Geometric Top-k Processing: Updates since MDM'16 [Advanced Seminar]

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

***Abstract*—The top-k query has been studied extensively, and is considered the norm for multi-criteria decision making in large databases. In recent years, research has considered several complementary operators to the traditional top-k query, drawing inspiration (both in terms of problem formulation and solution design) from the geometric nature of the top-k processing model. In this seminar, we will present advances in that stream of work, focusing on updates since the preliminary seminar on the same topic in MDM'16.**

I. Introduction

In today's digital world, users increasingly make daily decisions via mobile apps and online portals, like Yelp or HungryGoWhere (for restaurants), Booking.com (for hotels), etc. Choosing among available options (e.g., restaurants) involves multiple, possibly conflicting criteria, such as service and value for money. In database research, the top-k query has emerged as the standard tool to support multi-criteria decisions, i.e., to shortlist the most suitable options for each user [7].

Traditional top-k processing assumes a dataset of *options* (e.g., restaurants) with d *attributes* (e.g., service, value for money, etc). User preferences are represented by a numeric weight per data attribute (i.e., per decision criterion), collectively forming a *weight vector* $\mathbf{w}$. The score of an option is the weighted sum of its attributes. The k (say, the 10) highest-scoring options are presented as a personalized recommendation to the user. The described linear scoring mechanism is the most commonly employed in the literature [2], [6], and shown by user studies to capture closely the way humans make multi-criteria decisions [14].

Interestingly, processing such top-k queries corresponds to the sweeping of the data space with a hyperplane that is normal to $\mathbf{w}$. The sweeping direction is from the top corner of the data space (i.e., point (1, 1, ..., 1)) towards the origin. The order in which the options are encountered indicates their rank with respect to $\mathbf{w}$. That is, the top-k result comprises the k options encountered first [18]. This observation has led to computational geometric approaches for the top-k query and its variants, often exploiting a dual version of the problem in the *preference space*, i.e., the domain of $\mathbf{w}$, like for example [5], [15], [24], [25]. Such methods were reviewed in a preliminary seminar at MDM'16 [10]. Our objective in MDM'19 is to present advances and updates in that ongoing stream of work.

We will start with necessary background on the problem, and especially on its representation in the preference space. We will draw important links to traditional concepts/problems in computational geometry, such as the convex hull, hyperplane arrangements, and the k-level. We will then present recent systems that exploit these geometric fundamentals, focusing mostly on methods that (i) produce exact solutions, and (ii) are not limited to 2 dimensions only[1]. We will conclude with a discussion on dimensionality, backed up with empirical results, demonstrating that (unless the data are seriously skewed) the traditional top-k query, and thus its derivatives too, lose their meaning for more than a handful of dimensions.

II. Target Audience and Relevance to MDM'19

Due to the great 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 has become ubiquitous in the software and mobile application industry. Enhancing these systems with functionality that extends further than basic top-k reporting is underway due to the need for stronger decision support and deeper decision analytics, as we will demonstrate with pragmatic application scenarios in the seminar.

The topic is closely related to *spatial databases*, *multi-criteria decision making*, and *decision analytics*. The spatial database community, which comprises a large fraction of the MDM audience, has a special interest in applied geometry and geometric reasoning at scale, which lies at the core of this seminar. Given that the application domain of the seminar centers on preference-based querying and top-k processing, the topic is directly related to audiences interested in multi-criteria decisions and decision analytics too.

Specific areas from the MDM'19 call for papers that match the seminar's topic are:

- *Indexing, Optimisation and Query Processing for Moving Objects/Users*
- *Mobile Recommendation Systems*
- *Data Stream Processing in Mobile/Sensor Networks*

[1] As we will explain, 2-dimensional solutions address degenerate versions of preference-based ranking, which may be simpler to process, but lack generality.

III. Seminar Outline

The seminar's main content is divided into 4 themes: monochromatic reverse top-k querying; processing when user preferences are described by a region R in preference space (instead of a specific weight vector **w**); creation/improvement of options so that they belong to the top-k result; and result stability considerations in rank-aware processing.

