Geometric Approaches for Top-k Queries

Kyriakos Mouratidis
School of Information Systems
Singapore Management University
80 Stamford Road, Singapore 178902
kyriakos@smu.edu.sg

ABSTRACT

Top-k processing is a well-studied problem with numerous applications that is becoming increasingly relevant with the growing availability of recommendation systems and decision making software. The objective of this tutorial is twofold. First, we will delve into the geometric aspects of top-k processing. Second, we will cover complementary features to top-k queries, with strong practical relevance and important applications, that have a computational geometric nature. The tutorial will close with insights in the effect of dimensionality on the meaningfulness of top-k queries, and interesting similarities to nearest neighbor search.

1. INTRODUCTION

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 [10]. 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 [16]. 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 query 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 [6] or the processing of *continuous* top-k queries [19].

Furthermore, insights in the properties of the query 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 query space) where the result remains the same [14, 20]. 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) [12]. This also entails the calculation of the exact regions of the query 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].

The first objective of this tutorial is to shed light to the connection between top-k processing and fundamental computational geometry problems. In particular, we will review the key geometric concepts of (i) convex hull, (ii) half-space range reporting, (iii) hyperplane arrangement, and (iv) k-level, and we will explain how they can help us support top-k queries, variants and auxiliary features.

The tutorial will continue with systems that exploit the aforementioned geometric fundamentals to efficiently process top-k and related queries. More than 10 papers will be discussed, with a focus on methods that (i) are not limited to 2 dimensions only, and (ii) produce exact solutions. Regarding item (i), as we will explain, 2-dimensional solutions address degenerate versions of preference-based ranking, which may simplify processing, but also sacrifice the generality of the key ideas. Regarding item (ii), approximate methods may be plausible for some problems, but they largely dismiss the geometric aspects of the problems in order to simplify them.

We will conclude with insights into the effect of dimensionality on the behaviour and the usefulness of top-k ranking/processing. Connections to the standard *nearest neighbor query* (NN) will be drawn, and surprising similarities to its behaviour with dimensionality will be demonstrated.

Proceedings of the VLDB Endowment, Vol. 10, No. 8
Copyright 2017 VLDB Endowment 2150-8097/17/04.

Due to its close connection to computational geometry, the tutorial will be very "visual", with a multitude of drawings used for illustration; this will hopefully make it eye-catching and fun (in addition to useful).

2. RELEVANCE AND TARGET AUDIENCE

Due to the great and world-wide diffusion of the internet, the number of options available to cover a user's needs far exceeds her capacity to exhaustively browse through all of them. For that reason, preference-based querying and recommendations systems have become ubiquitous in the software and mobile application industry. Enhancing these systems with functionality that extends further than basic top-k reporting is underway, since features auxiliary to it offer stronger decision support and deeper decision analytics, as we will demonstrate with pragmatic application scenarios in the tutorial.

On the research front, the majority of the papers covered are very recent, which attests to the timeliness of the topic. Computational geometry is a branch of theoretical computer science. However, its application to large scale datasets in the context of practical, multi-criteria decision making is a challenging and intriguing topic for database researchers, especially in the sub-communities of spatial databases, recommendation systems, and data analytics.

Specific sub-communities aside, the tutorial is meant for the broader VLDB audience. In terms of geometry, only basic knowledge will be required, since the essential computational geometric concepts will be abstracted and presented from scratch, chiefly with the use of visual examples. Basic algorithmic and indexing background will be necessary, e.g., branch-and-bound search, Quad-tree and R-tree indices, etc.

3. LENGTH, SCOPE AND STRUCTURE

The tutorial is designed for 90 minutes, and comprises 3 parts. The first covers standard computational geometry concepts and their relevance to top-k processing, namely, convex hull, half-space range reporting, hyperplane arrangement, and k-level. The second (which is the main part) covers existing database work that exploits the connection of these problems to top-k processing. The third part concludes with insights about the effect of dimensionality on the usefulness of top-k and related queries, and draws a parallel to the traditional NN query. Below we summarize the systems to be reviewed in the main part of the tutorial.

The first work covered is [6], where Das et al. consider the evaluation of ad-hoc top-k queries over a data stream in the sliding window model. The main idea is that only a small subset of the options in the sliding window could appear in the top-k result w.r.t. any query vector. To identify (and maintain) this small subset they rely on a geometric representation of the top-k query and a notion of duality, where options and queries are mapped into lines and rays, respectively. Although tailored to 2 dimensions explicitly, this work involves fundamental principles that are key to the tutorial, especially those regarding geometric arrangements.

Yu et al. [19] extend the principles of [6] to higher dimensions, targeting this time *continuous* top-k queries. At the core of their approach lies the effective maintenance of the *query response surface*, which encodes the score and identity of the k-th result option for any query vector, and is therefore very relevant to k-levels.

