R. Bělohlávek, L. Urbanová, V. Vychodil: Sensitivity Analysis for Similarity-Based Relational Databases

Department of Computer Science, Palacky University, Olomouc (17. listopadu 12, CZ–77146 Olomouc, Czech Republic)

Introduction

We study different similarity issues in the extension of Codd's relational model of data, where each domain is equipped with similarity relation, and each tuple (row in the data table) has assigned a rank, which is a degree that express how much the tuple match a given query. We will show:

- :: How to assess similarity of two ranked tables, $\mathcal{D}_1$ and $\mathcal{D}_2$,
- :: Several formulas for estimation of similarity of output (data tables after query) based on similarity of inputs (data tables prior to query)
- :: How much is the truth degree of functional dependency $A \Rightarrow B$ in $\mathcal{D}$ dependent on the truth degree appearing in A, B?

Model of Data

- :: Y: set of all attributes, relation schemes = subsets $R \subseteq Y$
- :: For $R \subseteq Y$ and domains D_y: $\mathcal{T}_R = \Pi_{y \in R} D_y$
- :: Each domain D_y equipped with binary $\mathbf{L}$-relation $\approx_y$ satisfying:
  (Ref) for each $u \in D_y$: $u \approx_y u = 1$,
  (Sym) for each $u, v \in D_y$: $u \approx_y v = v \approx_y u$.
- :: Domain with similarity $= \langle D_y, \approx_y \rangle$
- :: A ranked data table (RDT) on R over $\{\langle D_y, \approx_y \rangle \mid y \in R\}$ is any finite $\mathbf{L}$-set in $\mathcal{T}_R$.

Similarity-based query: "Show me all cars which are sold for (approximately) $7\,000$ € and are (approximately) 3 years old."

$\mathcal{D}(t)$	brand	model	price	year	km
0.97	Renault	Scénic	$6\,940$	2009	$115\,556$
0.8	Renault	Scénic	$7\,200$	2010	$101\,478$
0.75	Opel	Zafira	$7\,500$	2008	$130\,656$
0.54	Citroen	Picasso	$7\,925$	2009	$109\,015$
0.1	Volkswagen	Caddy	$8\,600$	2008	$122\,855$

Notes on Generalization of Codd's Model of Data

similarities = to express "approximate matches"
ranks = degrees to which "tuples match queries"
ranks have comparative meaning
degrees of similarities and ranks - taken form the same scale L (complete residuated lattice)

Classic (Codd's) Model results by:
taking two-valued Boolean algebra for $\mathbf{L}$ (all ranks become 1 (match or equality) and 0 (no match or difference))
considering each $\approx_y$ to be identity relation on D_y.

Similarity of RDTs

Q: "Do similar input RDTs yield similar query results?"
A: "Yes, under suitable notions of similarity or RDTs."

Rank-based similarity

$\mathcal{D}_1$ is similar to $\mathcal{D}_2$ if ranks to which a tuple belongs to $\mathcal{D}_1$ and $\mathcal{D}_2$ are similar.

$$E(\mathcal{D}_1, \mathcal{D}_2) = \textstyle\bigwedge_{r \in \mathcal{T}_R} \big(\mathcal{D}_1(r) \leftrightarrow \mathcal{D}_2(r)\big).$$

Formalization using subsethood degrees:

$$S(\mathcal{D}_1, \mathcal{D}_2) = \textstyle\bigwedge_{r \in \mathcal{T}_R} \big(\mathcal{D}_1(r) \rightarrow \mathcal{D}_2(r)\big).$$
$$E(\mathcal{D}_1, \mathcal{D}_2) = S(\mathcal{D}_1, \mathcal{D}_2) \wedge S(\mathcal{D}_2, \mathcal{D}_1).$$

