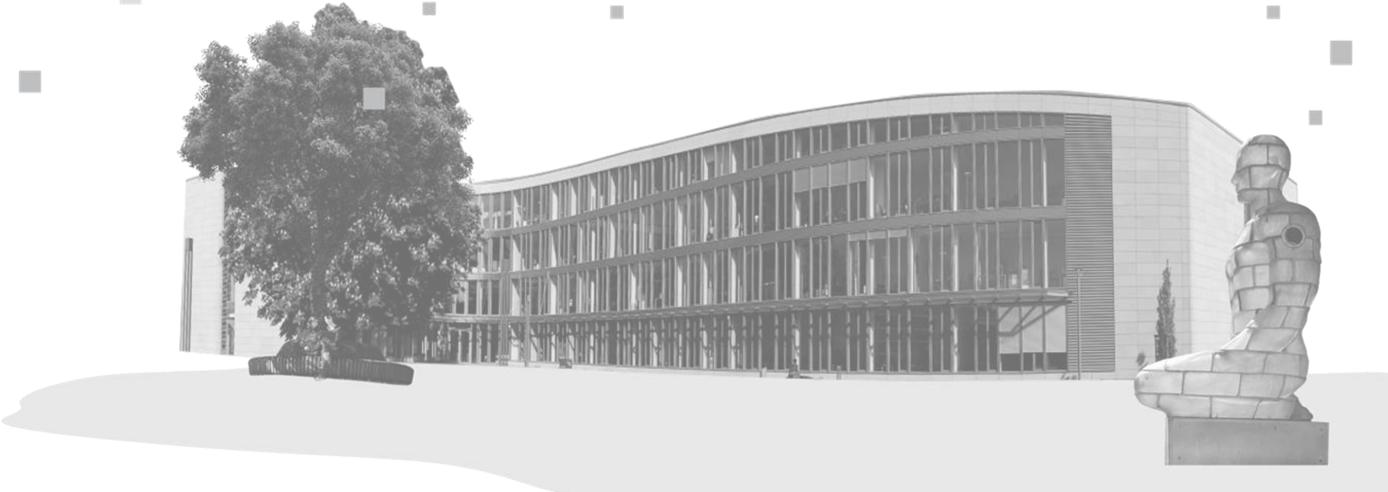
Advanced Data Profiling

Introduction to Order Dependencies

WS 2023/2024
Sebastian Schmidl



Design IT.
Create Knowledge.

www.hpi.de

Overview

1. Intuitions
2. Notation and Running Example
3. Order Specification
4. Lexicographical Order Operator
5. Bidirectional Order Dependency (list-based)
6. Axioms (list-based)
7. Set-based Canonical Form
8. Axioms (set-based)
9. Mapping between Set-based and List-based bODs

Intuition: Bidirectional Order Dependency (bOD)

salary ↦ tax

salary	tax
5k	1k
6k	1.5k
8k	2k
10k	3k

age↑ ↦ year-of-birth↓

age	yob
19	2001
25	1995
25	1995
31	1989
45	1975

[cases↑, r0↓] ↦ [dt↑]

cases	r0	doubling time
46	1.5	4 d
57	1.8	7 d
102	1.7	10 d
102	1.4	12 d
188	1.4	14 d

- Lexicographical ordering
- Think SQL:
  `... ORDER BY cases asc, r0 desc` ⇔ `... ORDER BY „doubling time“ asc`
- Default order direction: asc = ↑

Notation

Consistent with [1,7]

R	relation
r	relational instance of R
$t, s, t_0, t_1, \ldots$	tuples of r
$\mathrm{A}, \mathrm{B}, \mathrm{C}$	attributes of R
$\mathbf{X}, \mathbf{Y}, \mathbf{Z}$	list of attributes
$\mathbf{Y} = [\mathrm{A} \mid \mathbf{T}]$	A is the head element of list $\mathbf{Y}$ and $\mathbf{T}$ the tail list
X, Y, Z	se… ibutes

Definition box

Defined concept name is bold

*An **example definition** is a list of attributes denoted as* $\mathbf{X} = [\mathrm{A} \mid \mathrm{A} \in \mathrm{R}]$.

Lists and sets with the same name reference the same data:
Set X contains all distinct elements of list $\mathbf{X}$

Running Example [7]

ID	Month	Day	OCode	OFips	OState	DepDelay	ArrDelay	ArrDelGrp
t_0	12	3	CLE	39	Ohio	80	72	4
t_1	12	14	ORD	17	Illinois	-4	-10	-1
t_2	12	14	JFK	36	New York	-6	0	0
t_3	12	16	ORD	17	Illinois	13	59	3
t_4	12	20	CLE	39	Ohio	$\bot$	$\bot$	$\bot$
t_5	12	24	ORD	17	Illinois	5	-2	-1
t_6	12	27	BWI	24	Maryland	-5	-23	-2
t_7	12	28	ORD	17	Illinois	57	82	5
t_8	12	30	SGF	29	Missouri	2	0	0
t_9	12	30	ORD	17	Illinois	-5	-14	-1

$\bot$ = null values

Order Specification [1,7,8]

- Compare to SQL ORDER BY
  - `ORDER BY A asc, B desc`

*An **order specification** is a list of marked attributes denoted as* $\overline{\mathbf{X}} = [\overline{\mathrm{A}}\,|\mathrm{A} \in \mathrm{R}]$.

