Doktorandský den ’05

Ústav informatiky
Akademie věd České republiky

Hosty – Týn nad Vltavou

5. – 10. říjen 2005

vydavatelství Matematicko-fyzikální fakulty
University Karlovy v Praze

Publikaci ”Doktorandský den ’05” sestavil a připravil k tisku František Hakl
Ústav Informatiky AV ČR, Pod Vodárenskou věží 2, 182 07 Praha 8

ISBN *– not yet –*

Obsah

Relational Data Model with Preferences

Post-Graduate Student:
ING. RADIM NEDBAL

Institute of Computer Science
Academy of Sciences of the Czech Republic
Pod Vodárenskou věží 2

182 07 Prague 8

radned@seznam.cz

Supervisor:
ING. JÚLIUS ŠTULLER, CSC.

Institute of Computer Science
Academy of Sciences of the Czech Republic
Pod Vodárenskou věží 2

182 07 Prague 8

stuller@cs.cas.cz

Field of Study:
Mathematical Engineering
Classification: X11

Abstract

The paper proposes to extend the classical relational data model by the notion of preference realized through a partial ordering on the set of relation instances. This extension involves not only data representation but also data manipulation. As the theory of the majority of query languages for the relational model is based on the relational algebra and because of its fundamental nature, the algebra can be regarded as a basic measure of expressive power for database languages in general. To reach this expressive power in the proposed – semantically reacher – extended relational data model, the relational algebra operators need to be generalized. Simultaneously, it is desirable to preserve their usual properties. To sum up, the proposed extension of the relational model should be as minimal as possible, in the sense that the formal basis of the relational model is preserved. At the same time, the extended model should be fundamental enough to provide a sound basis for the investigation of new possible applications.

1. Introduction

Preference modelling is used in a great variety of fields. The purpose of this article is to present fundamental ideas of preference modelling in the framework of relational data model.

In this section, a brief overview of related research work and fundamentals of the proposed, extended relational model are presented. In the second section, the notion of preference and methods of its realization through ordering is introduced. In particular, order representation through $\langle P, I \rangle$ *preference structure* on attribute domains and ordering on instances of a given relation are explored. The third section discusses an effective implementation method through a generalized Hasse diagram notation. The last section summarizes the solutions pursued and the the approach potential.

1.1. Related Research Work

Recent work in AI and related fields has led to new types of preference models and new problems for applying preference structures.

Preference modelling fundamental notions as well as some recent results present Öztürk et al. [6]. The authors discuss different reasons for constructing a model of preference and number of issues that influence the construction of preference models. Information used when such models are established is analyzed, and different sources and types of uncertainty are introduced. Also, different formalisms, such as classical and nonclassical logics, and fuzzy sets, that can be used in order to establish a preference model are discussed, and different types of preference structures reflecting the behavior of a decision maker: classical, extended and valued ones, are presented. The concepts of thresholds and minimal representation are also introduced. Finally, the concept of deontic logic (logic of preference) and other formalisms associated with "compact representation of preferences", introduced for special purposes, are explored.

As ordering is inherent to the underlying data structure in database applications, Ng [5] proposes to extend the relational data model to incorporate partial orderings into data domains. Within the extended model, the partially ordered relational algebra (the PORA) is defined by allowing the ordering predicate to be used in formulae of the selection operator. The development of Ordered SQL (OSQL) as a query language for ordered databases is justified. Also, ordered functional dependencies (OFDs) on ordered databases are studied.

Nedbal [4] allows actual values of an arbitrary attribute to be partially ordered. Accordingly, relational algebra operators, aggregation functions, and arithmetic are redefined. Thus, on one side, the expressive power of the classical relational model is preserved, and, at the same time, as the new operators operate on and return ordered relations, information of preference, which is represented by a partial ordering, can be handled. Nevertheless, the redefinition of the relational operators causes loss of some of their common properties. For instance, $A = A \setminus (A \setminus B)$ does not hold. To rectify this weak point, more general concept is needed.

1.2. Extended Relational Data Model

The model proposed is a generalization of the one introduced in [4]. It extends the classical relational model both on the data representation and data manipulation levels. On the data representation level, the extension is based on incorporating an ordering into the set $\mathscr{I}(R)$ of all possible instances R^* of a relation R. Consequently, on the data manipulation level, the operators of relational algebra need to be generalized to enable handling the new information represented by the ordering. Considering the minimal set of relational algebra operators, at least, five operators: union, difference, cartesian product, selection, and projection, need to be generalized.

