Choosing the Right Thing: Cooperative Trade-Off Enhanced Skyline Queries

Von der Carl-Friedrich-Gauß-Fakultät
Technische Universität Carolo-Wilhelmina zu Braunschweig

zur Erlangung des Grades
Doktor Naturwissenschaften (Dr. rer. nat.)

genehmigte **kumulative Dissertation**
von **Christoph Lofi**
geboren am 08.01.1981 in Birkenfeld

Eingereicht am: 18.01.2011
Mündliche Prüfung am: 21.03.2011

Referent:
Prof. Dr. *Wolf-Tilo Balke*, Technische Universität Braunschweig
Korreferentin:
Prof. *Seung-won Hwang, Ph.D.*, POSTECH University, Pohang, Korea

Druckjahr 2011

Inhaltsverzeichnis

Zusammenfassung

In den letzten 10 Jahren sind Skyline-Anfragen als Technik aus dem Bereich der Anfrage-Personalisierung vor allem durch ihre einfache und elegante Anfrageformalisierung unter Datenbanken-Wissenschaftlern populär geworden. Sie basieren auf dem aus den Wirtschaftswissenschaften bekannten Prinzip der Pareto-Optimalität und füllen damit die Lücke zwischen nicht personalisierten, klassisch Mengen-basierten Anfragesprachen wie beispielsweise SQL und personalisierbaren, aber komplexen Rang-basierten Verfahren wie z.B. *top-k queries*.

Trotz ihrer weiten Verbreitung im wissenschaftlichen Bereich blieben größere praktische Erfolge von Skyline-Anfragen bisher aus. Die Ursachen hierfür liegen vornehmlich in den inhärenten Problemen von Skyline-Anfragen, besonders in dem als *„curse of dimensionality“* bekannten Phänomen. Demnach wächst die Ergebnismenge von Skyline-Anfragen mit steigender Anzahl der beteiligten Anfrageattribute annähernd exponentiell – Ergebnisse, die bis zu 30 % oder 50 % der kompletten Datenbank beinhalten, sind daher keine Seltenheit. Derart große Ergebnismengen sind für den Benutzer zumeist unbrauchbar, da sie nicht mehr manuell durchgesehen werden können.

Im Zuge der vorliegenden Arbeit wurde daher ein Lösungsansatz für das Problem der zu großen und umfangreichen Skyline-Anfrageergebnisse entwickelt. Der Fortschritt der Arbeit wurde in mehreren Papieren auf internationalen Konferenzen und in Journalen publiziert. Das entwickelte Verfahren basiert darauf, dass von den Benutzern eines entsprechenden Informationssystems zusätzlich zur eigentlichen Anfrage Informationen in Form von Kompromissen (*„trade-offs“*) erhoben werden. Diese Kompromisse, die ein Benutzer bereit ist einzugehen, können nun dazu verwendet werden, die anfangs berechnete und meist zu umfangreiche Skyline zu fokussieren und individuell zu verkleinern. So kann die Zweckmäßigkeit der Ergebnismenge deutlich erhöht und damit letztendlich die Verwendbarkeit des Skyline-Paradigmas ermöglicht werden.

Im Nachfolgenden findet sich eine kurze Einführung in die Problematik von Skyline-Anfragen sowie Trade-Off-Skylines, gefolgt von einer Vorstellung der im Zuge dieser Promotion publizierten Forschungsarbeiten.

Abstract

Skyline queries are well-known for their intuitive query formalization and easy to understand semantics for selecting the most interesting data objects from large data sets. They naturally fill the gap between set-based queries using strict predicates and only few personalization options and rank-aware database retrieval, offering a high degree of personalization at the cost of very complex query formalization. Thus, skyline queries enjoyed great popularity in the database personalization research community. Unfortunately, the simplicity and elegance of the query paradigm come at high costs: skyline queries often suffer from a problem usually known as “curse of dimensionality”. With the increasing number of query attributes, the size of skyline result sets grows exponentially and the results are thus seldom useful or manageable by users –result sets containing 30%-50% are commonly heard of. This problem severely hinders the practical application of the skyline paradigm.

During the course of this thesis, the concept of trade-off skylines has been incrementally developed and successfully published on numerous international conferences and journals. Trade-off skylines approach the curse of dimensionality by eliciting additional user information in form of intuitive trade-offs. This additional information can be used to compensate between certain characteristics of database objects in order to focus the skyline result sets. Ultimately, this will lead to more manageable and useful query results, and thus alleviating one of the most severe problems of the Skyline paradigm.

In this cumulative doctoral thesis, the problem of unmanageable large Skyline query result sets is addressed and a solution based on cooperative user trade-offs is developed. In the following, after a short introduction to the area of Skyline queries, the relevant papers published during the course of this thesis are summarized and discussed.

1 Introduction

The ever growing amount of available information is one of the major problems of today's information systems. Besides solving the resulting performance issues, it is imperative to provide personalized and tailored access to the vast amount of information available in information and data systems in order to avoid flooding the user with unmanageably large query results.