Tuple-based similarity

$\mathcal{D}_1$ is similar to $\mathcal{D}_2$ if they contain similar tuples, i.e. tuples with pairwise similar values

$$S^{\approx}(\mathcal{D}_i, \mathcal{D}_j) = \textstyle\bigwedge_{r \in \mathcal{T}_R} \big(\mathcal{D}_i(r) \rightarrow \bigvee_{r' \in \mathcal{T}_R} \big(\mathcal{D}_j(r') \otimes r \approx r'\big)\big),$$
$$E^{\approx}(\mathcal{D}_i, \mathcal{D}_j) = S^{\approx}(\mathcal{D}_i, \mathcal{D}_j) \wedge S^{\approx}(\mathcal{D}_j, \mathcal{D}_i)$$

$S^{\approx}(\mathcal{D}_i, \mathcal{D}_j) =$ degree to which the following proposition is true : "If r belongs to $\mathcal{D}_i$, then there is r' which belongs to $\mathcal{D}_j$ and is similar to r."

Remarks:
1) $r \approx r' = \bigwedge_{y \in R} r(y) \approx_y r'(y) = r(R) \approx_{\mathcal{D}} t'(R)$
2) $S(\mathcal{D}_i, \mathcal{D}_j) \leq S^{\approx}(\mathcal{D}_i, \mathcal{D}_j)$.
2) Both approaches are a special case of general approach.

Relational operations

- :: Similarity-based restrictions (selection):
  $(\sigma_{y \approx d}(\mathcal{D}))(r) = \mathcal{D}(r) \otimes r(y) \approx_y d$, where $y \in R$, $d \in D_y$
- :: Projection: For $\mathcal{D}$ on R_1, $R_2 \subseteq R_1$:
  $(\pi_{R_2}(\mathcal{D}))(r_2) = \bigvee_{r_3 \in \mathcal{T}_{R_1 \setminus R_2}} \mathcal{D}(r_2 r_3)$, for each $r_2 \in \mathcal{T}_{R_2}$.
- :: Division: For $\mathcal{D}_1$ on R_1, $\mathcal{D}_2$ on $R_2 \subseteq R_1$, and $\mathcal{D}_3$ on $R_3 = R_1 \setminus R_2$ and for each $r_3 \in \mathcal{T}_{R_3}$.
  $\big(\mathcal{D}_1 \div^{\mathcal{D}_3} \mathcal{D}_2\big)(r_3) = \bigwedge_{r_2 \in \mathcal{T}_{R_2}} \big(\mathcal{D}_3(r_3) \otimes (\mathcal{D}_2(r_2) \rightarrow \mathcal{D}_1(r_2 r_3))\big)$
- :: Natural join: For $\mathcal{D}_1$ on $R_1 \cup R_3$, $\mathcal{D}_2$ on $R_2 \cup R_3$ such that $R_1 \cap R_2 = R_1 \cap R_3 = R_2 \cap R_3 = \emptyset$:
  $\big(\mathcal{D}_1 \bowtie \mathcal{D}_2\big)(r_1 r_2 r_3) = \mathcal{D}_1(r_1 r_3) \otimes \mathcal{D}_2(r_2 r_3)$
- :: Ternary residuum: For all $r \in \mathcal{T}_R$.
  $\big(\mathcal{D}_1 \rightharpoonup^{\mathcal{D}_3} \mathcal{D}_2\big)(r) = \mathcal{D}_3(r) \otimes \big(\mathcal{D}_1(r) \rightarrow \mathcal{D}_2(r)\big)$
- :: Residuated c-negation of $\mathcal{D}_1$ which ranges over $\mathcal{D}_2$
  $\big(\mathcal{D}_2 \boxslash_c \mathcal{D}_1\big)(r) = \mathcal{D}_1(r) \rightharpoonup^{\mathcal{D}_2(r)} c$
- :: ... and more: Residuated c-shift, Similarity-based equijoin (or a theta-join), union, idempotent/strong intersection

Preservation of Similarity

Relational operations in our model are robust, they are insensitive to slight changes in data.