*A **marked attribute*** $\overline{\mathrm{A}}$ *can have two directions: ascending* $\mathrm{A}\uparrow$ *or descending* $\mathrm{A}\downarrow$. *The default order direction is ascending (*$\uparrow$*).*

- Examples

`[cases↑, r0↓]`

`[salary↑]`

`cases↑, r0↓, doubling_time↑`

`OFips, DepDelay, ArrDelay`

`[OCode, Month, Day]`

Order Specification

- Example table ordered by $[OCode \uparrow, Day \downarrow]$:

ID	Month	Day	OCode	OFips	OState	DepDelay	ArrDelay	ArrDelGrp
t_6	12	27	BWI	24	Maryland	-5	-23	-2
t_4	12	20	CLE	39	Ohio	$\perp$	$\perp$	$\perp$
t_0	12	3	CLE	39	Ohio	80	72	4
t_2	12	14	JFK	36	New York	-6	0	0
t_9	12	30	ORD	17	Illinois	-5	-14	-1
t_7	12	28	ORD	17	Illinois	57	82	5
t_5	12	24	ORD	17	Illinois	5	-2	-1
t_3	12	16	ORD	17	Illinois	13	59	3
t_1	12	14	ORD	17	Illinois	-4	-10	-1
t_8	12	30	SGF	29	Missouri	2	0	0

Lexicographical Order Operator [1,7,8]

- Lexicographical ordering
  - Numbers ordered numerically
  - Strings ordered alphabetically
  - Timestamps ordered chronologically
- Examples

Let $\overline{\mathbf{X}}$ *be an order specification,* s *and* t *be two tuples in relation instance* r*. The* ***lexicographical order operator*** $\preccurlyeq_{\overline{\mathbf{X}}}$ *is defined as follows:*

$$s \preccurlyeq_{\overline{\mathbf{X}}} t = \begin{cases} \overline{\mathbf{X}} = [] \\ \overline{\mathbf{X}} = [\mathrm{A}\uparrow |\, \overline{\mathbf{T}}] \wedge s_{\mathrm{A}} < t_{\mathrm{A}} \\ \overline{\mathbf{X}} = [\mathrm{A}\downarrow |\, \overline{\mathbf{T}}] \wedge s_{\mathrm{A}} > t_{\mathrm{A}} \\ \overline{\mathbf{X}} = [\overline{\mathrm{A}} |\, \overline{\mathbf{T}}] \wedge s_{\mathrm{A}} = t_{\mathrm{A}} \wedge s \preccurlyeq_{\overline{\mathbf{T}}} t \end{cases}$$

$t_5 \preccurlyeq_{[Day]} t_9$

$t_5 \preccurlyeq_{[DepDelay\downarrow]} t_9$

$t_9 \preccurlyeq_{[ArrDelGrp, ArrDelay]} t_5$

$t_5 \preccurlyeq_{[ArrDelGrp, ArrDelay\downarrow]} t_9$

$t_5 \preccurlyeq_{[OCode]} t_9 \wedge t_9 \preccurlyeq_{[OCode]} t_5$

$\Rightarrow\ t_5 =_{[OCode]} t_9$

ID	Month	Day	OCode	OFips	OState	DepDelay	ArrDelay	ArrDelGrp
t_9	12	30	ORD	17	Illinois	-5	-14	-1
t_5	12	24	ORD	17	Illinois	5	-2	-1

Bidirectional Order Dependency [1,7,8]

- If a bOD $\overline{\mathbf{X}} \mapsto \overline{\mathbf{Y}}$ holds:
  - when the tuples are ordered by $\overline{\mathbf{X}}$, then they are also ordered by $\overline{\mathbf{Y}}$
  - but not necessarily vice versa!
- Simpler form: (unidirectional) order dependencies (ODs) $\mathbf{X} \mapsto \mathbf{Y}$
  - All attributes have the same order direction (either all ↑ or all ↓)
  - $\mathbf{Y}$'s values are monotonically non-decreasing with respect to the values in $\mathbf{X}$

*Let $\overline{\mathbf{X}}$ and $\overline{\mathbf{Y}}$ be order specifications over relation R. A **bidirectional order dependency (bOD)** $\overline{\mathbf{X}} \mapsto \overline{\mathbf{Y}}$ (read X orders Y) specifies that ordering a relation instance r by $\overline{\mathbf{X}}$ also orders r by $\overline{\mathbf{Y}}$.*

