

Sensitivity Analysis for Declarative Relational Query Languages with Ordinal Ranks^{***}

Radim Belohlavek, Lucie Urbanova, and Vilem Vychodil

DAMOL (Data Analysis and Modeling Laboratory)
Dept. Computer Science, Palacky University, Olomouc
17. listopadu 12, CZ-77146 Olomouc, Czech Republic
radim.belohlavek@acm.org, lucie.urbanova01@upol.cz, vychodil@acm.org

Abstract. We present sensitivity analysis for results of query executions in a relational model of data extended by ordinal ranks. The underlying model of data results from the ordinary Codd’s model of data in which we consider ordinal ranks of tuples in data tables expressing degrees to which tuples match queries. In this setting, we show that ranks assigned to tuples are insensitive to small changes, i.e., small changes in the input data do not yield large changes in the results of queries.

Keywords: declarative query languages, ordinal ranks, relational databases, residuated lattices

1 Introduction

Since its inception, the relational model of data introduced by E. Codd [10] has been extensively studied by both computer scientists and database systems developers. The model has become the standard theoretical model of relational data and the formal foundation for relational database management systems. Various reasons for the success and strong position of Codd’s model are analyzed in [14], where the author emphasizes that the main virtues of the model like logical and physical data independence, declarative style of data retrieval (database querying), access flexibility and data integrity are consequences of a close connection between the model and the first-order predicate logic.

This paper is a continuation of our previous work [4,5] where we have introduced an extension of Codd’s model in which tuples are assigned ordinal ranks. The motivation for the model is that in many situations, it is natural to consider not only the exact matches of queries in which a tuple of values either *does* or *does not* match a query Q but also approximate matches where tuples match queries to degrees. The degrees of approximate matches can usually be described verbally using linguistic modifiers like “not at all (matches)” “almost (matches)”, “more or less (matches)”, “fully (matches)”, etc. From the user’s

* Supported by grant no. P103/11/1456 of the Czech Science Foundation.

** The paper will appear in Proceedings of the 19th International Conference on Applications of Declarative Programming and Knowledge Management (INAP 2011).

point of view, each data table in our extended relational model consists of (i) an ordinary data table whose meaning is the same as in the Codd's model and (ii) ranks assigned to all tuples in the original data table. This way, we come up with a notion of a ranked data table (shortly, an RDT). The ranks in RDTs are interpreted as “goodness of match” and the interpretation of RDTs is the same as in the Codd's model—they represent answers to queries which are, in addition, equipped with priorities expressed by the ranks. A user who looks at an answer to a query in our model is typically looking for the best match possible represented by a tuple or tuples in the resulting RDT with the highest ranks (i.e., highest priorities).

In order to have a suitable formalization of ranks and to perform operations with ranked data tables, we have to choose a suitable structure for ranks. Since ranks are meant to be compared by users, the set L of all considered ranks should be equipped with a partial order \leq , i.e. $\langle L, \leq \rangle$ should be a poset. Moreover, it is convenient to postulate that $\langle L, \leq \rangle$ is a complete lattice [7], i.e., for each subset $A \subseteq L$, its least upper bound (a supremum) and greatest lower bound (an infimum) exist. This way, for any $A \subseteq L$, one can take the least rank in L which represents a higher priority (a better match) than all ranks from A . Such a rank is then the supremum of A (dually for the infimum). Since $\langle L, \leq \rangle$ is a complete lattice, it contains the least element denoted 0 (no match at all) and the greatest element denoted 1 (full match).

The set L of all ranks should also be equipped with additional operations for aggregation of ranks. Indeed, if tuple t with rank a is obtained as one of the results of subquery Q_1 and the same t with another rank b is obtained from answers to subquery Q_2 then we might want to express the rank to which t matches a compound conjunctive query “ Q_1 and Q_2 ”. A natural way to do so is to take a suitable binary operation $\otimes: L \times L \rightarrow L$ which acts as a conjunctive and take $a \otimes b$ for the resulting rank. Obviously, not every binary operation on L represents a (reasonable) conjunctive, i.e. we may restrict the choices only to particular binary operations that make “good conjunctors”. There are various ways to impose such restrictions. In our model, we follow the approach of using residuated conjunctions that has proved to be useful in logics based on residuated lattices [2,18,19]. Namely, we assume that $\langle L, \otimes, 1 \rangle$ is a commutative monoid (i.e., \otimes is associative, commutative, and neutral with respect to 1) and there is a binary operation \rightarrow on L such that for all $a, b, c \in L$:

$$a \otimes b \leq c \quad \text{if and only if} \quad a \leq b \rightarrow c. \quad (1)$$

Operations \otimes (a multiplication) and \rightarrow (a residuum) satisfying (1) are called *adjoint operations*. Altogether, the structure for ranks we use is a *complete residuated lattice* $\mathbf{L} = \langle L, \wedge, \vee, \otimes, \rightarrow, 0, 1 \rangle$, i.e., a complete lattice in which \otimes and \rightarrow are adjoint operations, and \wedge and \vee denote the operations of infimum and supremum, respectively. Considering \mathbf{L} as a basic structure of ranks brings several benefits. First, in multiple-valued logics and in particular fuzzy logics [18,19], residuated lattices are interpreted as structures of truth degrees and the relationship (1) between \otimes (a fuzzy conjunction) and \rightarrow (a fuzzy implication) is

derived from requirements on graded counterpart of the *modus ponens* deduction rule (currently, there are many strong-complete logics based on residuated lattices).

Remark 1. The graded counterpart of *modus ponens* [19,26] can be seen as a generalized deduction rule saying “from φ valid (at least) to degree $a \in L$ and $\varphi \Rightarrow \psi$ valid (at least) to degree $b \in L$, infer ψ valid (at least) to degree $a \otimes b$ ”. If if-part of (1) ensures that the rule is sound while the only-if part ensures that it is as powerful as possible, i.e., $a \otimes b$ is the highest degree to which we infer ψ valid provided that φ valid at least to degree a and $\varphi \Rightarrow \psi$ valid at least to degree $b \in L$. This relationship between \rightarrow (a truth function for logical connective implication \Rightarrow) and \otimes has been discovered in [17] and later used, e.g., in [16,26]. Interestingly, (1) together with the lattice ordering ensure enough properties of \rightarrow and \otimes . For instance, \rightarrow is antitone in the first argument and is monotone in the second one, condition $a \leq b$ iff $a \rightarrow b = 1$ holds for all $a, b \in L$, $a \rightarrow (b \rightarrow c)$ equals $(a \otimes b) \rightarrow c$ for all $a, b, c \in L$, etc. Since complete residuated lattices are in general weaker structures than Boolean algebras, not all laws satisfied by truth functions of the classic conjunction and implication are preserved by all complete residuated lattices. For instance, neither $a \otimes a = a$ (idempotency of \otimes) nor $(a \rightarrow 0) \rightarrow 0 = a$ (the law of double negation) nor $a \vee (a \rightarrow 0) = 1$ (the law of the excluded middle) hold in general. Nevertheless, complete residuated lattices are strong enough to provide a formal framework for relational analysis and similarity-based reasoning as it has been shown by previous results.

