SkyRec: Finding Pareto Optimal Groups

Jinfei Liu
Emory University and Georgia Institute of Technology
jinfei.liu@emory.edu

Li Xiong
Emory University
lxiong@emory.edu

Jian Pei
Simon Fraser University
jpei@cs.sfu.ca

Jun Luo
Machine Intelligence Center Lenovo Group Limited
jluo1@lenovo.com

Haoyu Zhang
Indiana University
hz30@umail.iu.edu

Si Zhang
Jianghan University
sz04eu@jhun.edu.cn

ABSTRACT

We present SkyRec (Skyline Recommender), a recommendation toolkit for finding optimal groups based on the notion of group skyline. Skyline computation, aiming at identifying a set of skyline points that are not dominated by any other point, is particularly useful for multi-criteria data analysis and decision-making. Traditional skyline computation, however, is inadequate to answer queries that need to analyze not only *individual* points but also *groups* of points. To address this gap, SkyRec finds Pareto optimal groups with two group skyline models: G-Skyline [3] and Sum-Skyline [2]. SkyRec returns Pareto optimal groups with group size k that are not dominated by any other group with the same group size. Users can examine the results of the group skyline based recommendation compared to traditional top-k and skyline based recommendation and how different group skyline notions differ from each other. Although we demonstrate SkyRec for hotel reservation in this paper, it can be applied to various decision-making applications.

ACM Reference Format:
Jinfei Liu, Li Xiong, Jian Pei, Jun Luo, Haoyu Zhang, and Si Zhang. 2019. SkyRec: Finding Pareto Optimal Groups. In *The 28th ACM International Conference on Information and Knowledge Management (CIKM '19), November 3–7, 2019, Beijing, China.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3357384.3357838

1 INTRODUCTION

Recommender systems have become increasingly popular in recent years, and are utilized in a variety of areas including hotels, restaurants, movies, music, news, books, and products in general. The assumption for most of the recommender systems is that we know users' preferences (weights) on attributes. However, the relative preferences for different attributes are hard to specify in advance for different users.

Figure 1 illustrates a hotel search example. Assume a conference will be held at Emory University, and the conference organizers want to reserve a hotel for the participants. Figure 2(a) illustrates a dataset $P = \{p_1, p_2, ..., p_{11}\}$, each representing a hotel with two attributes: distance to the destination and price. Figure 2(b) shows the corresponding points in the two-dimensional space, where x and y coordinates correspond to the attributes of distance to the destination and price, respectively. We can see that $p_3(14, 340)$ dominates $p_2(24, 380)$ as an example of dominance because both attributes of p_3 are better than p_2. The skyline contains p_1, p_6, and p_{11}. Suppose the organizers of a conference need to reserve *one* hotel considering both distance to the conference destination and price for participants. Skyline offers a set of best options or Pareto optimal solutions with various tradeoffs between distance and price: p_1 is the nearest to the destination, p_{11} is the cheapest, and p_6 provides a good compromise of the two factors. p_8 will not be considered as p_{11} is better than p_8 in both factors.

Suppose the organizers need to reserve *a group of* hotels (instead of one) considering both distance to the conference destination and price for participants. In contrast to the traditional skyline problem which finds Pareto optimal solutions where each solution is a single point, we are interested in finding Pareto optimal solutions where each solution is a group of points. One may use the traditional

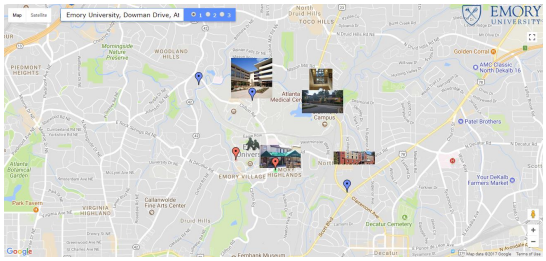
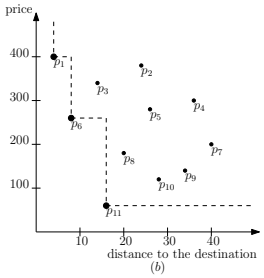


Figure 1: A hotel search example.

hotel	distance	price
p_1	4	400
p_2	24	380
p_3	14	340
p_4	36	300
p_5	26	280
p_6	8	260
p_7	40	200
p_8	20	180
p_9	34	140
p_{10}	28	120
p_{11}	16	60

(a)



(b)

Figure 2: A skyline example of hotels.

