Partial Match in Hierarchical Multidimensional Data Structures

Amalia Duch
Gustavo Lau
Conrado Martínez
Technical University of Catalonia–Barcelona Tech

ABSTRACT

We discuss the average-case analysis of partial match queries in general purpose hierarchical multidimensional data structures as k-d trees and quad trees. In particular, we define the problem, we give a quick description of the existing results and, we present our current lines of research to go further on its understanding.

KEYWORDS

Associative retrieval, partial match, k-d trees, quad trees, average-case analysis, multidimensional data structures.

ACM Reference format:
Amalia Duch, Gustavo Lau, and Conrado Martínez. 2017. Partial Match in Hierarchical Multidimensional Data Structures. In *Proceedings of ACM Celebration of Women in Computing womENcourage 2017, Barcelona, September 2017 (womENcourage 2017),* 1 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

Have you ever asked Google Maps for the closest gas station? Or TripAdvisor for good restaurants around your location area? These questions are examples of what we formally call the *Associative Retrieval* problem, a computing task frequent in applications.

In associative retrieval we consider a collection $\mathcal{F}$ of n records, where each *record* is an ordered k-tuple ($k \geq 2$) $x = (x_0, \ldots, x_{k-1})$ of values (the attributes or coordinates of the record) drawn from domain $D = \prod_{0 \leq j < k} D_j$, where each D_j is totally ordered.

A *query* over $\mathcal{F}$ is a retrieval of all records whose attributes satisfy some given conditions. The query is considered *associative* when it deals with at least two attributes. Examples of associative queries are: (i) *nearest-neighbor* queries, to retrieve the record in $\mathcal{F}$ closest to a given record under a given distance, (ii) *partial match* queries (PM queries), to retrieve all records in $\mathcal{F}$ that match the s (out of k) attributes of the query that are specified, or (iii) *range* queries, to retrieve all records in $\mathcal{F}$ that fall inside a given region.

In order to efficiently deal with associative queries the storage of the records in $\mathcal{F}$ is crucial. Thus, general purpose multidimensional data structures –such as k-d trees and quad trees– are adequate storage methods for supporting a wide range of associative queries. The correct election of the data structure (DS) that better fits an application requires a deep understanding of the DS's performance towards possible associative queries. Our research focused precisely in the average-case analysis of PM queries in hierarchical multidimensional DSs. Indeed, the study of PM queries is fundamental in this computing area: (i) because of their intrinsic interest and (ii) because the analysis of other associative queries (such as range queries) is based on it [5].

womENcourage 2017, Barcelona
2017. 978-x-xxxx-xxxx-x/YY/MM…$15.00
DOI: 10.1145/nnnnnnn.nnnnnnn

The study of PM queries started with the seminal paper of Flajolet and Puech [6] studying random partial match queries (contrary to fixed ones) in k-d trees and k-d tries. Currently, there are several results in the literature analyzing random and fixed PM queries. The cost of PM queries in general purpose hierarchical DSs is of the form $O(n^{\phi})$ where the sublinear exponent ϕ is specific of the DS under consideration. In our research work we have extensively studied random and fixed PM queries finding the value of ϕ (and sometimes the hidden constant) on a variety of DSs [2–4].

Our efforts to go further on the understanding of PM queries follows two different lines of research. The first approach is (following the research line in this area up to now) to analyze the expected performance of PM queries in as many hierarchical multidimensional DSs as possible trying to determine how the performance of PM varies with respect to the characteristics of every specific DS. Open problems in this approach are, for instance, (i) to determine a general form (if it exists) of the exponent ϕ and its dependence with respect to s, k, the kind of partial match query, and the characteristics of the specific tree, (ii) to obtain the expected cost of PM in k-d-t trees (with $t \geq 2$) and quad trees (with $k > 2$) –including the hidden factor in the asymptotic notation– and to determine how the cost fluctuates from the highest one of relaxed k-d trees to the best and optimal one of squarish k-d trees.

Our second line of research deals with the analysis of PM queries in quad k-d trees: a general framework for the joint study of hierarchical multidimensional DSs [1]. Quad trees are 2^k-ary trees (where each node discriminates by all the coordinates), k-d trees are 2-ary trees (where each node discriminates by one coordinate), and quad k-d trees are trees where every node has arity 2^i (for $0 < i \geq k$). The choice of i is given by an insertion heuristic (many are proposed [1]). Since quad k-d trees include quad trees and k-d trees –as well as a wide range of different hierarchical DSs– as particular cases, a succesful analysis of PM queries in quad k-d trees would imply a unifying analysis of PM in hierarchical DSs [4].

REFERENCES

[1] N. Bereczky, A. Duch, K. Németh, and S. Roura. 2016. Quad-k-d trees: A general framework for k-d trees and quad trees. *Theoretical Computer Science* 616 (2016), 126–140. DOI: https://doi.org/10.1016/j.tcs.2015.12.030
[2] A. Duch, R. M. Jiménez, and C. Martínez. 2014. Selection by rank in K-dimensional binary search trees. *Random Structures and Algorithms* 45, 1 (2014), 14–37.
[3] Amalia Duch and Gustavo Lau. 2017. Partial Match Queries in Relaxed K-dt trees. In *Proceedings of the Fourteenth Workshop on Analytic Algorithmics and Combinatorics, ANALCO 2017, Barcelona, Spain, Hotel Porta Fira, January 16-17, 2017.* 131–138. DOI: https://doi.org/10.1137/1.9781611974775.13
[4] A. Duch, G. Lau, and C. Martínez. 2016. On the Cost of Fixed Partial Match Queries in K-d Trees. *Algorithmica* 75, 4 (2016), 684–723. DOI: https://doi.org/10.1007/s00453-015-0097-4
[5] A. Duch and C. Martínez. 2002. On the Average Performance of Orthogonal Range Search in Multidimensional Data Structures. *Journal of Algorithms* 44, 1 (2002), 226–245. DOI: https://doi.org/{10.1016/S0196-6774(02)00213-4}
[6] Ph. Flajolet and C. Puech. 1986. Partial Match Retrieval of Multidimensional Data. *J. ACM* 33, 2 (1986), 371–407.