On MBR Approximation of Histories for Historical Queries: Expectations and Limitations

Reza Sherkat
Department of Computing Science
University of Alberta
reza@cs.ualberta.ca

Davood Rafiei
Department of Computing Science
University of Alberta
drafiei@cs.ualberta.ca

Abstract

Traditional approaches for efficiently processing historical queries, where a history is a multidimensional time-series, employ a two step filter-and-refine scheme. In the filter step, an approximation of each history often as a set of minimum bounding hyper-rectangles (MBRs) is organized using a spatial index structure such as R-tree. The index is used to prune redundant disk accesses and to reduce the number of pairwise comparisons required in the refine step. To improve the efficiency of the filtering step, a heuristic is used to decrease the expected number of MBRs that overlap with a query, by reducing the volume of empty space indexed by the index. The heuristic selects, among all possible splitting schemes of a history, the one which results to a set of MBRs with minimum total volume. Although this heuristic is expected to improve the performance of spatial and history based queries with small temporal and spatial extents, in many real settings, the performance of historical queries depends on the extent of the query. Moreover, the optimal approximation of a history is not always the one with minimum total volume. In this paper, we present the limitations of using volume as a criteria for approximating histories, specially in high dimensional cases, where it is not feasible to index MBRs by traditional spatial index structures.

1 Introduction

There are many applications where the history of changes to an object or an entity can be described as a d-dimensional time-series, referred to here as history for short. A history is a sequence of points in a domain specific d-dimensional feature space. In transportation systems, for instance, the history of mobile objects (such as cars or people) can be tracked using GPS systems and for each object at each time, the location and speed can be recorded. In the financial sector, the history of a stock may be tracked using indicators such as daily opening and closing prices, trading volume, etc. In health and medicine, changes to body temperature, blood pressure, heart beat rate and blood sugar may be recorded to monitor the recovery history of a patient. In meteorology, measurements such as temperature, precipitation, wind speed, pressure, moisture and snowfall are regularly collected (e.g. daily or hourly) for many earth surfaces by weather stations. Given a database of histories, Pfoser et al. [4] propose two types of historical queries, coordinate based queries and history based queries. Given a time interval and a spatial range as a query, coordinate based queries search for histories (or the number of histories) that overlap with the query. On the other hand, history based queries involve the whole or part of the history of an object. Similarity query is an example of history based queries, which finds application in exploratory analysis, clustering, and prediction. For instance, finding patients with similar recovery histories may be useful for treatments or the trial of a new drug. The demand for indexing histories increases with an increasing number of domains where historical data are generated and stored.

1.1 Indexing Histories for Efficient Retrieval

Several index structures have been proposed in the literature for historical queries. To index histories for coordinate based queries, time is often considered as another spatial dimension. One straightforward approach for indexing histories is to construct an MBR for each history and index all MBRs in a spatial index structure such as an R-tree. In this approach, each MBR corresponds to a single history and encloses the spatial and temporal extents of the history it approximates. This approach, in general, uses poor approximation of histories and it is generally expected to show a poor performance, except for queries that have a long temporal extent. To address this problem, Had-

jieleftheriou et al. [1] propose a refined approximation of histories with more than one MBR. Their approach is expected to improve the performance by reducing the amount of empty space indexed. However, the authors admit that this approach can inversely affect the performance of the index; according to the analysis done by Pagel et al. [3], the expected number of disk accesses for a spatial query is a function of the number of MBRs, the total volume of MBRs, the total surface of MBRs, and the spatial extent of the query. To mitigate the drawback of the number of MBRs on the performance, a multi-version structure which clusters MBRs with close temporal extent can be used to attenuate the effect of the number of MBRs. Given a fixed number of MBRs to approximate a history (or a database of histories), Hadjieleftheriou et al. [1] claim that an approximation which reduces total volume is expected to improve query performance. Their analysis is based on using a multi-version spatial index structure.