As a possible remedy to this problem, Skyline queries [1] have been proposed, filling the gap between set-based SQL queries and rank-aware database retrieval [2]. Due to the paradigm's elegance and simplicity it has stirred a lot of interest within the database community in the recent years. Especially, Skyline queries allow for *human-centered* queries: instead of relying on hard filter criterions or complex and non-intuitive compensation functions, Skyline queries focus on user *preferences*. Preferences simply encode a user's likes and dislikes (e.g. "I like blue better than red") and may easily and naturally be elicited directly from users or implicitly from user profiles or interaction histories. Especially, preferences do not require extensive a-priori knowledge of the actual content of a database, nor do they require explicit numeric statements on attribute weights (e.g. "color is 0.7 times more important than spend and 0.2 times more important than price"). In contrast, preference queries aim at returning those objects from the database which match the user's preferences *best*.

Skyline queries rely on the notion of *Pareto optimality* which is an established paradigm already serving as a building block for a multitude of economic theories; e.g. given the choice between two objects, with one object being better with respect to at least one attribute but at least being equal with respect to all other attributes, users will always prefer the first object over the second one (the first object is said to *dominate* the second one). This simple concept can be used to implement an intuitive personalizable data filter as dominated objects can be safely excluded from the data collection, resulting in the *Skyline set* of the query. The semantic justification of this filter is easy to see using an example: if two car dealers in the neighborhood offer the same car model (with same warranties, etc.) at different prices, why should one want to consider the more expensive car?

In order to compute the Skyline set in a personalized fashion, the user needs only to provide so-called *ceteris paribus* ("all other being equal") preferences on each individual

attribute (e.g. “lower prices are better than higher prices given that all other attributes are equal”). Although, many works on skyline queries only consider numerical domains and preferences [1,3,4], skylining generally can also be extended to qualitative categorical preferences (e.g. on colors, “given two cars with free color choice, a black car would be better than a red car”) which are usually modeled as partial or weak orders [5,6]. Furthermore, many of these preferences don’t require any user input during elicitation as they can be deducted from common information in the collection of user profiles (e.g. preferences on price; no reasonable user would prefer the same object for a higher price).

The focus on individual attribute domains and the complete fairness of the Pareto paradigm are the major advantages of skyline queries: they are easy to specify and the algorithm will only remove definitely suboptimal objects. However, these characteristics also directly lead to the paradigms major shortcomings: Skyline queries completely lack the ability to relate attribute domains to each other and thus prevent compensation, weighting or ranking *between* attribute domains. This often results in most objects being incomparable to each other and thus generally causes Skyline sets to be rather large, especially in the quite common case of anti-correlated attribute dimensions. This effect is often referred to as “curse of dimensionality”. It has been shown (under certain assumptions on e.g. specific data distribution) that the skyline size grows roughly exponential with the number of query attributes [7,8]. However, there is still no reliable and accurate algorithm for predicting skyline sizes given arbitrary database instances and user preferences. Experimentally, it has been validated that already for only 5 to 10 attributes, skylines can easily contain 30% or more of the entire database instance [1,9,10] which is clearly unmanageable for most users and thus rendering the skyline paradigm inapplicable for many real-world problems.

Accordingly, reducing the size of result sets by choosing the most interesting or most relevant objects from the skyline is a major and prominent research problem. However, “interestingness” is usually an individual perception and is specific for each user and is thus hard to formalize. Nevertheless, for rendering the skyline paradigm useful for real world scenarios, such techniques are mandatorily required. Accordingly, in the recent years an impressive number of approaches have been developed introducing various heuristics for capturing the semantics of “interesting” in order to choose meaningful and manageable subsets from skylines in an efficient manner.

In this thesis, trade-off skylines are presented and developed. Trade-off skylines rely on a different base idea: instead of employing structural or statistical heuristics with

unclear semantics for capturing the concept of “interestingness”, trade-off skylines interactively elicit additional user feedback to cooperatively steer the selection of Skyline tuples from large result sets. This additional feedback is provided in the form of trade-offs which are especially designed to allow for a strictly qualitative compensation between individual attribute domains. Especially, this technique closely resembles the concept of natural compromises which are part of each person’s every day’s decision processes (e.g. “I am willing to pay more for better quality”). The semantics of trade-offs go well beyond the possibilities of strict Pareto semantics. While Skyline queries find the best objects from a database with respect to some attribute preferences represented by the Pareto efficiency frontier, trade-offs allow to focus within the wide selection of those “optimal” objects. Especially, qualitative compensation between multiple attribute dimensions is allowed. This empowers users to obtain a selection from the skyline set which is truly personalized and is not determined by some user oblivious heuristic.

