Algebraic Optimization of Relational Queries with Various Kinds of Preferences *

Radim Nedbal

Institute of Computer Science, Academy of Sciences of the Czech Republic
Pod Vodárenskou věží 2, 182 07 Prague 8, Czech Republic
radned@seznam.cz

Abstract. Preferences can be used for information filtering and extraction to deliver the most relevant data to the user. Therefore the efficient integration of querying with preferences into standard database technology is an important issue. The paper resumes a logical framework for formulating preferences and their embedding into relational algebra through a single *preference operator* parameterized by a set of user preferences of sixteen various kinds and returning only the most preferred subsets of its argument relation. Most importantly, preferences between sets of elements can be expressed. To make a relational query language with the preference operator useful for practical applications, formal foundation for algebraic optimization, applying heuristics like *push preference*, has to be provided. Therefore abstract properties of the preference operator and a variety of algebraic laws describing its interaction with other relational algebra operators are presented.

Key words: logic of preference, relational query, optimization.

1 Introduction

If users have requirements that are to be satisfied completely, their database queries are characterized by *hard constraints*, delivering exactly the required objects if they exist and otherwise empty result. This is how traditional database query languages treat all the requirements on the data. However, requirements can be understood also in the sense of wishes: in case they are not satisfied, database users are usually prepared to accept worse alternatives and their database query is characterized by *soft constraints*. Requirements of the latter type are called preferences.

Preferences are ubiquitous in our daily lives, which suggests that database query languages should support both views of requirements, characterized by

* This work was supported by the project 1ET100300419 of the Program Information Society (of the Thematic Program II of the National Research Program of the Czech Republic) "Intelligent Models, Algorithms, Methods and Tools for the Semantic Web Realization", and by the Institutional Research Plan AV0Z10300504 "Computer Science for the Information Society: Models, Algorithms, Applications".

hard or soft constraints. The research on preferences is extensive and encompasses preference logic, preference reasoning, non-monotonic reasoning, and, recently, preferences also attracted attention in database community (see Sect. 5).

Building on a logical framework for formulating preferences and on their embedding into relational algebra (RA) through a single *preference operator* (PO) to combat the empty result and the flooding effects, this paper presents an approach to algebraic optimization of relational queries with various kinds of preferences. The PO selects from its argument relation the *best-matching alternatives* with regard to user preferences, but *nothing worse*.[1] Preferences are specified using a propositional logic notation and their semantics is related to that of a disjunctive logic program. The language for expressing preferences i) is declarative, ii) includes various kinds of preferences, iii) is rich enough to express preferences between sets of elements, iv) and has an intuitive, well defined semantics allowing for conflicting preferences.

In Sect. 2, the above mentioned framework for formulating preferences and in Sect. 3 an approach to their embedding into RA are revisited. Presenting a variety of algebraic laws that describe interaction with other RA operators to provide a formal foundation for algebraic optimization, Sect. 4 provides the main contribution of this paper.[2] A brief overview of related work in Sect. 5 and conclusions in Sect. 6 end this paper. All the nontrivial proofs are given.

To improve the readability, $\succeq (x, y) \wedge \neg \succeq (y, x)$ and $\succeq (x, y) \wedge \succeq (y, x)$ is substituted by $\succ (x, y)$ and $= (x, y)$, respectively.

2 User Preferences

A user preference is expressed by a preference statement, e.g. "a is preferred to b", or symbolically by an appropriate preference formula (PF). PF's comprise a simple declarative language for expressing preferences. To capture its declarative aspects, model-theoretic semantics is defined: considering a set of states of affaires S and a set $W = 2^S$ of all its subsets – worlds, if $\mathcal{M} = \langle W, \succeq \rangle$ is an order $\succeq$ on W such that $w \succeq w'$ holds for some worlds w, w' from W, then $\mathcal{M}$ is termed a *preference model* (PM) of $w > w'$ – a preference of the world w over the world w', which we express symbolically as $\mathcal{M} \models w > w'$.

The basic differentiation between preferences is based on notions of optimism and pessimism. Defining a-world as a world in which a occurs, if we are optimistic about a and pessimistic about b for example, we expect some a-world to precede at least one b-world in each PM of a preference statement "a is preferred to b". This kind of preference is called *opportunistic*. By contrast, if we are pessimistic about a and optimistic about b, we expect every a-world to precede each b-world in each PM of a preference statement "a is preferred to b". This kind of preference is called *careful*. Alternatively, we might be optimistic or pessimistic about both a and b. Then we expect some a world to precede each b-world or each a-world

[1] A similar concept was proposed by Kießling et al. [1, 2] and Chomicki et al. [3] and, in a more restricted form, by Börzsönyi et al. [4] (for more detail refer to Sect. 5).

[2] The presented results correspond to those of [3].

to precede some b-world in each PM of a preference statement "a is preferred to b". This kind of preference is called *locally optimistic* or *locally pessimistic*, respectively. Locally optimistic, locally pessimistic, opportunistic and careful preferences are symbolically expressed by PF's of the form: $a\ {}^{M}{>}^{M}\ b$, $a\ {}^{m}{>}^{m}\ b$, $a\ {}^{M}{>}^{m}\ b$, and $a\ {}^{m}{>}^{M}\ b$, respectively.

Also, we distinguish between strict and non-strict preferences. For example, if w precedes w' strictly in a PM, then we strictly prefer w to w'.