Second, our extension of the Codd’s model results from the model by replacing the two-element Boolean algebra, which is the classic structure of truth values, by a more general structure of truth values represented by a residuated lattice, i.e. we make the following shift in (the semantics of) the underlying logic:

$$\text{two-element Boolean algebra} \quad \Longrightarrow \quad \text{a complete residuated lattice.}$$

Third, the original Codd’s model is a special case of our model for \mathbf{L} being the two-element Boolean algebra (only two borderline ranks 1 and 0 are available). As a practical consequence, data tables in the Codd’s model can be seen as RDTs where all ranks are either equal to 1 (full match) or 0 (no match; tuples with 0 rank are considered as not present in the result of a query). Using residuated lattices as structures of truth degrees, we obtain a generalization of Codd’s model which is based on solid logical foundations and has desirable properties. In addition, its relationship to residuated first-order logics is the same as the relationship of the original Codd’s model to the classic first-order logic. The formalization we offer can further be used to provide insight into several isolated approaches that have been provided in the past, see e.g. [8], [15], [23], [27], [28], [30], and a comparison paper [6].

A typical choice of \mathbf{L} is a structure with $L = [0, 1]$ (ranks are taken from the real unit interval), \wedge and \vee being minimum and maximum, \otimes being a left-continuous (or a continuous) t-norm with the corresponding \rightarrow , see [2,18,19]. For example, an RDT with ranks coming from such \mathbf{L} is in Table 1. It can be seen

Table 1. Houses for sale at \$200,000 with square footage 1200

| | <i>agent</i> | <i>id</i> | <i>sqft</i> | <i>age</i> | <i>location</i> | <i>price</i> |
|------|--------------|-----------|-------------|------------|-----------------|--------------|
| 0.93 | Brown | 138 | 1185 | 48 | Vestal | \$228,500 |
| 0.89 | Clark | 140 | 1120 | 30 | Endicott | \$235,800 |
| 0.86 | Brown | 142 | 950 | 50 | Binghamton | \$189,000 |
| 0.85 | Brown | 156 | 1300 | 85 | Binghamton | \$248,600 |
| 0.81 | Clark | 158 | 1200 | 25 | Vestal | \$293,500 |
| 0.81 | Davis | 189 | 1250 | 25 | Binghamton | \$287,300 |
| 0.75 | Davis | 166 | 1040 | 50 | Vestal | \$286,200 |
| 0.37 | Davis | 112 | 1890 | 30 | Endicott | \$345,000 |

as a result of similarity-based query “show all houses which are sold for (approximately) \$200,000 and have (approximately) 1200 square feet”. The left-most column contains ranks. The remaining part of the table is a data table in the usual sense containing tuples of values. At this point, we do not explain in detail how the particular ranks in Table 1 have been obtained (this will be outlined in further sections). One way is by executing a similarity-based query that uses additional information about similarity (proximity) of domain values which is also described using degrees from **L**. Note that the concept of a similarity-based query appears when human perception is involved in rating or comparing close values from domains where not only the exact equalities (matches) are interesting. For instance, a person searching in a database of houses is usually not interested in houses sold for a particular exact price. Instead, the person wishes to look at houses sold approximately at that price, including those which are sold for other prices that are sufficiently close. While the ranks constitute a “visible” part of any RDT, the similarities are not a direct part of RDT and have to be specified for each domain independently. They can be seen as an additional (background) information about domains which is supplied by users of the database system.

Let us stress the meaning of ranks as priorities. As it is usual in fuzzy logics in narrow sense, their meaning is primarily *comparative*, cf. [19, p. 2] and the comments on comparative meaning of truth degrees therein. In our example, it means that tuple $\langle \text{Clark}, 140, 1120, 30, \text{Endicott}, \$235,800 \rangle$ with rank 0.89 is a better match than tuple $\langle \text{Brown}, 142, 950, 50, \text{Binghamton}, \$189,000 \rangle$ whose rank 0.86 is strictly smaller. Thus, for end-users, the numerical values of ranks (if L is a unit interval) are not so important, the important thing is the relative ordering of tuples given by the ranks.

Note that our model which provides theoretical foundations for similarity-based databases [4,5] should not be confused with models for probabilistic databases [29] which have recently been studied, e.g. in [9,12,13,20,22,25], see also [11] for a survey. In particular, numerical ranks used in our model (if $L = [0, 1]$) cannot be interpreted as probabilities, confidence degrees of belief degrees as in case of probabilistic databases where ranks play such roles. In probabilistic databases, the tuples (i.e., the data itself) are uncertain and the ranks express probabilities that tuples appear in data tables. Consequently, a probabilistic database is formalized by a discrete probability space over the pos-

sible contents of the database [11]. Nevertheless, the underlying logic of the models is the classical two-valued first-order logic—only yes/no matches are allowed (with uncertain outcome). In our case, the situation is quite different. The data (represented by tuples) is absolutely certain but the tuples are allowed to match queries to degrees. This, translated in terms of logic, means that formulas (encoding queries) are allowed to be evaluated to truth degrees other than 0 and 1. Therefore, the underlying logic in our model is not the classic two-element Boolean logic as we have argued hereinbefore.

In [1], a report written by leading authorities in database systems, the authors say that the current database management systems have no facilities for either approximate data or imprecise queries. According to this report, the management of uncertainty and imprecision is one of the six currently most important research directions in database systems. Nowadays, probabilistic databases (dealing with approximate data) are extensively studied. On the contrary, it seems that similarity-based databases (dealing with imprecise queries) have not yet been paid full attention. This paper is a contribution to theoretical foundations of similarity-based databases.

2 Problem Setting

The issue we address in this paper is the following. In our model, we can get two or more RDTs (as results of queries) which are not exactly the same but which are perceived (by users) as being similar. For instance, one can obtain two RDTs containing the same tuples with numerical values of ranks that are almost the same. A question is whether such similar RDTs, when used in subsequent queries, yield similar results. In this paper, we present a preliminary study of the phenomenon of similarity of RDTs and its relationship to the similarity of query results obtained by applying queries to similar input data tables. We present basic notions and results providing formulas for computing estimations of similarity degrees. The observations we present provide a formal justification for the phenomenon discussed in the previous section—slight changes in ranks do not have a large impact on the results of (complex) queries. The results are obtained for any complete residuated lattice taken as the structure of ranks (truth degrees). Note that the basic query systems in our model are (extensions of) domain relational calculus [5,24] and relational algebra [4,24]. We formulate the results in terms of operations of the relational algebra but due to its equivalence with the domain relational calculus [5], the results pertain to both the query systems. Thus, based on the domain relational calculus, one may design a declarative query language preserving similarity in which execution of queries is based on transformations to expressions of relational algebra in a similar way as in the classic case [24].