The general approach for processing history based queries is to translate each query into one or more coordinate queries; the smaller queries are used to retrieve relevant histories that intersect with the smaller queries. Often a similarity (or distance) measure such as $\mathcal{L}_p$-norm, DTW, and longest common subsequence (LCSS) is used to rank histories based on their closeness to the query history. Similar to coordinate based queries, a history is approximated as a set of MBRs organized using a spatial index structure. The index is probed to find histories with MBRs that overlap with the MBRs of the query. To reduce the number of times a history is compared against the query, an under(over) estimation of the distance(similarity) function is evaluated on the MBR approximation of the history and the query [2, 8]. The performance of history based queries depends on two factors; the expected number of disk accesses and the number of times the distance function is evaluated. The first factor depends on the expected number of disk accesses required for each coordinate based query, which is a function of the query and the criteria used to derive at optimal approximations of a history. The second factor, however, depends on the pruning power of the lower-(upper-) bound, which in turn is affected by how close a set of MBRs approximate the corresponding history. Most proposed solutions for history based queries only take the first factor into consideration and ignore the second factor. For instance, Lee et al. [2] use an estimation of the expected number of disk accesses as a function of the query, the history, and an experimentally estimated parameter. Vlachos et al. [8], consider minimizing the total volume to reduce the number of intersecting MBRs and claim that this yields a better pruning power.

Most experiments on indexing histories have been conducted on either time-series or trajectories of mobile objects, where $d \leq 3$. In this paper, we study the limitation of minimizing total volume on indexing histories with higher dimensionality when histories are approximated using a set of MBRs. Since each MBR contains both temporal and spatial extents of a history segment that it represents, each segment of a d-dimensional history becomes a $2(d+1)$ vector. The performance of traditional spatial access methods start to degrade rapidly somewhere above 16 dimensions [6], i.e. $d > 7$ in our case. In our study we consider two cases; when $d \leq 7$, for which it is feasible to index MBRs, and when $d > 7$. Although an index on MBRs is infeasible for the latter case, still MBR approximation can be used to derive an efficient to compute estimate for costly distance functions used to compare histories (see for instance [8]).

2 Limitation of Minimizing Total Volume

An approximation of a history using k MBRs is represented by a set $B^k = \{B_1, \ldots, B_k\}$, such that each $B_i \in B^k$ is a multidimensional interval represented as

$$[s_i, e_i] \times [l_i^1, h_i^1] \times \ldots \times [l_i^d, h_i^d]$$

where $[s_i, e_i]$ is the temporal extent of B_i and $[l_i^j, h_i^j]$ is the spatial extent for dimension j. By construction, $s_1 = 1$, $e_k = n$, and $s_{i+1} = 1 + e_i$ since MBRs are assumed to be consecutive and non-overlapping. The total volume of B^k is evaluated as

$$\sum_{i=1}^{k} (e_i - s_i) \prod_{j=1}^{d} (h_i^j - l_i^j). \tag{1}$$

2.1 Minimizing Total Volume vs. Expected Performance

Pagel et al. [3] give an analysis for the performance of spatial queries; this analysis would also apply to historical queries on an R-tree that organizes history MBRs. For brevity, here we assume that points of histories are distributed in a unit hyper-cube and that a query is within a unit hyper-cube with temporal range equal to t_q and spatial range equal to $q_j \leq 1$ for dimension j of the query. The performance measure is proportional to the probability that a query MBR intersects the MBRs of a history. Given that the MBRs of a history are mutually exclusive, and that points are distributed in unit hyper-cube, the performance measure for a data history with k MBRs is stated as the probability that the MBRs of the history overlap with an MBR of the query. This probability is equal to

$$\sum_{i=1}^{k} (e_i - s_i + t_q) \prod_{j=1}^{d} (h_i^j - l_i^j + q_j). \tag{2}$$

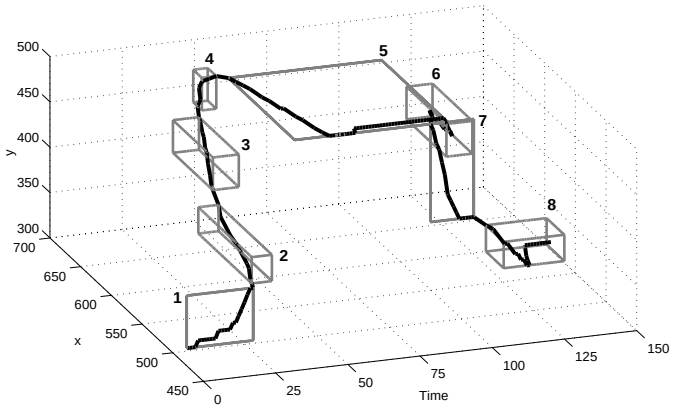

Figure 1. Approximating a 2-dimensional history using a set of MBRs which minimizes total volume.