In addition, we distinguish between preferences with and without *ceteris paribus* proviso – a notion introduced by von Wright [5] and generalized by Doyle and Wellman [6] by means of contextual equivalence relation – an equivalence relation on W.[3] For example, a PM of a preference statement "a is *carefully* preferred to b *ceteris paribus*" is such an order on W that a-worlds precede b-worlds in the same contextual equivalence class. Specifically, the preference statement "I prefer playing tenis to playing golf ceteris paribus" might express by means of an contextual equivalence that I prefer playing tenis to playing golf only if the context of weather is the same, i.e., it is not true that I prefer playing tenis in strong winds to playing golf during a sunny day.

Next, we revisit the basic definitions introducing syntax and model-theoretic semantics of the language for expressing user preferences:

Definition 1 (Language). Propositional formulas *are defined inductively:*

Given a finite set of propositional variables $p, q, \ldots$ i) every propositional variable is a propositional formula; ii) if φ, ψ are propositional formulas then so are $\varphi \wedge \psi$ and $\neg\varphi$.

PF's are expressions $\varphi\ {}^{x}{>}^{y}\ \psi$ and $\varphi\ {}^{x}{\geq}^{y}\ \psi$ for $x, y \in \{m, M\}$, where φ, ψ are propositional variables.

If we identify propositional variables with tuples over a relation schema R, then we get *PF's over* R. A relation instance $I(R)$, i.e., a set of tuples over R, creates a *world* w, an element of a set W.

The PM is defined so that any set of (possibly conflicting) preferences is consistent: the partial pre-order, i.e., a binary relation which is reflexive and transitive, in the definition of the PM, enables to express some kind of conflict by incomparability:

Definition 2 (Preference model). *A PM $\mathcal{M} = \langle W, \succeq \rangle$ over a relation schema R is a couple in which W is a set of worlds, relation instances of R, and $\succeq$ is a partial pre-order over W, the* preference relation *over R.*

A set of user preferences of various kinds can by represented symbolically by a *preference specification* (PS), which corresponds to an appropriate complex PF in the above defined language.

[3] As it has been shown [7] that any preference with contextual equivalence specification can be expressed by a set of preferences without contextual specification, we can restrict ourselves only to preferences without ceteris paribus proviso.

Definition 3 (Preference specification). *Let R be a relation schema and $\mathcal{P}_\rhd$ a set of PF's over R of the form $\{\varphi_i \rhd \psi_i : i = 1, \ldots, n\}$. A PS $\mathcal{P}$ over R is a tuple $\langle \mathcal{P}_\rhd | \rhd \in \{ \ ^x\!>^y, \ ^x\!\geq^y \ | x, y \in \{m, M\}\}\rangle$, and $\mathcal{M}$ is its model, i.e., a* PS model*, iff it models all elements $\mathcal{P}_\rhd$ of the tuple. Interpreting*

$$\mathcal{M} \models \mathcal{P}_\rhd \iff \forall(\varphi_i \rhd \psi_i) \in \mathcal{P}_\rhd : \mathcal{M} \models \varphi_i \rhd \psi_i \ .$$

3 Preference Operator

To embed preferences into RQL, the PO $\omega_\mathcal{P}$ returning only the best sets of tuples in the sense of user preferences $\mathcal{P}$ is defined:

Definition 4 (Preference operator). *The PO ω is a mapping from a powerset into itself. Specifically, if R is a relation schema, $\mathcal{P}$ a PS over R, and $\mathscr{M}$ the set of its models; then the PO $\omega_\mathcal{P}$ is defined for all sets $\{I_1(R), \ldots, I_n(R)\}$ of instances of R as follows:*

$$\begin{aligned}&\omega_\mathcal{P}(\{I_1(R), \ldots, I_n(R)\}) = \big\{w \in \{I_1(R), \ldots, I_n(R)\}\big| \\ &\exists \mathcal{M}_k = \langle W, \succeq_k\rangle \in \mathscr{M}, \forall w' \in \{I_1(R), \ldots, I_n(R)\} : \succeq_k (w', w) \Rightarrow \succeq_k (w, w')\big\} \ .\end{aligned}$$

Remark 1 (Preference operator notation). For brevity, when writing the argument without braces, e.g., $\omega_\mathcal{P}(I(R))$, then the unabbreviated notation is $\omega_\mathcal{P}(\{2^{I(R)}\})$, showing that the argument is the powerset of $I(R)$.

3.1 Basic Properties.

The following propositions are essential for investigation of algebraic properties describing interaction of the PO with other RA operations:

Proposition 1. *Given a relation schema R and a PS $\mathcal{P}$ over R, for all instances $I(R)$ of R the following properties hold:*

$$\begin{aligned}\omega_\mathcal{P}(I(R)) &\subseteq 2^{I(R)} \ , \\ \omega_\mathcal{P}\left(\{\omega_\mathcal{P}(I(R))\}\right) &= \omega_\mathcal{P}(I(R)) \ , \\ \omega_{\mathcal{P}_{empty}}(I(R)) &= 2^{I(R)} \ ,\end{aligned}$$

where $\mathcal{P}_{empty}$ is the empty PS, i.e., containing no preference.