2 Trade-Off Skylines

For motivating trade-off skylines, consider the following two database objects representing cars: let object A be a ‘blue metallic’ car for $18,000 and object B be a ‘blue’ car for $17,000, accompanied by a preference favoring cheaper cars and metallic colors. Looking at the ranking on attribute level, both cars are incomparable with respect to the Pareto order: one car is cheaper; the other car has the more preferred color. In this scenario, a natural question of a real-life car dealer would be, whether the customer is willing to compromise on those attributes, i.e. if she is willing to pay the additional $1,000 for a metallic paint job for that particular car (such a compromise is called a *trade-off*). If the answer is yes, then object A is the better choice for the user and should dominate object B with respect to a trade-off enhanced Pareto order. However, if some object C like a ‘blue’ car for $15,000 exists, A and C would still be incomparable as the premium for the metallic color on that car C is larger than the $1,000 the user is willing to pay. When providing a strong trade-off, many skyline objects can now be removed and thus the skyline is focused consistently with the refined trade-off enhanced user preferences. At the same time, the approach abstains from assuming arbitrary user agnostic heuristics for selecting objects from a too large skyline.

During the course of the research performed for this thesis, eleven topic-relevant papers have been published which make up this thesis. The following section summarizes

these papers and puts them into context with respect to the whole thesis. Mainly, these papers can be categorized into five topics; each will be discussed in its own section:

- Theoretical Foundations
- Consistency Checking
- Simplified Approaches
- Trade-Off Skyline Computation
- Alternative Approaches

2.1 Theoretical Foundations

In the early phases of the research related to this thesis, the theoretical foundations of trade-offs skylines had to be integrated into the already existing theory of Skyline computation.

Skyline queries can easily be described from an order-theoretical point of view. For actually formalizing a query, users provide so-called *base preferences* on attributes. Each base preference (also called attribute preference) encodes a user's likes and dislikes regarding the values of a given attributes. Base preferences are often provided as total or weak orders (as, e.g. introduced by [1], allowing for preferences like "the lower the price, the better" or "the faster the car, the better"). While this allowed for very efficient query evaluation, the preferences' expressiveness was rather limited [11]. [12] and [13] popularized the notion of user preferences as *strict partial orders* which can easily encode intuitive statements like "I like A better than B" (see figure 1). Usually, using partial orders increases the complexity of Skyline algorithms and even further aggravates the problem of

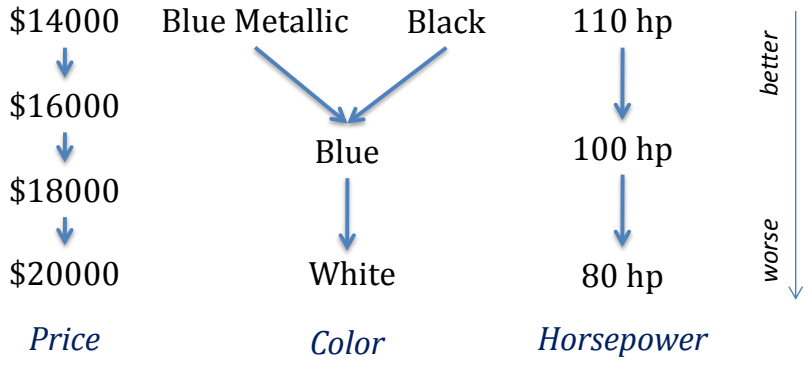

Fig. 1. Example partial order preferences for the domain of cars.

overly large result sets. Thus, trade-off skylines have been specifically adapted to also deal with the more expressive case of partial orders.

Given a set of base preferences, the Skyline set can now be derived using the Pareto semantics as known from the field of economy [13]: some object o_1 *dominates* an object o_2, if and only if o_1 is preferred over o_2 with respect to any attribute and o_1 is preferred over or is equivalent to o_2 with respect to all other attributes. The implied semantics are quite clear considering some examples, e.g. given two completely identical objects with one being slightly cheaper than the other (and also all other properties being equal, e.g. warranties, payment options, etc.), why should one consider choosing the more expensive object? This dominated object may thus safely be removed. On the other hand, given two nearly identical objects where the slightly more expensive one is also of slightly higher quality, no domination relationships between those two objects can be established.

Given the definition of Pareto domination, the so called *full product order* can be defined using the concept of Pareto *aggregation*. This product order materializes all domination relationships between all possible database objects. The Skyline set then can be defined as all those objects *existing* in the database which *are not dominated by any other existing objects* with respect to the product order (i.e. a database object o_1 is in the Skyline if there is no database object o_2 such that (o_2, o_1) is an element of the product order). Unfortunately, this order-theoretical definition of Skyline sets does not directly lead to practically viable algorithms as materializing the product order is prohibitively expensive. But usually, materialization of the product order can be avoided by Skyline algorithms anyway: this can be attributed to the *seperability characteristics* [13] of product orders directly resulting from a Pareto aggregation. This characteristic describes the fact that a given product order can be losslessly decomposed into its base preferences. As a consequence, Skyline algorithms can be build which only need to compare existing database objects attribute-wise without the need of the product order at all.

