VLDB
Very Large Data Bases
28 August 2012 - 30 August 2012
ISTANBUL
TURKEY

Reviews For Paper

Track	Research -> September 2011
Paper ID	224
Title	Skyline Groups

Masked Reviewer ID: Assigned_Reviewer_1

Review:

Question	
Overall Recommendation	2nd chance, revision requested
Summary of the paper (what is being proposed and in what context; brief justification of your overall recommendation). One paragraph	The paper introduces the problem of mining skyline groups, that are characterized by dominating other groups of tuples by means of an aggregate measure on their tuple attributes. The problem has apparently not been studied so far. The work introduces concepts for pruning the input and search space and algorithms exploiting those pruning techniques. Experimental evaluations for comparing different algorithm variants are provided.
Three (or more) strong points about the paper (Please be precise and explicit; clearly explain the value and nature of the contribution).	- novel problem - interesting non-trivial ideas on search pruning and search algorithms - significant improvement over baseline brute force search
Three (or more) weak points about the paper (Please indicate clearly whether the paper has any mistakes, missing related work, or results that cannot be considered a contribution; write it so that he authors can understand what is seen as negative aspects	- application motivation not very strong - questionable whether this is a "database" problem, looks more like an algorithmic search problem - evaluation results not coherent and incomplete
Relevant for VLDB2012	NO

Novelty (Please give a high novelty ranking to papers on new topics, opening new fields, or proposing truly new ideas; assign medium ratings for delta papers and papers on well known topics but still with some valuable contribution; low novelty ratings are	Novel
Rationale for novelty rating	The problem definition and in particular the ideas on pruning are novel. The authors do a good job in analyzing the various ramifications of the problem definition.
Significance	Improvement over existing work
Technical Depth and Quality of Content	Solid work
Experiments, Repeatability	Ok, but certain claims are not covered by the experiments
Presentation	Reasonable: improvements needed
Would you "champion" for acceptance of the paper in a discussion with the peer reviewers?	Well, not really
Detailed Evaluation (Contribution, Pros/Cons, Errors); please number each point	The problem definition and its analysis is an interesting intellectual contribution. Whether it is of practical relevance is hard to say, the examples provided look more like toy problems, e.g., it is not clear whether NBA teams would be selected in the way described. The main issue I have is with this work is that I would not consider it is a real database problem. The algorithms are already very expensive for relatively small datasets (most of the evaluations are done with a 500 tuple dataset), and are designed for main memory processing. Thus it appears to be an algorithmic search problem. The experiments do not really help to that respect. Not much is said about the scalability of the method, and looking e.g. at figure 4(b) or 5(b) one might have the impression that also the cost of the proposed algorithms grows exponentially in the input size. Unfortunately the report on the evaluation on a somewhat more significant, though also not spectacular dataset (35002 tuples from a stock quote) is

	extremely slim. Several observations that might raise some doubts are not commented, e.g. - why does the cost suddenly decrease when the input size grows (e.g. Fig 9(b)) - why are the costs for the larger datasets seemingly significantly lower than for the NBA dataset, if you compare absolute numbers and extrapolate the trend from the experiments on the NBA data. This raises at least doubts. Some minor errors: page 3, second column first para: It should be G not G_1 page 5, section 3.3, second para: I guess it hsould be h>=k and not h<=k
If 2nd chance with revision required, list specific revisions you seek from the Authors	The authors need to discuss aspects related to scalability of the approach. The also need to provide more detailed results on the second experiment, including comments on some of the observed effects which are hard to explain as such. Finally they should comment on how the methods would work for large datasets where the algorithm can no more be executed within a single main memory.

