Information Sciences 324 (2015) 146–162

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Efficient computation for probabilistic skyline over uncertain preferences

Arun K. Pujari [a], Venkateswara Rao Kagita [a], Anubhuti Garg [b], Vineet Padmanabhan [a,*]

[a] Artificial Intelligence Lab, School of Computer & Information Sciences, University of Hyderabad, Hyderbad 500046, Andhra Pradesh, India
[b] LNM Institute of Information Technology, Jaipur, India

ARTICLE INFO

Article history:
Received 16 October 2014
Revised 9 May 2015
Accepted 22 June 2015
Available online 27 June 2015

Keywords:
Skyline query
Skyline computation
Uncertain preferences

ABSTRACT

Efficient computation of skyline probability over uncertain *preferences* has not received much attention in the database community as compared to skyline probability computation over uncertain *data*. All known algorithms for probabilistic skyline computation over uncertain preferences attempt to find inexact value of skyline probability by resorting to sampling or to approximation scheme. Exact computation of skyline probability for database with uncertain preferences of moderate size is not possible with any of the existing algorithms. In this paper, we propose an efficient algorithm that can compute skyline probability exactly for reasonably large database. The *inclusion–exclusion principle* is used to express skyline probability in terms of joint probabilities of all possible combination. In this regard we introduce the concept of *zero-contributing set* which has zero effect in the signed aggregate of joint probabilities. Our algorithm employs a *prefix-based k-level absorption* to identify zero-contributing sets. It is shown empirically that only a very small portion of exponential search space remains after level wise application of prefix-based absorption. Thus it becomes possible to compute skyline probability with respect to large datasets. Detailed experimental analysis for real and synthetic datasets are reported to corroborate this claim. We also propose an incremental algorithm to compute skyline probability in dynamic scenarios wherein objects are added incrementally. Moreover, the theoretical concepts developed in this paper help to devise an efficient technique to compute skyline probability of all objects in the database. We show that the exponential search space is pruned once and then for each individual object skyline probability can be derived by inspecting a portion of the pruned lattice. We also use a concept of *revival of absorbed pairs*. We believe that this process is more efficient than computing the skyline probability individually.

1. Introduction

Skyline computation aims at retrieving skyline-objects in a multi-attribute (equivalently, multidimensional) database. A skyline object is an object that is not *dominated* by any other object. An object p *dominates* an object q if p is strictly better than q on at least one dimension and p is better than or equal to q on the remaining dimensions. The domains of dimensions are

* Corresponding author. Tel.: +919989897205.
E-mail addresses: akpcs@uohyd.ernet.in (A.K. Pujari), 585venkat@gmail.com (V.R. Kagita), anubhuti.grg@gmail.com (A. Garg), vineetcs@uohyd.ernet.in, vcpnair73@gmail.com (V. Padmanabhan).

http://dx.doi.org/10.1016/j.ins.2015.06.041
0020-0255/

assumed to be ordered. For instance, given a set of points P in $\mathbb{R}^d$, a point $p \in P$ is *on the skyline* of P if for every other point $q \in P$ at least one coordinate of p is larger than that of q [1]. Skyline computation exists for both *certain* and *uncertain* datasets. Scenarios wherein a *certain* dataset is given, skyline computation returns a subset of tuples called the *skyline set* that are not dominated by any other tuples when all dimensions are taken into consideration. For example, suppose that we want to analyze NBA players using multiple technical statistics criteria like *number of points*, *number of assists* and *number of rebounds*. Though it is impossible to find a perfect player who can achieve the best performance in all aspects, a skyline analysis would disclose the trade-off among the merits of multiple aspects. We can say that a player P_1 is in the skyline if there exists no other player P_2 such that P_2 is better than P_1 in one aspect and is not worse than P_1 in all other aspects. On the other hand in the case of *uncertain* datasets a point(object) U in the dataset is represented by a set of *instances* (in the discrete case) such that each instance $u \in U$ has a probability p_u to appear. Two assumptions that holds for uncertain objects are that they are *independent* (instance of an object does not depend on the instances of any other objects) and each instance carries the same probability to happen. In such cases (probabilistic) skyline computation finds for each object the probability of its being in the skyline set, or its skyline probability. For example, in the case of an online shopping scenario a product might receive multiple evaluations that may vary from one another. Measuring the potential popularity of a product is almost impossible and therefore the product can be modelled as an uncertain object where each evaluation is regarded as an instance. Taking into consideration the products rating on price, quality etc. evaluation can also be multi-dimensional.