CIKM '19, November 3–7, 2019, Beijing, China

ACM ISBN 978-1-4503-6976-3/19/11…$15.00
https://doi.org/10.1145/3357384.3357838

skyline definition, and return all subsets from the skyline points p_1, p_6, and p_{11}. If the desired group size is 2, group $\{p_1, p_6\}$, $\{p_1, p_{11}\}$, and $\{p_6, p_{11}\}$ can be returned. However, we show that this definition does not capture all the best groups. For example, $\{p_{11}, p_{10}\}$ should clearly be considered a Pareto optimal group to users who use price as the main criterion, e.g., Ph.D. students with a low travel budget since p_{11} provides the best price and p_{10} the second best price. Hence no other groups are better than this group in terms of price.

G-Skyline, for finding Pareto optimal groups, is defined in [3, 5, 6]. Given two different groups G and G' with k points, we say G g-dominates G', if for any point p'_i in G', we can find a distinct point p_i in G, such that p_i dominates p'_i or $p_i = p'_i$, and for at least one i, p_i dominates p'_i. The G-Skyline are those groups that are not g-dominated by any other group with the same size. Intuitively, if we consider the points in each group as a set of dimensions orthogonal to the attributes of each point, the definition of G-Skyline groups with the group dominance is in spirit similar to the skyline definition, in that a group is a skyline group if no permutation of any other group exists that is better for at least one point and at least as good for every other point.

The most related works to G-Skyline are [1, 2, 7]. They formulated and investigated the problem of computing skyline groups. However, the notion of dominance between groups in these works is defined by the dominance relationship between an "aggregate" or "representative" point of each group. More specifically, they calculate for each group a single aggregate point, whose attribute values are aggregated over the corresponding attribute values of all points in the group. The groups are then compared by their aggregate points using the traditional point dominance. While many aggregate functions can be considered in calculating aggregate points, they focus on several functions commonly used in database applications, such as SUM, MIN, and MAX. In fact, the result of skyline groups under SUM dominance [1, 2, 7] is a subset of G-Skyline groups.

In this demonstration, we present SkyRec, a toolkit for finding Pareto optimal groups. The core of SkyRec is G-Skyline and Sum-Skyline definitions. The application implements and extends our recent work [3] as well as Sum-Skyline work.

Contributions. We briefly summarize our contributions as follows.

- SkyRec implements G-Skyline [3] for finding Pareto optimal groups.
- SkyRec involves state-of-the-art definition related to group skyline, i.e., Sum-Skyline [2].
- SkyRec provides an easy-to-use web-based interface. We employ the Google Maps Javascript API which provides an access to world-wide geolocation datasets. Based on users' input, SkyRec returns the Pareto optimal groups for users in the practical scenario.

2 SYSTEM OVERVIEW

In this section, we introduce the SkyRec framework and its key components as well as the brief algorithm of G-Skyline. As shown in Figure 3, for the algorithm of G-Skyline, we first build *skyline layers*, and then compute Directed Skyline Graph (*DSG*). Finally, we process the *pruning* on enumeration tree based on DSG.

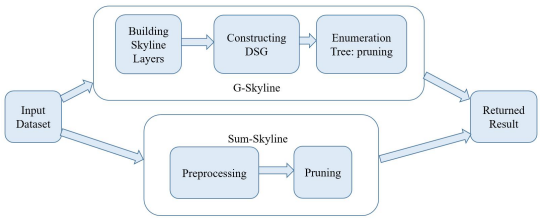


Figure 3: SkyRec framework.

2.1 Definitions

Definition 2.1. (**Skyline**) [4]. Given a dataset P of n points in d-dimensional space. Let p and p' be two different points in P, p dominates p', denoted by $p \prec p'$, if for all i, $p[i] \leq p'[i]$, and for at least one i, $p[i] < p'[i]$, where $p[i]$ is the i^{th} dimension of p and $1 \leq i \leq d$. The skyline points are those points that are not dominated by any other point in P.

Definition 2.2. (**G-Skyline**). Given a dataset P of n points in a d-dimensional space. Let $G = \{p_1, p_2, ..., p_k\}$ and $G' = \{p'_1, p'_2, ..., p'_k\}$ be two different groups with k points of P, we say group G **g-dominates** group G', denoted by $G \prec_g G'$, if we can find two permutations of the k points for G and G', $G = \{p_{u_1}, p_{u_2}, ..., p_{u_k}\}$ and $G' = \{p'_{v_1}, p'_{v_2}, ..., p'_{v_k}\}$, such that $p_{u_i} \prec p'_{v_i}$ or $p_{u_i} = p'_{v_i}$ for all i $(1 \leq i \leq k)$ and $p_{u_i} \prec p'_{v_i}$ for at least one i. The k-point G-Skyline consists of those groups with k points that are not g-dominated by any other group with the same size.

