Opportunities for Spatial Database Research in the Context of Preference Queries [Keynote Speech]

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

ABSTRACT

This is the outline of the keynote speech at LocalRec@ACM SIGSPATIAL 2023. The main objective of the talk is to point out opportunities for spatial database researchers in the area of preference-based querying. We will commence with an overview of the standard queries for multi-objective decision making, and demonstrate their direct connection to recommendations and to market analysis. In this context, there is a number of specific decision criteria, and user preferences are represented as vectors with as many dimensions. We will demonstrate how and why this type of preferences are natural to actual applications and practical for the support of real users in their choices and decisions. Next, we will illustrate that the principles which underlie preference-based querying are actually computational geometric in nature and, for the goal of practicality, they enable the use of spatial data management techniques, such as multi-dimensional indices and geometric reasoning for search space reduction (akin to traditional pruning). To showcase the potential of approaching preference querying challenges via spatial database techniques, we will use three recent studies as examples. The talk will conclude with a recap of the potential to apply a skillset typical to SIGSPATIAL attendees to a new domain, that of preference querying.

CCS CONCEPTS

• **Information systems → Data access methods**; **Decision support systems**; **Top-k retrieval in databases**.

KEYWORDS

Top-k query; Skyline; Multi-dimensional datasets

ACM Reference Format:
Kyriakos Mouratidis. 2023. Opportunities for Spatial Database Research in the Context of Preference Queries [Keynote Speech]. In *7th ACM SIGSPATIAL Workshop on Location-based Recommendations, Geosocial Networks and Geoadvertising (LocalRec '23), November 13, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3615896.3628418

1 OUTLINE OF THE TALK

In the era of ubiquitous access to the Internet, and more so in the current post-pandemic situation, users are increasingly covering their everyday needs by online purchases. Along with convenience, e-shopping offers them access to a large number of alternatives, way greater than any physical go-to-shop experience. Choosing from the available options generally entails the consideration of multiple, often conflicting aspects. For example, consumer electronics website CNET.com has identified the main decision criteria in choosing a laptop (i.e., design, features, performance, and battery) and provides ratings on these criteria for models available in the market. Similarly, TripAdvisor.com maintains hotel ratings on 4 principal criteria (location, cleanliness, service, value), etc. Letting d be the number of aspects, an option is characterized by its d per-aspect ratings. Identifying and rating the decision aspects, however, is not enough. Due to the number of alternatives typically available, it is tedious for the user (if possible at all) to individually examine each of them. Hence, to render effective support to the user, it is essential to shortlist the most promising options.

The first matter we will elaborate is that preferences and recommendations do not only occur in recommender system research. Traditionally, the aim of the latter is to predict the score a user would give to an option [15]. The prediction could be based on the scores given to the option by similar users [5] or the user's own scores for similar options [7, 16]. Typically, the preferences dealt with in that area are represented as high-dimensional vectors where dimensions do not need to have a particular interpretation for the user. Instead, there is a multitude of applications (like the CNET and TripAdvisor examples) where (i) there are pre-defined decision aspects with concrete meaning for the user, (ii) there is only a handful of decisions aspects (dimensions) and thus the options and preference vectors are low-dimensional, and (iii) the options' attributes are given and no information for other users is required. In multi-objective querying, the options' scores and/or dominance relationships can be derived directly from their attributes. The challenge is to suitably filter a potentially large option-set in a computationally-optimized manner, to permit real-time display of the shortlisted options without requiring the explicit evaluation of all scores and/or dominance relationships across all options.

The two traditional paradigms to produce the shortlist in such multi-objective decisions are based on *dominance* and *ranking by utility*, respectively. The first paradigm considers that an option dominates another if all its attributes are more preferable (e.g., if its ratings are higher on all aspects), giving rise to the well-known *skyline* operator [2], and its generalization, the k-skyband [13]. The second paradigm, i.e., ranking by utility, associates each option with a utility score via a (user-specific) function over the option's attributes, and shortlists the top-k. Most commonly, the utility function is a weighted sum of the option's coordinates, where the d weights capture the user's preferences and together form the user's

LocalRec '23, November 13, 2023, Hamburg, Germany

ACM ISBN 979-8-4007-0358-4/23/11.
https://doi.org/10.1145/3615896.3628418

weight vector $\boldsymbol{w}$. This linear type of scoring has been shown by user studies to effectively model the way humans assess tradeoffs in real-life multi-objective decisions [14]. Furthermore, there are sophisticated preference learning approaches to extract (an estimate of) the user's preference vector, like those reviewed in [3].