Table r over the relation R satisfies a bOD $\overline{\mathbf{X}} \mapsto \overline{\mathbf{Y}}$, where $X \subset \mathrm{R}$ and $Y \subset \mathrm{R}$, if

$$\forall s, t \in r: s \preccurlyeq_{\overline{\mathbf{X}}} t \Longrightarrow s \preccurlyeq_{\overline{\mathbf{Y}}} t.$$

Order Equivalence (OE) ↔

[1,7,8]

Two order specifications $\overline{\mathbf{X}}$ *and* $\overline{\mathbf{Y}}$ *are* ***order equivalent****, denoted as* $\overline{\mathbf{X}} \leftrightarrow \overline{\mathbf{Y}}$*, iff* $\overline{\mathbf{X}} \mapsto \overline{\mathbf{Y}}$ *and* $\overline{\mathbf{Y}} \mapsto \overline{\mathbf{X}}$*.*

$$[OFips] \leftrightarrow [\text{OState}]$$

ID	Month	Day	OCode	OFips	OState	DepDelay	ArrDelay	ArrDelGrp
t_1	12	14	ORD	17	Illinois	-4	-10	-1
t_3	12	16	ORD	17	Illinois	13	59	3
t_5	12	24	ORD	17	Illinois	5	-2	-1
t_7	12	28	ORD	17	Illinois	57	82	5
t_9	12	30	ORD	17	Illinois	-5	-14	-1
t_6	12	27	BWI	24	Maryland	-5	-23	-2
t_8	12	30	SGF	29	Missouri	2	0	0
t_2	12	14	JFK	36	New York	-6	0	0
t_0	12	3	CLE	39	Ohio	80	72	4
t_4	12	20	CLE	39	Ohio	⊥	⊥	⊥

Order Compatibility (OC) ~ [1,7,8]

> *Two order specifications $\overline{\mathbf{X}}$ and $\overline{\mathbf{Y}}$ are* ***order compatible****, denoted as $\overline{\mathbf{X}} \sim \overline{\mathbf{Y}}$, iff $\overline{\mathbf{XY}} \leftrightarrow \overline{\mathbf{YX}}$.*

$$[ArrDelGrp] \sim [ArrDelay]$$

ID	Month	Day	OCode	OFips	OState	DepDelay	ArrDelay	ArrDelGrp
t_6	12	27	BWI	24	Maryland	-5	-23	-2
t_9	12	30	ORD	17	Illinois	-5	-14	-1
t_1	12	14	ORD	17	Illinois	-4	-10	-1
t_5	12	24	ORD	17	Illinois	5	-2	-1
t_2	12	14	JFK	36	New York	-6	0	0
t_8	12	30	SGF	29	Missouri	2	0	0
t_3	12	16	ORD	17	Illinois	13	59	3
t_0	12	3	CLE	39	Ohio	80	72	4
t_7	12	28	ORD	17	Illinois	57	82	5
t_4	12	20	CLE	39	Ohio	⊥	⊥	⊥

Violation of ODs (Splits and Swaps)

[1,7,8]

$\overline{\mathbf{X}} \mapsto \overline{\mathbf{Y}}$ is valid only iff $\overline{\mathbf{X}} \mapsto \overline{\mathbf{XY}}$ and $\overline{\mathbf{X}} \sim \overline{\mathbf{Y}}$ are valid.

Functional Dependency (FD)

Order Compatibility (OC)

Example:
$[C1] \mapsto [C2]$

ID	C1	C2	C3
t_0	1	4	3
t_1	1	4	3
t_2	2	2	5
t_3	2	3	4

- 2 ways for violation of an OD $\overline{\mathbf{X}} \mapsto \overline{\mathbf{Y}}$
  1. Violation of the FD-part $\overline{\mathbf{X}} \mapsto \overline{\mathbf{XY}}$

*A **split** w.r.t. a bOD $\overline{\mathbf{X}} \mapsto \overline{\mathbf{XY}}$ is a pair of tuples t and s in relational instance r, such that $t_{\mathbf{X}} = s_{\mathbf{X}}$ but $t_{\mathbf{Y}} \neq s_{\mathbf{Y}}$.*

$[C1] \mapsto [C1, C2]$ (FD: $C1 \rightarrow C2$)
→ Split in t_2 and t_3

  2. Violation of the OC-part $\overline{\mathbf{X}} \sim \overline{\mathbf{Y}}$

*A **swap** w.r.t. a bOD $\overline{\mathbf{X}} \sim \overline{\mathbf{Y}}$ ($\overline{\mathbf{XY}} \leftrightarrow \overline{\mathbf{YX}}$) is a pair of tuples t and s in relational instance r, such that $t <_{\overline{\mathbf{X}}} s$ but $s <_{\overline{\mathbf{Y}}} t$.*

$[C1] \sim [C2]$ ($[C1, C2] \leftrightarrow [C2, C1]$)
→ Swaps in t_0 and t_2,
t_0 and t_3,
t_1 and t_2,
t_1 and t_3