2. Preference on a Relation

Let us start with the following illustrative and motivating example introduced in [4]:

Example 1 (Partially ordered domain) How could we express our intention to find employees if we prefer those who speak English to those who speak German, who are preferred to those speaking any other germanic language? At the same time, we may similarly have preference for Spanish or French speaking employees to those speaking any other romanic language. To sum up, we have the following preferences:

A. Germanic languages:
   1. English,
   2. German,
   3. other germanic languages.

B. Romanic languages:
   1. Spanish of French,
   2. other romanic languages.

These preferences can be formalized by an ordering, in a general case by a partial ordering on equivalence classes. The situation is depicted in the following figure. The relation `R(NAME, POSITION, LANGUAGE)` of employees is represented by a table and the above preferences by means of the standard Hasse diagram notation.

Marie is preferred to David as she speaks English and David speaks "just" German. Analogically, Patrik is preferred to Andrea due to his knowledge of French. However, Patrik and David, for instance, are "incomparable" as we have expressed no preference order between German and French. Similarly, Roman is "incomparable" to any other employee as Russian is in the preference relation with no other language. □

Dominik	President	English
Marie	Manager	English
David	Manager	German
Petr	Manager	Swedish
Adam	Manager	German
Filip	Programmer	Dutch
Martina	Programmer	English
Patrik	Programmer	French
Rudolf	Programmer	Italian
Ronald	Programmer	Spanish
Andrea	Programmer	Portuguese
Roman	Programmer	Russian

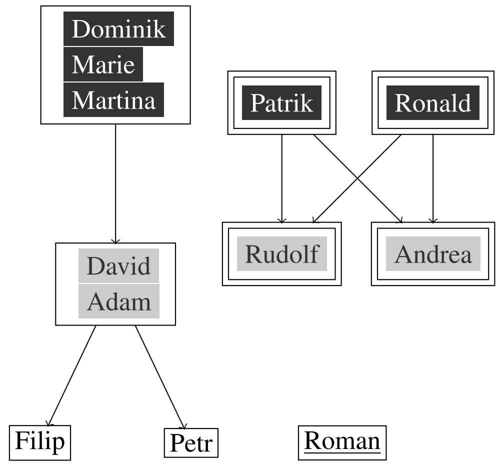


Figure 1: Partially ordered relation

Remark 1 The ordering representing the preference from the above example can be formally described by means of $\langle P, I \rangle$ *preference structure* ([6]).

Definition 1 (Preference Structure) A preference structure is a collection of binary relations defined on the set A and such that for each couple $a, b \in A$ exactly one relation is satisfied

Definition 2 ($\langle \mathbf{P}, \mathbf{I} \rangle$ preference structure) A $\langle P, I \rangle$ preference structure on the set A is a pair $\langle P, I \rangle$ of relations on A such that:

- P is asymmetric,
- I is reflexive, symmetric.

Remark 2 Any instance, R^*, of arbitrary relation, R, with an ordering, $\preceq^{R^*}$, represented by a $\langle P, I \rangle$ preference structure, $\preceq^{R^*} = P \cup \Delta_R$, determines, through a mapping $\mathcal{P}$:

$$\{[A; \preceq^A] | A \text{ is arbitrary set}\} \rightarrow \mathscr{P}(\sqcap(A)),$$

a set,

$$\mathcal{P}([R^*; \preceq^{R^*}]) = \{R^{F*} | t_i \preceq^{R^*} t_j \Rightarrow \mu_R(t_i^F) \leq \mu_R(t_j^F)\},$$

of fuzzy instances, R^{F*}, (of R) whose tuples,

$$t_i^F = \big(a_1, \ldots, a_n, \mu_R(t_i^F)\big),$$

have membership degrees, $\mu_R(t_i^F) \in \langle 0; 1 \rangle$, consistent with the ordering $\preceq^{R^*}$, i.e.

