6-2018

Applying spatial database techniques to other domains: A case study on top-k and computational geometric operators

Kyriakos MOURATIDIS
Singapore Management University, kyriakos@smu.edu.sg

DOI: https://doi.org/10.1145/3210272.3226094

Applying Spatial Database Techniques to Other Domains: a Case Study on Top-k and Computational Geometric Operators

Kyriakos Mouratidis
School of Information Systems
Singapore Management University
kyriakos@smu.edu.sg

ABSTRACT

In this seminar, we will explore how processing rich spatial data is not the only practical (and research-wise promising) application domain for traditional spatial database techniques. An equally promising direction, possibly with low-hanging fruits for research innovation, may be to apply the spatial data management expertise of our community to non-spatial types of queries, and to extend standard, more theoretical operators to large scale datasets with the objective of practical solutions (as opposed to favorable asymptotic complexity alone). As a case study, we will review spatial database work on top-k-related operators (i.e., non-spatial problems) and how it integrates fundamental computational geometric operators with spatial indexing/pruning to produce efficient solutions to practical problems.

1 INTRODUCTION

The objective of this seminar is to showcase that processing rich data is not the only promising direction for modern spatial database research, and that looking into traditional, theoretical problems from a spatial database point of view, may reveal new challenges and practical application domains for the expertise of our community. Such an example is computational geometric problems. Computational Geometry [2, 9] is a discipline under theoretical computer science that studies algorithms for (queries that can be framed as) geometric problems. By this definition, its affinity to spatial databases becomes obvious. However, having emerged from theoretical algorithm design and analysis, and although scalability is a key objective in that area, the latter is usually approached in terms of theoretical analysis and asymptotic complexity. We argue that there is space for research that will approach traditional computational geometric problems from a more applied point of view. In this seminar, we will demonstrate how traditional computational geometric operators can be incorporated/enhanced/applied in tandem with standard notions of spatial indexing and pruning to process large amounts of data in reasonable time for practical, not necessarily spatial problems. Specifically, we will focus on queries related to the top-k operator in multi-criteria settings, and show that their inherent geometric nature allows for fast processing.

GeoRich'18, June 10, 2018, Houston, TX, USA

ACM ISBN 978-1-4503-5832-3/18/06.
https://doi.org/10.1145/3210272.3226094

Consider a dataset that contains a large number of *options* (e.g., restaurants, hotels, etc). Each option $\mathbf{r}$ has d *attributes*. In an example where the dataset contains hotels, the attributes could correspond to the ratings of the hotels on d aspects, such as service, sleep quality, convenience of location, etc. The top-k query is a common means to shortlist the k best options according to the user's preferences on the d data attributes. Specifically, in the most prevalent top-k model, the user specifies a *query vector* $\mathbf{q}$ which comprises a numeric weight for each attribute [5]. The score of an option is defined as the weighted sum of its attributes (equivalently, the dot product $\mathbf{r} \cdot \mathbf{q}$), which in turn imposes a ranking among the available options. The k highest ranking options form the top-k result and are reported to the user.

Despite its algebraic definition, top-k processing has a geometric nature and a connection to fundamental computational geometry problems. For example, if the options are treated as points in a d-dimensional space, top-k computation can be seen as a sweeping of the data space from its top corner to the origin with a hyper-plane (normal to the query vector q) until k options are swept [12]. In addition to ideas for query processing, this parallel reveals important properties of the problem, such as the fact that the top option for any query vector lies on the *convex hull* of the dataset [3].

Things become more interesting when variants or auxiliary features to top-k processing are considered in the *preference space*, i.e., the space where the query vector may lie. Geometric properties in that space, and particularly the concept of *k-levels* from computational geometry, can be used for the efficient processing of ad-hoc top-k queries over data streams [4], the processing of *continuous* top-k queries [13], and the identification of all possible top-k results when the query vector may lie anywhere in a region of the preference space [7].