The PO is not *monotone* or *antimonotone* with respect to its relation argument. However, partial antimonotonicity holds:

Proposition 2 (Partial antimonotonicity). *Given a relation schema R and a PS $\mathcal{P}$ over R, for all instances $I(R), I'(R)$ of R the following property holds:*

$$I(R) \subseteq I'(R) \Rightarrow 2^{I(R)} \cap \omega_\mathcal{P}(I'(R)) \subseteq \omega_\mathcal{P}(I(R)) \ .$$

Proof. Assume $w \in 2^{I(R)} \cap \omega_{\mathcal{P}}(I'(R))$. It follows that $w \subseteq I(R)$ and from the definition (Def. 4) of the PO $w \subseteq I'(R) \wedge \exists \mathcal{M}_k \in \mathscr{M}$ s.t. $\forall w' \in W : w' \subseteq I'(R) \wedge \succeq_k (w', w) \Rightarrow \succeq_k (w, w')$. As $I(R) \subseteq I'(R)$, we can conclude that $\exists \mathcal{M}_k \in \mathscr{M}$ s.t. $\forall w' \in W : w' \subseteq I(R) \wedge \succeq_k (w', w) \Rightarrow \succeq_k (w, w')$, which together with $w \subseteq I(R)$ implies $w \in \omega_{\mathcal{P}}(I(R))$. □

The following theorems enable to reduce cardinality of an argument relation of the PO without changing the return value and ensure that the empty query result effect is successfully eliminated:

Theorem 1 (Reduction). *Given a relation schema R, a PS $\mathcal{P}$ over R, for all instances $I(R), I'(R)$ of R the following property holds:*

$$I(R) \subseteq I'(R) \wedge \omega_{\mathcal{P}}(I'(R)) \subseteq 2^{I(R)} \Rightarrow \omega_{\mathcal{P}}(I(R)) = \omega_{\mathcal{P}}(I'(R)) \; .$$

Proof. $\subseteq$: Assume $w \in \omega_{\mathcal{P}}(I(R))$. Then, it follows from the definition of the PO $w \subseteq I(R) \wedge \exists \mathcal{M}_k \in \mathscr{M}$ s.t. $\forall w' \subseteq I(R) : \succeq_k (w', w) \Rightarrow \succeq_k (w, w')$. The assumption $\omega_{\mathcal{P}}(I'(R)) \subseteq 2^{I(R)}$ implies $\forall w' \in 2^{I'(R)} - 2^{I(R)} : \neg \succeq_k (w', w)$, and we can conclude $\forall w' \subseteq I'(R) : \succeq_k (w', w) \Rightarrow \succeq_k (w, w')$, which together with the assumption $I(R) \subseteq I'(R)$ implies $w \subseteq I'(R) \wedge \exists \mathcal{M}_k \in \mathscr{M}$ s.t. $\forall w' \subseteq I'(R) : \succeq_k (w', w) \Rightarrow \succeq_k (w, w')$, the definition of $w \in \omega_{\mathcal{P}}(I'(R))$.
$\supseteq$: Immediately follows from Prop. 2. □

Theorem 2 (Non-emptiness). *Given a relation schema R, a PS $\mathcal{P}$ over R, then for every finite, nonempty instance $I(R)$ of R, $\omega_{\mathcal{P}}(I(R))$ is nonempty.*

3.2 Multidimensional Composition.

In multidimensional composition, we have a number of PS defined over several relation schemas, and we define PS over the Cartesian product of those relations: the most common ways are Pareto and lexicographic composition.

Definition 5 (Pareto composition). *Given two relation schemas R_1 and R_2, PS's $\mathcal{P}_1$ over R_1 and $\mathcal{P}_2$ over R_2, and their sets of models $\mathscr{M}_1$ and $\mathscr{M}_2$, the Pareto composition $P(\mathcal{P}_1, \mathcal{P}_2)$ of $\mathcal{P}_1$ and $\mathcal{P}_2$ is a PS $\mathcal{P}_0$ over the Cartesian product $R_1 \times R_2$, whose set of models $\mathscr{M}_0$ is defined as:*

$$\forall \mathcal{M}_m = \langle W_1 \times W_2, \succeq_m \rangle \in \mathscr{M}_0, \\ \exists \mathcal{M}_k = \langle W_1, \succeq_k \rangle \in \mathscr{M}_1, \exists \mathcal{M}_l = \langle W_2, \succeq_l \rangle \in \mathscr{M}_2 \text{ s.t.}$$

$$\forall w_1, w_1' \in W_1, \forall w_2, w_2' \in W_2 : \\ \succeq_m (w_1 \times w_2, w_1' \times w_2') \equiv \succeq_k (w_1, w_1') \wedge \succeq_l (w_2, w_2') \; .$$

Definition 6 (Lexicographic composition). *Given two relation schemas R_1 and R_2, PS's $\mathcal{P}_1$ over R_1 and $\mathcal{P}_2$ over R_2, and their sets of models $\mathscr{M}_1$ and*