For establishing the theoretical foundations of trade-off skylines, the effects of trade-offs on the product order need to be examined and the resulting implications researched. This important ground work was laid in the early phases of this thesis by the papers [14] and [15], which for the first time, formalized the problem of trade-off skylines and showed their theoretical feasibility. [14] established the effects of trade-offs on skylines from a purely order-theoretical point of view and was published in April 2007 at the renowned Conference on Database Systems for Advanced Applications (DASFAA) in Bangkok,

Thailand. This conference can be considered to be the most significant database conferences in the Asian-Pacific area. Especially, [14] showed how trade-offs induce new domination relationships into the full product. It was shown that these new relationships are *local* and *transitive* within the product order. Also, it could be shown that skylines enhanced by trade-offs can be computed incrementally during the user interaction. The paper also introduced the concept of *preference compatibility* which shows that there are certain trade-offs which will lead to inconsistent or even infinite product orders. Preferences and trade-offs which are not compatible must thus be detected during computation of a trade-off skylines. One of the most significant results was that using trade-offs in Skylines may easily break the seperability characteristic of a product order, thus new computation algorithms are required. Especially, these algorithms will have to rely on at least partially materializing some parts of the product order. Thus, [14] can be considered as a major milestone for incorporating trade-offs into the established theories of Skyline computation.

> [14] W.-T. Balke, U. Güntzer, and C. Lofi, "Eliciting Matters - Controlling Skyline Sizes by Incremental Integration of User Preferences", *12th International Conference on Database Systems for Advanced Applications (DASFAA)*, Bangkok, Thailand: 2007.

In [15], the process of incrementally computing trade-off enhanced skylines is further examined and was published at the IEEE Conference on Research Challenges in Information Science (RCIS). This paper additionally received the "*Best Paper*" award. [15] formalizes the domination relationships which are induced by each iteration and provides a representation with Hasse diagrams. These Hasse diagrams can now be used for a locally re-computing the product order. Additionally, the paper suggests a representation scheme for trade-offs using the ontology language OWL and provided first drafts for a potential user interfaces. While this paper did provide theoretical results which could directly be implemented as a working algorithm, the performance of this algorithm is prohibitive as it still relied on the materializing the impractically large product order. However, the paper successfully supplemented the theoretical results of [14] and paved the way for optimized solutions which avoid materializing the full product order.

> [15] W.-T. Balke, C. Lofi, and U. Güntzer, "User Interaction Support for Incremental Refinement of Preference-Based Queries", *1st International IEEE Conference on Research Challenges in Information Science (RCIS)*, Ouarzazate, Morocco: 2007. *Best Paper Award*.

An additional journal paper unifying and expanding [14] and [15] was published, featuring a clearer and more compact notation for theorems and proofs and also

containing further elaborations on the iterative computation of trade-off skylines. The resulting paper [16] was also published in 2007 in the International Journal of Computer Science & Applications (IJCSA).

> [16] W.-T. Balke, C. Lofi, and U. Güntzer, "Incremental Trade-Off Management for Preference Based Queries", *International Journal of Computer Science & Applications (IJCSA)*, vol. 4, 2007, pp. 75-91.

2.2 Consistency Checks for Trade-Offs

As already discovered in [14], trade-offs need to be *compatible* with the provided user preferences, i.e. there may be *inconsistent* trade-offs which will lead to preference cycles in the product order (i.e. a given database object would then be preferred to itself). Unfortunately, specifying inconsistent trade-offs is a mistake which may happen to users easily. This can be explained by users providing preference or trade-off statements in a local and isolated fashion, just considering a small excerpt of the full preference space. Additional psychological effects will then often result in contradicting information [13] when aggregating the elicited information using the strict and transitive semantics which usually underlie preference-based information systems. Thus, natural user preferences and trade-offs are often unintentionally inconsistent.

Checking user provided trade-offs for consistency is a mandatory step in the process of computing trade-off skylines. However, this problem is far from trivial. Inconsistencies result in cycles within the full product order. However, for rendering trade-off skylines applicable in real world scenarios, any algorithms involved must abstain from materializing the full product order. Unfortunately, [14] showed that by introducing trade-offs, the product order loses its *seperability* characteristics (e.g. the product order cannot be losslessly deconstructed into its respective base preferences and trade-offs). Seperability allows for designing efficient algorithms which avoid materializing the product order altogether. Thus, at least some parts of the full trade-off order must be considered when checking trade-off for consistency. In [17], a technique was introduced to detect cycles (and thus inconsistencies) by materializing just small and generic representatives of relevant parts of the full product order. The paper was published at the IEEE RCIS conference in 2008. The presented technique relied on generating a tree-shaped data structure representing *trade-off chains* (i.e. all possible generic combinations of the specified user trade-offs). Using a simple criterion, inconsistencies could easily be detected without the need to consider the actual database at all.

