Significant-Presence Range Queries in Categorical Data

(extended abstract)

Mark de Berg[*] Herman J. Haverkort[†]

Abstract

In traditional colored range-searching problems, one wants to store a set of n objects with m distinct colors for the following queries: report all colors such that there is at least one one object of that color intersecting the query range. Such an object, however, could be an 'outlier' in its color class. Therefore we consider a variant of this problem where one has to report only those colors such that at least a fraction τ of the objects of that color intersects the query range, for some parameter τ. We present efficient data structures for such queries with orthogonal query ranges in sets of colored points, and for point stabbing queries in sets of colored rectangles.

1 Introduction

Motivation. The range-searching problem is one of the most fundamental problem in computational geometry. In this problem we wish to construct a data structure on a set S of objects in $\mathbb{R}^d$, such that we can quickly decide for a query range which of the input objects it intersects. The range-searching problem comes in many flavors, depending on the type of objects in the input set S, on the type of allowed query ranges, and on the required output (whether one wants to report all intersected objects, to count the number of intersected objects, etc.). The range-searching problem is not only interesting because it is such a fundamental problem, but also because it arises in numerous applications in areas like databases, computer graphics, geographic information systems, and virtual reality. Hence, it is not surprising that there is an enormous literature on the subject—see for instance the surveys by Agarwal [1], Agarwal and Erickson [2], and Nievergelt and Widmayer [4].

In this paper, we are interested in range searching in the context of databases. Here one typically wants to be able to answer questions like: given a database of customers, report all customers whose ages are between 20 and 30, and whose income is between \$50,000 and \$75,000. In this example, the customers can be represented as points in $\mathbb{R}^2$, and the query range is an axis-parallel rectangle.[1] This is called the (planar) *orthogonal range-searching problem*, and it has been studied extensively.

There are situations, however, where the data points are not all of the same type but fall into different categories. Suppose, for instance, that we have a database of stocks. Each stock falls into a certain category, namely the industry sector it belongs to—IT, energy, banking, food, chemicals, etc. Then it can be interesting for an analyst to get answers to questions like: "In which sectors companies had a 10–20% increase in their stock values over the past year?" In this simple example, the input data can be seen as points in 1D (namely for each stock its increase in value), and the query is a one-dimensional orthogonal range-searching query.

Now we are no longer interested in reporting all the points in the range, but only the categories that have points in the range. This means that we would like to have a data structure whose query time is not sensitive to the total number of points in the range, but to the total number of

[*] Department of Computer Science, TU Eindhoven, P.O.Box 513, 5600 MB Eindhoven, the Netherlands.

[†] Institute of Information and Computing Science, Utrecht University, P.O.Box 80.089, 3508 TB Utrecht, the Netherlands. The work by H.H. is supported by the Netherlands' Organization for Scientific Research (NWO).

[1] From now on, whenever we use terms like "rectangle" or "box" we implicitly assume these are axis-parallel.

categories in the range. This can be achieved by building a suitable data structure for each category separately, but this is inefficient if the number of categories is large. This has led researchers to study so-called *colored range-searching problems*: store a given set of colored objects—the color of an object represents its category—such that one can efficiently report those colors that have at least one object intersecting a query range [3, 5, 6].

We believe, however, that this is not always the correct abstracted version of the range-searching problem in categorical data. Consider for instance the stock example sketched earlier. The standard colored range-searching data structures would report all sectors that have *at least one* company whose increase in stock value lies in the query range. But this does not necessarily say anything about how the sector is performing: a given sector could be doing very badly in general, but contain a single 'outlier' whose performance has been good. It is much more natural to ask for all sectors for which *most* stocks, or at least a significant portion of them, had their values increase in a certain way. Therefore we propose a different version of the colored range-searching problem: given a fixed threshold parameter τ, with $0 < \tau < 1$, we wish to report all colors such that at least a fraction τ of the objects of that color intersect the query range. We call this a *significant-presence query*, as opposed to the standard *presence query* that has been studied before. (We also have some results on the case where τ is not fixed beforehand, but part of the query. Due to lack of space, these results are omitted.)