The rest of the paper is organized as follows. Section 3 presents a short survey of notions. Section 4 contains results on sensitivity analysis, an illustrative example, and a short outline of future research. Because of the limited scope of the paper, proofs are sketched or omitted.

3 Preliminaries

In this section, we recall basic notions of RDTs and relational operations we need to provide insight into the sensitivity issues of RDTs in Section 4. Details can be found in [2,4,6]. In the rest of the paper, \mathbf{L} always refers to a complete residuated lattice $\mathbf{L} = \langle L, \wedge, \vee, \otimes, \rightarrow, 0, 1 \rangle$, see Section 1.

3.1 Basic Structures

Given \mathbf{L} , we make use of the following notions: An \mathbf{L} -set A in universe U is a map $A : U \rightarrow L$, $A(u)$ being interpreted as “the degree to which u belongs to A ”. If \mathbf{L} is the two-element Boolean algebra, then $A : U \rightarrow L$ is an indicator function of a classic subset of U , $A(u) = 1$ ($A(u) = 0$) meaning that u belongs (does not belong) to that subset. In our approach, we tacitly identify sets with their indicator functions. In a similar way, a binary \mathbf{L} -relation B on U is a map $B : U \times U \rightarrow L$, $B(u_1, u_2)$ interpreted as “the degree to which u_1 and u_2 are related according to B ”. Hence, B is an \mathbf{L} -set in universe $U \times U$.

3.2 Ranked Data Tables over Domains with Similarities

We denote by Y a set of *attributes*, any subset $R \subseteq Y$ is called a *relation scheme*. For each attribute $y \in Y$ we consider its *domain* D_y . In addition, each D_y is equipped with a binary \mathbf{L} -relation \approx_y on D_y satisfying reflexivity ($u \approx_y u = 1$) and symmetry $u \approx_y v = v \approx_y u$ (for all $u, v \in D_y$). Each binary \mathbf{L} -relation \approx_y on D_y satisfying (i) and (ii) shall be called a *similarity*. Pair $\langle D_y, \approx_y \rangle$ is called a *domain with similarity*.

Tuples contained in data tables will be considered as usual, i.e., as elements of Cartesian products of domains. Recall that a Cartesian product $\prod_{i \in I} D_i$ of an I -indexed system $\{D_i \mid i \in I\}$ of sets D_i ($i \in I$) is a set of all maps $t : I \rightarrow \bigcup_{i \in I} D_i$ such that $t(i) \in D_i$ holds for each $i \in I$. Under this notation, a *tuple* over $R \subseteq Y$ is any element from $\prod_{y \in R} D_y$. For brevity, $\prod_{y \in R} D_y$ is denoted by $\text{Tupl}(R)$. Following the example in Table 1, tuple $\langle \text{Brown}, 142, 950, 50, \text{Binghamton}, \$189,000 \rangle$ is a map $r \in \text{Tupl}(R)$ for $R = \{\text{agent}, \text{id}, \dots, \text{price}\}$ such that $r(\text{agent}) = \text{Brown}$, $r(\text{id}) = 142$, etc.

A *ranked data table* on $R \subseteq Y$ over $\{\langle D_y, \approx_y \rangle \mid y \in R\}$ (shortly, an RDT) is any (finite) \mathbf{L} -set \mathcal{D} in $\text{Tupl}(R)$. The degree $\mathcal{D}(r)$ to which r belongs to \mathcal{D} is called a *rank* of tuple r in \mathcal{D} . According to its definition, if \mathcal{D} is an RDT on R over $\{\langle D_y, \approx_y \rangle \mid y \in R\}$ then \mathcal{D} is a map $\mathcal{D} : \text{Tupl}(R) \rightarrow L$. Note that \mathcal{D} is an n -ary \mathbf{L} -relation between domains D_y ($y \in Y$) since \mathcal{D} is a map from $\prod_{y \in R} D_y$ to L . In our example, $\mathcal{D}(r) = 0.86$ for r being the tuple with $r(\text{id}) = 142$.

3.3 Relational Operations with RDTs

Relational operations we consider in this paper are the following: For RDTs \mathcal{D}_1 and \mathcal{D}_2 on T , we put $(\mathcal{D}_1 \cup \mathcal{D}_2)(t) = \mathcal{D}_1(t) \vee \mathcal{D}_2(t)$ and $(\mathcal{D}_1 \cap \mathcal{D}_2)(t) =$

$\mathcal{D}_1(t) \wedge \mathcal{D}_2(t)$ for each $t \in \text{Tupl}(T)$; $\mathcal{D}_1 \cup \mathcal{D}_2$ and $\mathcal{D}_1 \cap \mathcal{D}_2$ are called the *union* and the \wedge -*intersection* of \mathcal{D}_1 and \mathcal{D}_2 , respectively. Analogously, one can define an \otimes -*intersection* $\mathcal{D}_1 \otimes \mathcal{D}_2$. Hence, \cup , \cap , and \otimes are defined componentwise based on the operations of the complete residuated lattice \mathbf{L} .

Moreover, our model admits new operations that are trivial in the classic model. For instance, for $a \in L$, we introduce an a -*shift* $a \rightarrow \mathcal{D}$ of \mathcal{D} by $(a \rightarrow \mathcal{D})(t) = a \rightarrow \mathcal{D}(t)$ for all $t \in \text{Tupl}(T)$.

Remark 2. Note that if \mathbf{L} is the two-element Boolean algebra then a -shift is a trivial operation since $1 \rightarrow \mathcal{D} = \mathcal{D}$ and $0 \rightarrow \mathcal{D}$ produces a possibly infinite table containing all tuples from $\text{Tupl}(T)$. In our model, an a -shift has the following meaning: If \mathcal{D} is a result of query Q then $(a \rightarrow \mathcal{D})(t)$ is a “degree to which t matches query Q at least to degree a ”. This follows from properties of residuum, see [2,19]. Hence, a -shifts allow us to emphasize results that match queries at least to a prescribed degree a .

The remaining relational operations we consider represent counterparts of projection, selection, and join in our model. If \mathcal{D} is an RDT on T , the *projection* $\pi_R(\mathcal{D})$ of \mathcal{D} onto $R \subseteq T$ is defined by

$$(\pi_R(\mathcal{D}))(r) = \bigvee_{s \in \text{Tupl}(T \setminus R)} \mathcal{D}(rs),$$

for each $r \in \text{Tupl}(R)$. In our example, the result of $\pi_{\{\text{location}\}}(\mathcal{D})$ is a ranked data table with single column such that $\pi_{\{\text{location}\}}(\mathcal{D})(\langle \text{Binghamton} \rangle) = 0.86$, $\pi_{\{\text{location}\}}(\mathcal{D})(\langle \text{Vestal} \rangle) = 0.93$, and $\pi_{\{\text{location}\}}(\mathcal{D})(\langle \text{Endicott} \rangle) = 0.89$.