Lemma 2.3. *A point in a G-Skyline group cannot be dominated by a point outside the group.*

Skyline Layers. Motivated by Lemma 2.3, we present a structure representing the points and their dominance relationships based on the notion of skyline layers.

Definition 2.4. (**Skyline Layers**). Given a dataset P of n points in a d-dimensional space. The set of skyline layer $layer_1$ contains the skyline points of P, i.e., $layer_1 = skyline(P)$. The set of $layer_2$ contains the skyline points of $P \backslash layer_1$, i.e., $layer_2 = skyline(P \backslash layer_1)$. Generally, the set of $layer_j$ contains the skyline points of $P \backslash \bigcup_{i=1}^{j-1} layer_i$, i.e., $layer_j = skyline(P \backslash \bigcup_{i=1}^{j-1} layer_i)$. The above process is repeated iteratively until $P \backslash \bigcup_{i=1}^{j-1} layer_i = \emptyset$.

Directed Skyline Graph. We now present a definition of directed skyline graph, a data structure we use to represent the points from the first k skyline layers as well as their dominance relationships in order to compute k-point G-Skyline groups.

Definition 2.5. (**Directed Skyline Graph (DSG)**). A directed skyline graph is a graph where a node represents a point and an edge represents a dominance relationship. Each node has a structure as follows.

$$[layer\ index, point\ index, parents, children]$$

where layer index ranging from 1 to k indicates the skyline layer that the point lies on, point index ranging from 0 to $\mathbb{S}_k - 1$ uniquely identifies the point and $\mathbb{S}_k$ is the number of points in the first k skyline layers, parents include all the points that dominate this point, and children include all the points that are dominated by the point.

Definition 2.6. (**Sum-Skyline**) [2]. Given a dataset P of n points in a d-dimensional space. Let $G = \{p_1, p_2, ..., p_k\}$ and $G' = \{p'_1, p'_2, ..., p'_k\}$ be two different groups with k points of P. The new tuples $p_G = \{\sum_{i=1}^{i=k} p_i[1], \sum_{i=1}^{i=k} p_i[2], ..., \sum_{i=1}^{i=k} p_i[d]\}$ and $p_{G'} = \{\sum_{i=1}^{i=k} p'_i[1], \sum_{i=1}^{i=k} p'_i[2], ..., \sum_{i=1}^{i=k} p'_i[d]\}$, we say group G **sum-dominates** group G', denoted by $G \prec_{sum} G'$, if $p_G \prec p_{G'}$. The k-point Sum-Skyline consists of those groups with k points that are not sum-dominated by any other group with the same size.

2.2 Algorithm for Computing G-Skyline

Computing Skyline Layers. For two-dimensional space, the intuition of the algorithm is motivated by the monotonic property of skyline points, i.e., if we sort the skyline points with increasing x-coordinate, their y-coordinates decrease monotonically. Our key idea of the algorithm is to sort all the points in ascending x-coordinates if they are not sorted already and process them in that sequence by either adding them into an existing layer it belongs to or starting a new layer. Please see [3] for the high-dimensional space case.

Constructing Directed Skyline Graph. Once we build the skyline layers, we can build a DSG to capture all the dominance relationships between the layers which is detailed shown in [3].

Finding G-Skyline Groups. We already showed (in Lemma 2.3) that a point in a G-Skyline group cannot be dominated by a point outside the group. In other words, for a point in a G-Skyline group, its unit group must be in the group. This motivates our unit group-wise algorithm which expands candidate groups by unit groups, adding one unit group at a time. We can represent the entire search space of candidate groups as a set enumeration tree, where each node is a set of unit groups. And then dynamically generate the tree while pruning as many non-G-Skyline groups as possible.

3 SYSTEM DEMONSTRATION

This section describes how to run the demonstration application. In this demonstration, users are able to search for nearby hotels, and it is extendable to other points of interest, e.g., restaurants and cafes. By using the G-Skyline and Sum-Skyline algorithms, SkyRec returns Pareto optimal groups with k hotels.

3.1 Datasets

We use the Google Maps Javascript API which provides access to world-wide geolocation datasets. Based on users' location search, Google returns a list of up-to-date information about nearby hotels which is within a specified area (a circle whose radius is one mile). The hotel information required by the application is hotel name, hotel address, latitude and longitude (for computing the distance to the search location), star, and price. This can be easily extended to any dimensions. For the sake of simplicity, we assume all the attributes have inverted scores (the lower the better). This can be easily modified by taking the negation in practical systems. We note here that we need to pay for accessing the hotel price and star attributes, so we randomly generate the price and star values. We also note that SkyRec is capable with all major location-based service providers, e.g., Yelp, FourSquare, and Facebook Places.

