Maximal Reverse Skyline Query

Farnoush Banaei-Kashani
Computer Science Dept.
Univ. of Southern California
banaeika@usc.edu

Parisa Ghaemi
Computer Science Dept.
Univ. of Southern California
ghaemi@usc.edu

John P. Wilson
Spatial Sciences Inst.
Univ. of Southern California
jpwilson@usc.edu

ABSTRACT

Given a set S of sites and a set O of objects in a metric space, the Optimal Location (OL) problem is about computing a location in the space where introducing a new site (e.g., a retail store) maximizes the number of the objects (e.g., customers) that would choose the new site as their "preferred" site among all sites. However, the existing solutions for the optimal location problem assume that there is only one criterion to determine the preferred site for each object (i.e., the metric distance between objects and sites), whereas with numerous real-world applications multiple criteria are used as preference measures. In this paper, for the first time we develop an efficient and exact solution for the so-called *Multi-Criteria Optimal Location* (*MCOL*) problem that can scale with large datasets. Toward that end, first we formalize the MCOL problem as *maximal reverse skyline query* (*MaxRSKY*). Given a set of sites and a set of objects in a d-dimensional space, MaxRSKY query returns a location in the space where if a new site s is introduced, the size of the (bichromatic) reverse skyline set of s is maximal. To the best of our knowledge, this paper is the first to define and study MaxRSKY query. Accordingly, we propose a baseline solution for identification of the optimal location.

Categories and Subject Descriptors

H.2 [**Database Management**]: Systems

General Terms

Performance

Keywords

Optimal Location Query, Maximal Reverse Skyline Query

1. INTRODUCTION

The problem of "optimal location" is a common problem with many applications in spatial decision support systems and marketing tools. With this problem, given the sets S of sites and O of objects in a metric space, one must compute the optimal location where introducing a new site maximizes the number of the objects that would choose the new site as their *preferred* site among all sites. For instance, a city planner must solve an optimal location problem to answer questions such as "where is the optimal location to open a new public library such that the number of patrons in the proximity of the new library (i.e., the patrons that would perhaps prefer the new library as their nearest library among all libraries) is maximized?".

An important limitation of the existing solutions for the optimal location problem is due to a common simplifying assumption that there is only one criterion to determine the preferred site for each object, i.e., the metric distance between objects and sites. In other words, the preferred site for an object is always assumed to be the closest site to the object. However, there are numerous real-world applications with which one needs to consider *multiple* criteria (possibly including the distance) to choose the most preferred site for each object. The extension of the optimal location problem which allows for using multiple criteria in selecting the preferred site for each object is termed *Multi-Criteria Optimal Location* (or *MCOL*, for short).

For an instance of the MCOL problem, consider the following market analysis application. In order to decide on the ideal specifications of its next product to be released, a laptop manufacturer wants to identify the most preferred / desired combination of laptop specifications in the market. For example, the current most preferred combination of laptop specifications can be <5lb, 8GB, 2.3GHz, 14in>, where the numbers stand for weight, memory capacity, CPU speed, and display size of laptop, respectively. One can formulate this problem as an MCOL problem, where each site represents an existing laptop product in the market with known specifications, and each object represents a buyer in the market with known preferences on the specifications of his/her desired laptop (the preferences of the buyers can be obtained, for example, by collecting and compiling their web search queries). In this case, laptop specifications (i.e., weight, memory capacity, CPU speed, and display size) are the criteria that objects (buyers) use to determine their preferred site (laptop). Accordingly, by solving this MCOL problem, the manufacturer can identify the specifications of a new laptop product (i.e., the new site which is optimally located) such that the number of potential buyers is maximized. Similarly, a cell phone company can identify the features (e.g., the monthly voice service allowance in minutes, text service allowance in number of text messages, and data service allowance in GB) of a new cell phone plan that

SIGSPATIAL '14, November 04 - 07 2014, Dallas/Fort Worth, TX, USA

http://dx.doi.org/10.1145/2666310.2666435

would attract the largest number of potential subscribers with different usage statistics.