A. *Monochromatic Reverse Top-k Processing*

All the problems surveyed in this seminar (and in the preliminary version in MDM'16) are very closely related to the *monochromaric reverse top-k query*. Specified a focal option **p** in a set of alternatives, this query computes the parts of the preference space where the weight vector **w** should lie so that **p** belongs to the top-k result. In effect, this query identifies all possible user profiles (preferences) for which **p** ranks among the top-k options. Thus, it finds direct application in market impact analysis, potential customer identification, profile-based marketing, targeted advertising, etc. The query was first introduced by Vlachou et al. in [19], [20], however, the solution developed only applies to the degenerate case of 2 dimensions[2]. Recently, Tang et al. [16] solved the problem in higher dimensions too. They exploit a mapping of competing options into hyperplanes in the preference space, and work on the produced arrangement using a blend of computational geometric operations and linear programming.

Yang and Cai [22] treat each data option as a function and, given a common input to these functions, they compute the options (i.e., functions) that evaluate to the top-k highest values, to values in a desired range, or to the k values that are closest to a specific target value. Assuming that the option attributes represent coefficients in a weighted sum function, and the common input is a weight vector **w**, the problem translates to a reverse top-k variant. The authors pre-process the dataset such that a complete ordering of all options can be derived with little computational effort at query time for any input **w**. Essentially, they index the hyperplanes induced by each pair of options (excluding pairs where one dominates the other).

B. *Top-k for Approximate Preferences and Preference Regions*

The next theme draws from the observation that the top-k literature makes the long-standing assumption that an exact weight vector **w** is given as input. Although convenient in terms of query processing, the assumption/requirement that the exact **w** is known may not be realistic. Be it specified directly by the user or by a preference learning algorithm, it is inherently inaccurate. This has motivated a new stream of work, where the weight vector **w** could lie anywhere in a region R in preference space [3], [12], [4].

On that assumption, Ciaccia and Martinenghi [3] identify all possible top-1 options. In a similar setting, Mouratidis and Tang [12] compute all possible top-k sets for any **w** in R, together with the specific partition in R that produces each of these top-k sets. Setting an option **p** as the initial pivot, they derive the rank of **p** (and compute all options that score higher than it) in different parts of R. In those parts where the pivot's rank is smaller than k (i.e., only a prefix of the top-k result is known), a different pivot is chosen and the process is repeated recursively, until the entire top-k set is known at any position in region R.

In [4], Ciaccia and Martinenghi do not consider top-k per se, but compute a restricted k-skyband when the user's weight vector is bounded to lie in R. Under that constraint, dominance becomes more selective than the traditional sense, leading to fewer options in the restricted k-skyband compared to its traditional version.

In the conceptually inverse scenario, Qian et al. [14] aim to extract an approximation of the user's (unknown) weight vector **w**, following the iterative pairwise comparisons model [8]. The idea is to derive increasingly tighter approximations of the latent **w** (i.e., regions in preference space) via iterative polling rounds. In each of these rounds, the user is requested to make a choice between two alternative options, effectively eliminating a halfspace (in preference space) defined by this pair of alternatives. With intelligently selected pairs, a tight enough approximation (i.e., a small enough region R) can be derived in few rounds.

C. *Creating Top-ranking Options*

Assuming a setting where users browse options via top-k queries, studies under this theme take the standpoint of a business owner, and suggest optimal placement strategies to maximize the competitiveness of her product or service, be it an existing option (to be improved) or a new option (to be introduced into the market).

Specified a set of weight vectors Q and an existing option **p** that does not belong to any of their top-k results, Liu et al. [9] compute the minimum perturbation required in **p**, so that it is included in all their top-k results. Again for a known set of weight vectors Q, and a cost function, Yang and Cai [21] compute the minimum-cost improvement vector (that specifies an increment value for each dimension) for an existing option **p**, so that its enhanced version is in the top-k result of at least m vectors in Q. Given a set of weight vectors Q, Yang et al. [23] compute the attribute values a new option **p** should have, so that (i) it is the top-1 option in the dataset for at least m weight vectors, and (ii) the cost to create **p** is minimized. The cost function is monotonic to the attribute values, implying that the more competitive **p** is made, the more expensive it is to create.