$$\begin{aligned}
S(\mathcal{D}, \mathcal{D}') &\leq S(\sigma_{y \approx d_1}(\mathcal{D}), \sigma_{y \approx d_1}(\mathcal{D}')) \\
S(\mathcal{D}, \mathcal{D}') \otimes (d_1 \approx d_2) &\leq S(\sigma_{y \approx d_1}(\mathcal{D}), \sigma_{y \approx d_2}(\mathcal{D}')) \\
S(\mathcal{D}, \mathcal{D}') &\leq S(\pi_{R_2}(\mathcal{D}), \pi_{R_2}(\mathcal{D}')) \\
S(\mathcal{D}_1, \mathcal{D}'_1) \otimes S(\mathcal{D}'_2, \mathcal{D}_2) \otimes S(\mathcal{D}_3, \mathcal{D}'_3) &\leq S(\mathcal{D}_1 \div^{\mathcal{D}_3} \mathcal{D}_2, \mathcal{D}'_1 \div^{\mathcal{D}'_3} \mathcal{D}'_2) \\
S(\mathcal{D}_1, \mathcal{D}'_1) \otimes S(\mathcal{D}_2, \mathcal{D}'_2) &\leq S(\mathcal{D}_1 \bowtie \mathcal{D}_2, \mathcal{D}'_1 \bowtie \mathcal{D}'_2) \\
S(\mathcal{D}'_1, \mathcal{D}_1) \otimes S(\mathcal{D}_2, \mathcal{D}'_2) &\leq S(\mathcal{D}_2 \boxslash_c \mathcal{D}_1, \mathcal{D}'_2 \boxslash_c \mathcal{D}'_1) \\
S(\mathcal{D}'_1, \mathcal{D}_1) \otimes S(\mathcal{D}_2, \mathcal{D}'_2) \otimes (c \leftrightarrow c') &\leq S(\mathcal{D}_2 \boxslash_c \mathcal{D}_1, \mathcal{D}'_2 \boxslash_{c'} \mathcal{D}'_1)
\end{aligned}$$

$$\begin{aligned}
&S(\mathcal{D}_1, \mathcal{D}'_1) \otimes S(\mathcal{D}_2, \mathcal{D}'_2) \otimes (c \leftrightarrow c') \leq \\
&\qquad S(\mathcal{D}_1 \bowtie_{c/_{y_1 \approx y_2}} \mathcal{D}_2, \mathcal{D}'_1 \bowtie_{c'/_{y_1 \approx y_2}} \mathcal{D}'_2) \\
&S(\mathcal{D}'_1, \mathcal{D}_1) \otimes S(\mathcal{D}_2, \mathcal{D}'_2) \otimes S(\mathcal{D}_3, \mathcal{D}'_3) \leq \\
&\qquad S(\mathcal{D}_1 \rightharpoonup^{\mathcal{D}_3} \mathcal{D}_2, \mathcal{D}'_1 \rightharpoonup^{\mathcal{D}'_3} \mathcal{D}'_2)
\end{aligned}$$

Remarks:
1) All inequalities holds when S is replaced by E.

2) $(\mathcal{D}_1 \bowtie_{c/_{y_1 \approx y_2}} \mathcal{D}_2)(r_1 r_2)$ is join with partial emphasis on the similarity-based condition ("the y_1-value is similar to the y_2-value at least to degree c"), formally:

$$(\mathcal{D}_1 \bowtie_{c/_{y_1 \approx y_2}} \mathcal{D}_2)(r_1 r_2) = \mathcal{D}_1(r_1) \otimes \mathcal{D}_2(r_2) \otimes (c \rightarrow r_1(y_1) \approx_{y_1} r_2(y_2))$$

3) In the second inequality $\approx$ needs to be $\otimes$-transitive

Preservation of similarity for $S^{\approx}$:

- :: $S^{\approx}$ can be expressed using a similarity-based closure:
  $$\begin{aligned}
  (\mathrm{C}^{\approx}(\mathcal{D}))(r) &= \textstyle\bigvee_{r' \in \mathcal{T}_R} \big(\mathcal{D}(r') \otimes r' \approx r\big) \\
  S^{\approx}(\mathcal{D}_i, \mathcal{D}_j) &= \textstyle\bigwedge_{r \in \mathcal{T}_R} \big(\mathcal{D}_i(r) \rightarrow \mathrm{C}^{\approx}(\mathcal{D}_j)(r)\big) \\
  &= S(\mathcal{D}_i, \mathrm{C}^{\approx}(\mathcal{D}_j)).
  \end{aligned}$$
  $\mathrm{C}^{\approx}(\mathcal{D})$ is a fuzzy closure operator
- :: Only projection preserve similarity
  $$S^{\approx}(\mathcal{D}_1, \mathcal{D}_2) \leq S^{\approx}(\pi_R(\mathcal{D}_1), \pi_R(\mathcal{D}_2))$$
- :: Every relational operation can be modified to preserve $S^{\approx}$
- :: Modification for selection:
  New type of selection: composed operation - select results not only from tuples in $\mathcal{D}$ but also from tuples similar to those in $\mathcal{D}$.
  $$\begin{aligned}
  \big(\sigma^{\approx}_{y \approx d}(\mathcal{D})\big)(r) &= \sigma_{y \approx d}(\mathrm{C}^{\approx}(\mathcal{D})) \\
  &= \textstyle\bigvee_{r' \in \mathcal{T}_R} \big(\mathcal{D}(r') \otimes r' \approx r \otimes r(y) \approx_y d\big)
  \end{aligned}$$
  This new type of selection is compatible with $S^{\approx}$:
  $$S^{\approx}(\mathcal{D}_1, \mathcal{D}_2) \leq S^{\approx}(\sigma_{y \approx d}(\mathrm{C}^{\approx}(\mathcal{D}_1)), \sigma_{y \approx d}(\mathrm{C}^{\approx}(\mathcal{D}_2)))$$

Functional dependency

An expression: $A \Rightarrow B$, where $A, B \in L^Y$.
The degree to which $A \Rightarrow B$ is true in data table $\mathcal{D}$ is given by:

$$\begin{aligned}
||A \Rightarrow B||_{\mathcal{D}} = \bigwedge_{r_1, r_2 \in \mathcal{T}_R} \Big(&(r_1(A) \approx_{\mathcal{D}} r_2(A))^* \rightarrow \\
&\rightarrow (r_1(B) \approx_D r_2(B)) \Big)
\end{aligned}$$

$$\begin{aligned}
r_1(C) \approx_{\mathcal{D}} r_2(C) = (\mathcal{D}(r_1) \otimes \mathcal{D}(r_2)) &\rightarrow \\
\rightarrow \bigwedge_{y \in R} (C(y) &\rightarrow r_1(y) \approx_y r_2(y))
\end{aligned}$$

For every pair of tuples: "If they have very similar values on attributes from A, then they have similar values on attributes from B."

How is the similarity of $||A_1 \Rightarrow B_1||_{\mathcal{D}}$ and $||A_2 \Rightarrow B_2||_{\mathcal{D}}$ dependent on similarity of A_1 to A_2 and B_1 to B_2?

$$S(A_1, A_2)^* \otimes S(B_2, B_1) \otimes \|A_1 \Rightarrow B_1\|_{\mathcal{D}} \leq \|A_2 \Rightarrow B_2\|_{\mathcal{D}}$$

Conclusions

- :: Similarity of ranked data tables
- :: Lower estimates for similarities of query results (robust model)
- :: Similarity based closure - new nontrivial operation
- :: Similarity based functional dependency and its sensitivity to degrees appearing in antecedent and consequent

Future work

- :: More on general approach of similarity of RDTs
- :: Importance of similarity of data tables for the validity of similarity-based functional dependency

References

(1) Maier, David,Theory of Relational Databases, Computer Science Pr, Rockville, MD, USA, 1983
(2) Bělohlávek, Radim and Vychodil, Vilém, Codd's Relational Model from the Point of View of Fuzzy Logic, Journal of Logic and Computation, 21, 5, October 2011, p. 851-862
(3) Bělohlávek, Radim, Urbanová, Lucie, and Vychodil, Vilém: Similarity of query results in similarity-based databases, In: J. T. Yao, S. Ramanna, G. Wang, Z. Suraj (eds.): Rough Sets and Knowledge Technology, Lecture Notes in Computer Science 6954, 2011, pp. 258–267, Springer, 2011

iciu
International Center for Information and Uncertainty
SSIU 2013: International Summer School "Information and Uncertainty"
http://mcin.upol.cz/SSWIU-2013/

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