$\mathscr{M}_2$, the lexicographic composition $L(\mathcal{P}_1, \mathcal{P}_2)$ of $\mathcal{P}_1$ and $\mathcal{P}_2$ is a PS $\mathcal{P}_0$ over the Cartesian product $R_1 \times R_2$, whose set of models $\mathscr{M}_0$ is defined as:

$$\forall \mathcal{M}_m = \langle W_1 \times W_2, \succeq_m \rangle \in \mathscr{M}_0, \\ \exists \mathcal{M}_k = \langle W_1, \succeq_k \rangle \in \mathscr{M}_1, \exists \mathcal{M}_l = \langle W_2, \succeq_l \rangle \in \mathscr{M}_2 \text{ s.t.} \\ \forall w_1, w_1' \in W_1, \forall w_2, w_2' \in W_2 : \\ \succeq_m (w_1 \times w_2, w_1' \times w_2') \equiv \succ_k (w_1, w_1') \vee (=_k (w_1, w_1') \wedge \succeq_l (w_2, w_2')) .$$

4 Algebraic Optimization

As the PO extends RA, the optimization of queries with preferences can be realized as an extension of a classical relational query optimization. Most importantly, we can inherit all well known laws from RA, which, together with algebraic laws governing the commutativity and distributivity of the PO with respect to RA operations, constitute a formal foundation for rewriting queries with preferences using the standard strategies (*push selection*, *push projection*) aiming at reducing the sizes of intermediate relations.

Remark 2 (RA operators notation). In the following, RA selection and projection are generalized so that they can operate on set arguments, denoted by braces, e.g., $\sigma_\varphi(\{\omega_\mathcal{P}(I(R))\})$. The corresponding definitions are indicated by $\stackrel{\text{def}}{=}$.

4.1 Commuting with Selection

The following theorem identifies a sufficient condition under which the PO commutes with RA selection:

Theorem 3 (Commuting with selection). *Given a relation schema R, a PS $\mathcal{P}$ over R, the set of its PM's $\mathscr{M}$, and a selection condition φ over R, if*

$$\forall \mathcal{M}_k = \langle W, \succeq_k \rangle \in \mathscr{M}, \forall w, w' \in W : \quad \succ_k (w', w) \wedge w = \sigma_\varphi(w) \Rightarrow w' = \sigma_\varphi(w')$$

is a valid formula, then for any relation instance $I(R)$ of R:

$$\omega_\mathcal{P}(\sigma_\varphi(I(R))) = \sigma_\varphi(\{\omega_\mathcal{P}(I(R))\}) \stackrel{def}{=} \{w \in \omega_\mathcal{P}(I(R)) | \sigma_\varphi(w) = w\} .$$

Proof. Observe that:

$$w \in \omega_\mathcal{P}(\sigma_\varphi(I(R))) \equiv w \subseteq I(R) \wedge \sigma_\varphi(w) = w \wedge \\ \neg(\forall \mathcal{M}_k \in \mathscr{M}, \exists w' \subseteq I(R) : (\sigma_\varphi(w') = w' \wedge \succ_k (w', w)) . \\ w \in \sigma_\varphi(\{\omega_\mathcal{P}(I(R))\}) \equiv w \subseteq I(R) \wedge \sigma_\varphi(w) = w \wedge \\ \neg(\forall \mathcal{M}_k \in \mathscr{M}, \exists w' \subseteq I(R) : \succ_k (w', w)) ,$$

Obviously, the second formula implies the first. To see that the opposite implication also holds, we prove that $w \notin \sigma_\varphi(\{\omega_\mathcal{P}(I(R))\}) \Rightarrow w \notin \omega_\mathcal{P}(\sigma_\varphi(I(R)))$. There are three cases when $w \notin \sigma_\varphi(\{\omega_\mathcal{P}(I(R))\})$. If $w \nsubseteq I(R)$ or $\sigma_\varphi(w) \neq w$, it is immediately clear that $w \notin \omega_\mathcal{P}(\sigma_\varphi(I(R)))$. In the third case, $\forall \mathcal{M}_k \in \mathscr{M}, \exists w' \subseteq I(R) : \succ_k (w', w)$. However, due to the theorem assumption, $\forall \mathcal{M}_k \in \mathscr{M}, \exists w' \subseteq I(R) : \sigma_\varphi(w') = w' \wedge \succ_k (w', w)$, which completes the proof. □

4.2 Commuting with Projection

The following theorem identifies sufficient conditions under which the PO commutes with RA projection. To prepare the ground for the theorem, some definitions have to be introduced:

Definition 7 (Restriction of a preference relation). *Given a relation schema R, a set of attributes X of R, and a preference relation $\succeq$ over R, the restriction $\theta_X(\succeq)$ of $\succeq$ to X is a preference relation $\succeq_X$ over $\pi_X(R)$ defined using the following formula:*

$$\succeq_X (w_X, w'_X) \equiv \forall w, w' \in W : \pi_X(w) = w_X \wedge \pi_X(w') = w'_X \Rightarrow \succeq (w, w') .$$

Definition 8 (Restriction of the preference model). *Given a relation schema R, a set of relation attributes X of R, and a PM $\mathcal{M} = \langle W, \succeq \rangle$ over R, the restriction $\theta_X(\mathcal{M})$ of $\mathcal{M}$ to X is a PM $\mathcal{M}_X = \langle W_X, \succeq_X \rangle$ over $\pi_X(R)$ where $W_X = \{\pi_X(w) \mid w \in W\}$.*

