PŘÍRODOVĚDECKÁ FAKULTA UNIVERZITY PALACKÉHO
KATEDRA INFORMATIKY

PŘEPISY PŘEDNÁŠEK

Databázové systémy 1

KMI/DATA1

18. prosince 2012

Martin Rotter

Abstrakt

Tento dokument obsahuje přepisy přednášek, které vedl doc. RNDr Vilém Vychodil, Ph.D.. Uvedená práce (dílo) podléhá licenci Attribution-NonCommercial-NoDerivs 3.0 Unported, a to včetně vložených souborů, to neplatí pro loga jiných společností, jako je například logo PostgreSQL.

Obsah

Seznam obrázků

Seznam tabulek

Seznam zdrojových kódů

1 Přehled databázových systémů a jejich historie

Databázový systém (anglicky Database Management System (DBMS)) je soustava komplexního aplikačního vybavení, které je podloženo určitým teoretickým základem. Jeden bez druhého nemůže existovat (resp. může, což má ale za následek formální selhání DBMS jako takového), a tak je nutné znát obě strany pomyslné databázové barikády. Databázový systém tedy tvoří:

1. Aplikační software, který je obvykle používán jako rozhraní pro přístup k databázi samotné.

2. Teorie, která formálně podkládá návrh databáze, organizaci dat a obsahuje algoritmickou stránku věci.

Mějme na paměti, že obě částí databázových systémů se rozvíjely postupně a mnohdy metodou pokus - omyl. Obecně platí, že pokud selže teoretický základ, tak již ani sebelepší frontend nic nezmůže. V databázovém systému se objeví formální rozpor a jedinou cestou vpřed je začít znovu.

Hlavním úkolem DBMS je poskytovat *perzistentní uložení dat*, dále poskytnout svým uživatelům konzistentní rozhraní a případně nabídnout *transa-kční zpracování dat*. Nutnto podotknout, že poslední bod nemusí být takovou samozřejmostí, jak by se mohlo zdát.

1.1 Historie databázových systémů

Potřeba organizace dat je stará jako lidstvo samo. Již staří Egypťané si vedli podrobné záznamy o výběru daní, stavbách chrámů a jiných činnostech. Zde můžeme hovořit o databázích založených na souborech.

1.1.1 Databáze založené na souborech

Souborem může být například papyrus, hliněná destička nebo (lépe) papír. Na papíře můžete být napsány v řádcích nějaké záznamy. Například seznam dlužníků nějakého podnikatele. *Vytvoření* takového seznamu je vskutku *lehké*. Představme si, že dlužník uprostřed seznamu splatil svůj dluh a bude ze seznamu vyškrtnut. Místo něj v seznamu vznikne prázdné místo. Zbytek seznamu se následně musí zkonsolidovat (přepsat na nový papír), aby vypadal konzistentně. Odtud plyne určitá *těžkopádnost* vyplývající z definice souboru a z určité *nízkoúrovňovosti* práce s ním. Přitom souborem může být myšlen i soubor na disku počítače.

Představme si navíc, že daný podnikatel si vede další soubor se seznamem, kde si u každého dlužníka značí jeho adresu, aby jej mohl v případě nutnosti navštívit. Jméno dlužníka tedy uchovává hned na dvou seznamech, přitom je přirozeně jasné, že stačí dlužníka evidovat jednou a poté se na něj „odkazovat.“ *Redundace dat* je tedy zřejmá.

1.1.2 Databáze založené na síťovém modelu

Průkopníkem této oblasti byl Charles Bachman. Aktuálně se toto paradigma považuje za dávno překonané, nicméně pro časy budoucí poskytlo několik dobrých poznatků. Typicky „síťový“ databázový systém je tvoří mj. i:

1. *Záznamy*, které mají určitý *typ*. Ten deklaruje, jaké *atributy* daný zázna obsahuje. Atributy mohou nabývat různých hodnot.

2. *Odkazy*, které reprezentují *vztahy* mezi jednotlivými záznamy.

Obecně se ví, že databázové systémy tohoto typu mohou poskytovat velmi rychlé prostředky pro získávání jednoduchých dat, ale na druhou stranu na komplexnější data se dotazuje hůře. Tvorba dat není rovněž jednoduchá, protože práce s odkazy požaduje vyšší míru sofistikace.