A similarity-based selection is a counterpart to ordinary selection which selects from a data table all tuples which approximately match a given condition: Let \mathcal{D} be an RDT on T and let $y \in T$ and $d \in D_y$. Then, a *similarity-based selection* $\sigma_{y \approx d}(\mathcal{D})$ of tuples in \mathcal{D} matching $y \approx d$ is defined by

$$(\sigma_{y \approx d}(\mathcal{D}))(t) = \mathcal{D}(t) \otimes t(y) \approx_y d.$$

Considering \mathcal{D} as a result of query Q , the rank of t in $\sigma_{y \approx d}(\mathcal{D})$ can be interpreted as a degree to which “ t matches the query Q and the y -value of t is similar to d ”. In particular, an interesting case is $\sigma_{p \approx q}(\mathcal{D})$ where p and q are both attributes with a common domain with similarity.

Similarity-based joins are considered as derived operations based on Cartesian products and similarity-based selections. For $r \in \text{Tupl}(R)$ and $s \in \text{Tupl}(S)$ such that $R \cap S = \emptyset$, we define a concatenation $rs \in \text{Tupl}(R \cup S)$ of tuples r and s so that $(rs)(y) = r(y)$ for $y \in R$ and $(rs)(y) = s(y)$ for $y \in S$. For RDTs \mathcal{D}_1 and \mathcal{D}_2 on disjoint relation schemes S and T we define a RDT $\mathcal{D}_1 \times \mathcal{D}_2$ on $S \cup T$, called a *Cartesian product* of \mathcal{D}_1 and \mathcal{D}_2 , by $(\mathcal{D}_1 \times \mathcal{D}_2)(st) = \mathcal{D}_1(s) \otimes \mathcal{D}_2(t)$. Using Cartesian products and similarity-based selections, we can introduce *similarity-based θ -joins* such as $\mathcal{D}_1 \bowtie_{p \approx q} \mathcal{D}_2 = \sigma_{p \approx q}(\mathcal{D}_1 \times \mathcal{D}_2)$. Various other types of similarity-based joins can be introduced in our model, see [5].

4 Estimations of Sensitivity of Query Results

4.1 Rank-Based Similarity of Query Results

We now introduce the notion of similarity of RDTs which is based on the idea that RDTs \mathcal{D}_1 and \mathcal{D}_2 (on the same relation scheme) are similar iff for each tuple t , ranks $\mathcal{D}_1(t)$ and $\mathcal{D}_2(t)$ are similar (degrees from \mathbf{L}). Similarity of ranks can be expressed by biresiduum \leftrightarrow (a fuzzy equivalence [2,18,19]) which is a derived operation of \mathbf{L} such that $a \leftrightarrow b = (a \rightarrow b) \wedge (b \rightarrow a)$. Since we are interested in similarity of $\mathcal{D}_1(t)$ and $\mathcal{D}_2(t)$ for all possible tuples t , it is straightforward to define the similarity $E(\mathcal{D}_1, \mathcal{D}_2)$ of \mathcal{D}_1 and \mathcal{D}_2 by an infimum which goes over all tuples:

$$E(\mathcal{D}_1, \mathcal{D}_2) = \bigwedge_{t \in \text{Tuple}(T)} (\mathcal{D}_1(t) \leftrightarrow \mathcal{D}_2(t)). \quad (2)$$

An alternative (but equivalent) way is the following: we first formalize a degree $S(\mathcal{D}_1, \mathcal{D}_2)$ to which \mathcal{D}_1 is included in \mathcal{D}_2 . We can say that \mathcal{D}_1 is fully included in \mathcal{D}_2 iff, for each tuple t , the rank $\mathcal{D}_2(t)$ is at least as high as the rank $\mathcal{D}_1(t)$. Notice that in the classic (two-values) case, this is exactly how one defines the ordinary subsethood relation “ \subseteq ”. Considering general degrees of inclusion (subsethood), a degree $S(\mathcal{D}_1, \mathcal{D}_2)$ to which \mathcal{D}_1 is included in \mathcal{D}_2 can be defined as follows:

$$S(\mathcal{D}_1, \mathcal{D}_2) = \bigwedge_{t \in \text{Tuple}(T)} (\mathcal{D}_1(t) \rightarrow \mathcal{D}_2(t)). \quad (3)$$

It is easy to prove [2] that (2) and (3) satisfy:

$$E(\mathcal{D}_1, \mathcal{D}_2) = S(\mathcal{D}_1, \mathcal{D}_2) \wedge S(\mathcal{D}_2, \mathcal{D}_1). \quad (4)$$

Note that E and S defined by (2) and (3) are known as degrees of similarity and subsethood from general fuzzy relational systems [2] (in this case, the fuzzy relations are RDTs).

The following assertion shows that \cup , \cap , \otimes , and a -shifts preserve subsethood degrees given by (3). In words, the degree to which $\mathcal{D}_1 \cup \mathcal{D}_2$ is included in $\mathcal{D}'_1 \cup \mathcal{D}'_2$ is at least as high as the degree to which \mathcal{D}_1 is included in \mathcal{D}'_1 and \mathcal{D}_2 is included in \mathcal{D}'_2 . A similar verbal description can be made for the other operations.

Theorem 1. *For any $\mathcal{D}_1, \mathcal{D}'_1, \mathcal{D}_2$, and \mathcal{D}'_2 on relation scheme T ,*

$$S(\mathcal{D}_1, \mathcal{D}'_1) \wedge S(\mathcal{D}_2, \mathcal{D}'_2) \leq S(\mathcal{D}_1 \cup \mathcal{D}_2, \mathcal{D}'_1 \cup \mathcal{D}'_2), \quad (5)$$

$$S(\mathcal{D}_1, \mathcal{D}'_1) \wedge S(\mathcal{D}_2, \mathcal{D}'_2) \leq S(\mathcal{D}_1 \cap \mathcal{D}_2, \mathcal{D}'_1 \cap \mathcal{D}'_2), \quad (6)$$

$$S(\mathcal{D}_1, \mathcal{D}'_1) \otimes S(\mathcal{D}_2, \mathcal{D}'_2) \leq S(\mathcal{D}_1 \otimes \mathcal{D}_2, \mathcal{D}'_1 \otimes \mathcal{D}'_2), \quad (7)$$

$$S(\mathcal{D}_1, \mathcal{D}_2) \leq S(a \rightarrow \mathcal{D}_1, a \rightarrow \mathcal{D}_2). \quad (8)$$