Definition 9 (Restriction of the preference operator). *Given a relation schema R, a set of attributes X of R, a PS $\mathcal{P}$ over R, and the set $\mathscr{M}_X$ of its models restricted to X, the restriction $\theta_X(\omega_\mathcal{P})$ of the PO $\omega_\mathcal{P}$ to X is the PO $\omega^X_\mathcal{P}$ defined as follows:*

$$\omega^X_\mathcal{P}(\pi_X(I(R))) = \{w_X \subseteq \pi_X(I(R)) \mid \exists \mathcal{M}_X \in \mathscr{M}_X \text{ s.t. } \forall w'_X \subseteq \pi_X(I(R)) : \succeq_X (w'_X, w_X) \Rightarrow \succeq_X (w_X, w'_X)\} .$$

Theorem 4 (Commuting with projection). *Given a relation schema R, a set of attributes X of R, a PS $\mathcal{P}$ over R, and the set of its PM's $\mathscr{M}$, if the following formulae*

$$\forall \mathcal{M}_k \in \mathscr{M}, \forall w_1, w_2, w_3 \in W : \pi_X(w_1) = \pi_X(w_2) \wedge \pi_X(w_1) \neq \pi_X(w_3) \wedge \succeq_k (w_1, w_3) \Rightarrow \succeq_k (w_2, w_3) ,$$

$$\forall \mathcal{M}_k \in \mathscr{M}, \forall w_1, w_3, w_4 \in W : \pi_X(w_3) = \pi_X(w_4) \wedge \pi_X(w_1) \neq \pi_X(w_3) \wedge \succeq_k (w_1, w_3) \Rightarrow \succeq_k (w_1, w_4)$$

are valid, then for any relation instance $I(R)$ of R:

$$\omega^X_\mathcal{P}(\pi_X(I(R))) = \pi_X(\{\omega_\mathcal{P}(I(R))\}) \stackrel{def}{=} \{\pi_X(w) \mid w \in \omega_\mathcal{P}(I(R))\} .$$

Proof. We prove: $\pi_X(w) \notin \omega^X_\mathcal{P}(\pi_X(I(R))) \Longleftrightarrow \pi_X(w) \notin \pi_X(\{\omega_\mathcal{P}(I(R))\})$.

$\Rightarrow$: Assume $\pi_X(w_3) \notin \omega^X_\mathcal{P}(\pi_X(I(R)))$. The case $\pi_X(w_3) \nsubseteq \pi_X(I(R))$ is trivial. Otherwise, it must be the case that $\forall \mathcal{M}_X \in \mathscr{M}_X, \exists w_X \subseteq \pi_X(I(R)) : \succ_X (w_X, \pi_X(w_3))$, which implies $\forall \mathcal{M}_k \in \mathscr{M}, \forall w_1, w_4 \in W : \pi_X(w_1) = w_X \wedge \pi_X(w_4) = \pi_X(w_3) \Rightarrow \succ_k (w_1, w_4)$ and thus $\pi_X(w_3) \notin \pi_X(\{\omega_\mathcal{P}(I(R))\})$.

$\Leftarrow$: Assume $\pi_X(w_3) \notin \pi_X(\{\omega_\mathcal{P}(I(R))\})$. Then $\forall \mathcal{M}_k \in \mathscr{M}$ and $\forall w_4 \subseteq I(R)$ s.t. $\pi_X(w_4) = \pi_X(w_3)$, there is $w_1 \subseteq I(R)$ s.t. $\succ_k (w_1, w_4)$ and $\pi_X(w_1) \neq \pi_X(w_4)$. From the assumption of the theorem, it follows that $\forall w_2, w_4 \subseteq I(R) : \pi_X(w_2) = \pi_X(w_1) \wedge \pi_X(w_4) = \pi_X(w_3) \Rightarrow \succ_k (w_2, w_4)$, which implies $\theta_X(\succ_k)(\pi_X(w_1), \pi_X(w_3))$ and thus $\pi_X(w_3) \notin \omega^X_\mathcal{P}(\pi_X(I(R)))$. □

4.3 Distributing over Cartesian Product

For the PO to distribute over the Cartesian product of two relations, the PS, which is the parametr of the PO, needs to be decomposed into the PS's that will distribute into the argument relations. We obtain the same property for both Pareto and lexicographic composition:

Theorem 5 (Distributing over Cartesian product). *Given two relation schemas R_1 and R_2, and PS's $\mathcal{P}_1$ over R_1 and $\mathcal{P}_2$ over R_2, for any two relation instances $I(R_1)$ and $I(R_2)$ of R_1 and R_2:*

$$\omega_{\mathcal{P}_0}(I(R_1) \times I(R_2)) = \omega_{\mathcal{P}_1}(I(R_1)) \times \omega_{\mathcal{P}_2}(I(R_2)) \stackrel{def}{=} \{w_1 \times w_2 \mid w_1 \in \omega_{\mathcal{P}_1}(I(R_1)) \wedge w_2 \in \omega_{\mathcal{P}_2}(I(R_2))\} ,$$

where $\mathcal{P}_0 = P(\mathcal{P}_1, \mathcal{P}_2)$ is a Pareto composition of $\mathcal{P}_1$ and $\mathcal{P}_2$.