[17] C. Lofi, W.-T. Balke, and U. Güntzer, "Efficiently Performing Consistency Checks for Multi-Dimensional Preference Trade-Offs", *2nd International IEEE Conference on Research Challenges in Information Science (RCIS)*, Marakech, Morocco: 2008. *Best Paper Award Candidate.*

Unfortunately, while the technique presented in [17] showed good performance in the average case, the data structure could grow extremely large (or even prohibitively large) in some rare cases. In [18], an enhanced journal version of this technique was presented which *pruned* redundant branches from the underlying tree data structure. Also, a more efficient criterion for actually detecting an inconsistency in this data structure was presented. Those two enhancements significantly increased the worst case performance as well as average case performance (e.g., in realistically-sized simulations, it could be shown that the response time for performing a consistency check of a set of 20 trade-offs stayed below 1 second in nearly all cases with a median of 3 milliseconds).

[18] C. Lofi, W.-T. Balke, and U. Güntzer, "Consistency Check Algorithms for Multi-Dimensional Preference Trade-Offs", *International Journal of Computer Science & Applications (IJCSA)*, vol. 5, 2008, pp. 165-185.

2.3 Simplified Approaches

As mentioned in the previous section, the main reason for the need of an inefficient materialization of the product order is the loss of the order's *separability*. To design better performing algorithms, a fundamental technique for computing skylines is to rely as much as possible on basic component-wise attribute comparisons and avoiding the materialization of the object whenever possible.

In [19], an approach is presented which restricts the semantics of allowed trade-offs such that only very small parts of the product order need to be materialized while heavily relying on simple object comparisons similar to common standard Skyline algorithms. The underlying rationale is as follows: most problems encountered when computing trade-off enhanced skylines arise from *trade-off chains*, i.e. domination relationships which are induced by not one trade-off alone, but which are the results of the transitive closure of multiple trade-off induced domination relationships and ordinary Pareto domination relationships. Thus, a possibility for simplification of the trade-off computation problem is to restrict the complexity of the allowed trade-off chains. A good heuristic which works well with many real-world scenarios is to allow only trade-offs on pairs of two antagonistic attributes each (e.g. power and fuel efficiency for cars, or display size and weight for laptop computers). Furthermore, those attribute pairs must be disjoint. If both conditions

are fulfilled, all resulting trade-off chains are of a very simple nature and thus allow for algorithms showing high performance due to the possibility of primarily relying on attribute comparisons which can easily be implemented using SQL. This approach has shown a very good practical performance with respect to response times which are well below one second for most scenarios. Furthermore, the paper introduced a preference elicitation heuristics proposing trade-offs to the user who, in turn, may accept or dismiss the suggestions. The rationale of this heuristic is that a major cause for unmanageable large skyline result sets is object incomparability resulting from anti-correlated attributes. Accordingly, this heuristic analyzes the correlation and clustering properties of the database objects to suggest trade-offs which will minimize the incomparability between strongly anti-correlated attribute clusters.

> [19] C. Lofi, W.-T. Balke, and U. Güntzer, "Efficient Skyline Refinement Using Trade-Offs", *3rd International IEEE Conference on Research Challenges in Information Science (RCIS)*, Fès, Morocco: 2009. *Best Paper Award Candidate*.

An extended version of [19] was published in the IJCSA journal [20], additionally extending the approach by "don't care" attributes which allowed the user to be indifferent with respect to certain attribute values.

> [20] C. Lofi, W.-T. Balke, and U. Güntzer, "Efficient Skyline Refinement Using Trade-Offs Respecting Don't-Care Attributes", *International Journal of Computer Science and Applications (IJCSA)*, vol. 6, 2009, pp. 1-29.

2.4 The Complete Trade-Off Skyline Lifecycle

The simplified approach in [19] was well suited to implement a simple trade-off skyline system. However, the semantic restrictions are quite radical as only trade-offs between pairwise disjoint attribute are allowed. This drawback was finally remedied by [21], providing an efficient algorithm for computing trade-off skylines without any restrictions. Furthermore, this paper could be published at the prestigious Conference on Extending Database Technology (EDBT), held in Lausanne, Switzerland. This conference is among the Top-5 database conferences overall. The presented approach heavily modified and extended the tree structure and pruning technique presented in [18]. As a result, the data structure cannot only be used for checking trade-off consistency, but also to test for object dominance respecting any trade-off chain. Thus, this paper presented the first complete, unrestricted, and practically viable algorithm for computing trade-off skylines.

[21] C. Lofi, U. Güntzer, and W.-T. Balke, "Efficient Computation of Trade-Off Skylines," *13th International Conference on Extending Database Technology (EDBT)*, Lausanne, Switzerland: 2010.

Finally, the complete concept of trade-off skylines has been wrapped up and published in [22] at the annual workshop of the German database community. This paper focuses on briefly summarizing all preceding efforts and papers in developing the full trade-off skyline lifecycle.

[22] C. Lofi and W.-T. Balke, "Preference Trade-Offs – Towards Manageable Skylines," *22. GI-Workshop Grundlagen von Datenbanken (GvD)*, Bad Helmstedt, Germany: 2010.

A full summarization of this thesis with a focus on alternative approaches to the overly large skyline result set problem will be published and discussed at the PhD workshop of the International Conference On Data Engineering (ICDE). This conference is one of the three most important conferences on databases and information systems overall [22].