Masked Reviewer ID: Assigned_Reviewer_2

Review:

Question

Overall Recommendation	Reject
Summary of the paper (what is being proposed and in what context; brief justification of your overall recommendation). One paragraph	The paper studies the issue of computing the "best" subsets of a given relational table. This computation is defined in two steps. First, each set is summarized using an "aggregate vector" whose components are aggregate functions, e.g., MAX, applied to the set. Then the skyline of all those vectors is computed. All the subsets whose aggregate vectors are in the skyline are returned. The problem studied in the paper has already been addressed in the literature. The paper uses essentially the same framework as [29]. The contribution of the paper is to design new techniques for pruning the input and the search space (the latter using novel anti-monotonicity properties), and algorithms based on those properties. The scope of the proposed techniques is rather narrow.
Three (or more) strong points about the paper (Please be precise and explicit; clearly explain the value and nature of the contribution).	(S1) The paper proposes several new pruning techniques for best-subset generation and applies them in algorithms. The most sophisticated algorithm is based on dynamic programming. The novel pruning techniques are applicable to the case of the aggregation functions MAX and MIN (input pruning) and MAX/MIN/SUM (search space pruning). (S2) Some experimental results for small databases. (S3) The paper is readable.
Three (or more) weak points about the paper (Please indicate clearly whether the paper has any mistakes, missing related	(W1) Compared to [29], the conceptual framework of this paper is very limited. The notion of "aggregate vector" of a set used in the paper is significantly less general than that of "profile" [29]. Each component of the aggregate vector is a result of an aggregate function applied to the set. Although the definition (section 2.3.2) allows different functions for different components, all the technical results in the paper assume that all of those functions are the same,

work, or results that cannot be considered a contribution; write it so that he authors can understand what is seen as negative aspects	e.g., MAX. In [29] the components of a profile can be defined by simple, single-valued SQL queries that use aggregation and have a WHERE clause. Also, different components may be defined by different queries. Moreover, in [29] the best profiles may be defined by general preference relations, not just skylines. (W2) The authors themselves notice in section 5.2 that the applicability of their techniques in the MAX case may be very limited ("skyline groups are formed by a relatively small number of phenomenal players"). If they show how their techniques generalize to the case of aggregation vectors that contain a mix of MAX, MIN and SUM, their case would be much stronger. (W3) The input pruning techniques are similar to those achieved by superpreference [29]. Output compression is also studied in that paper. The current paper should provide a more in-depth comparison with [29], including experimental tests to compare the effectiveness of pruning and compression in both approaches. (W4) The paper contains some technical errors (see below).
Relevant for VLDB2012	YES
Novelty (Please give a high novelty ranking to papers on new topics, opening new fields, or proposing truly new ideas; assign medium ratings for delta papers and papers on well known topics but still with some valuable contribution; low novelty ratings are	With some new ideas
Rationale for novelty rating	The problem is not new; the proposed solutions are new but they apply only to very restricted cases.
Significance	No impact
Technical Depth and Quality of Content	Syntacticaly complete but with little contribution
Experiments, Repeatability	Ok, but certain claims are not covered by the experiments

Presentation	Reasonable: improvements needed
Would you "champion" for acceptance of the paper in a discussion with the peer reviewers?	No
Detailed Evaluation (Contribution, Pros/Cons, Errors); please number each point	(P1) The main contribution of the paper consists of new pruning techniques and algorithms to solve very restricted cases of the "best subsets" problem formulated in [29]. (P2) Two versions of the skyline order are used in the literature: "prefer smaller attribute values" and "prefer larger attribute values." The authors need to make it very clear throughout the paper which one is used. (P3) The paper identifies two basic properties of set comparison functions. (Incidentally, those functions are Boolean and are thus usually called binary relations - no need for new terminology.) The properties are quite intuitive. Nevertheless, it is not clear why other, equally intuitive properties were not considered. For example, transitivity of preferences is often assumed in the literature. The skyline order is transitive, but the function defined using the outdegrees in Fig.2 is not transitive. (P4) The NP-hardness result in section 3.2 is incorrect. First, a reduction should be FROM, not TO, an NP-complete problem. Second, a brute-force approach has time complexity which is polynomial in n. (It is not intuitive to consider the impact of k on the complexity as its value would be small.) (P5) Section 3.3: $h\geq k$. (P6) The explanation at the beginning of section 3.4.1 is incorrect. Consider $D=\{1,2,3\}, t=3, k=2, F=MAX, $G_k=\{1,3\}$. Then $G_{k-1}=\{1\}$ which is dominated by $\{2\}$ which does not contain 3. (P7) Include a proof of Theorem 1 with at least one case worked out. Give the counterexample in Theorem 3. (P8) The scalability of the proposed techniques to larger inputs should be studied. Also, provide precise asymptotic time complexity bounds for the algorithms.