Proof. We prove:

$$w_1 \times w_2 \notin \omega_{\mathcal{P}_0}(I(R_1) \times I(R_2)) \Longleftrightarrow w_1 \times w_2 \notin \omega_{\mathcal{P}_1}(I(R_1)) \times \omega_{\mathcal{P}_2}(I(R_2)) .$$

$\Rightarrow$: Assume $w_1 \times w_2 \notin \omega_{\mathcal{P}_0}(I(R_1) \times I(R_2))$. Then $\forall \mathcal{M}_m \in \mathscr{M}_0$, models of $\mathcal{P}_0$, there are $w_1' \subseteq I(R_1), w_2' \subseteq I(R_2)$ s.t. $\succ_m (w_1' \times w_2', w_1 \times w_2)$. Consequently, $\forall \mathcal{M}_k \in \mathscr{M}_1, \forall \mathcal{M}_l \in \mathscr{M}_2$, models of $\mathcal{P}_1$ and $\mathcal{P}_2$, there are $w_1' \subseteq I(R_1), w_2' \subseteq I(R_2)$ s.t. $\succ_k (w_1', w_1)$ or $\succ_l (w_2', w_2)$, which implies $w_1 \notin \omega_{\mathcal{P}_1}(I(R_1))$ or $w_2 \notin \omega_{\mathcal{P}_2}(I(R_2))$ and thus $w_1 \times w_2 \notin \omega_{\mathcal{P}_1}(I(R_1)) \times \omega_{\mathcal{P}_2}(I(R_2))$.

$\Leftarrow$: Assume $w_1 \times w_2 \notin \omega_{\mathcal{P}_1}(I(R_1)) \times \omega_{\mathcal{P}_2}(I(R_2))$. Then $w_1 \notin \omega_{\mathcal{P}_1}(I(R_1))$ or $w_2 \notin \omega_{\mathcal{P}_2}(I(R_2))$. Assume the first. Then $\forall \mathcal{M}_k \in \mathscr{M}_1$, models of $\mathcal{P}_1$, there must be $w_1' \subseteq I(R_1)$ s.t. $\succ_k (w_1', w_1)$. Consequently, $\forall \mathcal{M}_m \in \mathscr{M}_0$, models of $\mathcal{P}_0$, $\exists w_1' \subseteq I(R_1) : \succ_m (w_1' \times w_2, w_1 \times w_2)$, which implies $w_1 \times w_2 \notin \omega_{\mathcal{P}_0}(I(R_1) \times I(R_2))$. The second case is symmetric. □

Theorem 6 (Distributing over Cartesian product). *Given two relation schemas R_1 and R_2, and PS's $\mathcal{P}_1$ over R_1 and $\mathcal{P}_2$ over R_2, for any two relation instances $I(R_1)$ and $I(R_2)$ of R_1 and R_2:*

$$\omega_{\mathcal{P}_0}(I(R_1) \times I(R_2)) = \omega_{\mathcal{P}_1}(I(R_1)) \times \omega_{\mathcal{P}_2}(I(R_2)) \stackrel{def}{=} \{w_1 \times w_2 \mid w_1 \in \omega_{\mathcal{P}_1}(I(R_1)) \wedge w_2 \in \omega_{\mathcal{P}_2}(I(R_2))\} ,$$

where $\mathcal{P}_0 = L(\mathcal{P}_1, \mathcal{P}_2)$ is a lexicographic composition of $\mathcal{P}_1$ and $\mathcal{P}_2$.

Proof. We prove:

$$w_1 \times w_2 \notin \omega_{\mathcal{P}_0}(I(R_1) \times I(R_2)) \Longleftrightarrow w_1 \times w_2 \notin \omega_{\mathcal{P}_1}(I(R_1)) \times \omega_{\mathcal{P}_2}(I(R_2)) .$$

$\Rightarrow$: Assume $w_1 \times w_2 \notin \omega_{\mathcal{P}_0}(I(R_1) \times I(R_2))$. Then $\forall \mathcal{M}_m \in \mathscr{M}_0$, models of $\mathcal{P}_0$, there are $w_1' \subseteq I(R_1), w_2' \subseteq I(R_2)$ s.t. $\succ_m (w_1' \times w_2', w_1 \times w_2)$. Consequently, $\forall \mathcal{M}_k \in \mathscr{M}_1, \forall \mathcal{M}_l \in \mathscr{M}_2$, models of $\mathcal{P}_1$ and $\mathcal{P}_2$, there are $w_1' \subseteq I(R_1), w_2' \subseteq I(R_2)$ s.t. $\succ_k (w_1', w_1)$ or $=_k (w_1', w_1) \wedge \succ_l (w_2', w_2)$, which implies $w_1 \notin \omega_{\mathcal{P}_1}(I(R_1))$ or $w_2 \notin \omega_{\mathcal{P}_2}(I(R_2))$ and thus $w_1 \times w_2 \notin \omega_{\mathcal{P}_1}(I(R_1)) \times \omega_{\mathcal{P}_2}(I(R_2))$.