The 3 aforementioned studies consider a finite number of specific weight vectors. In the seminar, we will discuss possible problem formulations, with practical applications, where no exact query vector is known in advance, but only general descriptors of the intended customers are specified for the option to be improved or created [17].

[2] The main focus in that work is a different (i.e., bichromatic) query version, where a finite set of specific weight vectors are input, leading to a very different treatment compared to the dealing with infinite possible placements of **w** in the preference space.

D. Result Stability

The final theme of the seminar regards the stability of rankings derived by linear score ordering, i.e., rankings of options by the weighted sum of their values. An intuitive stability measure for a ranking is the probability that a random weight vector **w** would produce this ranking [15]. Given a top-k result, Zhang et al. [25] compute the maximal region in preference space where any weight vector produces the specific top-k result. That region is shown to be a convex polytope, called global immutable region (GIR). The sought probability is the ratio of the GIR volume to the volume of the entire preference space. Recently, Asudeh et al. [1] applied the GIR concept to define the stability of complete dataset rankings. Given a region R in preference space, they propose randomized sampling methods to compute the most stable complete rankings, i.e., the rankings with the largest GIR volumes.

The seminar will conclude with a discussion on dimensionality, demonstrating that as it grows, the scores of all options in the dataset tend to converge to the same value, rendering score-based ranking by linear functions (as in top-k and related problems) meaningless for more than a few dimensions.

IV. 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 and Lee Kong Chian Fellow 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 7 years has concentrated on complementary features to top-k queries (e.g., [11], [25], [13], [16], [12], [17]), for which he has received support by the Singapore Management University Lee Kong Chian Fellowship. A complete CV and publication list can be found at: http://www.mysmu.edu/faculty/kyriakos/

References

[1] A. Asudeh, H. V. Jagadish, G. Miklau, and J. Stoyanovich. On obtaining stable rankings. *PVLDB*, 12(3):237–250, 2018.

[2] 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.

[3] P. Ciaccia and D. Martinenghi. Reconciling skyline and ranking queries. *PVLDB*, 10(11):1454–1465, 2017.

[4] P. Ciaccia and D. Martinenghi. FA + TA < FSA: Flexible score aggregation. In *CIKM*, pages 57–66, 2018.

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

[6] 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.

[7] 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.

[8] K. G. Jamieson and R. D. Nowak. Active ranking using pairwise comparisons. In *NIPS*, pages 2240–2248, 2011.

[9] Q. Liu, Y. Gao, G. Chen, B. Zheng, and L. Zhou. Answering why-not and why questions on reverse top-k queries. *VLDB J.*, 25(6):867–892, 2016.

[10] K. Mouratidis. Geometric aspects and auxiliary features to top-k processing. In *MDM*, pages 1–3, 2016.

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

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

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

[14] L. Qian, J. Gao, and H. V. Jagadish. Learning user preferences by adaptive pairwise comparison. *PVLDB*, 8(11):1322–1333, 2015.

[15] 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.

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

[17] B. Tang, K. Mouratidis, M. L. Yiu, and Z. Chen. Creating top ranking options in the continuous option and preference space. *PVLDB*, 12: to appear, 2019.

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

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

[20] 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.

[21] G. Yang and Y. Cai. Querying improvement strategies. In *EDBT*, pages 294–305, 2017.

[22] G. Yang and Y. Cai. Querying a collection of continuous functions. *IEEE Trans. Knowl. Data Eng.*, 30(9):1783–1795, 2018.

[23] J. Yang, Y. Zhang, W. Zhang, and X. Lin. Influence based cost optimization on user preference. In *ICDE*, pages 709–720, 2016.

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

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