We will demonstrate that both aforementioned paradigms have a strong geometric nature, and thus the most effective methods to process them rely on spatial database tools. Skyline/skyband computation and top-k processing, however, have been exhaustively studied. We will show that there are many problem definitions centered on these two paradigms which have only recently started being explored, and which find application in effective shortlisting of options, but also in effective market analysis and product design/improvement. Using three specific examples, we will show that the design principles for such problems are largely geometric, and thus their efficient processing is amenable to spatial database techniques.

The first example is the *m-impact region* (mIR) problem [17]. In a context where users look for available products with top-k queries, mIR identifies the part of the option space that attracts the most attention from a given set of users (weight vectors). Specifically, mIR determines the kind of attribute values that lead a (new or existing) option to the top-k result for at least a fraction of the user population. The mIR problem has several important applications. First, it helps determine the "hottest" part of the market, i.e., the part that attracts most of the users' attention. Furthermore, mIR may guide the improvement (or the design) of products. For instance, a hotel's management who plan to upgrade their premises/services, would look for a placement in the region reported by mIR, i.e., they can determine which aspects they need to boost more aggressively to meet their goal (e.g., focus more on improving value than service).

The second example is motivated by the shortcomings of the two paradigms for shortlisting options in multi-objective scenarios. The skyline/skyband is intuitive and requires no preference input but, on the downside, it is not personalizable and offers no control over its output size which is often overwhelming [1, 4]. On the other hand, ranking by utility is personalizable and, in the form of a top-k query, has a controllable output size too. Its Achilles' heel, however, lies in specifying the "correct" weights, since a miniscule change in $\boldsymbol{w}$ can drastically alter the top-k result [6, 20]. Whether input directly by the user or mined by a preference learning technique, $\boldsymbol{w}$ is a mere estimate of the user's actual, latent preferences. We will overview the approach in [8], which provides an alternative shortlisting methodology. Specifically, it does take a weight vector $\boldsymbol{w}$ as personalized preference input, but allows it to relax (expand) in the preference domain, thus treating it only as a starting point in the shortlist generation process. Moreover, it allows the user/application to control exactly the output size, stopping the expansion when the desired number of options have been shortlisted.

The third example [9] focuses on the dataset itself, and aims to assess its competitiveness with regard to different possible preferences, i.e., different types of users. Assuming that the set of available options is represented as a multi-attribute dataset D, it defines measures of competitiveness, and represents them in the form of a heat-map in the preference domain. In particular, it defines two categories of measures: (i) *utility-based*, which capture how satisfied the different user types are expected to be with the options available in D; and (ii) *competition-based*, which quantify how steep the competition among alternative products is with regard to different preferences. After choosing the appropriate measure, it partitions the preference domain into cells, and computes the competitiveness value for each cell. The competitiveness-marked cells form a heat-map that represents how competitive the dataset is at different parts of the user spectrum, i.e., for different types of preferences. Applications of this heat-map include market analysis and business development. On the former front, the highest competitiveness cells indicate where the market's strength lies and what types of customers have attracted most of its efforts. This may lend a better understanding of supply dynamics or spark further investigation of the reasons behind the high competitiveness (or lack thereof) for certain types of users. On the front of business development, high competitiveness cells may indicate a saturation of the market for the respective parts of the user spectrum. When the distribution (or a representative sample) of the user preferences is known, the heat-map enables even stronger support in both aforementioned applications. To illustrate, we will present a case study based on actual hotel data and user preferences mined from real TripAdvisor reviews, revealing actionable market insights.

The talk will conclude with a recap of the potential to apply a skillset typical to SIGSPATIAL attendees to a new domain, that of preference querying.