$\Leftarrow$: Assume $w_1 \times w_2 \notin \omega_{\mathcal{P}_1}(I(R_1)) \times \omega_{\mathcal{P}_2}(I(R_2))$. Then $w_1 \notin \omega_{\mathcal{P}_1}(I(R_1))$ or $w_2 \notin \omega_{\mathcal{P}_2}(I(R_2))$. Assume the first. Then $\forall \mathcal{M}_k \in \mathscr{M}_1$, models of $\mathcal{P}_1$, there must be $w_1' \subseteq I(R_1)$ s.t. $\succ_k (w_1', w_1)$. Consequently, $\forall \mathcal{M}_m \in \mathscr{M}_0$, models of $\mathcal{P}_0$, there must be w_1' s.t. $\succ_m (w_1' \times w_2, w_1 \times w_2)$, which implies $w_1 \times w_2 \notin \omega_{\mathcal{P}_0}(I(R_1) \times I(R_2))$. The second case is symmetric. □

The equality $\omega_{\mathcal{P}_{\text{empty}}}(I(R)) = 2^{I(R)}$ and both Theorem 5 and Theorem 6 make it possible to derive the transformation rule that pushes the PO with a one-dimensional PS down the appropriate argument of the Cartesian product:

Corollary 1. *Given two relation schemas R_1 and R_2, a PS's $\mathcal{P}_1$ over R_1, and an empty PS $\mathcal{P}_2$ over R_2, for any two relation instances $I(R_1)$ and $I(R_2)$ of R_1 and R_2, the following property holds:*

$$\omega_{\mathcal{P}_0}(I(R_1) \times I(R_2)) = \omega_{\mathcal{P}_1}(I(R_1)) \times 2^{I(R_2)} \stackrel{def}{=} \{w_1 \times w_2 \mid w_1 \in \omega_{\mathcal{P}_1}(I(R_1)) \wedge w_2 \subseteq I(R_2)\} ,$$

where $\mathcal{P}_0 = P(\mathcal{P}_1, \mathcal{P}_2)$ is a Pareto of lexicographic composition of $\mathcal{P}_1$ and $\mathcal{P}_2$.

4.4 Distributing over Union

The following theorem shows how the PO distributes over RA union:

Theorem 7 (Distributing over union). *Given two compatible relation schemas[4] R and S, and a PS $\mathcal{P}$ over R (and S), if the following formula*

$$\omega_{\mathcal{P}}(I(R) \cup I(S)) \subseteq 2^{I(R)} \cup 2^{I(S)}$$

is valid for relation instances $I(R)$ and $I(S)$ of R and S, then:

$$\omega_{\mathcal{P}}(I(R) \cup I(S)) = \omega_{\mathcal{P}}(\{\omega_{\mathcal{P}}(I(R)) \cup \omega_{\mathcal{P}}(I(S))\}) .$$

Proof. It follows from Proposition 1 that $\omega_{\mathcal{P}}(I(R)) \cup \omega_{\mathcal{P}}(I(S)) \subseteq 2^{I(R)} \cup 2^{I(S)} \subseteq 2^{I(R) \cup I(S)}$. If we show that $\omega_{\mathcal{P}}(I(R) \cup I(S)) \subseteq \omega_{\mathcal{P}}(I(R)) \cup \omega_{\mathcal{P}}(I(S))$, then the theorem follows from Theorem 1.

If $w \in \omega_{\mathcal{P}}(I(R) \cup I(S))$, then it follows from the definition of the PO $w \subseteq I(R) \cup I(S) \wedge \exists \mathcal{M}_k \in \mathscr{M}$ s.t. $\forall w' \subseteq I(R) \cup I(S) : \succeq_k (w', w) \Rightarrow \succeq_k (w, w')$. As $w \subseteq I(R) \vee w \subseteq I(S)$ from the assumption of the theorem and $2^{I(R)} \cup 2^{I(S)} \subseteq 2^{I(R) \cup I(S)}$, we can conclude $(w \subseteq I(R) \vee w \subseteq I(S)) \wedge \forall w' \in 2^{I(R)} \cup 2^{I(S)} : \succeq_k (w', w) \Rightarrow \succeq_k (w, w')$, implying $w \in \omega_{\mathcal{P}}(I(R)) \cup \omega_{\mathcal{P}}(I(S))$. □

[4] We call two relation schemas *compatible* if they have the same number of attributes and the corresponding attributes have identical domains.

4.5 Distributing over Difference

Only in the trivial case, the the distribution over RA difference is possible:

Theorem 8 (Distributing over difference). *Given two compatible relation schemas R and S, and a PS $\mathcal{P}$ over R (and S), if the following formula*

$$\omega_{\mathcal{P}}(I(R)) \subseteq 2^{I(R)-I(S)} \cup 2^{I(S)}$$

is valid for relation instances $I(R) \neq I(S)$ of R and S, then:

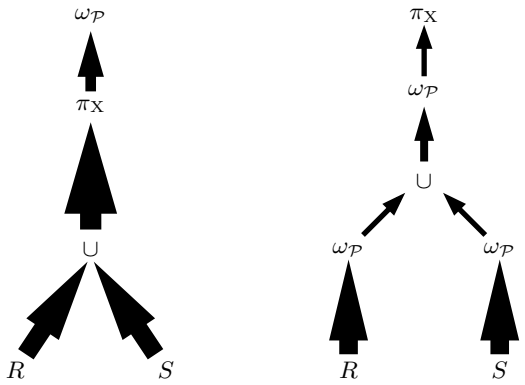
$$\omega_{\mathcal{P}}(I(R) - I(S)) = \omega_{\mathcal{P}}(I(R)) - \omega_{\mathcal{P}}(I(S))$$

iff the PS $\mathcal{P}$ is empty.

4.6 *Push Preference*

The question arises how to integrate the above algebraic laws into the classical, well-known hill-climbing algorithm. In particular, we want to add heuristic strategy of *push preference*, which is based on the assumption that early application of the PO reduces intermediate results. Indeed, the Theorem 1 provides a formal evidence that it is correct to pass exactly all the tuples that have been included in any world returned by the PO to the next operator in the operator tree. This leads to a better performance in subsequent operators.

