Relational Similarity-Based Databases I

Vilem Vychodil (Palacky University, Olomouc)

DATA ANALYSIS AND MODELING LAB
Palacky University, Olomouc, Czech Republic

european social fund in the czech republic
EUROPEAN UNION
MINISTRY OF EDUCATION, YOUTH AND SPORTS
OP Education for Competitiveness
INVESTMENTS IN EDUCATION DEVELOPMENT

Relational Similarity-Based Databases

general relational model of data:

- generalization of the classic RM (E. F. Codd)
- **similarity** relations on domains
- **ranks** assigned to tuples

motivation:

1. *similarity-based queries*
   "Show all houses that are sold for $600,000."
2. *approximate dependencies* in data
   "Do houses in similar locations have similar prices?"

goal:

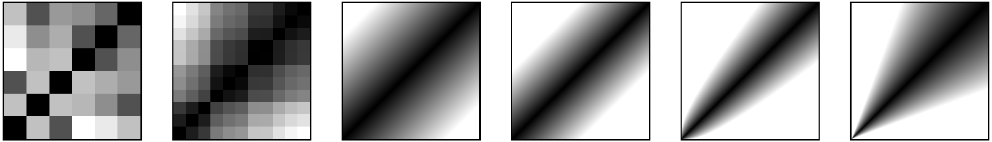
- rank-aware approach with solid logical foundations (logics of residuated structures)
- focus on all DB aspects (foundations, querying, dependencies, algorithms, ...)

Similarity-Based Query: An Example

id	*type*	*location*	*built*	*bdrm*	*sqft*	*price*
45	Single Family	Green St	1979	3	1180	$754,000
66	Ranch	Fulton St	1977	2	2400	$998,000
78	Single Family	Purdue Ave	1962	4	1360	$850,000
81	Residential	Hamilton Ave	1961	5	1450	$986,000
82	Condominium	Fulton St	1998	2	650	$540,000
87	Single Family	Bryant St	1927	3	1230	$854,000
95	Log Cabin	Schembri Ln	1936	2	750	$754,000
97	Penthouse	Cabrillo St	1984	1	932	$720,000

Similarity-Based Query: An Example

	id	*type*	*location*	*built*	*bdrm*	*sqft*	*price*
0.890	82	Condominium	Fulton St	1998	2	650	$540,000
0.595	97	Penthouse	Cabrillo St	1984	1	932	$720,000
0.535	87	Single Family	Bryant St	1927	3	1230	$854,000
0.487	66	Ranch	Fulton St	1977	2	2400	$998,000
0.472	45	Single Family	Green St	1979	3	1180	$754,000
0.277	81	Residential	Hamilton Ave	1961	5	1450	$986,000
0.213	95	Log Cabin	Schembri Ln	1936	2	750	$754,000



"Show houses located in Old Palo Alto and sold for $600,000."

Example (“Show houses with prices similar to $600,000” in SQL)

```
CREATE TABLE house (
  ⋮
  price NUMERIC NOT NULL
);

INSERT INTO house VALUES ···

CREATE FUNCTION sim (NUMERIC, NUMERIC) RETURNS NUMERIC AS
  'SELECT least (1, greatest (0, 1 + abs ($1 - $2) / -200000.0));'
  LANGUAGE SQL;

SELECT *, sim (price, 600000) AS rank
  FROM house
  ORDER BY sim DESC
  LIMIT 5;
```

Related work (1 of 2)

Fagin at al.

- R. Fagin. Combining fuzzy information: an overview.
  ACM SIGMOD Record 31(2):109–118, 2002.
- Natsev, Chang, Smith, Li, Vitter: Supporting incremental join queries on ranked inputs.
  In: *VLDB 2001*, pp. 281–290.
- Cohen, Sagiv: An incremental algorithm for computing ranked full disjunctions.
  In: *PODS 2005*, pp. 98–107.

RankSQL + related research

- Li, Chang, Ilyas, Song: RanSQL: Query Algebra and Optimization for Relational top-k queries.
  In: *ACM SIGMOD* 2005, pages 131–142, 2005.
- Illyas, Aref, Elmagarmid: Supporting top-k join queries in relational databases.
  The VLDB Journal 13:207–221, 2004.

Related work (2 of 2)

Extensions of Codd's model employing fuzzy logic

- several approaches (including "fuzzy data"), many papers
- Raju, Majumdar, Fuzzy functional dependencies and lossless join decomposition of fuzzy relational database systems.
  ACM Trans. Database Systems 13:129–166, 1988.

Extensions of Codd's model employing probability

- **different** both semantically and technically (degrees of belief $\neq$ degrees of truth)
- D. Dey and S. Sarkar S. A probabilistic relational model and algebra.
  ACM Trans. Dat. Syst. 21:339–369, 1996.
- Fuhr, Rölleke, A probabilistic relational algebra for the integration of information retrieval and database systems.
  ACM Trans. Information Systems 15:32–66, 1997.
- Dalvi, Ré, Suciu, Probabilistic databases: diamonds in the dirt.
  Communications of the ACM 52:86–94, 2009.

Preliminaries from the Classic RM

attributes = *names for columns of ranked data tables*

- Y: denumerable set of all attributes
- attributes denoted $y, y', y_1, y_2, \ldots$

relation schemes = finite subsets $R \subseteq Y$

- relation schemes determine table columns (as in the Codd model)

cartesian (direct) product =

- set $\prod_{i \in I} A_i$ of all maps $f\colon I \to \bigcup_{i \in I} A_i$ such that $f(i) \in A_i$ for all $i \in I$ (for given I-indexed set $\{A_i \mid i \in I\}$ of sets)

domains =

- sets of attribute values (D_y is domain of $y \in Y$)

tuples =

- elements of $\prod_{y \in R} D_y$ ($R \subseteq Y$)
- denoted $r \in \mathrm{Tupl}(R)$ (r is tuple on R over D_y's); $r(y)$ is called y-value of r

Motivation for Our Approach (1 of 3)

we want:

- *similarity-based queries* answered by *imprecise matches*

generalized RM:

- Shift from *two-element Boolean algebra* to (*complete*) *residuated lattices*
- Structure of matches in the classic RM $\Longrightarrow$ the generalized RM

starting with the classic RM: $\mathcal{D}$ on R can be viewed:

$$\mathcal{D} \colon \textstyle\prod_{y \in R} D_y \to \{0, 1\}$$

so that for only finitely many tuples $r \in \prod_{y \in R} D_y$: $\mathcal{D}(r) = 1$.

interpretation (if $\mathcal{D}$ is answer to Q)

$\mathcal{D}(r) = 1$ means "the tuple r matches the query Q"

$\mathcal{D}(r) = 0$ means "the tuple r does not match the query Q"

Motivation for Our Approach (2 of 3)

take a **partially ordered set** $\langle L, \leq, 0\rangle$ instead of $\langle\{0,1\}, \leq\rangle$:

$$\mathcal{D}\colon \prod_{y \in R} D_y \to L$$ (**ranked data table**, an RDT)

so that for only finitely many tuples $r \in \prod_{y \in R} D_y$: $\mathcal{D}(r) \neq 0$

desirable properties of L and $\leq$:

- lower and upper bound in L (0 for *no match*, 1 for *full match*),
- $\langle L, \leq\rangle$ is a complete lattice;
- additional operations on L to *aggregate degrees*.

conjunctive aggregator $\otimes$ motivated by *natural join* (for $\mathcal{D}_1$ on $R \cup S$ and $\mathcal{D}_2$ on $S \cup T$):

$$(\mathcal{D}_1 \bowtie \mathcal{D}_2)(rst) = \mathcal{D}_1(rs) \otimes \mathcal{D}_2(st),$$ (R, S, T are pairwise disjoint)

with $\otimes\colon \{0,1\}^2 \to \{0,1\}$ defined by $1 \otimes 1 = 1$ and $1 \otimes 0 = 0 \otimes 1 = 0 \otimes 0 = 0$

Motivation for Our Approach (3 of 3)

in our setting: $\otimes\colon L^2 \to L$ such that $\langle L, \otimes, 1\rangle$ is a commutative monoid and $\otimes$ is distributive w.r.t. $\bigvee$ (stronger condition than monotony):

$$a \otimes \bigvee_{i\in I} b_i = \bigvee_{i\in I}(a \otimes b_i)$$

which is equivalent to: $\langle L, \otimes, 1\rangle$ is a commutative monoid and there is (uniquely given) $\to\colon L^2 \to L$ such that

$$a \otimes b \leq c \quad \text{iff} \quad a \leq b \to c$$

(**adjointness property**)

altogether: $\mathbf{L} = \langle L, \wedge, \vee, \otimes, \to, 0, 1\rangle$ is a (**complete**) **residuated lattice**, i.e.

- $\langle L, \wedge, \vee, 0, 1\rangle$ … (complete) lattice,
- $\langle L, \otimes, 1\rangle$ … commutative monoid,
- $\langle \otimes, \to\rangle$ … adjoint pair ($a \otimes b \leq c$ iff $a \leq b \to c$).

Residuated Structures in Fuzzy Logics

- fuzzy logic in **broad sense**: any application of fuzzy approach in modeling
  - Zadeh L. A.: Fuzzy sets. *Inf. Control* (1965)
  - simple observations on handling of vagueness
- fuzzy logic in **narrow sense**: mathematical fuzzy logic
  - Hájek P.: *Metamathematics of Fuzzy Logic.* (1998)
  - Basic Logic (BL-logic), propositional/predicate; *logic of continuous t-norms*
  - Höhle, Esteva, Godo, Gottwald, Montagna, ...
  - various logical calculi (MTL-logic)

basic principles:

- adjointness derived from graded *modus ponens*
- propositions allowed to have "intermediate truth degrees", like:

$$||\text{value } x \text{ is similar to value } y||_{\mathbf{M}} = 0.9$$

- our case: $||\varphi||_{\mathbf{M},v}$ (φ formula; $\mathbf{M}$ database instance; v induced by tuples)

Domains with Similarities

similarity relations on domains (needed for approximate matches)
each domain D_y equipped with map $\approx_y : D_y \times D_y \to L$ satisfying:

(Ref) for each $d \in D_y$: $d \approx_y d = 1$,

(Sym) for each $d_1, d_2 \in D_y$: $d_1 \approx_y d_2 = d_2 \approx_y d_1$, and (optionally):

(Sep) for each $d_1, d_2 \in D_y$: $d_1 \approx_y d_2 = 1$ iff d_1 equals d_2, and

(Tra) for each $d_1, d_2, d_3 \in D_y$: $d_1 \approx_y d_2 \otimes d_2 \approx_y d_3 \leq d_1 \approx_y d_3$.

so-called **similarity relation**

domain with similarity $= \langle D_y, \approx_y \rangle$, where

- D_y is domain of attribute $y \in Y$;
- $\approx_y$ is similarity on D_y.

notes:

- interpretation: $u \approx_y v =$ degree to which u and v are similar
- boundary case: strict identity

Ranked Data Tables over Domains with Similarities

central notion to our model:

- formal counterpart to *relations on relation schemes* from Codd's model
- in mathematical fuzzy logic: interpretations of relation symbols

Definition (ranked data table)

Let $R \subseteq Y$ be a relation scheme and each $\langle D_y, \approx_y \rangle$ be a domain with similarity ($y \in R$). A **ranked data table on R over $\{\langle D_y, \approx_y \rangle \mid y \in R\}$** is any map $\mathcal{D}\colon \mathrm{Tupl}(R) \to L$ so that for only finitely many tuples $r \in \prod_{y \in R} D_y$: $\mathcal{D}(r) \neq 0$.

notes:

- RDTs are denoted $\mathcal{D}, \mathcal{D}', \mathcal{D}_1, \ldots$
- RDT on R over $\{\langle D_y, \approx_y \rangle \mid y \in R\}$ = fuzzy relation between D_y
- degree $\mathcal{D}(r)$ is called a **rank of r in $\mathcal{D}$**

Special Cases of RDTs

two important special cases:

Definition (RDTs on empty relation schemes)

For each $a \in L$, define $a_{\emptyset} = \{\langle \emptyset, a \rangle\}$.