$$t_i \preceq^{R^*} t_j \Rightarrow \mu_R(t_i^F) \leq \mu_R(t_j^F)$$

□

Consider a classical unary relational operator,

$$\mathcal{O} : \mathscr{I}(R) \to \mathscr{I}(Q),$$

operating on the set, $\mathscr{I}(R)$, of all possible instances, R^*, of a relation R. Despite each R^* being assigned a preference, $\preceq^{R^*}$, the operator, $\mathcal{O}$, returns an instance, Q^*, of a resulting relation, Q, ignoring the ordering $\preceq^{R^*}$:

$$\mathcal{O} : \{[R^*; \preceq^{R^*}] \mid R^* \in \mathscr{I}(R)\} \to \mathscr{I}(Q)$$

Therefore we would like the operator, $\mathcal{O}$, to be generalized so that its result contains ordering, based on the ordering, $\preceq^{R^*}$, of the corresponding operand R^*:

$$\mathcal{O}_G : \{[R^*; \preceq^{R^*}] \mid R^* \in \mathscr{I}(R)\} \to \{[Q_G; \preceq^{Q_G}]_1 \dots [Q_G; \preceq^{Q_G}]_n\}$$

Specifically, we would like the generalized operator, $\mathcal{O}_G$, to be consistent with respect to remark 2, i.e. to return a result, $\mathcal{O}_G([R^*; \preceq^{R^*}]) = [Q_G; \preceq^{Q_G}]$, that determines a set,

$$\mathcal{P}([Q_G; \preceq^{Q_G}]) = \{\mathcal{O}_{\mathcal{F}}(R^{F*}) \mid R^{F*} \in \mathcal{P}([R^*; \preceq^{R^*}])\},$$

containing all the fuzzy instances, $Q^{F*} = \mathcal{O}_{\mathcal{F}}(R^{F*})$, we obtain if we employ a fuzzy propositional calculi equivalent, $\mathcal{O}_{\mathcal{F}}$, of the operator $\mathcal{O}$ to fuzzy instances $R^{F*} \in \mathcal{P}([R^*; \preceq^{R^*}])$. Moreover, this consistence should be independent of a t-norm related to the fuzzy propositional calculi, $\mathcal{F}$, employed. Observe that Q_G is generally a set $Q_G \subseteq \mathscr{I}(Q)$. In brief, we are looking for such a generalized operator, O_G, that the mapping $\mathcal{P}$ is a homomorphism from algebra

$$\big(\{[R^*; \preceq^{R^*}] \mid R^* \in \mathscr{I}(R)\} \cup \{[Q_G; \preceq^{Q_G}] \mid Q_G \subseteq \mathscr{I}(Q)\}; \mathcal{O}_G\big)$$

into algebra

$$\big(\mathscr{I}_F(R) \cup \mathscr{I}_F(Q); \mathcal{O}_{\mathcal{F}}\big)$$

for any t-norm.

Intuition suggests considering tuples according to their preferences. That is to say, we always take into account the more preferred tuples ahead of those with lower preference. In addition, the other tuples that are explicitly not preferred less should also be taken into consideration. To sum up, with each tuple, t_i, we take into account a set, S_{t_i}, containing this tuple and all the tuples with higher preference:

$$S_{t_i} = \{t \mid t \in R^* \ \wedge \ t_i \preceq^{R^*} t\}$$

Then the relational operator, $\mathcal{O}$, is applied to all the elements, S_j, of $S = \{S_j = \cup_{t_i \in R^*_k} S_{t_i} | R^*_k \subseteq R^*\}$. Finally, the order, $\preceq^{Q_G}$, on the set

$$\{\mathcal{O}(S_j) | S_j \in S\} \subseteq \mathscr{I}(Q)$$

is to be determined.