Proof (sketch). (5): Using adjointness, it suffices to check that $(S(\mathcal{D}_1, \mathcal{D}'_1) \wedge S(\mathcal{D}_2, \mathcal{D}'_2)) \otimes (\mathcal{D}_1 \cup \mathcal{D}_2)(t) \leq (\mathcal{D}'_1 \cup \mathcal{D}'_2)(t)$ holds true for any $t \in \text{Tuple}(T)$. Using (3), the monotony of \otimes and \wedge yields $(S(\mathcal{D}_1, \mathcal{D}'_1) \wedge S(\mathcal{D}_2, \mathcal{D}'_2)) \otimes (\mathcal{D}_1 \cup \mathcal{D}_2)(t) \leq ((\mathcal{D}_1(t) \rightarrow \mathcal{D}'_1(t)) \wedge (\mathcal{D}_2(t) \rightarrow \mathcal{D}'_2(t))) \otimes (\mathcal{D}_1(t) \vee \mathcal{D}_2(t))$. Applying $a \otimes (b \vee c) = (a \otimes b) \vee (a \otimes c)$ to the latter expression, we get $((\mathcal{D}_1(t) \rightarrow \mathcal{D}'_1(t)) \wedge (\mathcal{D}_2(t) \rightarrow \mathcal{D}'_2(t))) \otimes (\mathcal{D}_1(t) \vee \mathcal{D}_2(t)) \leq ((\mathcal{D}_1(t) \rightarrow \mathcal{D}'_1(t)) \otimes \mathcal{D}_1(t)) \vee ((\mathcal{D}_2(t) \rightarrow \mathcal{D}'_2(t)) \otimes \mathcal{D}_2(t))$.

$\mathcal{D}'_2(t)) \otimes \mathcal{D}_2(t))$. Using $a \otimes (a \rightarrow b) \leq b$ twice, it follows that $((\mathcal{D}_1(t) \rightarrow \mathcal{D}'_1(t)) \otimes \mathcal{D}_1(t)) \vee ((\mathcal{D}_2(t) \rightarrow \mathcal{D}'_2(t)) \otimes \mathcal{D}_2(t)) \leq \mathcal{D}'_1(t) \vee \mathcal{D}'_2(t)$. Putting previous inequalities together, $(S(\mathcal{D}_1, \mathcal{D}'_1) \wedge S(\mathcal{D}_2, \mathcal{D}'_2)) \otimes (\mathcal{D}_1 \cup \mathcal{D}_2)(t) \leq (\mathcal{D}'_1 \cup \mathcal{D}'_2)(t)$ which proves (5). (6) can be proved analogously as (5); (7) can be proved analogously as (6) using monotony of \otimes ; (8) follows from the fact that $a \rightarrow b \leq (c \rightarrow a) \rightarrow (c \rightarrow b)$. \square

Using (4), we have the following consequence of Theorem 1:

Corollary 1. *For \diamond being \cap and \cup , we have:*

$$E(\mathcal{D}_1, \mathcal{D}'_1) \wedge E(\mathcal{D}_2, \mathcal{D}'_2) \leq E(\mathcal{D}_1 \diamond \mathcal{D}_2, \mathcal{D}'_1 \diamond \mathcal{D}'_2). \quad (9)$$

$$E(\mathcal{D}_1, \mathcal{D}'_1) \otimes E(\mathcal{D}_2, \mathcal{D}'_2) \leq E(\mathcal{D}_1 \otimes \mathcal{D}_2, \mathcal{D}'_1 \otimes \mathcal{D}'_2). \quad (10)$$

$$E(\mathcal{D}_1, \mathcal{D}_2) \leq E(a \rightarrow \mathcal{D}_1, a \rightarrow \mathcal{D}_2). \quad (11)$$

Proof (sketch). For \diamond being \cap , (6) applied twice yields: $S(\mathcal{D}_1, \mathcal{D}'_1) \wedge S(\mathcal{D}_2, \mathcal{D}'_2) \leq S(\mathcal{D}_1 \cap \mathcal{D}_2, \mathcal{D}'_1 \cap \mathcal{D}'_2)$ and $S(\mathcal{D}'_1, \mathcal{D}_1) \wedge S(\mathcal{D}'_2, \mathcal{D}_2) \leq S(\mathcal{D}'_1 \cap \mathcal{D}'_2, \mathcal{D}_1 \cap \mathcal{D}_2)$. Hence, (9) for \cap follows using (2). The rest is analogous. \square

Using the idea in the proof of Corollary 1, in order to prove that operation O preserves similarity, it suffices to check that O preserves (graded) subsethood. Thus, from now on, we shall only investigate whether operations preserve subsethood. In case of Cartesian products, we have:

Theorem 2. *Let \mathcal{D}_1 and \mathcal{D}'_1 be RDTs on relation scheme S and let \mathcal{D}_2 and \mathcal{D}'_2 be RDTs on relation scheme T such that $S \cap T = \emptyset$. Then,*

$$S(\mathcal{D}_1, \mathcal{D}'_1) \otimes S(\mathcal{D}_2, \mathcal{D}'_2) \leq S(\mathcal{D}_1 \times \mathcal{D}_2, \mathcal{D}'_1 \times \mathcal{D}'_2), \quad (12)$$

Proof (sketch). The proof is analogous to that of (7). \square

The following assertion shows that projection and similarity-based selection preserve subsethood degrees (and therefore similarities) of RDTs:

Theorem 3. *Let \mathcal{D} and \mathcal{D}' be RDTs on relation scheme T and let $y \in T$, $d \in D_y$, and $R \subseteq T$. Then,*

$$S(\mathcal{D}, \mathcal{D}') \leq S(\pi_R(\mathcal{D}), \pi_R(\mathcal{D}')), \quad (13)$$

$$S(\mathcal{D}, \mathcal{D}') \leq S(\sigma_{y \approx d}(\mathcal{D}), \sigma_{y \approx d}(\mathcal{D}')). \quad (14)$$

Proof (sketch). In order to prove (13), we check $S(\mathcal{D}, \mathcal{D}') \otimes (\pi_R(\mathcal{D}))(r) \leq (\pi_R(\mathcal{D}'))(r)$ for any $r \in \text{Tuple}(R)$. It means showing that

$$S(\mathcal{D}, \mathcal{D}') \otimes \bigvee_{s \in \text{Tuple}(T \setminus R)} \mathcal{D}(rs) \leq (\pi_R(\mathcal{D}'))(r).$$

Thus, it suffices to prove $S(\mathcal{D}, \mathcal{D}') \otimes \mathcal{D}(rs) \leq (\pi_R(\mathcal{D}'))(r)$ for all $s \in \text{Tuple}(T \setminus R)$. Using monotony of \otimes , we get $S(\mathcal{D}, \mathcal{D}') \otimes \mathcal{D}(rs) \leq (\mathcal{D}(rs) \rightarrow \mathcal{D}'(rs)) \otimes \mathcal{D}(rs) \leq \mathcal{D}'(rs)$, because $rs \in \text{Tuple}(T)$. Therefore, $S(\mathcal{D}, \mathcal{D}') \otimes \mathcal{D}(rs) \leq \mathcal{D}'(rs) \leq \bigvee_{s \in \text{Tuple}(T \setminus R)} \mathcal{D}'(rs) = (\pi_R(\mathcal{D}'))(r)$, which proves the first claim of (13). In case of (14), we proceed analogously. \square