Definition (singleton RDTs)

For each $y \in Y$ and $d \in D_y$, define $[y{:}d] = \{\langle \{\langle y, d \rangle\}, 1 \rangle\}$.

notes:

- $a_{\emptyset}$ is RDT on $R = \emptyset$ such that $a_{\emptyset}(\emptyset) = a$ (C. J. Date: $0_{\emptyset} = \texttt{TABLE\_DUM}$, $1_{\emptyset} = \texttt{TABLE\_DEE}$)
- $[y{:}d]$ is RDT on $R = \{y\}$ such that $[y{:}d](r) = \begin{cases} 1, & \text{if } r(y) = d, \\ 0, & \text{otherwise} \end{cases}$

Notes on Generalization of Codd's Model of Data

classic relational model results by:

- taking two-valued Boolean algebra for $\mathbf{L}$ (complete residuated lattice);
- considering each $\approx_y$ to be identity relation on D_y

consequence: all ranks become 1 (match) and 0 (no match)

nonranked RDT

- all ranks are from $\{0, 1\} \subseteq L$, $\mathbf{L}$ is arbitrary;
- stored data prior to querying;

Important feature of our model: stored data $=$ results of queries

RDTs represent both

- **stored data**, and
- **results of queries**.

Notes on Domain Similarities and Ranks

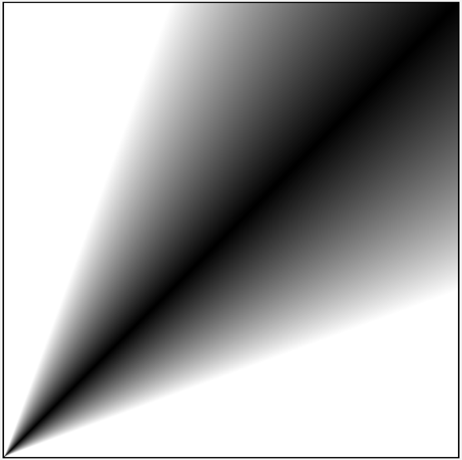
Where do similarities come from?

- can be assigned by an expert:
  - finite $\mathbf{L}$ or a finite subset of infinite linear $\mathbf{L}$;
  - Likert scale $L = \{1, \ldots, 5\}$ of degrees of satisfaction (Miller's 7 ± 2 phenomenon);
- can be determined based on "distance":
  - $\mathbf{L}$ on $[0, 1]$ with $\otimes$ being continuous Archimedean t-norm;
  - (pseudo)metric $\quad\Longmapsto\quad$ $\otimes$-transitive similarity;
- similarities are *purpose dependent*;
- implementation remark: can be stored (as data) / computed on demand.

Where do ranks come from?

- appear from nonranked data after performing similarity-based queries,
- can be assigned by experts,
- important aspect: *comparative meaning of truth degrees*.

Example (similarity on domain of "house prices")



$$d_1 \approx_{price} d_2 = s\big(|\log_b d_1 - \log_b d_2|\big)$$

$$b = 1 + 10^{-4}$$

$$s(x) = 1 - x \cdot 10^{-4}$$

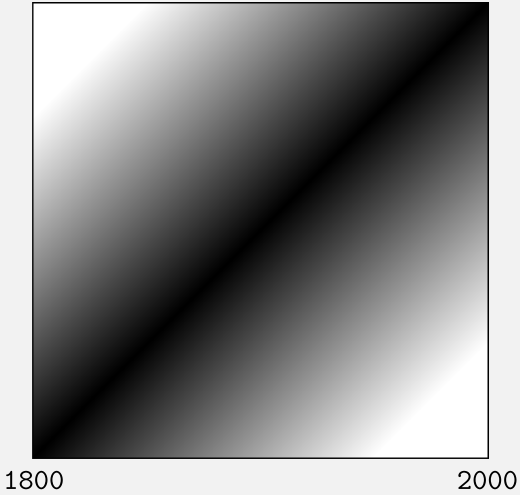
example:

$$\$1{,}000 \approx_{price} \$2{,}000 = 0.306$$

$$\$100{,}000 \approx_{price} \$101{,}000 = 0.990$$

$$\vdots \qquad \vdots$$

Example (similarity on domain of "construction years")



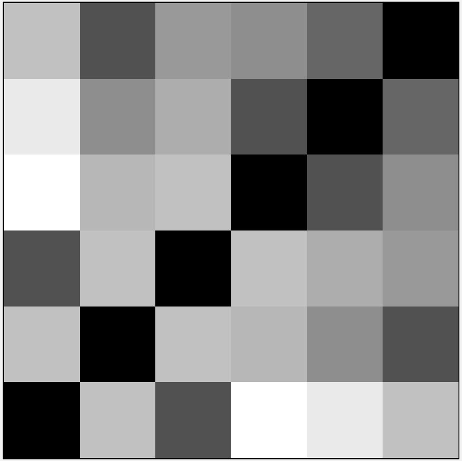
$$d_1 \approx_{year} d_2 = s\big(|d_1 - d_2|\big)$$

$$s(x) = 1 - x \cdot 150^{-1}$$

example:

$$1800 \approx_{year} 1840 = 0.733$$
$$1960 \approx_{year} 2000 = 0.733$$
$$\vdots \qquad \vdots$$

Example (similarity on domain of "property types")



Single Family

Residential

Ranch

Penthouse

Log Cabin

Condominium

Operations with RDTs

goal:

- propose set of (basic) operations with RDTs
- **purpose:** querying by performing operations with RTDs (relation algebra)
- **questions:** basic/derived operations, expressive power, …

groups of operations in our model:

- counterparts to boolean operations (union, intersection, residuum)
- natural join (and cross join)
- projection and residuated division
- similarity-based restrictions
- kernel and support
- renaming attributes

derived operations and extensions (II. part)