Axioms (Inference of ODs) [8]

- Sound and complete axioms:

OD1: *Reflexivity*

$$\mathbf{XY} \mapsto \mathbf{X}$$

OD2: *Prefix*

$$\frac{\mathbf{X} \mapsto \mathbf{Y}}{\mathbf{ZX} \mapsto \mathbf{ZY}}$$

OD3: *Normalization*

$$\mathbf{WXYXV} \leftrightarrow \mathbf{WXYV}$$

OD4: *Transitivity*

$$\frac{\mathbf{X} \mapsto \mathbf{Y} \quad \mathbf{Y} \mapsto \mathbf{Z}}{\mathbf{X} \mapsto \mathbf{Z}}$$

OD5: *Suffix*

$$\frac{\mathbf{X} \mapsto \mathbf{Y}}{\mathbf{X} \leftrightarrow \mathbf{YX}}$$

OD6: *Chain*

$$\frac{\mathbf{X} \sim \mathbf{Y}_1 \quad \forall_{i \in [1,n-1]} \mathbf{Y}_i \sim \mathbf{Y}_{i+1} \quad \mathbf{Y}_n \sim \mathbf{Z} \quad \forall_{i \in [1,n]} \mathbf{Y}_i\mathbf{X} \sim \mathbf{Y}_i\mathbf{Z}}{\mathbf{X} \sim \mathbf{Z}}$$

Axioms for bidirectional ODs (Inference of bODs) [1]

- Sound and complete axioms:

1. Reflexivity^L

$$\overline{\mathbf{XY}} \mapsto \overline{\mathbf{X}}$$

2. Prefix^L

$$\frac{\overline{\mathbf{X}} \mapsto \overline{\mathbf{Y}}}{\overline{\mathbf{ZX}} \mapsto \overline{\mathbf{ZY}}}$$

3. Transitivity^L

$$\frac{\overline{\mathbf{X}} \mapsto \overline{\mathbf{Y}} \quad \overline{\mathbf{Y}} \mapsto \overline{\mathbf{Z}}}{\overline{\mathbf{X}} \mapsto \overline{\mathbf{Z}}}$$

4. Normalization^L

$$\overline{\mathbf{WXYXV}} \leftrightarrow \overline{\mathbf{WXYV}}.$$

5. Suffix^L

$$\frac{\overline{\mathbf{X}} \mapsto \overline{\mathbf{Y}}}{\overline{\mathbf{X}} \leftrightarrow \overline{\mathbf{YX}}}$$

6. Chain^L

$$\frac{\overline{\mathbf{X}} \sim \overline{\mathbf{Y}}_1 \quad \forall_{i \in [1, n-1]} \overline{\mathbf{Y}}_i \sim \overline{\mathbf{Y}}_{i+1} \quad \overline{\mathbf{Y}}_n \sim \overline{\mathbf{Z}} \quad \forall_{i \in [1, n]} \overline{\mathbf{Y}}_i \overline{\mathbf{X}} \sim \overline{\mathbf{Y}}_i \overline{\mathbf{Z}}}{\overline{\mathbf{X}} \sim \overline{\mathbf{Z}}}$$

7. Reverse^L

$$\frac{\overline{\mathbf{X}} \mapsto \overline{\mathbf{Y}}}{\overline{\mathbf{X}}\updownarrow \mapsto \overline{\mathbf{Y}}\updownarrow}$$

8. Constant^L

$$\frac{\overline{\mathbf{XZ}} \mapsto \overline{\mathbf{XY}} \quad \overline{\mathbf{XZ}} \mapsto \overline{\mathbf{XY}}\updownarrow}{\overline{\mathbf{X}} \mapsto \overline{\mathbf{XY}}}$$

Set-based Canoncial Form [1,7]

- Equivalence class

> *An **equivalence class** $\mathcal{E}(t_X)$ w.r.t. a given attribute set $X \in \mathrm{R}$ groups tuples s and t together when their projection on X is equal: $\mathcal{E}(t_X) = \{s \in \mathrm{r} \mid s_X = t_X\}$.*

- Partition

> *A **partition** Π_X is a set of disjoint equivalence classes with the same set of attributes: $\Pi_X = \{\mathcal{E}(t_X) \mid t \in \mathrm{R}\}$.*

- Examples

$\Pi_{\{Month\}} = \{\{t_0, \dots, t_9\}\}$

$\mathcal{E}(12) = \{t_0, \dots, t_9\}$

ID	Month	Day	OCode	OFips	OState
t_0	12	3	CLE	39	Ohio
t_1	12	14	ORD	17	Illinois
t_2	12	14	JFK	36	New York
t_3	12	16	ORD	17	Illinois
t_4	12	20	CLE	39	Ohio
t_5	12	24	ORD	17	Illinois
t_6	12	27	BWI	24	Maryland
t_7	12	28	ORD	17	Illinois
t_8	12	30	SGF	29	Missouri
t_9	12	30	ORD	17	Illinois