Table 2. Alternative ranks for houses for sale from Table 1

| | <i>agent</i> | <i>id</i> | <i>sqft</i> | <i>age</i> | <i>location</i> | <i>price</i> |
|------|--------------|-----------|-------------|------------|-----------------|--------------|
| 0.93 | Brown | 138 | 1185 | 48 | Vestal | \$228,500 |
| 0.91 | Clark | 140 | 1120 | 30 | Endicott | \$235,800 |
| 0.87 | Brown | 156 | 1300 | 85 | Binghamton | \$248,600 |
| 0.85 | Brown | 142 | 950 | 50 | Binghamton | \$189,000 |
| 0.82 | Davis | 189 | 1250 | 25 | Binghamton | \$287,300 |
| 0.79 | Clark | 158 | 1200 | 25 | Vestal | \$293,500 |
| 0.75 | Davis | 166 | 1040 | 50 | Vestal | \$286,200 |
| 0.37 | Davis | 112 | 1890 | 30 | Endicott | \$345,000 |

Theorem 2 and Theorem 3 used together yield

Corollary 2. *Let \mathcal{D}_1 and \mathcal{D}'_1 be RDTs on relation scheme S and let \mathcal{D}_2 and \mathcal{D}'_2 be RDTs on relation scheme T such that $S \cap T = \emptyset$. Then,*

$$S(\mathcal{D}_1, \mathcal{D}'_1) \otimes S(\mathcal{D}_2, \mathcal{D}'_2) \leq S(\mathcal{D}_1 \bowtie_{p \approx q} \mathcal{D}_2, \mathcal{D}'_1 \bowtie_{p \approx q} \mathcal{D}'_2). \quad (15)$$

for any $p \in S$ and $q \in T$ having the same domain with similarity. \square

As a result, we have shown that important relational operations in our model (including similarity-based joins) preserve similarity defined by (2). Thus, we have provided a formal justification for the (intuitively expected but nontrivial) fact that similar input data yield similar results of queries.

Remark 3. In this paper, we have restricted ourselves only to a fragment of relational operations in our model. In [5], we have shown that in order to have a relational algebra whose expressive power is the same as the expressive power of the domain relational calculus, we have to consider additional operations of *residuum* (defined componentwise using \rightarrow) and *division*. Nevertheless, these two additional operations preserve E as well—it can be shown using similar arguments as in the proof of Theorem 1. As a consequence, the similarity is preserved by all queries that can be formulated in DRC [5].

4.2 Illustrative Example

Consider again the RDT from Table 1. The RDT can be seen as a result of querying a database of houses for sale where one wants to find a house which is sold for (approximately) \$200,000 and has (approximately) 1200 square feet. The attributes in the RDT are: real estate agent name (*agent*), house ID (*id*), square footage (*sqft*), house age (*age*), house location (*location*), and house price (*price*). In this example, the complete residuated lattice $\mathbf{L} = \langle L, \wedge, \vee, \otimes, \rightarrow, 0, 1 \rangle$ serving as the structure of ranks will be the so-called Lukasiewicz algebra [2,18,19]. That is, $L = [0, 1]$, \wedge and \vee are minimum and maximum, respectively, and the multiplication and residuum are defined as follows: $a \otimes b = \max(a + b - 1, 0)$ and $a \rightarrow b = \min(1 - a + b, 1)$ for all $a, b \in L$.

Intuitively, it is natural to consider similarity of values in domains of *sqft*, *age*, *location*, and *price*. For instance, similarity of prices can be defined by

$p_1 \approx_{price} p_2 = s(|p_2 - p_1|)$ using an antitone scaling function $s: [0, \infty) \rightarrow [0, 1]$ with $s(0) = 1$ (i.e., identical prices are fully similar). Analogously, a similarity of locations can be defined based on their geographical distance and/or based on their evaluation (safety, school districts, ...) by an expert. In contrast, there is no need to have similarities for *id* and *agents* because end-users do not look for houses based on (similarity of) their (internal) IDs which are kept as keys merely because of performance reasons. Obviously, there may be various reasonable similarity relations defined for the above-mentioned domains and their careful choice is an important task. In this paper, we neither explain nor recommend particular ways to do so because (i) we try to keep a general view of the problem and (ii) similarities on domains are purpose and user dependent.

Consider now the RDT in Table 2 defined over the same relation scheme as the RDT in Table 1. These two RDTs can be seen as two (slightly different) answers to the same query (when e.g., the domain similarities have been slightly changed) or answers to a modified query (e.g., “show all houses which are sold for (approximately) \$210,000 and ...”). The similarity of both the RDTs given by (2) is 0.98 (very high). The results in the previous section say that if we perform any (arbitrarily complex) query (using the relational operations we consider in this paper) with Table 2 instead of Table 1, the results will be similar at least to degree 0.98.

Table 3. Join of Table 1 and the table of customers

| | <i>agent</i> | <i>id</i> | <i>price</i> | <i>name</i> | <i>budget</i> |
|------|--------------|-----------|--------------|-------------|---------------|
| 0.91 | Brown | 138 | \$228,500 | Grant | \$240,000 |
| 0.89 | Brown | 138 | \$228,500 | Evans | \$250,000 |
| 0.89 | Brown | 138 | \$228,500 | Finch | \$210,000 |
| 0.88 | Clark | 140 | \$235,800 | Grant | \$240,000 |
| 0.86 | Clark | 140 | \$235,800 | Evans | \$250,000 |
| 0.84 | Brown | 156 | \$248,600 | Evans | \$250,000 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 0.16 | Davis | 112 | \$345,000 | Grant | \$240,000 |
| 0.10 | Davis | 112 | \$345,000 | Finch | \$210,000 |

For illustration, consider an additional RDT of customers over relation scheme containing two attributes: *name* (customer name) and *budget* (price the customer is willing to pay for a house). In particular, let $\langle \text{Evans}, \$250,000 \rangle$, $\langle \text{Finch}, \$210,000 \rangle$, and $\langle \text{Grant}, \$240,000 \rangle$ be the only tuples in the RDT (all with ranks 1). The answer to the following query

$$\pi_{\{agent, id, price, name, budget\}}(\mathcal{D}_1 \bowtie_{price \approx budget} \mathcal{D}_c),$$

where \mathcal{D}_1 stands for Table 1 and \mathcal{D}_c stands for the RDT of customers is in Table 3 (for brevity, some records are omitted). The RDT thus represents an answer to query “show deals for houses sold for (approximately) \$200,000 with (approximately) 1200 square feet and customers so that their budget is similar to the house price”. Furthermore, we can obtain an RDT of best agent-customer

Table 4. Results of agent-customer matching for Table 1 and Table 2

| | <i>agent</i> | <i>name</i> |
|------|--------------|-------------|
| 0.91 | Brown | Grant |
| 0.89 | Brown | Evans |
| 0.89 | Brown | Finch |
| 0.88 | Clark | Grant |
| 0.86 | Clark | Evans |
| 0.84 | Clark | Finch |
| 0.74 | Davis | Evans |
| 0.72 | Davis | Grant |
| 0.66 | Davis | Finch |

| | <i>agent</i> | <i>name</i> |
|------|--------------|-------------|
| 0.91 | Brown | Grant |
| 0.90 | Clark | Grant |
| 0.89 | Brown | Evans |
| 0.89 | Brown | Finch |
| 0.88 | Clark | Evans |
| 0.86 | Clark | Finch |
| 0.75 | Davis | Evans |
| 0.73 | Davis | Grant |
| 0.67 | Davis | Finch |

matching is we project the join onto *agent* and *name*:

$$\pi_{\{agent, name\}}(\mathcal{D}_1 \bowtie_{price \approx budget} \mathcal{D}_c).$$

The result of matching is in Table 4 (left). Due to our results, if we perform the same query with Table 2 instead of Table 1, the new result is guaranteed to be similar with the obtained result at least to degree 0.98. The result for Table 2 is shown in Table 4 (right).