Counterparts to Boolean Intersection and Union

Definition

For RDTs $\mathcal{D}_1$ and $\mathcal{D}_2$ on relation scheme R, we define

$$(\mathcal{D}_1 \cup \mathcal{D}_2)(r) = \mathcal{D}_1(r) \vee \mathcal{D}_2(r),$$
$$(\mathcal{D}_1 \cap \mathcal{D}_2)(r) = \mathcal{D}_1(r) \wedge \mathcal{D}_2(r),$$
$$(\mathcal{D}_1 \otimes \mathcal{D}_2)(r) = \mathcal{D}_1(r) \otimes \mathcal{D}_2(r),$$

for all tuples r on R. $\mathcal{D}_1 \cup \mathcal{D}_2$ is called a **union** of $\mathcal{D}_1$ and $\mathcal{D}_2$; $\mathcal{D}_1 \cap \mathcal{D}_2$ and $\mathcal{D}_1 \otimes \mathcal{D}_2$ are called the **$\wedge$-intersection** and **$\otimes$-intersection** of $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively.

idempotent vs. non-indempotent conjunction:

- RDT $\mathcal{D}$ on relation scheme R is called **idempotent (with respect to** $\otimes$**)** if $\mathcal{D} \otimes \mathcal{D} = \mathcal{D}$
- example: for $\mathcal{D}_1(r) = 0.5$ and $\mathcal{D}_2(r) = \cdots = \mathcal{D}_k(r) = 0.98$, we distinguish:
  - worst-match semantics: $(\mathcal{D}_1 \cap \cdots \cap \mathcal{D}_k)(r) = 0.5$ (also if $\mathcal{D}_2(r) = \cdots = \mathcal{D}_k(r) = 0.5$)
  - all-match semantics: $(\mathcal{D}_1 \otimes \cdots \otimes \mathcal{D}_k)(r) = 0.5 \cdot 0.98^{k-1}$ for Goguen $\otimes$
    $(\mathcal{D}_1 \otimes \cdots \otimes \mathcal{D}_k)(r) = 0.5^k \lll 0.5 \cdot 0.98^{k-1}$ if $\mathcal{D}_2(r) = \cdots = \mathcal{D}_k(r) = 0.5$

Operations Based on Residuated Implication

issues with finiteness:

- componentwise application of $\rightarrow$: $(\mathcal{D}_1 \rightarrow \mathcal{D}_2)(r) = \mathcal{D}_1(r) \rightarrow \mathcal{D}_2(r)$
- if at least one D_y is infinite: $(\mathcal{D}_1 \rightarrow \mathcal{D}_2)(r) = 1$ for *infinitely many* r

(one possible) solution: for arbitrary degrees $a, b, c \in L$, define $b \twoheadrightarrow^{a} c \in L$ as follows:

$$b \twoheadrightarrow^{a} c = a \otimes (b \rightarrow c)$$

(**a-residuum of $b \in L$ with respect to $c \in L$**)

Definition (residuum of RDTs)

For RDTs $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3$ on R, we put

$$\big(\mathcal{D}_1 \twoheadrightarrow^{\mathcal{D}_3} \mathcal{D}_2\big)(r) = \mathcal{D}_1(r) \twoheadrightarrow^{\mathcal{D}_3(r)} \mathcal{D}_2(r)$$

for all tuples r. $\mathcal{D}_1 \twoheadrightarrow^{\mathcal{D}_3} \mathcal{D}_2$ is a **residuum** of $\mathcal{D}_1$ with respect to $\mathcal{D}_2$ which ranges over $\mathcal{D}_3$.

note:

- $\mathcal{D}_1 \twoheadrightarrow^{\mathcal{D}_3} \mathcal{D}_2 \subseteq \mathcal{D}_3$ (result of $\twoheadrightarrow$ in an RDT)

Theorem (properties of $\rightharpoondown$)

1. $b \rightharpoondown^1 c = b \rightarrow c$,
2. $1 \rightharpoondown^a c = 1 \rightharpoondown^c a = a \otimes c$,
3. $0 \rightharpoondown^a c = b \rightharpoondown^a 1 = a$,
4. $b \rightharpoondown^0 c = b \rightharpoondown^b 0 = 1 \rightharpoondown^b 0 = 0$,
5. $b \rightharpoondown^a c \leq b \rightharpoondown^1 (a \otimes c)$,
6. $\rightharpoondown$ *is monotone in the first and in the third argument,*
7. $\rightharpoondown$ *is antitone in the second argument,*
8. $a \rightharpoondown^a b \leq a \wedge b$,
9. *if* $\mathbf{L}$ *is divisible, then* $a \rightharpoondown^a b = a \wedge b$,
10. *if* $b \leq c$, *then* $b \rightharpoondown^a c = a$,
11. *if* $\mathbf{L}$ *is a linear* Π*-algebra, then* $b \leq c$ *iff* $b \rightharpoondown^a c = a$ *for all* $a > 0$,
12. $b \rightharpoondown^b c = c$ *iff there is* $x \in L$ *such that* $1 \rightharpoondown^x b = c$,
13. $1 \rightharpoondown^a b \leq c$ *iff* $a \leq b \rightharpoondown^1 c$.

$\langle \otimes, \rightarrow \rangle$ vs. $\rightarrowtriangle$

Theorem

Let $\mathbf{L} = \langle L, \wedge, \vee, \rightarrowtriangle, 0, 1 \rangle$ *be a structure such that* $\langle L, \wedge, \vee, 0, 1 \rangle$ *is a bounded lattice and* $\rightarrowtriangle$ *be a ternary operation satisfying the following conditions:*

$$
\begin{aligned}
1 \rightarrowtriangle^a 1 &= a, \\
1 \rightarrowtriangle^a b &= 1 \rightarrowtriangle^b a, \\
1 \rightarrowtriangle^a (1 \rightarrowtriangle^b c) &= 1 \rightarrowtriangle^c (1 \rightarrowtriangle^a b), \\
1 \rightarrowtriangle^a b \leq c \quad &\textit{iff} \quad a \leq b \rightarrowtriangle^1 c
\end{aligned}
$$

for all $a, b, c \in L$. *Then,* $\mathbf{L}' = \langle L, \wedge, \vee, \otimes, \rightarrow, 0, 1 \rangle$, *where* $a \otimes b = 1 \rightarrowtriangle^a b$ *and* $a \rightarrow b = a \rightarrowtriangle^1 b$ *for all* $a, b \in L$, *is a residuated lattice.*

corollary:

The class of all bounded lattices with $\rightarrowtriangle$ *satisfying the conditions above is a variety which is term equivalent to the variety of residuated lattices.*

Theorem (properties of operations $\otimes$, $\cap$, $\cup$, $\rightarrowtriangle$)

1. $\mathcal{D}_1 \otimes (\mathcal{D}_2 \cup \mathcal{D}_3) = (\mathcal{D}_1 \otimes \mathcal{D}_2) \cup (\mathcal{D}_1 \otimes \mathcal{D}_3)$

If $\mathbf{L}$ *is prelinear or divisible, then*

2. $\mathcal{D}_1 \otimes (\mathcal{D}_2 \cap \mathcal{D}_3) = (\mathcal{D}_1 \otimes \mathcal{D}_2) \cap (\mathcal{D}_1 \otimes \mathcal{D}_3)$,
3. $\mathcal{D}_1 \cap (\mathcal{D}_2 \cup \mathcal{D}_3) = (\mathcal{D}_1 \cap \mathcal{D}_2) \cup (\mathcal{D}_1 \cap \mathcal{D}_3)$.