2 BIOGRAPHY OF THE SPEAKER

Kyriakos Mouratidis holds a B.Sc. from Aristotle University of Thessaloniki (AUTH), and a Ph.D. from Hong Kong University of Science and Technology (HKUST), both in Computer Science. He is a Professor of Computer Science at the School of Computing and Information Systems at 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 10 years has concentrated on complementary features to top-k queries (e.g., [8–12, 17–20]). A complete CV and publication list can be found at: www.mysmu.edu/faculty/kyriakos/

3 ACKNOWLEDGMENT

This work was supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 2 (Award No. MOE-T2EP20121-0002). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author and do not reflect the views of the Ministry of Education, Singapore.

REFERENCES

[1] Jon Louis Bentley, H. T. Kung, Mario Schkolnick, and Clark D. Thompson. 1978. On the Average Number of Maxima in a Set of Vectors and Applications. *J. ACM* 25, 4 (1978), 536–543.

[2] Stephan Börzsönyi, Donald Kossmann, and Konrad Stocker. 2001. The Skyline Operator. In *ICDE*. 421–430.
[3] Johannes Fürnkranz and Eyke Hüllermeier. 2010. *Preference Learning*. Springer US, Boston, MA, 789–795.
[4] Parke Godfrey. 2004. Skyline Cardinality for Relational Processing. In *FoIKS*. 78–97.
[5] David Goldberg, David A. Nichols, Brian M. Oki, and Douglas B. Terry. 1992. Using Collaborative Filtering to Weave an Information Tapestry. *Commun. ACM* 35, 12 (1992), 61–70.
[6] David Rios Insua and Simon French. 1991. A framework for sensitivity analysis in discrete multi-objective decision-making. *Eur. J. Oper. Res.* 54, 2 (1991), 176–190.
[7] Pasquale Lops, Marco de Gemmis, and Giovanni Semeraro. 2011. Content-based Recommender Systems: State of the Art and Trends. In *Recommender Systems Handbook*. Springer, 73–105.
[8] Kyriakos Mouratidis, Keming Li, and Bo Tang. 2021. Marrying Top-k with Skyline Queries: Relaxing the Preference Input while Producing Output of Controllable Size. In *SIGMOD Conference*. 1317–1330.
[9] Kyriakos Mouratidis, Keming Li, and Bo Tang. (to appear). Quantifying the competitiveness of a dataset in relation to general preferences. *VLDB J.* ((to appear)).
[10] Kyriakos Mouratidis and HweeHwa Pang. 2013. Computing Immutable Regions for Subspace top-k queries. In *PVLDB*. 73–84.
[11] Kyriakos Mouratidis and Bo Tang. 2018. Exact Processing of Uncertain Top-k Queries in Multi-criteria Settings. *PVLDB* 11, 8 (2018), 866–879.
[12] Kyriakos Mouratidis, Jilian Zhang, and HweeHwa Pang. 2015. Maximum Rank Query. *PVLDB* 8, 12 (2015), 1554–1565.
[13] Dimitris Papadias, Yufei Tao, Greg Fu, and Bernhard Seeger. 2005. Progressive skyline computation in database systems. *ACM Trans. Database Syst.* 30, 1 (2005), 41–82.
[14] Li Qian, Jinyang Gao, and H. V. Jagadish. 2015. Learning User Preferences By Adaptive Pairwise Comparison. *PVLDB* 8, 11 (2015), 1322–1333.
[15] Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor (Eds.). 2011. *Recommender Systems Handbook*. Springer.
[16] Badrul Munir Sarwar, George Karypis, Joseph A. Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *WWW*. ACM, 285–295.
[17] Bo Tang, Kyriakos Mouratidis, and Mingji Han. 2021. On m-Impact Regions and Standing Top-k Influence Problems. In *SIGMOD Conference*. 1784–1796.
[18] Bo Tang, Kyriakos Mouratidis, and Man Lung Yiu. 2017. Determining the Impact Regions of Competing Options in Preference Space. In *SIGMOD Conference*. 805–820.
[19] Bo Tang, Kyriakos Mouratidis, Man Lung Yiu, and Zhenyu Chen. 2019. Creating Top Ranking Options in the Continuous Option and Preference Space. *PVLDB* 12, 10 (2019), 1181–1194.
[20] Jilian Zhang, Kyriakos Mouratidis, and HweeHwa Pang. 2014. Global immutable region computation. In *SIGMOD Conference*. 1151–1162.