[23] C. Lofi, "Choosing the Right Thing: Cooperative Trade-Off Enhanced Skyline Queries," *PhD Workshop at the 28th International Conference On Data Engineering (ICDE)*, Hannover, Germany: 2011.

The papers presented in the previous sections represent the core efforts in establishing trade-off skylines, mainly focusing on theoretical or algorithmic aspects of the problem. Two additional works are currently under review, one presenting an extensive overview on alternative approaches to deal with the problem of large skyline results, and one paper dealing with additional user-interface and elicitation issues.

3 Alternative Approaches

Besides developing theoretical and algorithmic foundations of trade-off skylines, some alternative approaches have been explored during this thesis. Especially in mobile environments (e.g. the ever-growing smartphone market), approaches based on presenting large result sets or eliciting extensive user preferences are hard to use due to available screen sizes and limited interface capabilities. To tailor for the specific challenges of mobile devices, an approach [24] based on Bayesian retrieval techniques has been developed and published at the IEEE Conference on Commerce and Enterprise Computing (CEC). This paper also received the "Best Paper Award".

[24] C. Lofi, C. Nieke, and W.-T. Balke, “Mobile Product Browsing Using Bayesian Retrieval,” *12th IEEE Conference on Commerce and Enterprise Computing (CEC)*, Shanghai, China: 2010. *Best Paper Award.*

4 Related Work

As mentioned in the introduction section, the problem of unusable large skyline result setsis a major obstacle for the real-life success of the promising Skyline paradigm. Therefore, this problem has been addressed by numerous previous works. These works can roughly be classified into three groups and are briefly discussed in this section: approaches relaxing the Pareto semantics, summarizing approaches, and approaches which rely on statistical or structural properties to explicitly rank skyline objects.

Relaxation of Pareto Semantics: Considering the definition of Pareto semantics, it is obvious that the manageability problems of skylines are heavily aggravated by incomparable attribute values, especially when working with natural preferences which are modeled as partial orders [7,8]. As soon as two database items are incomparable with respect to even a single attribute, both objects are incomparable and may end up in the skyline. One could say that the Pareto semantics generally is ‘too fair’. Unfortunately, anti-correlated attributes are very common in real life scenarios (e.g. quality vs. price). Accordingly, the first group of approaches uses weaker variants of the Pareto semantics which less likely lead to incomparability between database objects. Notable works in this spirit are weak skylines [9] which replace the Pareto definition for domination with: “one object is better than another one when it is better with respect to one attribute and not worse with respect to any other”, and k-dominant skylines [10] which require only a user-given number of k attributes to fulfill the Pareto condition. Skylines resulting from relaxed Pareto definitions can be of significant smaller size. However, their semantics are often hard to justify and the implied heuristics have a strong “ad-hoc” character. For example, such skylines may easily remove objects which are highly interesting to the user, and thus rendering the practical application semantically difficult.

Summarization approaches aim at finding a subset of objects which serves optimally as a summarization of the full skyline. The summarizing set should still maintain the full diversity and characteristics of the original skyline, but should be of a more manageable size. The focus of these approaches is to enable the user to grasp a quick overview of the nature and contents of the skyline result set such that he is easily able to further refine his

preferences and / or is directly able to perform subsequent queries for narrowing down the results even further (e.g. appending a top-k query which ranks the skyline result, or provide some SQL constraints to remove unwanted data points). Notable approaches are approximately dominating representatives [11] which return a subset minimally covering all skyline objects with some ϵ-balls and statistical sampling approaches [6] with subsequent top-k ranking. Both approaches try to maintain the diversity of the original skyline. However, summarized skylines are only useful if they are intended to provide a quick overview and should be accompanied by additional succeeding queries which focus on the most interesting object from a user's perspective.

Weighting approaches try to induce a ranking on skyline items based on some structural properties of the data set. The Pareto skyline operator treats all skyline objects as being equal, i.e. it does not impose any ranking on the result set. However, weighting approaches claim that there are more interesting and less interesting skyline objects, and that "interestingness" can be captured by properties like e.g. the data distribution, the structure of the subspace skylines, or other statistical means. Usually, they explicitly quantify the "interestingness" of a skyline object numerically and return the k-most interesting objects.

Especially subspace analysis [12] has gained a lot of attention which was encouraged by the development of efficient algorithms for materializing the possibly $2^d - 1$ subspace skylines (see e.g. SkyCubes [13]). For example, subspace analysis can be used to define top-k frequent skylines [14] which capture "interestingness" counting the occurrence of an object in each of the non-empty subspace skylines, i.e. claiming that objects which are more frequent in subspaces are also more interesting. A more elaborate subspace based ranking is provided by SkyRank [15], which uses subspace domination relationships of the full space skyline objects to construct a so called skyline graph which is used for a subsequent link-analysis which provides the interestingness scores with a variant of PageRank.

Other approaches use the number of dominated object as a metric for interestingness, resulting in the k Most Representative Skyline [16], or elicit additional preferences expressing a precedence of the query attributes for constructing a ranked result set based on the subspace frequency of objects and the precedence of the attributes defining the subspace (e.g. Telescope [17]).