Example 2 Consider a set $\{[\mathcal{O}(S_i); S_i] | S_i \in S\} \subseteq \mathscr{I}(Q)$ with a relation, $\sqsubseteq$, implied by preference, $\preceq^{R^*}$, on R^* through the inclusion relation on S:

$$[\mathcal{O}(S_i); S_i] \sqsubseteq [\mathcal{O}(S_j); S_j] \Leftrightarrow S_i \subseteq S_j$$

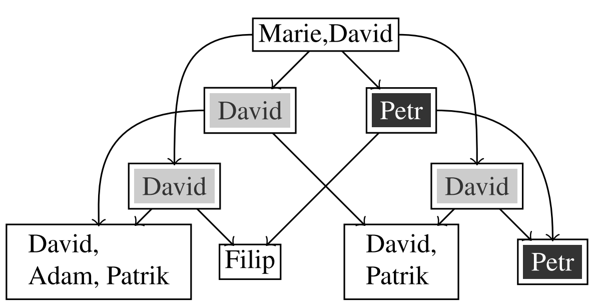


Figure 2: Projection $\{[\mathcal{O}(S_i); S_i; \sqsubseteq] | S_i \in S\}$ on $\{[\mathcal{O}(S_i); \sqsubseteq] | S_i \in S\}$

Notice that $\mathcal{O}$ generally is not an injection. In other words, $\mathcal{O}(S_i) = \mathcal{O}(S_j)$ for some $S_i \neq S_j$. In particular, $\mathcal{O}(S_i) = \mathcal{O}(S_j) =$ Petr and $\mathcal{O}(S_k) = \mathcal{O}(S_l) = \mathcal{O}(S_m) =$ David. To get an ordering, we need to resolve the duplicities:

- Firstly, as the occurrences of "Petr" are in the relation $\sqsubseteq$, we drop the less "preferred" one.
- In the case of the triplet of occurrences of "David", we are unable to determine the one with the highest "preference". Nevertheless, notice that:
  - The set $\{$Marie, David$\}$ is preferred to any of the occurrences of "David". In other words, whichever the most preferred occurrence of "David" is, it is less preferred then the set $\{$Marie, David$\}$.
  - There is a unique occurrence of "Filip", for which we can find an occurrence of "David" with a higher preference. In other words, whichever the most preferred occurrence of "David" is, it is preferred more then the occurrence of "Filip". The same rationale applies for the sets $\{$David, Adam, Patrik$\}$ and $\{$David, Patrik$\}$.

Thus, we get the resulting order, depicted in the following figure:

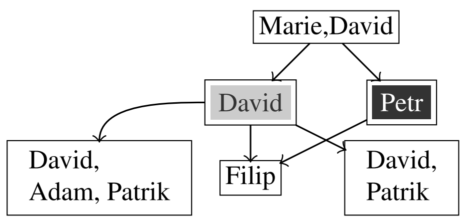


Figure 3: Ordering $\preceq^{Q_G}$ on $\{\mathcal{O}(S_i) | S_i \in S\} \subseteq \mathscr{I}(Q)$ $\square$

To sum up, the order $\preceq^{Q_G}$ on the set $\{\mathcal{O}(S_i) \mid S_i \in S\} \subseteq \mathscr{I}(Q)$ is defined as follows:

$$\mathcal{O}(S_i) \preceq^{Q_G} \mathcal{O}(S_j) \Leftrightarrow (\forall S_k \in S)\big([\mathcal{O}(S_k) = \mathcal{O}(S_i)] \Rightarrow (\exists S_l \in S)[\mathcal{O}(S_l) = \mathcal{O}(S_j) \wedge S_k \supseteq S_l]\big)$$

Approaching in this way a binary relational operator,

$$\mathcal{O} : \mathscr{I}(R) \times \mathscr{I}(R') \to \mathscr{I}(Q),$$

applied to a couple of relations, R, R', we get a set

$$\{\mathcal{O}(S_i, S'_k) \mid (S_i, S'_k) \in S \times S'\} \subseteq \mathscr{I}(Q)$$

and the order $\preceq^{Q_G}$ definition:

$$\begin{aligned}\mathcal{O}(S_i, S'_k) \preceq^{Q_G} \mathcal{O}(S_j, S'_l) \Leftrightarrow & \\ & [\forall (S_m, S'_p) \in S \times S']\big([\mathcal{O}(S_m, S'_p) = \mathcal{O}(S_i, S'_k)] \Rightarrow \\ & \quad [\exists (S_n, S'_q) \in S \times S'][\mathcal{O}(S_n, S'_q) = \mathcal{O}(S_j, S'_l) \ \wedge S_m \supseteq S_n \ \wedge \ S'_p \supseteq S'_q]\big)\end{aligned}$$

What are the consequences of this approach? Generally

$$\mathcal{O}(S_i) \preceq^{Q_G} \mathcal{O}(S_j) \Rightarrow \mathcal{O}(S_i) \supseteq \mathcal{O}(S_j)$$

does not hold. With respect to the relational property of *closure*, it is clear that the concept of defining preference through a $\langle P, I \rangle$ preference structure needs to be generalized.

