probabilities, and (3) threshold query: given any threshold τ as part of the query, return all points of P that lie in I with probabilities at least τ. We present data structures for these range queries with linear or nearly linear space and efficient query time.

1. Introduction

With a rapid increase in the number of application domains, such as data integration, information extraction, sensor networks, scientific measurements, etc., where uncertain data are generated in an unprecedented speed, managing, analyzing and query processing over such data have become a major challenge and have received significant attentions. We study an important problem in this domain, building data structures for uncertain data for efficiently answering certain range queries. The problem has been studied extensively with a wide range of applications [3,14,28,33,36,43,44]. We formally introduce the problem below.

Let $\mathbb{R}$ be any real line (e.g., the x-axis). In the (traditional) deterministic version of this problem, we are given a set P of n deterministic points on $\mathbb{R}$, and the goal is to build a data structure (also called "index" in database) such that given a range, specified by an interval $I \subseteq \mathbb{R}$, one point (or all points) in I can be retrieved efficiently. It is well known that a simple solution for this problem is a binary search tree over all points which is of linear size and can support logarithmic (plus output size) query time. However, in many applications, the location of each point may be uncertain and the uncertainty is represented in the form of probability distributions [4,6,14,43,44]. In particular, an *uncertain point* p is specified by its probability density function (pdf) $f_p : \mathbb{R} \to \mathbb{R}^+ \cup \{0\}$. Let P be the set of n uncertain points in $\mathbb{R}$ (with pdfs specified as input). Our goal is to build data structures to quickly answer range queries on P. In this paper, we consider the following three types of range queries, each of which involves a query interval $I = [x_l, x_r]$. For any point $p \in P$, we use $\Pr[p \in I]$ to denote the probability that p is contained in I.

[☆] A preliminary version of this paper appeared in the Proceedings of the 25th International Symposium on Algorithms and Computation (ISAAC 2014).
[*] Corresponding author.
E-mail addresses: lijian83@mail.tsinghua.edu.cn (J. Li), haitao.wang@usu.edu (H. Wang).

http://dx.doi.org/10.1016/j.tcs.2015.09.005
0304-3975/© 2015 Elsevier B.V. All rights reserved.

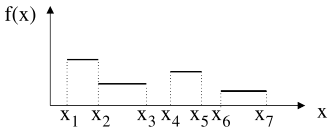

Fig. 1. The pdf of an uncertain point.

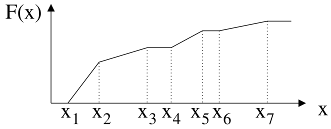

Fig. 2. The cdf of the uncertain point in Fig. 1.

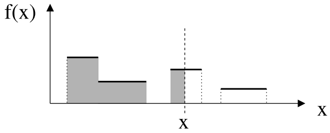

Fig. 3. Geometrically, $F_p(x)$ is equal to the sum of the areas of the shaded rectangles.

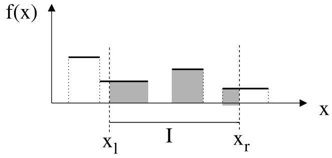

Fig. 4. Geometrically, the probability $\Pr[p \in I]$ is equal to the sum of the areas of the shaded rectangles.

Top-1 query: Return a point p of P such that $\Pr[p \in I]$ is the largest.
Top-k query: Given any integer k, $1 \le k \le n$, as part of the query, return the k points p of P such that $\Pr[p \in I]$ are the largest.
Threshold query: Given a threshold τ, as part of the query, return all points p of P such that $\Pr[p \in I] \ge \tau$.

Note that for each top-1 query, if P has more than one point p such that $\Pr[p \in I]$ is the largest, then an arbitrary such point is considered as a correct answer to the query. Similarly, we break ties arbitrarily for the top-k queries.

We assume f_p is a step function, i.e., a *histogram* consisting of at most c pieces (or intervals) for some integer $c \ge 1$ (e.g., see Fig. 1). More specifically, $f_p(x) = y_i$ for $x_{i-1} \le x < x_i$, $i = 1, \dots, c$, with $x_0 = -\infty$, $x_c = \infty$, and $y_1 = y_c = 0$. Throughout the paper, we assume c is a constant. The cumulative distribution function (cdf) $F_p(x) = \int_{-\infty}^{x} f_p(t)dt$ is a monotone piecewise-linear function consisting of c pieces (e.g., see Fig. 2). Note that $F_p(+\infty) = 1$, and for any interval $I = [x_l, x_r]$ the probability $\Pr[p \in I]$ is $F_p(x_r) - F_p(x_l)$. From a geometric point of view, each interval of f_p defines a rectangle with the x-axis, and the sum of the areas of all these rectangles of f_p is exactly one. Further, the cdf value $F_p(x)$ is the sum of the areas of the subsets of these rectangles to the left of the vertical line through x (e.g., see Fig. 3), and the probability $\Pr[p \in I]$ is the sum of the areas of the subsets of these rectangles between the two vertical lines through x_l and x_r, respectively (e.g., see Fig. 4).

As discussed in [3], the histogram model can be used to approximate most pdfs with arbitrary precision in practice. In addition, the *discrete* pdf where each uncertain point can appear in a few locations, each with a certain probability, can be viewed as a special case of the histogram model because we can use infinitesimal pieces around these locations.

We also study an important special case where the pdf f_p is a uniform distribution function, i.e., f is associated with an interval $[x_l(p), x_r(p)]$ such that $f_p(x) = 1/(x_r(p) - x_l(p))$ if $x \in [x_l(p), x_r(p)]$ and $f_p(x) = 0$ otherwise. Clearly, the cdf $F_p(x) = (x - x_l(p))/(x_r(p) - x_l(p))$ if $x \in [x_l(p), x_r(p)]$, $F_p(x) = 0$ if $x \in (-\infty, x_l(p))$, and $F_p(x) = 1$ if $x \in (x_r(p), +\infty)$. Uniform distributions have been used as a major representation of uncertainty in some previous work (e.g., [12,14,30]). We refer to this special case as the *uniform case* and the more general case where f_p is a histogram distribution function as the *histogram* case.

Throughout the paper, we will always use $I = [x_l, x_r]$ to denote the query interval. The query interval I is *unbounded* if either $x_l = -\infty$ or $x_r = +\infty$ (otherwise, I is *bounded*). For the threshold query, we will always use m to denote the output size of the query, i.e., the number of points p of P such that $\Pr[p \in I] \ge \tau$.

Range reporting on uncertain data has many applications [3,14,28,36,43,44], As shown in [3], our problems are also useful even in some applications that involve only deterministic data. For example, consider the movie rating system in IMDB where each reviewer gives a rating from 1 to 10. A top-k query on $I = [7, +\infty)$ would find "the k movies such that