The difference between Eq. 1 and Eq. 2 is negligible, only if the temporal and spatial range of the query is zero; otherwise, a decomposition that minimizes total volume is not guaranteed to be the one that optimizes the performance measure, depending on the spatial and temporal range of the query. On the other hand, optimizing Eq. 2 requires information about the query at indexing time. Rasetic et al. e.g. [5] consider replacing t_q and q_j, respectively, with their expected value and optimize query performance given the expected temporal and spatial range of queries.

Another problem with minimizing total volume arises when a projection of a history on a subset of its dimensions remains constant within a time interval. For any possible splitting of the history in that interval, the total volume is the same (i.e. zero). This is a serious problem as d increases; because each history can become more sparse and contains several time intervals in which there is at least one dimension that does not change. The real dimensionality of the history, for those intervals, becomes less than d. Figure 1 depicts an example of a two dimensional history which is approximated by eight MBRs with minimum total volume. Note that the history does not change along dimension x in two time intervals which coincides with the first and seventh MBR. Although the history changes along dimension y, minimizing total volume ignores the change in this interval. Similarly, no change is observed for dimension y, and a relatively large segment is approximated by fifth MBR. In real settings, such cases are not rare and it is expected that the number of such intervals increases with d. For such intervals, the trivial solution that optimizes Eq. 1 does not yield an optimal approximation for the subspaces that are active and change during these intervals.

2.2 Minimizing Total Volume and the Tightness of MBR Approximation

In order to efficiently prune costly comparison of histories, several lower-(upper-) bounds have been proposed for distance(similarity) measures between histories (e.g. [2, 8]). The tightness of such bounds depends on how closely the MBRs approximate histories. If the expected error of approximation is more accurately formalized, we are seeking an approximation that minimizes this error. This is the subject we are currently investigating [7]. Our preliminary experiments on various datasets including some of those used before(e.g. [8]) confirm that the tightness of the lower-bounds can be improved, resulting in a better pruning power.

3 Summary and Future Work

A few issues associated with minimizing total volume for indexing MBR approximation of histories have been studied. We are currently investigating other approximation alternatives and filtering techniques that can scale up to high dimensional histories, possibly avoiding the issues associated with total volume.

Acknowledgments

This work is supported by Natural Sciences and Engineering Research Council of Canada.

References

[1] M. Hadjieleftheriou, G. Kollios, V. J. Tsotras, and D. Gunopulos. Efficient indexing of spatiotemporal objects. In *Proc. of the EDBT*, pages 251–268, 2002.

[2] S.-L. Lee, S.-J. Chun, D.-H. Kim, J.-H. Lee, and C.-W. Chung. Similarity search for multidimensional data sequences. In *Proc. of the ICDE*, pages 599–608, 2000.

[3] B. U. Pagel, H. W. Six, H. Toben, and P. Widmayer. Towards an analysis of range query performance in spatial data structures. In *Proc of the PODS*, pages 214–221, 1993.

[4] D. Pfoser, C. S. Jensen, and Y. Theodoridis. Novel approaches in query processing for moving object trajectories. In *Proc of the VLDB*, pages 395–406, 2000.

[5] S. Rasetic, J. Sander, J. Elding, and M. A. Nascimento. A trajectory splitting model for efficient spatio-temporal indexing. In *Proc. of the VLDB*, pages 934–945, 2005.

[6] T. Seidl and H.-P. Kriegel. Optimal multi-step k-nearest neighbor search. In *Proc. of the SIGMOD*, pages 154–165, 1998.

[7] R. Sherkat and D. Rafiei. On efficiently searching historical archives. In *Submited for Publications*, 2006.

[8] M. Vlachos, M. Hadjieleftheriou, D. Gunopulos, and E. J. Keogh. Indexing multidimensional time-series. *VLDB Journal*, 15(1):1–20, 2006.