Problem statement and results. We study significant-presence queries in categorical data in two settings: orthogonal range searching where the data is a set of colored points in $\mathbb{R}^d$ and the query is a box, and stabbing queries where the data is a set of colored boxes in $\mathbb{R}^d$ and the query is a point. In this extended abstract, we only discuss our results on orthogonal range searching. We also omit several of the proofs.

Let $S = S_1 \cup \cdots \cup S_m$ be a set of n points in $\mathbb{R}^d$, where m is the number of different colors and S_i is the subset of points of color class i. Let τ be a fixed parameter with $0 < \tau < 1$. We are interested in answering significant-presence queries on S: given a query box Q, report all colors i such that $|Q \cap S_i| \geq \tau \cdot |S_i|$. For $d = 1$, we present a data structure that uses $O(n)$ storage, and that can answer significant-presence queries in $O(\log n + k)$ time, where k is the number of reported colors. Unfortunately, for $d \geq 2$, we have not been able to design a data structure using near-linear storage with logarithmic query time for this problem. As a data structure with quadratic or more storage is prohibitive in practice, we study an approximate version of the problem. More precisely, we study *ε-approximate significant-presence queries*: here we are required to report all colors i with $|Q \cap S_i| \geq \tau \cdot |S_i|$, but we are also allowed to report colors with $|Q \cap S_i| \geq (1-\varepsilon)\tau \cdot |S_i|$, where ε is a fixed positive constant. For such queries we have developed a data structure that uses only $O((1/(\tau\varepsilon)^{2d-1}m)^{1+\delta})$ storage, for any $\delta > 0$, and that can answer queries in $O(\log n + k)$ time, where k is the number of reported colors. Note that the amount of storage does not depend on n, the total number of points, but only on m, the number of colors. This should be compared to the results for the previously considered case of presence queries on colored points sets. Here the best known results are: $O(n)$ storage with $O(\log n + k)$ query time for $d = 1$ [6], $O(n \log^2 n)$ storage with $O(\log n + k)$ query time for $d = 2$ [6], $O(n \log^4 n)$ storage with $O(\log^2 n + k)$ query time for $d = 3$ [5], and $O(n^{1+\delta})$ storage with $O(\log n + k)$ query time for $d \geq 4$ [3]. Note that these bounds all depend on n, the total number of points; this is of course to be expected, since these results are all on the exact problem, whereas we allow ourselves approximate answers.

2 Orthogonal range queries

One of the difficulties in significant-presence queries is that the problem is not readily decomposable: we cannot decide whether a color is significantly present in a range Q if we just know whether or not certain subsets of that color are significantly present in Q. In this respect, standard presence queries are easier: a color is present in Q iff a subset of that color is present in Q. Hence, our approach is to first reduce significant-presence queries to standard presence queries. We do this by introducing so-called *test sets*.

Test sets for orthogonal range queries Let S be a set of n points in $\mathbb{R}^d$, and let τ be a fixed parameter with $0 < \tau < 1$. A set T of boxes—that is, axis-parallel hyperrectangles—is called a *τ-test set* for S if the following holds:

- any box from T contains at least τn points from S;
- any query box Q that contains at least τn points from S fully contains at least one box from T.

This means that we can answer a significant-presence query on S by answering a presence query on T: a query box Q contains at least τn points from S if and only if it contains at least one box from T. We did not yet reduce the problem to a standard presence-query problem, because T contains boxes instead of points. However, we can map the set T of boxes in $\mathbb{R}^d$ to a set of points in $Reals^{2d}$, and the query box Q to a box in $\mathbb{R}^{2d}$, in such a way that a box $b \in T$ is fully contained in Q if and only if its corresponding point in $Reals^{2d}$ is contained in the transformed query box.[2] This means we can apply the results from the standard presence queries on colored point sets.

It remains to find small test sets. As it turns out, this is not possible in general: below we show that there are point sets that do not admit test sets of near-linear size. Hence, after studying the case of exact test sets, we will turn our attention to approximate test sets.