If $\mathcal{D}$ *is nonranked, then*

4. $\mathcal{D}_1 \rightarrowtriangle^{\mathcal{D}} (\mathcal{D}_2 \rightarrowtriangle^{\mathcal{D}} \mathcal{D}_3) = \mathcal{D}_2 \rightarrowtriangle^{\mathcal{D}} (\mathcal{D}_1 \rightarrowtriangle^{\mathcal{D}} \mathcal{D}_3)$,
5. $(\mathcal{D}_1 \otimes \mathcal{D}_2) \rightarrowtriangle^{\mathcal{D}} \mathcal{D}_3 = \mathcal{D}_1 \rightarrowtriangle^{\mathcal{D}} (\mathcal{D}_2 \rightarrowtriangle^{\mathcal{D}} \mathcal{D}_3)$,
6. $\mathcal{D}_1 \rightarrowtriangle^{\mathcal{D}} \mathcal{D}_2 = ((\mathcal{D}_1 \rightarrowtriangle^{\mathcal{D}} \mathcal{D}_2) \rightarrowtriangle^{\mathcal{D}} \mathcal{D}_2) \rightarrowtriangle^{\mathcal{D}} \mathcal{D}_2$,
7. $\mathcal{D}_1 \rightarrowtriangle^{\mathcal{D}} (\mathcal{D}_2 \cap \mathcal{D}_3) = (\mathcal{D}_1 \rightarrowtriangle^{\mathcal{D}} \mathcal{D}_2) \cap (\mathcal{D}_1 \rightarrowtriangle^{\mathcal{D}} \mathcal{D}_3)$,
8. $(\mathcal{D}_1 \cup \mathcal{D}_2) \rightarrowtriangle^{\mathcal{D}} \mathcal{D}_3 = (\mathcal{D}_1 \rightarrowtriangle^{\mathcal{D}} \mathcal{D}_3) \cap (\mathcal{D}_2 \rightarrowtriangle^{\mathcal{D}} \mathcal{D}_3)$,
9. $(\mathcal{D}_1 \rightarrowtriangle^{\mathcal{D}} \mathcal{D}_2) \otimes (\mathcal{D}_2 \rightarrowtriangle^{\mathcal{D}} \mathcal{D}_3) \subseteq \mathcal{D}_1 \rightarrowtriangle^{\mathcal{D}} \mathcal{D}_3$.

If $\mathbf{L}$ *is prelinear, then*

10. $\mathcal{D}_1 \rightarrowtriangle^{\mathcal{D}} (\mathcal{D}_2 \cup \mathcal{D}_3) = (\mathcal{D}_1 \rightarrowtriangle^{\mathcal{D}} \mathcal{D}_2) \cup (\mathcal{D}_1 \rightarrowtriangle^{\mathcal{D}} \mathcal{D}_3)$,
11. $(\mathcal{D}_1 \cap \mathcal{D}_2) \rightarrowtriangle^{\mathcal{D}} \mathcal{D}_3 = (\mathcal{D}_1 \rightarrowtriangle^{\mathcal{D}} \mathcal{D}_3) \cup (\mathcal{D}_2 \rightarrowtriangle^{\mathcal{D}} \mathcal{D}_3)$.

Natural Join

Definition (equality-based natural join)

If $\mathcal{D}_1$ is an RDT on relation scheme $R \cup S$ and $\mathcal{D}_2$ is an RDT of relation scheme $S \cup T$ such that $R \cap S = R \cap T = S \cap T = \emptyset$ (i.e., R, S, and T are pairwise disjoint), then the (**equality-based**) **natural join** of $\mathcal{D}_1$ and $\mathcal{D}_2$ is an RDT $\mathcal{D}_1 \bowtie \mathcal{D}_2$ on relation scheme $R \cup S \cup T$ defined by

$$\big(\mathcal{D}_1 \bowtie \mathcal{D}_2\big)(rst) = \mathcal{D}_1(rs) \otimes \mathcal{D}_2(st),$$

for each $r \in \mathrm{Tupl}(R)$, $s \in \mathrm{Tupl}(S)$, and $t \in \mathrm{Tupl}(T)$.

special cases:

- *cross join*: special case for $S = \emptyset$
- $\otimes$-*intersection*: special case for $R = \emptyset$ and $T = \emptyset$

basic properties:

- $\bowtie$ is *commutative* and *associative* (not idempotent in general); notation $\bowtie_{i=1}^{n} \mathcal{D}_i$
- $0_\emptyset$ is *annihilator*; $1_\emptyset$ is *neutral element*

Notes on Natural Joins

size of natural and cross joins:

- $|\mathcal{D}_1 \bowtie \mathcal{D}_2| \leq |\mathcal{D}_1| \cdot |\mathcal{D}_2|$
- but the converse inequality does not hold in general (not even in case of RDTs on disjoint relaiton schemes)

equality-based restriction via natural joins:

$$(\mathcal{D} \bowtie [y{:}d])(r) = \begin{cases} \mathcal{D}(r), & \text{if } r(y) = d, \\ 0, & \text{otherwise} \end{cases}$$

for all $r \in \mathrm{Tupl}(R)$

consequences:

- $\mathcal{D} \bowtie [y{:}d]$ = **equality-based restriction** of $\mathcal{D}$ consisting of tuples with y-values d
- ranks of those tuples in $\mathcal{D}$ are preserved