1.1.3 Databáze založené na hierarchickém modelu

Jedná se o speciální typ síťového modelu, kde bylo úmyslem tento model zjednodušit. Výsledkem bylo vytvoření hierarchie v záznamech a odkazech ve formě stromu[1].

1.1.4 Databáze založené na objektovém modelu

V objektovém modelu jsou hlavní prvky, jenž reprezentují data, *objekty*, tery serializované entity, které lze interpretovat pomocí nějakého vyššího programovacího jazyka. Prvním jazykem podporujícím takovéto pojetí byl Common Lisp.

1.1.5 Databáze založené na relačním modelu

Jedná se o aktuálně používaný model, který přináší určité výhody. Primárním nosičem informace v tomto modelu jsou *relace* (poněkud amatérsky je můžeme nazývat také tabulkami). Základy tohoto modelu položil Edgar Codd v roce 1970.

2 Relační model dat

Model má 3 základní komponenty:

1. *Struktury*, které uchovávají data a reprezentují výsledky dotazů.

2. *Integritní omezení*, která popisují vztahy mezi daty, které musí být splněny. Integritní omezení jsou v podstatě formule. Tato omezení slouží k popisu toho, jak mají data vypadat, slouži k odstranění jejich chyb.

3. *Manipulativní formalismy*, které říkají jak z uložených dat získat jiná data určitými operacemi.

Relační mode je konkrétním modelem dat, avšak modely jako takové lzde rozdělit do dvou skupin, tedy:

[1] Strom je jednou ze základních grafových struktur. Jedná se o (neorientovaný) graf bez kružnic. Více napoví [1, str. 91 – 96].

Tabulka 1: Záhlaví tabulky

ID	JMÉNO	ADRESA
15	Vyacheslav Drsoň	Peklo, 666
⋮	⋮	⋮

1. Abstraktní modely dat, které poskytují vyčerpávající logické definice datových struktur nebo operací s daty. Ty dohromady tvoří abstraktní formalismus. Tedy jakýsi abstraktní stroj, se kterým může uživatel operovat.
2. Konkrétní modely, pracující s persistentními daty.

3 Struktura relačních dat

Základním kamenem dat jsou tzv. „datové tabulky“, které obsahují záhlaví a další data.

1. Záhlaví (viz. tabulka 1) je množina atributů a jejich typů. U tabulky se seznamem zamě-stnanců může být atributem například jméno zaměstnance atp.
2. Vlastní data (neboli tělo) tabulky by u takovéto tabulky představovala data samotných zaměstnanců.

3.1 Požadavky na tabulky

1. Řádky by neměly mít žádné pořadí. Tabulka by měla být chápána jako množina řádků. A v množině na pořadí nezáleží.
2. To samé platí pro sloupce.
3. V definice množiny také vyplývá, že tabulka by neměla obsahovat identické řádky. Tento požadavek není v mnoha DBMS splněn.
4. Všechny atributy jsou regulární.

Tyto požadavky formuloval C. J. Date. Pokud tabulka splňuje body 1 až 4, pak je v 1. normální formě.

Zavedli jsme si několik pojmů jako *atribut* a tak dále. Formalizujme je nyní přesněji:

Atribut popisuje účel daného sloupce v tabulce, je to jméno sloupce.

Typ říká, jakých hodnot může atribut nabývat.

Doména je množinou všech možných hodnot daného typu. Tedy například nějaký pomyslný typ `BOOL` by mohl mít doménu ve tvaru:

$$\mathrm{dom}(BOOL) = \{true, false\}$$

3.2 Relační schéma

Již dříve jsme si uvedli pojem *záhlaví tabulky*, avšak ten byl poněkud vágní. Relační schéma je jeho analogií s přesnější definicí.

Definice 3.1 (Relační schéma)
Mějme množinu atributů Y a množinu typů T Následně platí, relační schéma R je množina definovaná jako:

$$R = \{\langle y, t\rangle \mid y \in Y, t \in T\}$$

Proveďme dohodu, že typ atributu y budeme zapisovat také jako typ(y) a analogicky doménu pro každý typ budeme značit jako dom(y).