In fact, we need to express the preference structure on powerset $\mathscr{I}(R)$ of all possible instances R^* of a relation R. This structure can be viewed through the model-theoretic approach as disjunction of conjuncts, where each conjunct, corresponding to an instance R^* of a relation R, has a given preference.

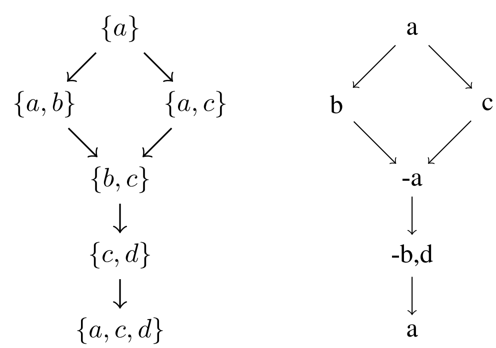
If we go further on in generalizing this structure, we get a preference structure on powerset $\mathscr{I}(DB)$ of all instances of a given database DB. It can be shown that such a structure generalizes the so-called *sure component* of M-table data structure (see Appendix) introduced by [2].

3. Sketch of Implementation

An important task to solve is the implementation of the proposed relational model. The so-called *generalized Hasse diagram* notation is suggested.

Example 3 Consider a set $S = \{a, b, c, d\}$ and its powerset, $\mathscr{P}(S)$, with an order, $\preceq^{\mathscr{P}(S)}$, represented by means of the standard Hasse diagram notation. The order on the powerset, $\mathscr{P}(S)$, can be represented as a relation $\mathcal{R}$ on S by means of the generalized Hasse diagram notation. There is one-to-one mapping between these two representations.

The generalization is based on the occurrences of “negative” elements, i.e. elements with a dash in front of them. Going through the diagram arrow-wise, they cancel a precedent occurrence of their “positive” equivalents (see figure 4). Moreover, all the elements depicted in the diagram are preferred to those that are absent in the diagram.



(a) Relation $\preceq^{\mathscr{P}(S)}$ on $\mathscr{P}(S)$ (b) Relation $\mathcal{R}$ on S

Figure 4: Standard and generalized Hasse diagram notation

Employing the generalized Hasse diagram notation, it is possible to develop effective algorithms for proposed, generalized relational algebra operations. Their description, however, is beyond the scope of this article. Their complexity is studied.

4. Conclusion

Methods of preference realization through $\langle P, I\rangle$ *preference structure* on attribute domains and through ordering on instances of a given relation have been discussed. Using the second approach, it has been proposed generalizing relational algebra operators in compliance with intuition. Also, a relationship of the proposed model with *M-table data structure* has been mentioned. Finally, the *generalized Hasse diagram* notation has been introduced as a means of effective implementation of the proposed model.

The proposed generalization of relational operators is necessary for a user of DBMS to be able to handle new information represented by preference. In the same way, the aggregation functions and arithmetics can be generalized.

It is possible to show that associativity and commutativity of the original union, product, restrict, and project operators are retained. Specifically for the generalized restrict operator, the following equivalences, which hold for the classical restrict operator, are retained:

$$\begin{aligned} R(\varphi_1 \vee \varphi_2) &\equiv R(\varphi_1) \cup R(\varphi_2) \\ R(\varphi_1 \wedge \varphi_2) &\equiv R(\varphi_1) \cap R(\varphi_2) \\ R(\neg\varphi) &\equiv R \setminus R(\varphi) \end{aligned}$$

Using the proposed approach, other relational operators (intersect, join, and divide), also, retain the usual properties of their classical relational counterparts:

$$\begin{aligned} R \cap S &\equiv R \setminus (R \setminus S) \\ R \div S &\equiv R[A-B] \setminus \big((R[A-B] \times S) \setminus R\big)[A-B] \\ R \bowtie S &\equiv (R \times S)(\varphi)[A] \end{aligned}$$