$\Pi_{\{OState\}} = \{\{t_0, t_4\}, \{t_1, t_3, t_5, t_7, t_9\}, \{t_2\}, \{t_6\}, \{t_8\}\}$

$\mathcal{E}(Ohio) = \{t_0, t_4\}$

$\mathcal{E}(New\ York) = \{t_2\}$

$\mathcal{E}(Missouri) = \{t_8\}$

$\mathcal{E}(Illinois) = \{t_1, t_3, t_5, t_7, t_9\}$

$\mathcal{E}(Maryland) = \{t_6\}$

Set-based Canoncial Form [1,7]

- Constant bOD from $X: [\,] \mapsto \overline{\mathrm{A}}$

> *A **constant bOD** $X: [\,] \mapsto \overline{\mathrm{A}}$ is a marked attribute $\overline{\mathrm{A}}$ that is constant within each equivalence class w.r.t. the set of attributes in the context $X \in \mathrm{R}$.*

- Order compatible bOD form: $X: \overline{\mathrm{A}} \sim \overline{\mathrm{B}}$

> *An **order compatible bOD** $X: \overline{\mathrm{A}} \sim \overline{\mathrm{B}}$ are two marked attributes $\overline{\mathrm{A}}$ and $\overline{\mathrm{B}}$ that are order compatible within each equivalence class w.r.t. the set of attributes in the context $X \in \mathrm{R}$.*

Set-based Canoncial Form

	cases	r0	doubling time
t_1	46	1.5	4 d
t_2	57	1.8	7 d
t_3	102	1.7	10 d
t_4	102	1.4	12 d
t_5	188	1.4	14 d

Constant bOD:

{cases, r0} : [] ↦ dt↑

Order compatible bOD:

{cases} : r0↓ ~ dt↑

context

Axioms (Inference for Set-based bODs) [1]

- Sound and complete axioms:

1. Reflexivity^S

$$\mathcal{X}\text{: } [\,] \mapsto \overline{\mathsf{A}},\ \forall \mathsf{A} \in \mathcal{X}$$

2. Identity^S

$$\mathcal{X}\text{: } \overline{\mathsf{A}} \sim \overline{\mathsf{A}}$$

3. Commutativity^S

$$\frac{\mathcal{X}\text{: } \overline{\mathsf{A}} \sim \overline{\mathsf{B}}}{\mathcal{X}\text{: } \overline{\mathsf{B}} \sim \overline{\mathsf{A}}}$$

4. Strengthen^S

$$\frac{\mathcal{X}\text{: } [\,] \mapsto \overline{\mathsf{A}} \qquad \mathcal{X}\mathsf{A}\text{: } [\,] \mapsto \overline{\mathsf{B}}}{\mathcal{X}\text{: } [\,] \mapsto \overline{\mathsf{B}}}$$

5. Propagate^S

$$\frac{\mathcal{X}\text{: } [\,] \mapsto \overline{\mathsf{A}}}{\mathcal{X}\text{: } \overline{\mathsf{A}} \sim \overline{\mathsf{B}}}$$

6. Augmentation-I^S

$$\frac{\mathcal{X}\text{: } [\,] \mapsto \overline{\mathsf{A}}}{\mathcal{Z}\mathcal{X}\text{: } [\,] \mapsto \overline{\mathsf{A}}}$$

7. Augmentation-II^S

$$\frac{\mathcal{X}\text{: } \overline{\mathsf{A}} \sim \overline{\mathsf{B}}}{\mathcal{Z}\mathcal{X}\text{: } \overline{\mathsf{A}} \sim \overline{\mathsf{B}}}$$

8. Chain^S

$$\frac{\mathcal{X}\text{: } \overline{\mathsf{A}} \sim \overline{\mathsf{B}_1} \qquad \forall_{i \in [1,n-1]}, \mathcal{X}\text{: } \overline{\mathsf{B}_i} \sim \overline{\mathsf{B}_{i+1}} \qquad \mathcal{X}\text{: } \overline{\mathsf{B}_n} \sim \overline{\mathsf{C}} \qquad \forall_{i \in [1,n]}, \mathcal{X}\mathsf{B}_i : \overline{\mathsf{A}} \sim \overline{\mathsf{C}}}{\mathcal{X}\text{: } \overline{\mathsf{A}} \sim \overline{\mathsf{C}}}$$

9. Reverse–I^S

$$\frac{\mathcal{X}\text{: } [\,] \mapsto \overline{\mathsf{A}}}{\mathcal{X}\text{: } [\,] \mapsto \overline{\mathsf{A}}{\updownarrow}}$$

10. Reverse–II^S

$$\frac{\mathcal{X}\text{: } \overline{\mathsf{A}} \sim \overline{\mathsf{B}}}{\mathcal{X}\text{: } \overline{\mathsf{A}}{\updownarrow} \sim \overline{\mathsf{B}}{\updownarrow}}$$