Projection

captures: existentially quantified queries (some A is B)

Definition (projection)

If $\mathcal{D}$ is an RDT on T, the **projection** $\pi_R(\mathcal{D})$ of $\mathcal{D}$ onto $R \subseteq T$ is defined by

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

for each $r \in \mathrm{Tupl}(R)$.

special cases:

- $\big(\pi_\emptyset(\mathcal{D})\big)(\emptyset) = \bigvee_{t \in \mathrm{Tupl}(T)} \mathcal{D}(t)$
- $\pi_T(\mathcal{D}) = \mathcal{D}$ (if $\mathcal{D}$ is RDT on relation scheme T)

Theorem (selected properties of projection)

For any $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}$ *on* R*:*

1. *if* $R_1 \subseteq R_2$*, then* $\pi_{R_1}(\pi_{R_2}(\mathcal{D})) = \pi_{R_1}(\mathcal{D})$*,*
2. $\pi_R(\mathcal{D}_1 \cup \mathcal{D}_2) = \pi_R(\mathcal{D}_1) \cup \pi_R(\mathcal{D}_2)$*,*
3. $\pi_R(\mathcal{D}_1 \cap \mathcal{D}_2) \subseteq \pi_R(\mathcal{D}_1) \cap \pi_R(\mathcal{D}_2)$*,*
4. $\pi_R(\mathcal{D}_1 \otimes \mathcal{D}_2) \subseteq \pi_R(\mathcal{D}_1) \otimes \pi_R(\mathcal{D}_2)$*,*

Let $\mathcal{D}_1$ *and* $\mathcal{D}_2$ *be RDTs on relation schemes* $R \cup S$ *and* $S \cup T$ *such that* $R \cap S = R \cap T = S \cap T = \emptyset$*. Furthermore, let* $\{\mathcal{D}_i \,|\, i \in I\}$ *be a finite set of RDTs on* R_i $(i \in I)$*, and let* $\mathcal{D}$ *be an RDT on* $R = \bigcup_{i \in I} R_i$*. Then,*

5. $\pi_{R \cup S}(\mathcal{D}_1 \bowtie \mathcal{D}_2) = \mathcal{D}_1 \bowtie \pi_S(\mathcal{D}_2)$*,*
6. $\pi_{R_i}(\bowtie_{j \in I} \mathcal{D}_j) \subseteq \mathcal{D}_i$ *for all* $i \in I$*,*
7. $\mathcal{D}^{|I|} \subseteq \bowtie_{i \in I} \pi_{R_i}(\mathcal{D})$*,*
8. *if* $\mathcal{D}$ *is idempotent, then* $\mathcal{D} \subseteq \bowtie_{i \in I} \pi_{R_i}(\mathcal{D})$*.* □

semijoin: $\mathcal{D}_1 \ltimes \mathcal{D}_2 = \pi_{R \cup S}(\mathcal{D}_1 \bowtie \mathcal{D}_2) = \mathcal{D}_1 \bowtie \pi_S(\mathcal{D}_2)$ ($\otimes$ is distributive over $\bigvee$)

Residuated Division

captures: universaly quantified queries (all A's are B's)

Definition (residuated division)

Let $\mathcal{D}_1$ be an RDT on R, let $\mathcal{D}_2$ be an RDT on $S \subseteq R$, and let $\mathcal{D}_3$ be an RDT on $T = R \setminus S$. Then, a **division** $\mathcal{D}_1 \div^{\mathcal{D}_3} \mathcal{D}_2$ of $\mathcal{D}_1$ by $\mathcal{D}_2$ which ranges over $\mathcal{D}_3$ is an RDT on T defined by

$$\big(\mathcal{D}_1 \div^{\mathcal{D}_3} \mathcal{D}_2\big)(t) = \textstyle\bigwedge_{s \in \mathrm{Tupl}(S)} \big(\mathcal{D}_2(s) \rightarrowtriangle^{\mathcal{D}_3(t)} \mathcal{D}_1(st)\big),$$

for each $t \in \mathrm{Tupl}(T)$.

meaning:

$\mathcal{D}_2$ reliable suppliers, $\mathcal{D}_3$ solvent customers, $\mathcal{D}_1$ suppliers frequently used by customers, result = solvent customers frequently using all reliable suppliers

special cases:

- **graded containment**: $\big(\mathcal{D}_1 \div^{1_\emptyset} \mathcal{D}_2\big)(\emptyset) = \textstyle\bigwedge_{r \in \mathrm{Tupl}(R)} \big(\mathcal{D}_2(r) \rightarrow \mathcal{D}_1(r)\big)$

Derived Notions

subsethood and **similarity** degrees (note the role of $a_\emptyset$ and $a \in L$):

$$S(\mathcal{D}_1, \mathcal{D}_2) = \big(\mathcal{D}_2 \div^{1_\emptyset} \mathcal{D}_1\big)(\emptyset)$$
$$E(\mathcal{D}_1, \mathcal{D}_2) = S(\mathcal{D}_1, \mathcal{D}_2) \wedge S(\mathcal{D}_2, \mathcal{D}_1)$$

degrees of joinability:

Let $\mathcal{D}_i$ be RDTs on relation schemes R_i ($i \in I$ for finite I). Then

$$Jnd(\{\mathcal{D}_i \,|\, i \in I\}) = \textstyle\bigwedge_{i \in I} S\big(\mathcal{D}_i, \pi_{R_i}(\bowtie_{j \in I} \mathcal{D}_j)\big)$$

is a **degree of joinability** of RDTs $\mathcal{D}_i$ ($i \in I$);
RDTs $\mathcal{D}_i$ ($i \in I$) **join completely** if $Jnd(\{\mathcal{D}_i \,|\, i \in I\}) = 1$

degrees of decomposability:

Let $\mathcal{D}$ be an RDT on relation schemes $R = \bigcup_{i \in I} R_i$ where I is finite. Then

$$Dcd(\mathcal{D}, \{R_i \,|\, i \in I\}) = E\big(\mathcal{D}, \bowtie_{i \in I} \pi_{R_i}(\mathcal{D})\big)$$

is a **degree of decomposability** of $\mathcal{D}$ with respect to R_i ($i \in I$);
$\mathcal{D}$ has a **nonloss decomposition** if $Dcd(\mathcal{D}, \{R_i \,|\, i \in I\}) = 1$