Exact test sets. It is easy to see that any minimal box containing at least τn points from S—that is, any box b containing at least τn points from S such that there is no box $b' \neq b$ with $b' \subset b$ and containing τn or more points—must be a box in T, and that the collection of all such minimal boxes forms a τ-test set. Hence, the smallest possible test set consists exactly of these minimal boxes. In the 1-dimensional case a box is a segment, and a minimal segment is uniquely defined by the point from S that is its left endpoint. This means that any set of n points on the real line has a test set of size $(1 - \tau)n + 1$. Unfortunately, the size of test sets increases rapidly with the dimension, as the next lemma shows.

Lemma 2.1 *For any set S of n points in $\mathbb{R}^d$, there is a τ-test set of size $O(\tau^{d-1}n^{2d-1})$. Moreover, there are sets S for which any τ-test set has size $\Omega(\tau^{d-1}n^{2d-1})$. (proof omitted from this abstract)*

Approximate test sets. The worst-case bound from Lemma 2.1 is quite disappointing. Therefore we now turn our attention to approximate test sets: a set T of boxes is called an *ε-approximate τ-test set* for a set S of n points if

- any box from T contains at least $(1 - \varepsilon)\tau n$ points from S;
- any query box Q that contains at least τn points from S fully contains at least one box from T.

This means we can answer ε-approximate significant-presence queries on S by answering a presence query on T.

Lemma 2.2 *For any set S of n points in $\mathbb{R}^d$ $(d > 1)$ and any ε with $0 < \varepsilon < 1/2$, there is an ε-approximate τ-test set of size $(2d - 1)^{2d-1}/(\varepsilon^{2d-1}\tau^{2d-2})$. Moreover, there are sets S for which any ε-approximate τ-test set has size $\Omega((1/\varepsilon)^{2d-1}(1/\tau)^d)$.*

Proof. We will prove an upper bound of $((2d - 1)/(\varepsilon\tau))^{2d-1}$ here; an improvement of a factor τ can be attained with a divide-and-conquer variation of the construction described below. Due to lack of space, we omit the details of the divide-and-conquer approach from this extended abstract.

To prove the upper bound of $((2d - 1)/(\varepsilon\tau))^{2d-1}$, we proceed as follows. We construct a collection H_1 of $(2d-1)/(\varepsilon\tau)$ hyperplanes orthogonal to the x_1-axis, such that there are $\varepsilon\tau n/(2d-1)$ points of S between any pair of consecutive hyperplanes.[3] We do the same for the other axes, obtaining sets $H_2, \ldots, H_d$ of hyperplanes. From these collections of hyperplanes we construct our test set as follows. Take any possible subset H^* of $2d - 1$ hyperplanes from $H_1 \cup \cdots \cup H_d$ such that H_1 up to H_{d-1} each contribute exactly two hyperplanes to H^*, and H_d contributes one hyperplane. Let $b(H^*)$ be the smallest box that is bounded by the hyperplanes from H^*, contains

[2] In fact, the transformed box is unbounded to one side along each coordinate-axis, so it is a d-dimensional 'octant'.

[3] If there are more points with the same x_1-coordinate, we have to be a bit careful. The details are omitted from this extended abstract.

exactly $(1-\varepsilon)\tau n$ points from S, and lies above the hyperplane we picked from H_d. If $b(H^*)$ exists, we add it to our test set T. Clearly, the size of T is at most $((2d-1)/(\varepsilon\tau))^{2d-1}$.