11. Constant^S

$$\frac{\mathcal{X}\mathsf{A}\text{: } [\,] \mapsto \overline{\mathsf{B}} \qquad \mathcal{X}\text{: } \overline{\mathsf{A}} \sim \overline{\mathsf{B}} \qquad \mathcal{X}\text{: } \overline{\mathsf{A}} \sim \overline{\mathsf{B}}{\updownarrow}}{\mathcal{X} : [\,] \mapsto \overline{\mathsf{B}}}$$

Axioms (Inference for Set-based bODs) [1]

- Examples

ID	Region	Name
1	AFRICA	Algeria
2	AFRICA	Angola
3	AFRICA	Benin
…	…	…
54	AFRICA	Zimbabwe
55	AMERICA	Argentina
…	…	…
193	OCEANIA	Vanuatu

3. Commutativity^S

$$\frac{\mathcal{X}\colon \overline{\mathsf{A}} \sim \overline{\mathsf{B}}}{\mathcal{X}\colon \overline{\mathsf{B}} \sim \overline{\mathsf{A}}}$$

{name}: [] ↦ region
{name}: [] ↦ id
{id}: [] ↦ name
{id}: [] ↦ region

{}: id↑ ~ region↑
{region}: id↑ ~ name↑

3 → **{}: region↑ ~ id↑**

Axioms (Inference for Set-based bODs) [1]

- Examples

ID	Region	Name
1	AFRICA	Algeria
2	AFRICA	Angola
3	AFRICA	Benin
…	…	…
54	AFRICA	Zimbabwe
55	AMERICA	Argentina
…	…	…
193	OCEANIA	Vanuatu

5. Propagate^S

$$\frac{\mathcal{X}\colon [\,] \mapsto \overline{\mathsf{A}}}{\mathcal{X}\colon \overline{\mathsf{A}} \sim \overline{\mathsf{B}}}$$

{name}: [] ↦ region
{name}: [] ↦ id
{id}: [] ↦ name
{id}: [] ↦ region

{}: id↑ ~ region↑
{region}: id↑ ~ name↑

5 →

{id}: name↑ ~ region↑
{id}: name↑ ~ region↓
{id}: name↑ ~ id↑
…

Axioms (Inference for Set-based bODs) [1]

- Examples

ID	Region	Name
1	AFRICA	Algeria
2	AFRICA	Angola
3	AFRICA	Benin
…	…	…
54	AFRICA	Zimbabwe
55	AMERICA	Argentina
…	…	…
193	OCEANIA	Vanuatu

6. Augmentation-I^S

$$\frac{\mathcal{X}\colon [\,] \mapsto \overline{\mathsf{A}}}{\mathcal{Z}\mathcal{X}\colon [\,] \mapsto \overline{\mathsf{A}}}$$

{name}: [] ↦ region
{name}: [] ↦ id
{id}: [] ↦ name
{id}: [] ↦ region

{}: id↑ ~ region↑
{region}: id↑ ~ name↑

→ 6 →

{id,region}: [] ↦ name
{id,name}: [] ↦ name
{id,region,name}: [] ↦ name

Axioms (Inference for Set-based bODs) [1]

- Examples

ID	Region	Name
1	AFRICA	Algeria
2	AFRICA	Angola
3	AFRICA	Benin
…	…	…
54	AFRICA	Zimbabwe
55	AMERICA	Argentina
…	…	…
193	OCEANIA	Vanuatu

7. Augmentation-II^S

$$\frac{\mathcal{X}\colon \overline{\mathsf{A}} \sim \overline{\mathsf{B}}}{\mathcal{ZX}\colon \overline{\mathsf{A}} \sim \overline{\mathsf{B}}}$$

{name}: [] ↦ region
{name}: [] ↦ id
{id}: [] ↦ name
{id}: [] ↦ region

{}: id↑ ~ region↑
{region}: id↑ ~ name↑

→ 7 →

{name}: id↑ ~ region↑
{name,id}: id↑ ~ region↑
…

Axioms (Inference for Set-based bODs) [1]

- Examples

ID	Region	Name
1	AFRICA	Algeria
2	AFRICA	Angola
3	AFRICA	Benin
…	…	…
54	AFRICA	Zimbabwe
55	AMERICA	Argentina
…	…	…
193	OCEANIA	Vanuatu

9. Reverse–I^S

$$\frac{\mathcal{X}\colon [\,] \mapsto \overline{\mathsf{A}}}{\mathcal{X}\colon [\,] \mapsto \overline{\mathsf{A}}\updownarrow}$$

{name}: [] ↦ region
{name}: [] ↦ id
{id}: [] ↦ name
{id}: [] ↦ region

{}: id↑ ~ region↑
{region}: id↑ ~ name↑

9 → **{name}: [] ↦ id↓**

Axioms (Inference for Set-based bODs) [1]