Concept-Forming Operators Induced by RDTs

Definition

For an RDT $\mathcal{D}_1$ on R; $S \subseteq R$, $T = R \setminus S$; and nonranked RDTs $\mathcal{D}_y$ on $\{y\}$ ($y \in R$), put

$$f^{S,T}_{\mathcal{D}_1,\{\mathcal{D}_y \,|\, y \in R\}}(\mathcal{D}_2) = \mathcal{D}_1 \div^{\bowtie_{y \in T} \mathcal{D}_y} \mathcal{D}_2$$

for any $\mathcal{D}_2$ on S.

notes:

- $\mathcal{D}_1$ and $\mathcal{D}_y$ ($y \in R$) induce $f^{S,T}_{\mathcal{D}_1,\{\mathcal{D}_y \,|\, y \in R\}}$ with respect to S and T (in this order)
- **dyadic case**: for $R = \{x, y\}$, $\mathcal{D}_x$, $\mathcal{D}_y$, $\mathcal{D} \subseteq \mathcal{D}_x \bowtie \mathcal{D}_y$, $\mathcal{D}_A \subseteq \mathcal{D}_x$, and $\mathcal{D}_B \subseteq \mathcal{D}_y$:

$$f^{\{x\},\{y\}}_{\mathcal{D},\{\mathcal{D}_x,\mathcal{D}_y\}}(\mathcal{D}_A) = \mathcal{D} \div^{\mathcal{D}_y} \mathcal{D}_A, \qquad f^{\{y\},\{x\}}_{\mathcal{D},\{\mathcal{D}_x,\mathcal{D}_y\}}(\mathcal{D}_B) = \mathcal{D} \div^{\mathcal{D}_x} \mathcal{D}_B,$$

  express concept-forming operators (denoted by $^{\uparrow}$ and $^{\downarrow}$) used in the dyadic FCA of object-attribute relational data with graded attributes (generalizes to n-adic case)

Similarity-Based Restriction

Definition (similarity-based restriction)

For any attributes $y_1, y_2 \in R$ with the same domains with similarity we define the **similarity-based restriction** $\sigma_{y_1 \approx y_2}(\mathcal{D})$ of $\mathcal{D}$ by $y_1 \approx y_2$ which is an RDT on R defined by

$$(\sigma_{y_1 \approx y_2}(\mathcal{D}))(r) = \mathcal{D}(r) \otimes r(y_1) \approx_{y_1} r(y_2),$$

for all $r \in \mathrm{Tupl}(R)$.

representation by natural joins: $\sigma_{y_1 \approx y_2}(\mathcal{D}) = \mathcal{D} \bowtie \mathcal{D}_{y_1 \approx y_2}$, where for all $r \in \mathrm{Tupl}(R)$,

$$\mathcal{D}_{y_1 \approx y_2}(r(\{y_1, y_2\})) = \begin{cases} r(y_1) \approx_{y_1} r(y_2), & \text{if } \mathcal{D}(r) > 0, \\ 0, & \text{otherwise.} \end{cases}$$

restriction based on domain values:

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

derived operation:

$$\sigma_{y \approx d}(\mathcal{D}) = \pi_R(\sigma_{y \approx y'}(\mathcal{D} \bowtie [y'{:}d])).$$

Theorem (properties of similarity-based restrictions)

The following are true (if both left and right-hand sides exist):

1. $\pi_S(\sigma_{y\approx z}(\mathcal{D})) = \sigma_{y\approx z}(\pi_S(\mathcal{D}))$ *if* $\mathcal{D}$ *is an RDT on* R *and* $R \cap \{y, z\} \subseteq S$,
2. $\sigma_{y\approx z}(\mathcal{D}_1 \bowtie \mathcal{D}_2) = \sigma_{y\approx z}(\mathcal{D}_1) \bowtie \mathcal{D}_2$ *if* $\mathcal{D}_2$ *is an RDT on* R_2 *and* $\{y, z\} \cap R_2 = \emptyset$,
3. $\sigma_\theta(\mathcal{D}_1 \cup \mathcal{D}_2) = \sigma_\theta(\mathcal{D}_1) \cup \sigma_\theta(\mathcal{D}_2)$,
4. $\sigma_\theta(\mathcal{D}_1 \cap \mathcal{D}_2) \subseteq \sigma_\theta(\mathcal{D}_1) \cap \mathcal{D}_2$,
5. $\sigma_\theta(\mathcal{D}_1 \otimes \mathcal{D}_2) = \sigma_\theta(\mathcal{D}_1) \otimes \mathcal{D}_2$,
6. $\mathcal{D}_1 \rightarrowtriangle^{\sigma_\theta(\mathcal{D}_3)} \mathcal{D}_2 = \sigma_\theta(\mathcal{D}_1 \rightarrowtriangle^{\mathcal{D}_3} \mathcal{D}_2)$.

If $\mathbf{L}$ *is prelinear or divisible, then*

7. $\sigma_\theta(\mathcal{D}_1 \cap \mathcal{D}_2) = \sigma_\theta(\mathcal{D}_1) \cap \sigma_\theta(\mathcal{D}_2)$,
8. $\mathcal{D}_1 \div^{\sigma_\theta(\mathcal{D}_3)} \mathcal{D}_2 = \sigma_\theta(\mathcal{D}_1 \div^{\mathcal{D}_3} \mathcal{D}_2)$.