We now argue that T is an ε-approximate τ-test set for S. By construction, every box in T contains at least $(1-\varepsilon)\tau n$ points, so it remains to show that every box Q that contains at least τn points from S fully contains at least one box from T. To see this, observe that for any i with $1 \leq i \leq d$, there must be a hyperplane $h_1^{(i)} \in H_i$ that intersects Q and has at most $\varepsilon\tau n/(2d-1)$ points from $Q \cap S$ 'below' it. Similarly, there is a hyperplane $h_2^{(i)} \in H_i$ intersecting Q with at most $\varepsilon\tau n/(2d-1)$ points from $Q \cap S$ 'above' it. Note that $h_2^{(i)} \neq h_1^{(i)}$. Now consider the box b bounded by the hyperplanes in the set $H^* := \{h_1^{(1)}, h_2^{(1)}, \ldots, h_1^{(d-1)}, h_2^{(d-1)}, h_1^{(d)}\}$ and unbounded in the positive x_d-direction. The number of points from S in $b \cap Q$ is at least $|Q \cap S| - (2d-1) \cdot \varepsilon\tau n/(2d-1) \geq (1-\varepsilon)\tau n$. Hence, the box $b(H^*) \in T$ is not larger than $b \cap Q$ and, hence, it is contained in Q.

The lower-bound construction is omitted in this extended abstract. □

Putting it all together. To summarize, the construction of our data structure for ε-approximate significant-presence queries on $S = S_1 \cup \cdots \cup S_m$ is as follows. We construct an ε-approximate τ-test set T_i for each color class S_i. This gives us a collection of $O(1/(\varepsilon^{2d-1}\tau^{2d-2})m)$ boxes in $\mathbb{R}^d$. We map these boxes to a set P of colored points in $\mathbb{R}^{2d}$, and construct a data structure for the standard colored range-searching problem (that is, presence queries) on P, using the techniques of Agarwal et al. [3]. Their structure was designed for searching on a grid, but using the standard trick of normalization—replace every coordinate by its rank, and transform the query box to a box in this new search space in $O(\log n)$ time before running the query algorithm—we can employ their results in our setting.

The same technique works for exact queries, if we use exact test sets. This gives a good result for $d = 1$, if we use the results from Gupta *et al.* [5] on quadrant range searching.

Theorem 2.1 *Let $S = S_1 \cup \cdots \cup S_m$ be a colored point set in $\mathbb{R}^d$, and τ a fixed constant with $0 < \tau < 1$. For $d = 1$, there is a data structure that uses $O(n)$ storage such that exact significant-presence queries can be answered in $O(\log n + k)$ time, where k is the number of reported colors. For $d > 1$, there is, for any ε with $0 < \varepsilon < 1/2$ and any $\delta > 0$, a data structure for S that uses $O((1/(\varepsilon^{2d-1}\tau^{2d-2})m)^{1+\delta})$ storage such that ε-approximate significant-presence queries on S can be answered in $O(\log n + k)$ time.*

Acknowledgements We thank Joachim Gudmundsson and Jan Vahrenhold for inspiring discussions about the contents of this paper.

References

[1] P.K. Agarwal. Range Searching. In: J. Goodman and J. O'Rourke (Eds.), *CRC Handbook of Computational Geometry*, CRC Press, pages 575–598, 1997.

[2] P.K. Agarwal and J. Erickson. Geometric range searching and its relatives. In: B. Chazelle, J. Goodman, and R. Pollack (Eds.), *Advances in Discrete and Computational Geometry*, Vol. 223 of *Contemporary Mathematics*, pages 1–56, American Mathematical Society, 1998.

[3] P.K. Agarwal, S. Govindarajan, and. S. Muthukrishnan. Range searching in categorical data: colored range searching on a grid. In *Proc. 10th Annu. European Sympos. Algorithms* (ESA 2002), pages 17–28, 2002.

[4] J. Nievergelt and P. Widmayer. Spatial data structures: concepts and design choices. In: J.-R. Sack and J. Urrutia (Eds.) *Handbook of Computational Geometry*, pages 725–764, Elsevier Science Publishers, 2000.

[5] J. Gupta, R. Janardan, and M. Smid. Further results on generalized intersection searching problems: counting, reporting, and dynamization. In *Proc. 3rd Workshop on Algorithms and Data Structures*, LNCS 709, pages 361–373, 1993.

[6] R. Janardan and M. Lopez. Generalized intersection searching problems. *Internat. J. Comput. Geom. Appl.* 3:39–70 (1993).