Example 1. Consider a simple query expressed in RA as: $\omega_{\mathcal{P}}(\pi_X(R \cup S))$. After applying the preference strategy, we get: $\pi_X(\omega_{\mathcal{P}}(\{\omega_{\mathcal{P}}(R) \cup \omega_{\mathcal{P}}(S)\}))$. The corresponding expression trees are depicted in Fig. 1, where data flow between the computer's main memory and secondary storage is represented by line width.



(a) Before pushing (b) After pushing

Fig. 1: Improving the query plan by pushing PO down the expression tree

We have supposed that relations R and S are too big to fit into main memory. Using the number of the secondary storage I/O's as our measure of cost for an operation, it can be seen that the strategy of pushing the PO improves the performance in this case significantly.

5 Related Work

The study of preferences in the context of database queries has been originated by Lacroix and Lavency [8]. They, however, don't deal with algebraic optimization.

Following their work, *preference datalog* was introduced in [9], where it was shown that the concept of preference provides a modular and declarative means for formulating optimization and relaxation queries in deductive databases.

Nevertheless, only at the turn of the millennium this area attracted broader interest again. Kießling [1] and Chomicki et al. [3] have pursued independently a similar, *qualitative* approach within which preferences between tuples are specified directly, using binary *preference relations*. The embedding into RQL they have used is similar to ours: they have defined an operator returning only the best preference matches. However, they, by contrast to the approach presented in this paper, don't consider preferences between *sets* of elements and are concerned only with one type of preference. Moreover, the relation to a preference logic unfortunately is unclear. On the other hand, both Chomicki et. al. [3] and Kießling [2, 10] have laid the foundation for preference query optimization that extends established query optimization techniques: preference queries can be evaluated by extended – preference RA. While some transformation laws for queries with preferences have been presented in [2, 10], the results presented in [3] are mostly more general.

A special case of the same embedding represents *skyline operator* introduced by Börzsönyi et al. [4]. Some examples of possible rewritings for skyline queries are given but no general rewriting rules are formulated.

In [11], actual values of an arbitrary attribute were allowed to be partially ordered according to user preferences. Accordingly, RA operations, aggregation functions and arithmetic were redefined. However, some of their properties were lost, and the the query optimization issues were not discussed.

6 Conclusions

We build on the framework of embedding preferences into RQL through the PO that is parameterized by user preferences expressed in a declarative, logical language containing sixteen kinds of preferences and that returns the most preferred sets of tuples of its argument relation. Most importantly, the language is suitable for expressing preferences between sets of elements and its semantics allows for conflicting preferences.

The main contribution of the paper consists in presenting basic properties of the PO and a number of algebraic laws describing its interaction with other

RA operators. Particularly, sufficient conditions for commuting the PO with RA selection or projection and for distributing over Cartesian product, set union, and set difference have been identified. Thus key rules for rewriting the preference queries using the standard algebraic optimization strategies like *push selection* or *push projection* have been established. Moreover, a new optimization strategy of *push preference* has been suggested.

Future work directions include identifying further algebraic properties and finding the best possible ordering of transformations to compose an effective hill-climbing algorithm for optimization of RA statements with the PO. Also, expressiveness of RA including the PO and complexity issues have to be addressed in detail.

References

1. Kießling, W.: Foundations of Preferences in Database Systems. In: Proceedings of the 28th VLDB Conference, Hong Kong, China (2002) 311–322
2. Kießling, W., Hafenrichter, B.: Algebraic optimization of relational preference queries. Technical Report 2003-01, Institute of Computer Science, University of Augsburg (February 2003)
3. Chomicki, J.: Preference Formulas in Relational Queries. ACM Trans. Database Syst. **28**(4) (2003) 427–466
4. Börzsönyi, S., Kossmann, D., Stocker, K.: The skyline operator. In: Proceedings of the 17th International Conference on Data Engineering, Washington, DC, USA, IEEE Computer Society (2001) 421–430
5. von Wright, G.: The logic of preference. Edinburgh University Press, Edinburgh (1963)
6. Doyle, J., Wellman, M.P.: Representing preferences as ceteris paribus comparatives. In: Decision-Theoretic Planning: Papers from the 1994 Spring AAAI Symposium, AAAI Press, Menlo Park, California (1994) 69–75
7. Kaci, S., van der Torre, L.W.N.: Non-monotonic reasoning with various kinds of preferences. In Ronen I.Brafman, Junker, U., eds.: IJCAI-05 Multidisciplinary Workshop on Advances in Preference Handling. (August 2005) 112–117
8. Lacroix, M., Lavency, P.: Preferences; Putting More Knowledge into Queries. In Stocker, P.M., Kent, W., Hammersley, P., eds.: VLDB, Morgan Kaufmann (1987) 217–225
9. Govindarajan, K., Jayaraman, B., Mantha, S.: Preference datalog. Technical Report 95-50 (1, 1995)
10. Hafenrichter, B., Kießling, W.: Optimization of relational preference queries. In: CRPIT '39: Proceedings of the sixteenth Australasian conference on Database technologies, Darlinghurst, Australia, Australia, Australian Computer Society, Inc. (2005) 175–184
11. Nedbal, R.: Relational Databases with Ordered Relations. Logic Journal of the IGPL **13**(5) (2005) 587–597