4.3 Tuple-Based Similarity and Further Topics

While the rank-based similarity from Section 4.1 can be sufficient in many cases, there are situations where one wants to consider a similarity of RDTs based on ranks and (pairwise) similarity of tuples. For instance, if we take the RDT from Table 1 and make a new one by taking all tuples (keeping their ranks) and increasing the prices by one dollar, we will come up with an RDT which is, according to rank-based similarity, very different from the original one. Intuitively, one would expect to have a high degree of similarity of the RDTs because they differ only by a slight change in price. This issue can be solved by considering the following tuple-based degree of inclusion:

$$S^{\approx}(\mathcal{D}_1, \mathcal{D}_2) = \bigwedge_{t \in \text{Tuple}(T)} (\mathcal{D}_1(t) \rightarrow \bigvee_{t' \in \text{Tuple}(T)} (\mathcal{D}_2(t') \otimes t \approx t')), \quad (16)$$

where $t \approx t' = \bigwedge_{y \in T} t(y) \approx_y t'(y)$ is a similarity of tuples t and t' over T , cf. [6]. In a similar way as in (4), we may define E^{\approx} using S^{\approx} instead of S .

Remark 4. By an easy inspection, $S(\mathcal{D}_1, \mathcal{D}_2) \leq S^{\approx}(\mathcal{D}_1, \mathcal{D}_2)$, i.e. (16) yields an estimate which is at least as high as (3) and analogously for E and E^{\approx} . Note that (16) has a natural meaning. Indeed, $S^{\approx}(\mathcal{D}_1, \mathcal{D}_2)$ can be understood as a degree to which the following statement is true: “If t belongs to \mathcal{D}_1 , then there is t' which is similar to t and which belongs to \mathcal{D}_2 ”. Hence, $E^{\approx}(\mathcal{D}_1, \mathcal{D}_2)$ is a degree to which for each tuple from \mathcal{D}_1 there is a similar tuple in \mathcal{D}_2 and *vice versa*. If \mathbf{L} is a two-element Boolean algebra and each \approx_y is an identity, then $E^{\approx}(\mathcal{D}_1, \mathcal{D}_2) = 1$ iff \mathcal{D}_1 and \mathcal{D}_2 are identical (in the usual sense).

For tuple-based inclusion (similarity) and for certain relational operations, we can prove analogous preservation formulas as in Section 4.1. For instance,

$$S^\approx(\mathcal{D}_1, \mathcal{D}'_1) \wedge S(\mathcal{D}_2, \mathcal{D}'_2) \leq S^\approx(\mathcal{D}_1 \cup \mathcal{D}_2, \mathcal{D}'_1 \cup \mathcal{D}'_2), \quad (17)$$

$$S^\approx(\mathcal{D}_1, \mathcal{D}'_1) \otimes S(\mathcal{D}_2, \mathcal{D}'_2) \leq S^\approx(\mathcal{D}_1 \times \mathcal{D}_2, \mathcal{D}'_1 \times \mathcal{D}'_2), \quad (18)$$

$$S^\approx(\mathcal{D}, \mathcal{D}') \leq S^\approx(\pi_R(\mathcal{D}), \pi_R(\mathcal{D}')). \quad (19)$$

On the other hand, similarity-based selection $\sigma_{y \approx d}$ (and, as a consequence, similarity-based join $\bowtie_{p \approx q}$) does not preserve S^\approx in general which can be seen as a technical complication. This issue can be overcome by introducing a new type of selection $\sigma_{y \approx d}^\approx$ which is *compatible* with S^\approx . Namely, we can define

$$(\sigma_{y \approx d}^\approx(\mathcal{D}))(t) = \bigvee_{t' \in \text{Tupl}(T)} (\mathcal{D}(t') \otimes t' \approx t \otimes t(y) \approx_y d). \quad (20)$$

For this notion, we can prove that $S^\approx(\mathcal{D}, \mathcal{D}') \leq S^\approx(\sigma_{y \approx d}^\approx(\mathcal{D}), \sigma_{y \approx d}^\approx(\mathcal{D}'))$. Similar extension can be done for any relational operation which does not preserve S^\approx directly. Detailed description of the extension is postponed to a full version of the paper because of the limited scope.

4.4 Unifying Approach to Similarity of RDTs

In this section, we outline a general approach to similarity of RDTs that includes both the approaches from the previous sections. Interestingly, both (3) and (16) have a common generalization using truth-stressing hedges [19,21]. Truth-stressing hedges represent unary operations on complete residuated lattices (denoted by $*$) that serve as interpretations of logical connectives like “very true”, see [19]. Two boundary cases of hedges are (i) identity, i.e. $a^* = a$ ($a \in L$); (ii) globalization: $1^* = 1$, and $a^* = 0$ if $a < 1$. The globalization [31] is a hedge which can be interpreted as “fully true”.

Let $*$ be truth-stressing hedge on \mathbf{L} . For RDTs $\mathcal{D}_1, \mathcal{D}_2$ on T , we define the degree $S_*^\approx(\mathcal{D}_1, \mathcal{D}_2)$ of inclusion of \mathcal{D}_1 in \mathcal{D}_2 (with respect to $*$) by

$$S_*^\approx(\mathcal{D}_i, \mathcal{D}_j) = \bigwedge_{t \in \text{Tupl}(T)} (\mathcal{D}_i(t) \rightarrow \bigvee_{t' \in \text{Tupl}(T)} (\mathcal{D}_j(t') \otimes (t \approx t')^*)). \quad (21)$$

Now, it is easily seen that for $*$ being the identity, (21) coincides with (16); if \approx is separating (i.e., $t_1 \approx t_2 = 1$ iff t_1 is identical to t_2) and $*$ is the globalization, (21) coincides with (3). Thus, both (3) and (16) are particular instances of (21) resulting by a choice of the hedge. Note that identity and globalization are two borderline cases of hedges. In general, complete residuated lattices admit other nontrivial hedges that can be used in (21). Therefore, the hedge in (21) serves as a parameter that has an influence on how much emphasis we put on the fact that two tuples are similar. In case of globalization, we put full emphasis, i.e., the tuples are required to be equal to degree 1 (exactly the same if \approx is separating).