Furthermore, insights in the properties of the preference space have given rise to very useful, complementary features (and measures) relevant to top-k processing. An example is the association of the top-k result with a region around the query vector $\mathbf{q}$ (in preference space) where the result remains the same [10, 14]. The volume of that region can be used as a measure for result sensitivity, while the region itself as a means for computation sharing among different top-k queries (result caching), for exploratory analysis, etc. Another example is the computation of the maximum possible rank that an option could achieve, given the competition (i.e., the alternative options in the data set) [8]. This also entails the calculation of the exact regions of the preference space where the maximum rank is achieved, which can be used for market impact analysis and customer profiling. The problem is related to *hyper-plane arrangements*, a very powerful concept in computational geometry [1, 2].

In this seminar, we will survey some key geometric concepts that underlie the aforementioned top-k-related problems (such as convex hull, half-space range reporting, hyperplane arrangement, and k-level), and we will review the specific approaches taken in the respective database papers. Along the way, we will point our some computational geometric operators, whose adaptation to large scale spatial datasets has many potential applications, but has not been resolved completely or for which the only known results are purely theoretical. Computational geometry aside, we will also list other domains that could serve as inspiration for spatial database research.

2 BIOGRAPHY OF THE PRESENTER

Kyriakos Mouratidis holds a B.Sc. in Computer Science from Aristotle University of Thessaloniki (AUTH), and a Ph.D. in Computer Science and Engineering from Hong Kong University of Science and Technology (HKUST). He is an Associate Professor at the School of Information Systems of Singapore Management University (SMU). His main research area is spatial databases, with a focus on continuous query processing, road network databases, and spatial optimization problems. His work in the last 4 years has concentrated on complementary features to top-k queries, like for example [6–8, 11, 14]. A compete CV and publication list can be found at: http://www.mysmu.edu/faculty/kyriakos/

REFERENCES

[1] P. K. Agarwal and M. Sharir. Arrangements and their applications. In *Handbook of Computational Geometry*, pages 49–119. Elsevier, 1998.

[2] M. D. Berg, O. Cheong, M. V. Kreveld, and M. Overmars. *Computational geometry: algorithms and applications*. Springer, 2008.

[3] Y.-C. Chang, L. Bergman, V. Castelli, C.-S. Li, M.-L. Lo, and J. R. Smith. The onion technique: Indexing for linear optimization queries. In *SIGMOD*, pages 391–402, 2000.

[4] G. Das, D. Gunopulos, N. Koudas, and N. Sarkas. Ad-hoc top-k query answering for data streams. In *VLDB*, pages 183–194, 2007.

[5] I. F. Ilyas, G. Beskales, and M. A. Soliman. A survey of top-k query processing techniques in relational database systems. *ACM Comput. Surv.*, 40(4):11:1–11:58, 2008.

[6] K. Mouratidis and H. Pang. Computing immutable regions for subspace top-k queries. In *PVLDB*, pages 73–84, 2013.

[7] K. Mouratidis and B. Tang. Exact processing of uncertain top-k queries in multi-criteria settings. *PVLDB*, 11(8):866–879, 2018.

[8] K. Mouratidis, J. Zhang, and H. Pang. Maximum rank query. *PVLDB*, 8(12):1554–1565, 2015.

[9] F. P. Preparata and M. I. Shamos. *Computational Geometry - An Introduction*. Texts and Monographs in Computer Science. Springer, 1985.

[10] M. A. Soliman, I. F. Ilyas, D. Martinenghi, and M. Tagliasacchi. Ranking with uncertain scoring functions: semantics and sensitivity measures. In *SIGMOD*, pages 805–816, 2011.

[11] B. Tang, K. Mouratidis, and M. L. Yiu. Determining the impact regions of competing options in preference space. In *SIGMOD*, pages 805–820, 2017.

[12] P. Tsaparas, T. Palpanas, Y. Kotidis, N. Koudas, and D. Srivastava. Ranked join indices. In *ICDE*, pages 277–288, 2003.

[13] A. Yu, P. K. Agarwal, and J. Yang. Processing a large number of continuous preference top-k queries. In *SIGMOD*, pages 397–408, 2012.

[14] J. Zhang, K. Mouratidis, and H. Pang. Global immutable region computation. In *SIGMOD*, pages 1151–1162, 2014.