3.2 Basic Functionalities

SkyRec is an easy-to-use web application that guides users to find the Pareto optimal groups. The SkyRec interface allows users to input the destination, group size k (k=1 for traditional skyline definition), interested attributes (price, star, or rating), and the preferred group skyline definition, i.e., G-Skyline or Sum-Skyline. SkyRec returns the Pareto optimal groups in the interface and allows users to visualize the result on the map. Furthermore, a user can click the hotel for more detailed information about this hotel, e.g., web link, address, rating, price, distance to the destination, and star. For the input parameters, we will engage the CIKM 2019 attendees to input their interested destination, group size k, and interested attributes to see if the returned results are appealing for them.

3.3 Visualization of G-Skyline

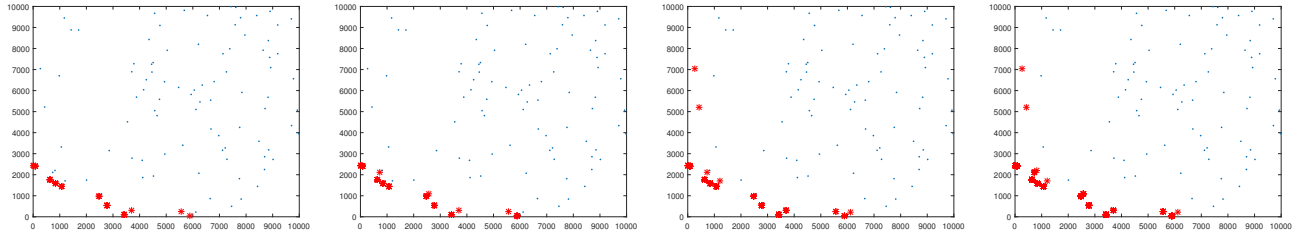
Given the dataset, users specify their input parameters (destination, group size k, and interested attributes), and examine the results. Figure 5 provides a visualization example of G-Skyline definition. Assume that an organizer needs to reserve a group of hotels near Emory University. The organizer needs to make a decision based on three criterions: distance to the conference destination, price, and star. The desired group size k is 2 in this case. We use circles with same color to represent hotels in the same Pareto optimal group. Furthermore, we use a line to connect these two circles which have the same color and radius in order to conveniently match them. The corresponding G-Skyline query result is shown in Figure 6. Based on the distance, price, and star, SkyRec returns four Pareto optimal groups with group size k=2 as shown in Figures 5 and 6.

3.4 Visualization of Sum-Skyline

Figure 7 provides a visualization example of Sum-Skyline definition. The corresponding Sum-Skyline query result is shown in Figure 8. Based on distance to the destination, price, and star, SkyRec returns two Pareto optimal groups as shown in Figures 7 and 8.

Analysis on Hotel Dataset. In fact, the result of Sum-Skyline is a subset of G-Skyline groups. If group G is dominated by G' in G-Skyline definition, then G must be dominated by G' in Sum-Skyline definition, but not vice versa. As shown in Figures 5 and 6, there are four groups returned by G-Skyline. However, in Figures 7 and 8, there are only two groups returned by Sum-Skyline because the second (third) group is dominated by the first (fourth) group in Sum-Skyline definition where the second group is shown in black line (circles) and the third group is shown in orange line (circles) in Figure 5. However, we cannot conclude that the first (fourth) group is better than the second (third) group because the assumption of skyline is that we do not know the users' weights on attributes in advance. For users who consider distance, price, and star in their interested attributes, they may prefer the second group and the third group, because they both provide a good compromise of distance, price, and star. Hence, some Pareto optimal solutions are not captured by Sum-Skyline definition.

Take the second group in Figure 6 as an example. In this case, the second group should be considered a "good" group because no other group can g-dominate it. However, if we only build groups from skyline hotels, this group will not be captured because the first hotel in the second group is dominated by the second hotel



(a) Sum-Skyline with group size k=2 (b) G-Skyline with group size k=2 (c) Sum-Skyline with group size k=3 (d) G-Skyline with group size k=3

Figure 4: Performances on synthetic dataset.