While the MCOL problem is previously studied in the operations research community, the existing solutions for this problem are not only approximate solutions without any guaranteed error bound, but also more importantly, unscalable solutions that can merely apply to very small site and object datasets (Section 2.1 reviews such solutions); hence inapplicable to real-world applications. In this paper, for the first time we focus on developing an efficient and exact solution for MCOL that can scale with large datasets containing thousands of sites and objects.

Toward that end, first we formalize the MCOL problem as *maximal reverse skyline query* (*MaxRSKY*). Given a set of sites and a set of objects in a d-dimensional space, MaxRSKY query returns a location in the d-dimensional space where if a new site s is introduced, the size of the (bichromatic) reverse skyline set of s is maximal. To the best of our knowledge, this paper is the first to define and study MaxRSKY query. Second, we develop a baseline solution for MaxRSKY which derives an answer for the query by 1) computing the skyline set and the corresponding skyline region for every object (the skyline region for an object is a region where if a new site is introduced it will become a skyline site for the object), and 2) for each subset of the set of skyline regions computed for all objects, overlap all regions in the subset to identify the maximum overlap region in the subset (i.e., the region where the largest number of skyline regions intersect in the subset). One can observe that among all maximum overlap regions identified for all subsets of skyline regions, the one with largest number of overlapping regions is where if a new site is introduced, its reverse skyline set is maximal. We call this region the maximal overlap region.

The remainder of this paper is organized as follows. Section 2 reviews the related work. Section 3 formally defines the MCOL problem and formalizes this problem as MaxRSKY query. In Section 4, we present a baseline solution. Finally, Section 5 concludes the paper and discusses our directions for future research.

2. RELATED WORK

In this section, we review the related work under two main categories. First, we discuss the previous work on the problem of optimal location. Thereafter, we present a review of the skyline query processing literature.

2.1 Optimal Location

Among other variations of the optimal location problem, the multi-criteria optimal location (MCOL) a.k.a. multi-objective or multi-attribute optimal location, has been widely studied by researchers in the operations research (OR) community [9, 8, 13, 16, 5, 19]. However, given the computational complexity of the MCOL problem most of the existing solutions 1) resort to the use of heuristics that can only approximate the optimal location without any guaranteed error bounds, and more importantly 2) fail to scale with real datasets that often consist of thousands of sites and objects (rather than tens of sites and objects usually assumed by the existing solutions).

Given similar scalability and accuracy issues with the existing solutions from the OR community for the general family of optimal location problems, recently the database community has shown interest in developing efficient and exact solutions for these problems. However, so far all proposed solutions from this community have focused on the basic (single-criterion) optimal location problem. In particular, Wong et al. [21] and Du et al. [7] formalized the basic optimal location problem as maximal reverse nearest neighbor (MaxRNN) query, and presented two scalable approaches to solve the problem in p-norm space (assuming L2-norm and L1-norm, respectively). Thereafter, Ghaemi et al. [10, 11, 12] and Xiao et al. [23] continued the previous studies and proposed solutions for MaxRNN assuming network distance. Finally, Zhou et al. [24] presented an efficient solution to solve the extended MaxRkNN problem, which computes the optimal location where introducing a new site maximizes the number of objects that consider the new site as one of their k nearest sites. To the best of our knowledge, we are the first to tackle the MCOL problem toward developing an efficient and exact solution that can scale with large datasets containing thousands of sites and objects.

2.2 Skyline Queries

Skyline queries were first theoretically studied as maximal vectors [15, 2, 1]. However, it was Borzsonyi et al. [3] who first introduced the concept to the database community and showed the need for scalable solutions to process skyline queries on large datasets. Since then, numerous efficient algorithms have been proposed for processing static and dynamic skyline queries, such as BNL [3], D&C [3], Bitmap [20], SFS [4], Index [20], NN [14], and BBS [18]. Moreover, several variations of skyline queries have been proposed and studied, among which the reverse skyline query is most relevant to this paper. The reverse skyline of a query object q returns the objects whose dynamic skyline contains q. Dellis et al. [6] first introduced reverse skyline queries in a monochromatic context (involving a single dataset). Lian et al. [17] extended the definition of reverse skyline to a bichromatic scenario, and proposed an algorithm for efficient computation of reverse skyline on uncertain data. Later, Wu et al. [22] further studied bichromatic reverse skyline queries and proposed the most efficient query processing solution known so far assuming certain datasets.