There has been considerable research interest in the area of skyline computation for *uncertain* data over the last decade. In 2007, Pei et al. [2] proposed a method of skyline computation where the database consists of uncertain attributes and developed bounding-pruning-refining techniques to deal with it efficiently. Efficient skyline computation against sliding windows on uncertain data streams was proposed by Zhang et al. [3]. In addition to conventional skyline operator, numerous skyline query variants have also been explored in the literature including skyline query for databases with uncertainties [2–7]. Lian and Chen [6] studied efficient reverse skyline query processing on uncertain data in both monochromatic and bi-chromatic fashion. Atallah et al. [4] investigated asymptotic time complexity of computing all skyline probabilities for data with discrete uncertainty. Other proposals dealing with skyline query processing can be found in [8–14]. Two recent publications that are in a way related to the present work is that of [15] and [5]. In the first work [15] an algorithm is proposed to compute exact skyline probabilities of all objects in a given uncertain dataset. The algorithm uses a dominance relation of attribute *values* that exhibit transitive property but cannot be applied in the context of uncertain *preferences*. The plus point is that it shows an exact way of computing skyline probabilities of all objects. The second one [5] introduces a novel concept of *P-domination* for tuples with probabilistic relations. In this setting, skyline of a probabilistic relation r is the set of tuples that are not P-dominated. One may consider *preference* as a dominance relation and instead of data uncertainties, it is pertinent to have uncertainty of preferences. Thus skyline computation can also be relevant when the attribute *preferences* are uncertain. *To the best of our knowledge all the above mentioned works deal with efficient computation of skyline probability in an uncertain environment assuming that uncertainty lies only in* `attribute values` *and does not deal explicitly with uncertainty in* `attribute preferences`.

To make things clear as to the importance of *uncertain preferences*, let us consider the problem of selecting interesting TV programs. TV programs have different attributes such as *category, genre, duration* etc. Different categories of TV programs are *family-serial, reality-show, movies* etc., and it is difficult to define an ordering of different categories, of genre or of other attributes. On the other hand, it is possible to specify a probability that 'movies' is preferred to 'reality show' and a probability of 'reality show' is preferred to 'movies', based on past record of usage. We can arrive at a preference probability of one TV program over other as the product of attribute-wise preference probabilities. Skyline probability of a TV program is then the probability that no other program is preferred to this one on all attributes. It is of interest to identify TV programs with high skyline probability.

Two works that deal with skyline computation on uncertain preferences are [16,17]. The *independent object dominance* assumption adopted in [16] to compute skyline probability on uncertain preferences was shown not valid in [17]. This leaves [17] as the only other work that deals with skyline probability computation over uncertain preferences. The algorithm given in [17] (named here as ZYLZ (initials of the authors)), the only known method to compute skyline probability of a given object, relies on *inclusion–exclusion principle* and computes joint probabilities of all elements of the power set. It essentially involves exponential number of terms and hence, skyline object cannot be determined even for a database of moderate size. An approximation scheme using Monte Carlo estimation is proposed in [17] to overcome performance shortfall. A preprocessing step is also proposed to avoid redundant computation. Notwithstanding these modifications, the algorithm cannot compute exact value of skyline probability for a database of reasonable size. We illustrate working of ZYLZ with the help of an example to support our motivation to explore alternative methods. Consider a database of five objects (Table 1) with four attributes A, B, C and D. Given two distinct attribute values a and b, we use the notation $a \preceq b$ to denote that a is preferred to b. All pairwise preference probabilities for values of attributes A, B, C and D are assumed to be 0.5. For instance, preference probabilities of values of A are $\Pr(a1 \preccurlyeq a2) = \Pr(a2 \preccurlyeq a1) = 0.5$. The problem is to find skyline probability of O, $\mathrm{sky}(O)$, with respect to $\{Q^1, Q^2, Q^3, Q^4, Q^5\}$ which as given in [17] can be reformulated for the above example as follows:(The formal definition of $\mathrm{Sky}(O)$, Pr is given in Section 2).

$$\mathrm{sky}(O) = 1 - \sum_{i=1}^{5} \Pr(e_i) + \sum_{\substack{i,j=1 \\ i<j}}^{5} \Pr(e_i \cap e_j) - \sum_{\substack{i,j,k=1 \\ i<j<k}}^{5} \Pr(e_i \cap e_j \cap e_k) + \sum_{\substack{i,j,k,l=1 \\ i<j<k<l}}^{5} \Pr(e_i \cap e_j \cap e_k \cap e_l) - [\Pr(e_1 \cap e_2 \cap e_3 \cap e_4 \cap e_5)].$$