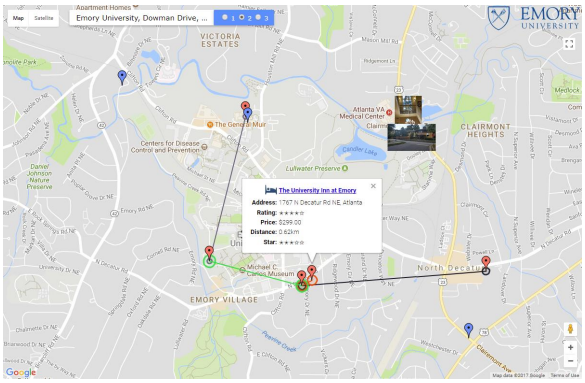


Figure 5: Query result of G-Skyline shown in map.

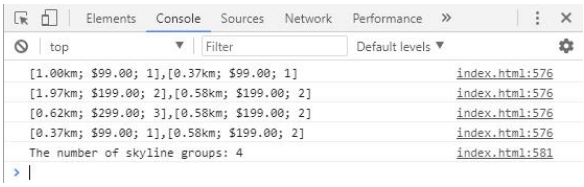


Figure 6: Query result of G-Skyline.

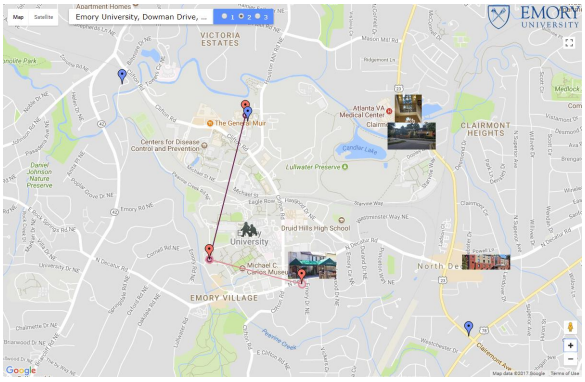


Figure 7: Query result of Sum-Skyline shown in map.

in the same group. In essence, taking only skyline hotels will not capture those groups which may include non-skyline hotels which are only dominated by another hotel in the same group but are not dominated by any other hotel outside the group.

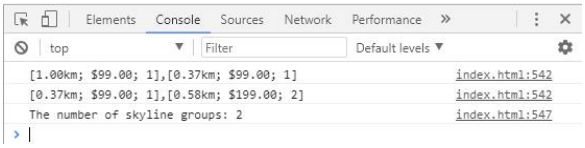


Figure 8: Query result of Sum-Skyline.

Analysis on Synthetic Dataset. To further demonstrate the difference between G-Skyline and Sum-Skyline for helping users to choose the appropriate group skyline model, we generated an independent and identically distributed random dataset of 100 points in two-dimensional space. The experimental results are shown in Figures 4(a)(b)(c)(d). Figures 4(a)(c) show the Sum-Skyline result in two-dimensional space, there are 11 (16) points belonging to one or more skyline groups when group size k = 2(3). Similarly, Figures 4(b)(d) show the G-Skyline result in two-dimensional space, there are 13 (18) points belonging to one or more skyline groups when group size k = 2(3). We can see that the result of Sum-Skyline is a subset of G-Skyline groups. Furthermore, for both G-Skyline and Sum-Skyline, the set of points belonging to one or more skyline groups with smaller group size is the subset of such set of points with bigger group size. Similar to the observation in the analysis on real hotel dataset, some non-skyline points are included in the group skyline.

ACKNOWLEDGEMENT

This work is supported in part by NSF under CNS-1618932 and AFOSR DDDAS Program under FA9550-12-1-0240.

REFERENCES

[1] H. Im and S. Park. Group skyline computation. *Inf. Sci.*, 188:151–169, 2012.
[2] C. Li, N. Zhang, N. Hassan, S. Rajasekaran, and G. Das. On skyline groups. In *CIKM*, pages 2119–2123, 2012.
[3] J. Liu, L. Xiong, J. Pei, J. Luo, and H. Zhang. Finding pareto optimal groups: Group-based skyline. *PVLDB*, 8(13):2086–2097, 2015.
[4] J. Liu, L. Xiong, and X. Xu. Faster output-sensitive skyline computation algorithm. *Inf. Process. Lett.*, 114(12), 2014.
[5] W. Yu, J. Liu, J. Pei, L. Xiong, X. Chen, and Z. Qin. Efficient contour computation of group-based skyline. *CoRR*, abs/1905.00700, 2019.
[6] W. Yu, Z. Qin, J. Liu, L. Xiong, X. Chen, and H. Zhang. Fast algorithms for pareto optimal group-based skyline. In *CIKM*, pages 417–426, 2017.
[7] N. Zhang, C. Li, N. Hassan, S. Rajasekaran, and G. Das. On skyline groups. *IEEE Trans. Knowl. Data Eng.*, 26(4):942–956, 2014.