However, it is important to note that reverse skyline query and maximal reverse skyline query (MaxRSKY) are two orthogonal problems. While with our focus problem (i.e., MaxRSKY) we can leverage any efficient solution for reverse skyline computation, as we show in Section 4 our main challenge is to identify a location which is on the reverse skyline set of a *maximal* number of objects.

3. PROBLEM DEFINITION

In this section, we first formally define the problem of Multi-Criteria Optimal Location (MCOL). Then, we formalize this problem as maximal reverse skyline query (MaxRSKY).

3.1 Multi-Criteria Optimal Location (MCOL)

Suppose we have a set S of sites $s(s^1, s^2, ..., s^d)$ where s^i is the value of the i-th attribute for the site s, as well as a set O of objects $o(o^1, o^2, ..., o^d)$ in the same d-dimensional space where o^i indicates the preference of o on the i-th attribute. For example, considering our laptop market analysis example from Section 1, each laptop is a site with four attributes, namely, weight, memory capacity, CPU speed, and display size. Similarly, each potential buyer is represented by an object with four preferences corresponding to the four afore-

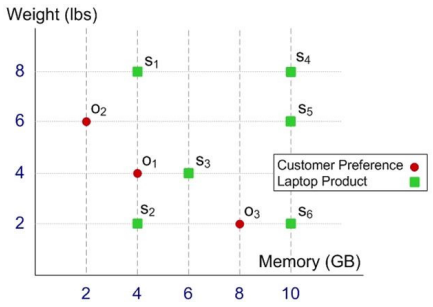

Figure 1: Example site and object datasets in 2-dimensional space

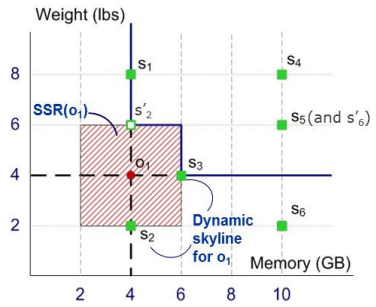

Figure 2: SSR region for object o_1

mentioned attributes. Figure 1 illustrates six sites/laptops s_1 to s_6 each characterized by two attributes, weight and memory capacity (for simplicity of presentation, hereafter we consider a 2-dimensional space without loss of generality). In the same figure, three objects/buyers o_1 to o_3 are shown by indicating their preferences on weight and memory capacity of laptops in the same 2-dimensional space.

Accordingly, we define the MCOL problem as follows. Given a set S of sites with d attributes and a set O of objects with d preferences corresponding to the same attributes, the *multi-criteria optimal location* problem seeks a location/region (or set of locations/regions) in the d-dimensional space such that introducing a new site in this location maximizes the number of objects that each considers the new site among its set of "preferred" sites. Intuitively, a site s is a preferred site for object o if given the preferences of o, there is no other site s' in S that is more "preferred" by o as compared to s; in other words, intuitively for an object o we say a site s is more preferred as compared to a site s' if considering its preferences collectively, o has no reason to choose s' over s. For example, in Figure 1 the set of preferred sites for the object o_1 is $\{s_2, s_3\}$; note that while for o, s_2 and s_3 are not preferred over each other, they both are preferred as compared to all other sites s_1, s_4, s_5, and s_6.

3.2 Maximal Reverse Skyline (MaxRSKY)

In this section, for the sake of self-containment we first review the formal definitions of dynamic skyline query and bichromatic reverse skyline query. Thereafter, we define maximal reverse skyline (MaxRSKY) query, which is equivalent to and formalizes the MCOL problem.

DEFINITION 1 (DYNAMIC SKYLINE QUERY): *Given a set S of sites with d attributes and a query object o in the same d-dimensional space, the dynamic skyline query with respect to o, termed DSL(o), returns all sites in S that are not "dominated" by other sites with respect to o. We say a site $s_1 \in S$ dominates a site $s_2 \in S$ with respect to o iff 1) for all $1 \leq i \leq d$, $\left|s_1^i - q^i\right| \leq \left|s_2^i - q^i\right|$, and 2) there exists at least one j $(1 \leq j \leq d)$ such that $\left|s_1^j - q^j\right| < \left|s_2^j - q^j\right|$* □

For example, as shown in Figure 2, the skyline set for the object o_1 is $DSL(o_1) = \{s_2, s_3\}$. Note that s'_2 and s'_6 are proxies of the sites s_2 and s_6 transformed to the first quarter with respect to the reference point o_1, respectively.