With respect to retaining of the above properties and equivalencies, we can conclude that the expressive power of the ordinary relational data model is maintained, and, at the same time, as the new operators operate on and return ordered relations, new information of preference represented by an ordering can be handled. This results in the ability to retrieve more accurate data.[1]

Appendix

Definition 3 (M-table) An M-table scheme, MR, is a finite list of relation schemes $\langle R_1, \ldots, R_k \rangle, k \geq 1$, where k is referred to as the *order* of the M-table. An M-table over the M-table scheme, $MR = \langle R_1, \ldots, R_k \rangle$, is a pair $T = \langle T_{sure}, T_{maybe} \rangle$ where

$$T_{sure} \subseteq \{(t_1, \ldots, t_k) \mid (\forall i)[1 \leq i \leq k \Rightarrow t_i \in \mathscr{I}(R_i)] \wedge (\exists i)[1 \leq i \leq k \wedge t_i \neq \emptyset]\}$$

$$T_{maybe} \in \{(r_1, \ldots, r_k) \mid (\forall i)[1 \leq i \leq k \Rightarrow r_i \in \mathscr{I}(R_i)]\},$$

Remark 3 We can associate k predicate symbols, $\tilde{R}_1, \ldots, \tilde{R}_k$, with an M-table of order k. An M-table consists of two components.

- Sure component, which consists of *mixed tuple sets*, whose elements

  $$\big\{\{t_1^1, \ldots, t_{n_1}^1\}, \ldots, \{t_1^k, \ldots, t_{n_k}^k\}\big\}$$

  represent definite and indefinite kind of information and correspond to the following logical formula

  $$[\tilde{R}_1(t_1^1) \vee \ldots \vee \tilde{R}_1(t_{n_1}^1)] \vee \ldots \vee [\tilde{R}_k(t_1^k) \vee \ldots \vee \tilde{R}_k(t_{n_k}^k)]$$

  That is, the sure component can be viewed as a conjunction of disjunctive formulas.

- Maybe component, which consists of *maybe tuples*, representing uncertain information. They may or may not correspond to the truth of the real world. Some of them may have appeared in the past in mixed tuple sets and there is more reason to expect them to be the truth of the real world than others that have not been mentioned anywhere.

[1] To my best knowledge, there is no similar study described in the literature.

References

[1] M. Ehrgott, J. Figuiera, and X. Gandibleux, editors. *State of the Art in Multiple Criteria Decision Analysis*. Springer Verlag, Berlin, 2005.

[2] K.-C. Liu and R. Sunderraman. A Generalized Relational Model for Indefinite and Maybe Information. *IEEE Transaction on Knowledge and Data Engineering*, 3(1):65–77, March 1991.

[3] R. Nedbal. Fuzzy database systems – concepts and implementation. Master's thesis, Czech Technical University, Faculty of Nuclear Sciences and Physical Engineering, Prague, June 2003.

[4] R. Nedbal. Relational databases with ordered relations. *Logic Journal of the IGPL*, 2005. Yet not published.

[5] W. Ng. An extension of the relational data model to incorporate ordered domains. *ACM Transactions on Database Systems*, 26(3):344–383, September 2001.

[6] M. Öztürk, A. Tsoukiàs, and P. Vincke. *Preference Modelling*, pages 27–72. In , [1], 2005.

List of Symbols

$\mathscr{P}(A)$ powerset of A,

$\Pi(R^*)$ $\{R^*_F \mid R^*_F \text{ is a fuzzy subset of } R^*\}$,

$\mathscr{I}(R)$ set of all possible instances, R^*, of a relation R,

$\mathscr{I}_F(R)$ set of all possible fuzzy instances, R^{F*}, of a relation R,

Δ_R $\{(t, t')|t \in R\}$

Ústav Informatiky AV ČR
DOKTORANDSKÝ DEN ’05

Vydal
MATFYZPRESS
vydavatelství
Matematicko-fyzikální fakulty
University Karlovy
Sokolovská 83, 186 75 Praha 8
jako svou *– not yet –* . publikaci

Obálku navrhl František Hakl

Z předloh připravených v systému LaTeX
vytisklo Reprostředisko MFF UK
Sokolovská 83, 186 75 Praha 8

Vydání první

Praha 2005

ISBN *– not yet –*