However, these presented approaches break the absolute fairness of Pareto semantics and replace it with some heuristics for removing “unwanted” objects. While each of those approaches has benefits and advantages on their own right, the imposed heuristics all rely on some “ad-hoc” assumptions on what makes a skyline point more interesting than others. However, the “correctness” and usefulness of these assumptions with respect to the real information needs of a given, individual user is very subjective and thus hard to determine. Therefore, trade-off skylines can be considered to be first approach which is completely focused on the user and thus delivers a truly personalized solution for addressing the curse of dimensionality.

5 Summary and Outlook

During this thesis, the complete theoretical and algorithmic framework for successfully integrating the natural concept of trade-offs into the Skyline paradigm has been established. In the early stages, the required theoretical foundations have been developed and explored [14-16]. As a result, two major challenges could be identified which have been mastered by later works:

a) The challenge of detecting *inconsistent trade-offs* which has been extensively discussed in [17,18].

b) The challenge of actually computing the trade-off skyline. This challenge is further aggravated by the fact that trade-offs will break the seperability of the underlying product order. Thus, novel skylines algorithms are necessary which minimally materialize at least some parts of the product order. A simplified approach additionally restricting the semantics of trade-offs was presented in [19,20]. Finally, an unrestricted and complete algorithm was presented in [22].

Beside the core publications establishing theoretical or algorithmic foundations, two summarizing papers have been published [21,23] as well as a paper presenting an alternative approach for mobile environments [24]. All in all, this thesis incorporates six papers published in proceeding of international conferences (two received a “Best Paper Award”), three papers published in international journals, and two papers published and discussed in proceedings of one national and one international workshop. Furthermore, it can be observed that these papers did have a notable impact on the research community (e.g. according to Google Scholar [14] is cited 17 times, and [16] is cited 11 times).

6 Bibliography

[1] S. Börzsönyi, D. Kossmann, and K. Stocker, "The Skyline Operator," *Int. Conf. on Data Engineering (ICDE)*, Heidelberg, Germany: 2001.

[2] R. Fagin, A. Lotem, and M. Naor, "Optimal aggregation algorithms for middleware," *Symposium on Principles of Database Systems (PODS)*, Santa-Barbara, California, USA: 2001.

[3] D. Kossmann, F. Ramsak, and S. Rost, "Shooting stars in the sky: an online algorithm for skyline queries," *Int. Conf. on Very Large Data Bases (VLDB)*, Hongkong, China: 2002.

[4] D. Papadias, Y. Tao, G. Fu, and B. Seeger, "An optimal and progressive algorithm for skyline queries," *International Conference on Management of Data (SIGMOD)*, San Diego, USA: 2003.

[5] M. Lacroix and P. Lavency, "Preferences: Putting More Knowledge into Queries," *Int. Conf. on Very Large Data Bases (VLDB)*, Brighton, UK: 1987.

[6] C.-Y. Chan, P.-K. Eng, and K.-L. Tan, "Stratified computation of skylines with partially-ordered domains," *International Conference on Management of Data (SIGMOD)*, Baltimore, USA: 2005.

[7] 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," *Journal of the ACM (JACM)*, vol. 25, 1978.

[8] S. Chaudhuri, N. Dalvi, and R. Kaushik, "Robust Cardinality and Cost Estimation for Skyline Operator," *22nd Int. Conf. on Data Engineering (ICDE)*, Atlanta, Georgia, USA: 2006.

[9] P. Godfrey, "Skyline cardinality for relational processing. How many vectors are maximal?," *Symp. on Foundations of Information and Knowledge Systems (FoIKS)*, Vienna, Austria: 2004.

[10] W.-T. Balke, J.X. Zheng, and U. Güntzer, "Approaching the Efficient Frontier: Cooperative Database Retrieval Using High-Dimensional Skylines," *Int. Conf. on Database Systems for Advanced Applications (DASFAA)*, Beijing, China: 2005.

[11] P. Fishburn, “Preference structures and their numerical representations,” *Theoretical Computer Science*, vol. 217, Apr. 1999, pp. 359-383.

[12] W. Kießling, “Foundations of preferences in database systems,” *28th int. conf. on Very Large Data Bases (VLDB)*, Hong Kong, China: 2002.

[13] Sven Ove Hansson, “Preference Logic,” *Handbook of Philosophical Logic*, vol. 4, 2002, pp. 319-393.

7 Bibliography of Papers Published as Part of this Thesis

[14] W.-T. Balke, U. Güntzer, and C. Lofi, "Eliciting Matters - Controlling Skyline Sizes by Incremental Integration of User Preferences," *12th International Conference on Database Systems for Advanced Applications (DASFAA)*, Bangkok, Thailand: 2007.

[15] W.-T. Balke, C. Lofi, and U. Güntzer, "User Interaction Support for Incremental Refinement of Preference-Based Queries," *1st International IEEE Conference on Research Challenges in Information Science (RCIS)*, Ouarzazate, Morocco: 2007.