DEFINITION 2 (BICHROMATIC REVERSE SKYLINE QUERY): *Let S and O be the sets of sites and objects in a d-dimensional space, respectively. Given a query site $s \in S$, the bichromatic reverse skyline query with respect to s returns all objects $o \in O$ such that s is in the dynamic skyline set of o, i.e., $s \in DSL(o)$* □

For instance, in Figure 1 the reverse skyline set of s_2 is $\{o_1\}$, because $DSL(o_1) = \{s_2, s_3\}$, $DSL(o_2) = \{s_1, s_4, s_5\}$, $DSL(o_3) = \{s_6\}$, and therefore, s_2 only belongs to $DSL(o_1)$.

DEFINITION 3 (MAXIMAL REVERSE SKYLINE QUERY (MaxRSKY)): *Let S and O be the sets of sites and objects in a d-dimensional space, respectively. The MaxRSKY query returns a location in this d-dimensional space where if a new site s is introduced, the size of the (bichromatic) reverse skyline set of s is maximal* □

It is easy to observe that MaxRSKY query and MCOL problem are equivalent, because maximizing the reverse skyline set of the newly introduced site s equivalently maximizes the number of objects whose sets of preferred sites include s.

4. BASELINE SOLUTION

Central to the solution for maximizing the reverse skyline is the concept of *Skyline Search Region (SSR)* [6]. The skyline search region for object o(or *SSR (o)*) is part of the data space containing points that can dominate at least one of the skyline sites of the object o. For instance, considering the running example in Figure 2 with skyline points $\{s_2, s_3\}$ for object o_1, the skyline search region $SSR(o_1)$ is the shaded area bounded by the skyline points and the two axes. Note that *SSR* does not include the skyline points themselves since a skyline point does not dominate itself.

LEMMA 1. (see [6] for proof) *For a given object point o, let $DSL(o)$ be the set of dynamic skyline sites for o. Let q be a query point. If $q \in SSR(o)$, then o is in reverse skyline of q.*

Accordingly, we propose our two-step *baseline* solution for maximal reverse skyline query as follows:

1. Compute the dynamic skyline set $DSL(o) \in S$ of sites for each object $o \in O$. Subsequently, construct the corresponding *SSR* for each object $o \in O$. This step produces $|O|$ regions.

2. Intersect the *SSRs* generated at the previous step to compute the maximal reverse skyline region. Given

$|O|$ SSRs, this step involves computing the overlap region for each of the $2^{|O|}$ combinations of SSR regions. Among the computed overlap regions for all combinations of SSRs, the overlap region(s) that involves the maximum number of $SSRs$ constitute the maximal reverse skyline region.

However, while correct, the proposed baseline approach suffers from two corresponding computational complexities that render its use impractical given the often large sizes of the site and object datasets:

1. Given the computational complexity of computing skyline query on the one hand, and the large size of the object and site datasets on the other hand, computing SSR for all objects is costly.

2. Computing the overlap region for all combinations of $SSRs$ is exponentially time complex.

5. CONCLUSIONS

In this study, for the first time we proposed a solution for the problem of maximal reverse skyline query computation. Accordingly, we proposed a baseline solution for computation of MaxRSKY.

One can observe that among the computational complexities associated with the baseline solution, computing the overlap region for all combinations of $SSRs$ is the dominating complexity. Therefore, in the future, we focus on reducing the computational complexity of the second step of the baseline solution, i.e., overlap computation, by leveraging index-based pruning mechanisms.

6. REFERENCES

[1] J. L. Bentley, K. L. Clarkson, and D. B. Levine. Fast linear expected-time alogorithms for computing maxima and convex hulls. In *Proceedings of the first annual ACM-SIAM symposium on Discrete algorithms*, SODA '90, pages 179–187, 1990.