If we consider properties needed to prove analogous estimation formulas for general S_*^\approx as we did in case of S and S^\approx , we come up with the following important property:

$$(r \approx s)^* \otimes (s \approx t)^* \leq (r \approx t)^*, \quad (22)$$

for every $r, s, t \in \text{Tupl}(T)$ which can be seen as transitivity of \approx with respect to \otimes and $*$. Consider the following two cases in which (22) is satisfied:

- Case 1: $*$ is globalization and \approx is separating. If the left hand side of (22) is nonzero, then $r \approx s = 1$ and $s \approx t = 1$. Separability implies $r = s = t$, i.e. $(r \approx t)^* = 1^* = 1$, verifying (22).
- Case 2: \approx is transitive. In this case, since $a^* \otimes b^* \leq (a \otimes b)^*$ (follows from properties of hedges by standard arguments), transitivity of \approx and monotony of $*$ yield $(r \approx s)^* \otimes (s \approx t)^* \leq ((r \approx s) \otimes (s \approx t))^* \leq (r \approx t)^*$.

The following lemma shows that S_*^\approx and consequently E_*^\approx have properties that are considered natural for (degrees of) inclusion and similarity:

Lemma 1. *If \approx satisfies (22) with respect to $*$ then*

- (i) S_*^\approx is a reflexive and transitive **L**-relation, i.e. an **L**-quasiorder.
- (ii) E_*^\approx defined by $E_*^\approx(\mathcal{D}_1, \mathcal{D}_2) = S_*^\approx(\mathcal{D}_1, \mathcal{D}_2) \wedge S_*^\approx(\mathcal{D}_2, \mathcal{D}_1)$ is a reflexive, symmetric, and transitive **L**-relation, i.e. an **L**-equivalence.

Proof. The assertion follows from results in [2, Section 4.2] by taking into account that \approx^* is reflexive, symmetric, and transitive with respect to \otimes . \square

5 Conclusion and Future Research

We have shown that an important fragment of relational operation in similarity-based databases preserves various types of similarity. As a result, similarity of query results based on these relational operations can be estimated based on similarity of input data tables before the queries are executed. Furthermore, the results of this paper have shown a desirable important property of the underlying similarity-based model of data: slight changes in input data do not produce huge changes in query results. Future research will focus on the role of particular relational operations called similarity-based closures that play an important role in tuple-based similarities of RDTs. An outline of results in this direction is presented in [3].

References

1. S. Abiteboul *et al.* The Lowell database research self-assessment. *Communications of the ACM* 48(5):111-118, 2005.
2. R. Belohlavek. *Fuzzy Relational Systems: Foundations and Principles*. Kluwer, Academic/Plenum Publishers, New York, 2002.
3. R. Belohlavek, L. Urbanova, and V. Vychodil. Similarity of query results in similarity-based databases (*in preparation*).
4. R. Belohlavek and V. Vychodil. Logical foundations for similarity-based databases. *DASFAA 2009 Workshops*, LNCS 5667:137-151, 2009.
5. R. Belohlavek and V. Vychodil. Query systems in similarity-based databases: logical foundations, expressive power, and completeness. In: *Proc. ACM SAC* 2010, pp. 1648-1655.

6. R. Belohlavek and V. Vychodil. Codd's relational model from the point of view of fuzzy logic. *J. Logic and Computation* (to appear, doi: 10.1093/logcom/exp056).
7. G. Birkhoff: *Lattice theory*. First edition. American Mathematical Society, Providence, 1940.
8. B. P. Buckles and F. E. Petry. Fuzzy databases in the new era. ACM SAC 1995, pages 497–502, Nashville, TN, 1995.
9. R. Cavallo and M. Pittarelli. The theory of probabilistic databases. VLDB 1987, pp. 71–81.
10. E. F. Codd. A Relational Model of Data for Large Shared Data Banks. *Communications of the ACM* 13(6):377–387, 1970.
11. N. Dalvi, C. Ré and D. Suciu. Probabilistic databases: diamonds in the dirt. *Communications of the ACM* 52:86–94, 2009.
12. N. Dalvi and D. Suciu. Efficient query evaluation on probabilistic databases. *The VLDB Journal* 16:523–544, 2007.
13. N. Dalvi and D. Suciu. Management of probabilistic data: foundations and challenges. ACM PODS 2007, pp. 1–12.
14. C. J. Date. *Database Relational Model: A Retrospective Review and Analysis*. Addison Wesley, 2000.
15. R. Fagin. Combining fuzzy information: an overview. *ACM SIGMOD Record* 31(2):109–118, 2002.
16. G. Gerla. *Fuzzy Logic. Mathematical Tools for Approximate Reasoning*. Kluwer, Dordrecht, 2001.
17. J. A. Goguen. The logic of inexact concepts. *Synthese* 18:325–373, 1968–9.
18. S. Gottwald. Mathematical fuzzy logics. *Bulletin for Symbolic Logic* 14(2):210–239, 2008.
19. P. Hájek. *Metamathematics of Fuzzy Logic*. Kluwer, Dordrecht, 1998.
20. T. Imieliński, W. Lipski. Incomplete information in relational databases. *Journal of the ACM* 31:761–791, 1984.
21. P. Hájek. On very true. *Fuzzy Sets and Syst.* 124:329–333, 2001.
22. C. Koch. On query algebras for probabilistic databases. *SIGMOD Record* 37(4):78–85, 2008.
23. C. Li, K. C.-C. Chang, I. F. Ilyas, and S. Song. RankSQL: Query Algebra and Optimization for Relational top-k queries. ACM SIGMOD 2005, pp. 131–142.
24. D. Maier. *The Theory of Relational Databases*. Comp. Sci. Press, Rockville, 1983.
25. D. Olteanu, J. Huang, C. Koch. Approximate confidence computation in probabilistic databases. *IEEE ICDE 2010*, pp. 145–156.
26. J. Pavelka: On fuzzy logic I, II, III. *Z. Math. Logik Grundlagen Math.* 25:45–52, 25:119–134, 25:447–464, 1979.
27. K. V. S. V. N. Raju and A. K. Majumdar. Fuzzy functional dependencies and loss-less join decomposition of fuzzy relational database systems. *ACM Trans. Database Systems* Vol. 13, No. 2:129–166, 1988.
28. S. Shenoi and A. Melton. Proximity relations in the fuzzy relational database model. *Fuzzy Sets and Syst.* 100:51–62, 1999.
29. D. Suciu, D. Olteanu, C. Ré, C. Koch. *Probabilistic Databases*. Synthesis Lectures on Data Management, Morgan & Claypool Publishers, 2011.
30. Y. Takahashi. Fuzzy database query languages and their relational completeness theorem. *IEEE Trans. Knowledge and Data Engineering* 5:122–125, February 1993.
31. G. Takeuti and S. Titani. Globalization of intuitionistic set theory. *Annals of Pure and Applied Logic* 33: 195–211, 1987.