- Examples

ID	Region	Name
1	AFRICA	Algeria
2	AFRICA	Angola
3	AFRICA	Benin
…	…	…
54	AFRICA	Zimbabwe
55	AMERICA	Argentina
…	…	…
193	OCEANIA	Vanuatu

10. Reverse–II^S

$$\frac{\mathcal{X}\colon \overline{\mathsf{A}} \sim \overline{\mathsf{B}}}{\mathcal{X}\colon \overline{\mathsf{A}}\updownarrow \sim \overline{\mathsf{B}}\updownarrow}$$

{name}: [] ↦ region
{name}: [] ↦ id
{id}: [] ↦ name
{id}: [] ↦ region

{}: id↑ ~ region↑
{region}: id↑ ~ name↑

10 → **{}: id↓ ~ region↓**

Axioms (Inference for Set-based bODs) [1]

- Examples

ID	Region	Name
1	AFRICA	Algeria
2	AFRICA	Angola
3	AFRICA	Benin
…	…	…
54	AFRICA	Zimbabwe
55	AMERICA	Argentina
…	…	…
193	OCEANIA	Vanuatu

11. Constant^S

$$\frac{\mathcal{X}\text{A}: [\,] \mapsto \overline{\text{B}} \qquad \mathcal{X}: \overline{\text{A}} \sim \overline{\text{B}} \qquad \mathcal{X}: \overline{\text{A}} \sim \overline{\text{B}}\updownarrow}{\mathcal{X}: [\,] \mapsto \overline{\text{B}}}$$

$X = \{\}$
A = id
B = region

{name}: [] ↦ region
{name}: [] ↦ id
{id}: [] ↦ name
{id}: [] ↦ region

{}: id↑ ~ region↑
{region}: id↑ ~ name↑

{}: id↑ ~ region ↓ missing

Mapping between Set-based and List-based ODs

- List-based bODs and set-based bODs are equivalent

→ there exists a mapping between the representations

$$\overline{\mathbf{X}} \mapsto \overline{\mathbf{Y}}$$

$$\mathcal{X}: [\,] \mapsto \overline{\mathrm{A}}$$
$$\mathcal{X}: \overline{\mathrm{A}} \sim \overline{\mathrm{B}}$$

- But NO 1:1-mapping

Mapping from List-based to Set-based ODs [1,7]

$\overline{\mathbf{X}} \mapsto \overline{\mathbf{Y}}$ ⟷ $X\colon [\,] \mapsto \overline{\mathrm{A}}$, $X\colon \overline{\mathrm{A}} \sim \overline{\mathrm{B}}$

$\overline{\mathbf{X}} \mapsto \overline{\mathbf{Y}}$ *is valid iff* $\overline{\mathbf{X}} \mapsto \overline{\mathbf{XY}}$ *and* $\overline{\mathbf{X}} \sim \overline{\mathbf{Y}}$ *are valid.*

No *splits*

$\forall \overline{\mathrm{A}} \in \overline{\mathbf{Y}}$:
$X\colon [\,] \mapsto \overline{\mathrm{A}}$ *must be valid*

➕

No *swaps*

$\forall i \in \{1, \ldots, |\overline{\mathbf{X}}|\}$ and $\forall j \in \{1, \ldots, |\overline{\mathbf{Y}}|\}$:
$\{X_1, \ldots, X_{i-1}, Y_1, \ldots, Y_{j-1}\}\colon \overline{X_i} \sim \overline{Y_j}$ *must be valid*

- Examples

$[\mathrm{A}\uparrow, \mathrm{B}\downarrow] \mapsto [\mathrm{C}\uparrow, \mathrm{D}\uparrow]$

$\{\mathrm{A},\mathrm{B}\}\colon [\,] \mapsto \mathrm{C}\uparrow$
$\{\mathrm{A},\mathrm{B}\}\colon [\,] \mapsto \mathrm{D}\uparrow$

$\{\,\}\colon \mathrm{A}\uparrow \sim \mathrm{C}\uparrow$
$\{\mathrm{A}\}\colon \mathrm{B}\downarrow \sim \mathrm{C}\uparrow$
$\{\mathrm{C}\}\colon \mathrm{A}\uparrow \sim \mathrm{D}\uparrow$
$\{\mathrm{A},\mathrm{C}\}\colon \mathrm{B}\downarrow \sim \mathrm{D}\uparrow$

$[\mathrm{B}\downarrow, \mathrm{A}\uparrow] \mapsto [\mathrm{C}\uparrow]$

$\{\mathrm{A},\mathrm{B}\}\colon [\,] \mapsto \mathrm{C}\uparrow$
$\{\,\}\colon \mathrm{B}\uparrow \sim \mathrm{C}\uparrow$
$\{\mathrm{B}\}\colon \mathrm{A}\downarrow \sim \mathrm{C}\uparrow$