Kernel and Support

Definition (kernel and support)

For any RDT $\mathcal{D}$ on relation scheme R, the **kernel** $\Delta\mathcal{D}$ and **support** $\nabla\mathcal{D}$ of $\mathcal{D}$ are RDTs on R defined by

$$(\Delta\mathcal{D})(r) = \begin{cases} 1, & \text{if } \mathcal{D}(r) = 1, \\ 0, & \text{otherwise,} \end{cases} \qquad (\nabla\mathcal{D})(r) = \begin{cases} 1, & \text{if } \mathcal{D}(r) > 0, \\ 0, & \text{otherwise,} \end{cases}$$

for all $r \in \mathrm{Tupl}(R)$.

notes:

- express non-ranked RDT from general ones
- notation by M. Baaz (projections and relativizations)
- *kernel* (interior operator); $\Delta\mathcal{D}$ is the greatest nonranked RDT such that $\Delta\mathcal{D} \subseteq \mathcal{D}$
- *support* (closure operator); $\nabla\mathcal{D}$ is the least nonranked RDT such that $\mathcal{D} \subseteq \nabla\mathcal{D}$
- two borderline cases of other possibilities (monotone and indepotent operators)

Theorem (properties of Δ and ∇)

The following are true (if both left and right-hand sides exist):

1. $\Delta\mathcal{D}_1 \otimes \mathcal{D}_2 = \Delta\mathcal{D}_1 \otimes \Delta\mathcal{D}_2$, $\nabla\mathcal{D}_1 \otimes \mathcal{D}_2 \subseteq \nabla\mathcal{D}_1 \otimes \nabla\mathcal{D}_2$,
2. $\Delta\mathcal{D}_1 \cap \mathcal{D}_2 = \Delta\mathcal{D}_1 \cap \Delta\mathcal{D}_2$, $\nabla\mathcal{D}_1 \cap \mathcal{D}_2 \subseteq \nabla\mathcal{D}_1 \cap \nabla\mathcal{D}_2$,
3. $\Delta\mathcal{D}_1 \cup \mathcal{D}_2 \supseteq \Delta\mathcal{D}_1 \cup \Delta\mathcal{D}_2$, $\nabla\mathcal{D}_1 \cup \mathcal{D}_2 = \nabla\mathcal{D}_1 \cup \nabla\mathcal{D}_2$,
4. $\Delta\mathcal{D}_1 \rightharpoonup^{\mathcal{D}_3} \mathcal{D}_2 \subseteq \Delta\mathcal{D}_1 \rightharpoonup^{\Delta\mathcal{D}_3} \Delta\mathcal{D}_2$,
   $\Delta\mathcal{D}_1 \rightharpoonup^{\mathcal{D}_3} \mathcal{D}_2 \subseteq \nabla\mathcal{D}_1 \rightharpoonup^{\Delta\mathcal{D}_3} \nabla\mathcal{D}_2 \subseteq \nabla\mathcal{D}_1 \rightharpoonup^{\nabla\mathcal{D}_3} \nabla\mathcal{D}_2$
5. $\Delta\mathcal{D}_1 \bowtie \mathcal{D}_2 = \Delta\mathcal{D}_1 \bowtie \Delta\mathcal{D}_2$, $\nabla\mathcal{D}_1 \bowtie \mathcal{D}_2 \subseteq \nabla\mathcal{D}_1 \bowtie \nabla\mathcal{D}_2$,
6. $\Delta\pi_R(\mathcal{D}) \supseteq \pi_R(\Delta\mathcal{D})$, $\nabla\pi_R(\mathcal{D}) = \pi_R(\nabla\mathcal{D})$,
7. $\Delta\mathcal{D}_1 \div^{\mathcal{D}_3} \mathcal{D}_2 \subseteq \Delta\mathcal{D}_1 \div^{\Delta\mathcal{D}_3} \Delta\mathcal{D}_2$,
   $\Delta\mathcal{D}_1 \div^{\mathcal{D}_3} \mathcal{D}_2 \subseteq \nabla\mathcal{D}_1 \div^{\Delta\mathcal{D}_3} \nabla\mathcal{D}_2 \subseteq \nabla\mathcal{D}_1 \div^{\nabla\mathcal{D}_3} \nabla\mathcal{D}_2$
8. $\Delta\sigma_\theta(\mathcal{D}) \subseteq \sigma_\theta(\Delta\mathcal{D})$, $\Delta\sigma_\theta(\mathcal{D}) = \Delta\sigma_\theta(\Delta\mathcal{D})$.

If $\mathbf{L}$ *is linear, then*

9. $\nabla\mathcal{D}_1 \cap \mathcal{D}_2 = \nabla\mathcal{D}_1 \cap \nabla\mathcal{D}_2$, $\Delta\mathcal{D}_1 \cup \mathcal{D}_2 = \Delta\mathcal{D}_1 \cup \Delta\mathcal{D}_2$,
10. $\Delta\pi_R(\mathcal{D}) = \pi_R(\Delta\mathcal{D})$.

□

Renaming

usual operation of renaming attributes:

Definition (renaming attributes)

For an RDT $\mathcal{D}$ on R and an injective map $h\colon R \to Y$ such that for all $y \in R$, the attributes $h(y)$ and y have identical domains with equalities, we define a **renaming** $\rho_h(\mathcal{D})$ of $\mathcal{D}$ by h as an RDT on $h(R) = \{h(y) \mid y \in R\}$ by $(\rho_h(\mathcal{D}))(h(r)) = \mathcal{D}(r)$, where $h(r) \in \mathrm{Tupl}(h(R))$ such that $(h(r))(h(y)) = r(y)$ for each attribute $y \in R$.

notation: $\rho_{h(y_1),\ldots,h(y_n)\leftarrow y_1,\ldots,y_n}(\mathcal{D})$ means $\rho_h(\mathcal{D})$ if $R = \{y_1, \ldots, y_n\}$

- we omit ith component in $y_1, \ldots, y_n \leftarrow h(y_1), \ldots, h(y_n)$ whenever $h(y_i) = y_i$

References

BELOHLAVEK, R. 2002.
Fuzzy Relational Systems: Foundations and Principles.
Kluwer Academic Publishers, Norwell, MA, USA.

DATE, C. J. AND DARWEN, H. 2006.
Databases, Types, and The Relational Model: The Third Manifesto, 3rd ed.
Addison-Wesley.

GOGUEN, J. A. 1979.
The logic of inexact concepts.
Synthese 19, 325–373.

HÁJEK, P. 1998.
Metamathematics of Fuzzy Logic.
Kluwer Academic Publishers, Dordrecht, The Netherlands.

MAIER, D. 1983.
Theory of Relational Databases.
Computer Science Pr, Rockville, MD, USA.

To Be Continued . . .

second part:

- types, domains, database instances
- formalization of queries
- relation algebra as query language
- domain relational calculus
- relational completeness
- derived operations
- further extensions
- notes