Masked Reviewer ID: Assigned_Reviewer_3

Review:

Question

Overall Recommendation	Reject
Summary of the paper (what is being proposed and in what	The paper identifies and solves the problem of finding k tuples as a group such that the group is not dominated by other k-tuple groups. The problem itself is somewhat interesting in applying algorithmic techniques and heuristics, mainly about using aggregate functions to define the dominance relation. However, I

context; brief justification of your overall recommendation). One paragraph	am not convinced that the problem is useful in practice. Moreover, the discussion pairwise comparison based functions may have some holes.
Three (or more) strong points about the paper (Please be precise and explicit; clearly explain the value and nature of the contribution).	S1: a literarily new problem S2: some heuristics S3: a well-written paper in general
Three (or more) weak points about the paper (Please indicate clearly whether the paper has any mistakes, missing related work, or results that cannot be considered a contribution; write it so that he authors can understand what is seen as negative aspects	W1: the problem proposed looks artificial W2: the discussion on the pairwise comparison based functions seems to have some holes W3: the aggregate based functions are confined to only SUM, MAX, and MIN. It is unclear whether the sophisticated techniques can be extended and generalized to other aggregate functions W4: the data set used in the experiments is either tiny or small. This conference is about very large databases W5: the format requirement seems to be broken by reduced inline space and very small fonts in figures
Relevant for VLDB2012	YES
Novelty (Please give a high novelty ranking to papers on new topics, opening new fields, or proposing truly new ideas; assign medium ratings for delta papers and papers on well known topics but still with some valuable contribution; low novelty ratings are	Novelty unclear
Significance	No impact
Technical Depth and Quality of Content	Syntacticaly complete but with little contribution

Experiments, Repeatability	Ok, but certain claims are not covered by the experiments
Presentation	Excellent: careful, logical, elegant
Would you "champion" for acceptance of the paper in a discussion with the peer reviewers?	No
Detailed Evaluation (Contribution, Pros/Cons, Errors); please number each point	I find the problem proposed is highly artificial. The paper does not provide concrete and practically meaningful and significant examples to motivate the problem. The second principle of reasonable pairwise comparison based functions does not sound right. For example, consider the dominance relation defined as follows: G dominates G' if every tuple in G' is dominated by at least one tuple in G. This relation violates the 2nd principle. For two groups G_1 and G_2, if tuples in G_1 and G_2 do not dominate each other, except for one tuple in G_1 is dominated by a tuple in G_1, then, according to the 2nd principle, G_2 should be dominated by G_1. Moreover, the method finding all skyline groups with respect to pairwise comparison based functions seems not always correct. For example, considering again the dominance relation defined as follows: G dominates G' if every tuple in G' is dominated by at least one tuple in G, which violates the 2nd principle. The method cannot find all skyline groups. Why are the SUM, MIN, and MAX functions important and of high priority in defining aggregate based functions? Without a systematic discussion on computing skyline groups with respect to aggregate based functions in general, there seems an unfilled gap between the general problem and the instances of the three types of functions. Again, I do not find those three aggregate functions are well motivated in the context. The NBA data set used in the experiments is way too small. I understand that the paper uses those two data sets to evaluate the runtime. However, it would be nice if the skyline groups found in those data sets are meaningful in practice. Some case studies may help. Please do not reduce the inline space. Moreover, the fonts in the figures should be legible. Reading this paper is not that enjoyable partially because space is reduced and many figures are hard to read. There are quite a few typos. A thorough proofreading would help. Some examples are as follows. P3, the right column, end of paragraph 1: G_1 should be G. Section 3.4.1, the first paragraph: "is not be" should be "is not".