[16] W.-T. Balke, C. Lofi, and U. Güntzer, "Incremental Trade-Off Management for Preference Based Queries," *International Journal of Computer Science & Applications (IJCSA)*, vol. 4, 2007, pp. 75-91.

[17] C. Lofi, W.-T. Balke, and U. Güntzer, "Efficiently Performing Consistency Checks for Multi-Dimensional Preference Trade-Offs," *2nd International IEEE Conference on Research Challenges in Information Science (RCIS)*, Marakech, Morocco: 2008.

[18] C. Lofi, W.-T. Balke, and U. Güntzer, "Consistency Check Algorithms for Multi-Dimensional Preference Trade-Offs," *International Journal of Computer Science & Applications (IJCSA)*, vol. 5, 2008, pp. 165-185.

[19] C. Lofi, W.-T. Balke, and U. Güntzer, "Efficient Skyline Refinement Using Trade-Offs," *3rd International IEEE Conference on Research Challenges in Information Science (RCIS)*, Fès, Morocco: 2009.

[20] C. Lofi, W.-T. Balke, and U. Güntzer, "Efficient Skyline Refinement Using Trade-Offs Respecting Don't-Care Attributes," *International Journal of Computer Science and Applications (IJCSA)*, vol. 6, 2009, pp. 1-29.

[21] C. Lofi, U. Güntzer, and W.-T. Balke, "Efficient Computation of Trade-Off Skylines," *13th International Conference on Extending Database Technology (EDBT)*, Lausanne, Switzerland: 2010.

[22] C. Lofi and W.-T. Balke, "Preference Trade-Offs – Towards Manageable Skylines," *22. GI-Workshop Grundlagen von Datenbanken (GvD)*, Bad Helmstedt, Germany: 2010.

[23] C. Lofi, “Choosing the Right Thing: Cooperative Trade-Off Enhanced Skyline Queries,” *PhD Workshop at the 28th International Conference On Data Engineering (ICDE)*, Hannover, Germany: 2011.

[24] C. Lofi, C. Nieke, and W.-T. Balke, “Mobile Product Browsing Using Bayesian Retrieval,” *12th IEEE Conference on Commerce and Enterprise Computing (CEC)*, Shanghai, China: 2010.

8 Appendix

In the following appendix, all publications being part of this cumulative doctoral theses have been reprinted in the following order:

C. Lofi, U. Güntzer, and W.-T. Balke, "Efficient Computation of Trade-Off Skylines," *13th International Conference on Extending Database Technology (EDBT)*, Lausanne, Switzerland: 2010.

C. Lofi, C. Nieke, and W.-T. Balke, "Mobile Product Browsing Using Bayesian Retrieval," *12th IEEE Conference on Commerce and Enterprise Computing (CEC)*, Shanghai, China: 2010.

C. Lofi, "Choosing the Right Thing: Cooperative Trade-Off Enhanced Skyline Queries," *PhD Workshop at the 28th International Conference On Data Engineering (ICDE)*, Hannover, Germany: 2011.

C. Lofi and W.-T. Balke, "Preference Trade-Offs – Towards Manageable Skylines," *22. GI-Workshop Grundlagen von Datenbanken (GvD)*, Bad Helmstedt, Germany: 2010.

C. Lofi, W.-T. Balke, and U. Güntzer, "Efficient Skyline Refinement Using Trade-Offs Respecting Don't-Care Attributes," *International Journal of Computer Science and Applications (IJCSA)*, vol. 6, 2009, pp. 1-29.

C. Lofi, W.-T. Balke, and U. Güntzer, "Efficient Skyline Refinement Using Trade-Offs," *3rd International IEEE Conference on Research Challenges in Information Science (RCIS)*, Fès, Morocco: 2009.

C. Lofi, W.-T. Balke, and U. Güntzer, "Consistency Check Algorithms for Multi-Dimensional Preference Trade-Offs," *International Journal of Computer Science & Applications (IJCSA)*, vol. 5, 2008, pp. 165-185.

C. Lofi, W.-T. Balke, and U. Güntzer, "Efficiently Performing Consistency Checks for Multi-Dimensional Preference Trade-Offs," *2nd International IEEE Conference on Research Challenges in Information Science (RCIS)*, Marakech, Morocco: 2008.

W.-T. Balke, C. Lofi, and U. Güntzer, "Incremental Trade-Off Management for Preference Based Queries," *International Journal of Computer Science & Applications (IJCSA)*, vol. 4, 2007, pp. 75-91.

W.-T. Balke, C. Lofi, and U. Güntzer, "User Interaction Support for Incremental Refinement of Preference-Based Queries," *1st International IEEE Conference on Research Challenges in Information Science (RCIS)*, Ouarzazate, Morocco: 2007.

W.-T. Balke, U. Güntzer, and C. Lofi, "Eliciting Matters - Controlling Skyline Sizes by Incremental Integration of User Preferences," *12th International Conference on Database Systems for Advanced Applications (DASFAA)*, Bangkok, Thailand: 2007.