Vlachou et al. [17, 18] study *reverse top-k queries*. Even though their main focus is on the *bichromatic* version of the query (whose definition and treatment deviates from the geometric focus of this tutorial), they introduce the *monochromatic* version too, and propose a geometric solution for 2 dimensions. Specified a focal option $\mathbf{p}$ in a set of alternatives, they compute the parts of the query space where the query vector should lie so that $\mathbf{p}$ belongs to the top-k result. While applicable to 2 dimensions only, the proposed solution is an interesting use of k-levels for a top-k-related problem. Very recently, Tang et al. [15] solved the problem (i.e., monochromatic reverse top-k processing) for higher dimensions too. They exploit a mapping of competing options into hyperplanes in the (transformed) query space, and work on the produced arrangement using a blend of computational geometric operations and linear programming.

Soliman et al. [14] consider uncertain scoring functions and identify the most "representative" top-k result, under different definitions. First, they compute the most likely top-k result if the query vector is randomly chosen. Next, they compute the top-k result that is least dissimilar to all possible alternative results. Finally, they introduce sensitivity measures for a given top-k result. Their approach relies on geometric insights and on operations involving hyperplane representations in query space.

Zhang et al. [20] introduce the concept of the *global immutable region* (GIR). The GIR is the maximal locus around a query vector $\mathbf{q}$ where the top-k result remains the same. It is shown to be a convex polytope (in query space), produced by half-space intersection. To offer scalability, the authors rely on properties of the convex hull, while their most efficient algorithm borrows ideas from Clarkson's algorithm, one of the most common methods for convex hull computation. The related *view cover* problem in [8, 9] will also be reviewed, where given a query vector $\mathbf{q}$, the region in query space is computed, wherein any query vector is guaranteed to have its top-k options among the top-m of $\mathbf{q}$ (where k and m are input parameters, and $k < m$).

Mouratidis et al. [12] propose the *maximum rank query* (*MaxRank*). Given a focal option in a set of alternatives, *MaxRank* computes the highest rank this option may achieve under any possible user preference. It additionally reports all the regions in the query space where that rank is achieved. Its applications include market impact analysis, customer profiling, targeted advertising, etc. The proposed solution relies on hyperplane arrangements and a coverage counting problem therein.

He and Lo [7] consider *why-not* top-k queries. Given a query vector $\mathbf{q}$ and an option $\mathbf{p}$ that does not belong to the top-k result, the why-not query computes the smallest perturbation required in the query vector and/or in value k so that $\mathbf{p}$ is included in the result. Event though the proposed solution relies on sampling (and is therefore approximate), it involves interesting geometric insights and relies on hyperplane arrangements in query space.

In addition to the above papers, we will make mention to studies of practical relevance, which however deviate from our focus, i.e., offer approximate answers, are bound to 2 dimensions, or have no geometric solutions. Examples include bichromatic reverse top-k [18], reverse k-ranks [21] (which can be seen as the bichromatic version of *MaxRank*), external k-level computation in 2 dimensions [4], k-regret minimizing sets [5], k-hit [13], etc.

4. 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 [11, 20, 12, 15]. A compete CV and publication list can be found at: http://www.mysmu.edu/faculty/kyriakos/

5. 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] M. A. Cheema, Z. Shen, X. Lin, and W. Zhang. A unified framework for efficiently processing ranking related queries. In *EDBT*, pages 427–438, 2014.

[5] S. Chester, A. Thomo, S. Venkatesh, and S. Whitesides. Computing k-regret minimizing sets. *PVLDB*, 7(5):389–400, 2014.

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

[7] Z. He and E. Lo. Answering why-not questions on top-k queries. *IEEE Trans. Knowl. Data Eng.*, 26(6):1300–1315, 2014.

[8] V. Hristidis, N. Koudas, and Y. Papakonstantinou. PREFER: A system for the efficient execution of multi-parametric ranked queries. In *SIGMOD*, pages 259–270, 2001.

[9] V. Hristidis and Y. Papakonstantinou. Algorithms and applications for answering ranked queries using ranked views. *VLDB J.*, 13(1):49–70, 2004.

[10] 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.

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

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

[13] P. Peng and R. C. Wong. k-hit query: Top-k query with probabilistic utility function. In *SIGMOD*, pages 577–592, 2015.

[14] 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.

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

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

[17] A. Vlachou, C. Doulkeridis, Y. Kotidis, and K. Nørvåg. Reverse top-k queries. In *ICDE*, pages 365–376, 2010.

[18] A. Vlachou, C. Doulkeridis, Y. Kotidis, and K. Nørvåg. Monochromatic and bichromatic reverse top-k queries. *IEEE Trans. Knowl. Data Eng.*, 23(8):1215–1229, 2011.

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

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

[21] Z. Zhang, C. Jin, and Q. Kang. Reverse k-ranks query. *PVLDB*, 7(10):785–796, 2014.