List-based ODs can share set-based ODs in their mapping

- Polynomial complexity ($|\overline{\mathbf{X}}| \times |\overline{\mathbf{Y}}|$)
- # set-based ODs is not necessarily larger than # list-based ODs

Mapping from Set-based to List-based ODs [1,7]

$\overline{\mathbf{X}} \mapsto \overline{\mathbf{Y}}$ ⇄ $X: [\,] \mapsto \overline{\mathrm{A}}$, $X: \overline{\mathrm{A}} \sim \overline{\mathrm{B}}$

- Constant bODs:

$$X: [\,] \mapsto \overline{\mathrm{A}}$$

$$\downarrow$$

$$\forall\ permutations\ \overline{\mathbf{X}}'\ of\ \overline{\mathbf{X}}:$$
$$\overline{\mathbf{X}'} \mapsto \overline{\mathbf{X}'\mathrm{A}}\ must\ be\ valid$$

- Example

$$\{\mathrm{A},\mathrm{B}\}: [\,] \mapsto \mathrm{C}\uparrow$$

$$[A\uparrow, B\uparrow] \mapsto [A\uparrow, B\uparrow, C\uparrow] \quad [B\uparrow, A\uparrow] \mapsto [B\uparrow, A\uparrow, C\uparrow]$$
$$[A\uparrow, B\downarrow] \mapsto [A\uparrow, B\downarrow, C\uparrow] \quad [B\uparrow, A\downarrow] \mapsto [B\uparrow, A\downarrow, C\uparrow]$$
$$[A\downarrow, B\uparrow] \mapsto [A\downarrow, B\uparrow, C\uparrow] \quad [B\downarrow, A\uparrow] \mapsto [B\downarrow, A\uparrow, C\uparrow]$$
$$[A\downarrow, B\downarrow] \mapsto [A\downarrow, B\downarrow, C\uparrow] \quad [B\downarrow, A\downarrow] \mapsto [B\downarrow, A\downarrow, C\uparrow]$$

- Factorial complexity ($|X|$)
- Set-based is more consice representation

- Order compatible bODs:

$$X: \overline{\mathrm{A}} \sim \overline{\mathrm{B}}$$

$$\downarrow$$

$$\forall\ permutations\ \overline{\mathbf{X}}'\ of\ \overline{\mathbf{X}}:$$
$$\overline{\mathbf{X}'\mathrm{A}} \sim \overline{\mathbf{X}'\mathrm{B}}\ must\ be\ valid$$

- Example

$$\{\mathrm{A}\}: \mathrm{B}\uparrow \sim \mathrm{C}\downarrow$$

$$[A\uparrow, B\uparrow] \sim [A\uparrow, C\downarrow]$$
$$[A\downarrow, B\uparrow] \sim [A\downarrow, C\downarrow]$$

References

[1] Szlichta, J.; Godfrey, P.; Golab, L.; Kargar, M.; Srivastava, D.: **Effective and complete discovery of bidirectional order dependencies via set-based axioms.** Proceedings of the VLDB Endowment, 2018.

[2] Langer, P.; Naumann, F.: **Efficient order dependency detection.** The VLDB Journal, 2016.

[3] Consonni, C.; Sottovia, P.; Montresor, A.; Velegrakis, Y.: **Discovering Order Dependencies through Order Compatibility.** Proceedings of the International Conference on Extending Database Technology (EDBT), 2019.

[4] Saxena, H.; Golab, L.; Ilyas, I. F.: **Distributed Implementations of Dependency Discovery Algorithms**. Proceedings of the VLDB Endowment, 2019.

[5] Szlichta, J.; Parke, G.; Lukasz, G.; Mehdi, K.; Divesh, S.: **Erratum for Discovering Order Dependencies through Order Compatibility.** *Proceedings of the International Conference on Extending Database Technology (EDBT)*, 2020.

[6] Jin, Y.; Lin, Z.; Zijing, T.: **Efficient Bidirectional Order Dependency Discovery.** *Proceedings of the International Conference on Data Engineering (ICDE)*, 2020.

[7] Schmidl, S.; Papenbrock, T.: **Efficient Distributed Discovery of Bidirectional Order Dependencies.** The VLDB Journal, 2022.

[8] Szlichta, J.; Godfreym, P.; Gryz J.: **Fundamentals of Order Dependencies**. In: Proceedings of the VLDB Endowment, 2012.

[9] Szlichta, J.; Parke G.; Jarek G.; Wenbin M.; Weinan Q.; Calisto Z.: **Business-Intelligence Queries with Order Dependencies in DB2.** *Proceedings of the International Conference on Extending Database Technology (EDBT)*, 2014.

[10] Szlichta, J.; Parke G.; Jarek G.; Wenbin M.; Przemyslaw P.; Calisto Z.: **Queries on Dates: Fast yet Not Blind.** Proceedings of the International Conference on Extending Database Technology, 2011.

Document icon made by Kiranshastry from www.flaticon.com.