[2] J. L. Bentley, H. T. Kung, M. Schkolnick, and C. D. Thompson. On the average number of maxima in a set of vectors and applications. *J. ACM*, 25:536–543, October 1978.

[3] S. Börzsönyi, D. Kossmann, and K. Stocker. The skyline operator. In *Proceedings of the 17th International Conference on Data Engineering*, pages 421–430, 2001.

[4] J. Chomicki, P. Godfrey, J. Gryz, and D. Liang. Skyline with presorting. In *In ICDE*, pages 717–719, 2003.

[5] J. L. Cohon. *Multiobjective programming and planning*. Mathematics in science and engineering. ; 140. Acad. Press, New York [u.a.], 1978.

[6] E. Dellis and B. Seeger. Efficient computation of reverse skyline queries. In *Proceedings of the 33rd international conference on Very large data bases*, VLDB '07, pages 291–302, 2007.

[7] Y. Du, D. Zhang, and T. Xia. The optimal location query. In *Proceedings of Advances in Spatial and Temporal Databases*, pages 163–180, 2005.

[8] R. Farahani and M. Hekmatfar. *Facility Location: Concepts, Models, Algorithms and Case Studies*. Contributions to Management Science. Physica-Verlag HD, 2011.

[9] R. Z. Farahani, M. SteadieSeifi, and N. Asgari. Multiple criteria facility location problems: A survey. *Applied Mathematical Modelling*, 34:1689–1709, 2010.

[10] P. Ghaemi, K. Shahabi, J. P. Wilson, and F. Banaei-Kashani. Optimal network location queries. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2010.

[11] P. Ghaemi, K. Shahabi, J. P. Wilson, and F. Banaei-Kashani. Continuous maximal reverse nearest query on spatial networks. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2012.

[12] P. Ghaemi, K. Shahabi, J. P. Wilson, and F. Banaei-Kashani. A comparative study of two approaches for supporting optimal network location queries. *GeoInformatica*, 18(2), 2014.

[13] M. Hekmatfar and M. SteadieSeifi. *Multi-Criteria Location Problem*. Contributions to Management Science. Physica-Verlag HD, 2009.

[14] D. Kossmann, F. Ramsak, and S. Rost. Shooting stars in the sky: An online algorithm for skyline queries. In *In VLDB*, pages 275–286, 2002.

[15] H. T. Kung, F. Luccio, and F. P. Preparata. On finding the maxima of a set of vectors. *Journal of the ACM*, 22:469–476, 1975.

[16] O. Larichev and D. L. Olson. *Multiple Criteria Analysis in Strategic Siting Problems*. Kluwer Academic Publishers, 2001.

[17] X. Lian and L. Chen. Monochromatic and bichromatic reverse skyline search over uncertain databases. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, SIGMOD '08, pages 213–226, 2008.

[18] D. Papadias, G. Fu, J. M. Chase, and B. Seeger. Progressive skyline computation in database systems. *ACM Trans. Database Syst*, 30:2005, 2005.

[19] F. Szidarovszky, M. Gershon, and L. Duckstein. *Techniques for multiobjective decision making in systems management*. Advances in industrial engineering. Elsevier, 1986.

[20] K. Tan, P. Eng, and B. C. Ooi. Efficient progressive skyline computation. In *Proceedings of the 27th International Conference on Very Large Data Bases*, VLDB '01, pages 301–310, 2001.

[21] R. C. Wong, M. T. Ozsu, P. S. Yu, A. W. Fu, and L. Liu. Efficient method for maximizing bichromatic reverse nearest neighbor. In *In VLDB*, pages 1126–1149, 2009.

[22] X. Wu, Y. Tao, R. C. Wong, L. Ding, and J. X. Yu. Finding the influence set through skylines. In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, EDBT '09, pages 1030–1041, 2009.

[23] X. Xiao, B. Yao, and F. Li. Optimal location queries in road network databases. In *Proceedings 27th ICDE Conference*, 2011.

[24] Z. Zhou, W. Wu, X. Li, M. Lee, and W. Hsu. Maxfirst for maxbrknn. In *ICDE'11*, pages 828–839, 2011.