3.3 Obecné kartézské součiny

Mějme následující systém množin:

$$\{A_i \mid i \in I\}$$

kde I je tzv. indexní (a libovolná) množina a i je index z této množiny. Kartézský součin $A_i (i \in I)$ se značí $\prod_{i\in I} A_i$ a je definován jako množina zobrazení $f : I \to \bigcup_{i\in I} A_i$, kde $f_i \in A_i$ pro každý $i \in I$.

Prvky $\prod_{i\in I} A_i$ jsou zobrazení, nezáleží v nich na pořadí indexů.

Příklad 3.1 (Kartézský součin)
Vypočítejme kartézský součin množin A a B, tedy $A \times B$ se zadáním:

$$A = \{a, b, c\}, \quad B = \{10, 10\}, \quad I = \{1, 2\}, \quad A_1 = A, \quad A_2 = B$$

Následně řešení:

$$\prod_{i\in\{1,2\}} A_i = \{\langle a, 10\rangle, \langle b, 10\rangle.\langle a, 20\rangle, \langle b, 20\rangle\}$$

Příklad 3.2 (Kartézský součin s tečkovou indexní množinou)
Mějme $I = \{\bullet\}$ a $A_\bullet$. Pak platí, že $\prod_{i\in I} A_\bullet = A_\bullet$ s funkcí $f : \{\bullet\} \to A_\bullet$.

Příklad 3.3 (Kartézský součin žádné množiny)
Nechť $I = \emptyset$. Pak $\prod_{i\in I} A_i = \{\emptyset\}$ s funkcí $f : \emptyset \to \emptyset$.

3.4 Relace nad relačním schématem

Definice 3.2 (Relace nad relačním schématem)
Mějme relační schéma R, pak relace nad R je libovolná konečná podmnožina na kartézském součinu domén atributů z R, tedy

$$\mathcal{D} \subseteq \prod_{y\in R} \text{dom}(y)$$

Relace v takovémto pojetí přibližně odpovídá vágnějšmu pojmu *tabulka*. Součin lze pochopitelně také rozepsat. Obsahuje-li relační schéma R celkem n atributů, pak platí, že:

$$\mathcal{D} \subseteq \mathrm{dom}(A_1) \times \cdots \times \mathrm{dom}(A_n)$$

Pozorný čtenář určitě pozoruje, že každá relace (tabulka) může mít určitý maximáln možný počet řádků. Matematicky:

$$|\mathcal{D}| = |dom(A_1)| \times \cdots \times |dom(A_n)|$$

Relaci lze také označit jako dvojici $\langle \mathcal{D}, R \rangle$. Tato dvojice takřak kompletně popisuje tabulku.

Každý prvek relace (tabulky) se nazývá n-tice. To je analogické vzhledem k matematickému pojetí relaci jako takových. Matematické relace též obsahují n-tice.

3.4.1 Speciální případy relací

1. Prázdná relace (tabulka) nad R, tedy $\langle \emptyset, R \rangle$.
2. Tabulka „dum“, tedy $\langle \emptyset, \emptyset \rangle$.
3. Tabulka „dee“, tedy $\langle \{\emptyset\}, \emptyset \rangle$.

4 Tutorial D

Programovací jazyk Tutorial D formuje opozici vůči mnohem známějšímu a v praxi používanějšímu Structured Query Language (SQL). Tutorial D má jedinou funkční a (víceméně) použitelnou implementaci známou jako Rel. Frontend Rel je naprogram v jazyce Java a jeho prostředí vypadá bohužel hrůzostrašně. Rel je silně staticky typovaný, nemá konverze typů. Obsahuje typy:

Skalární , které jsou opozitem hodnotových typů ze známějších jazyků a patří sem například `BOOLEAN`, `INTEGER` nebo `STRING`.

N-ticové , které jsou dané jmény atributů a jejich typy.

Dále Rel disponuje *relačním* typem, který představuje instance tabulek. Rel pracuje s proměnnými konstruktivně, proměnné nemohou měnit svůj obsah (hodnotu), místo toho se vytváří nová proměnná s novou hodnotou. Každá proměnná má tedy svůj typ.

4.1 Výrazy versus příkazy

Výrazem v Tutorial D (resp. v Rel) je například prostá logická rovnost hodnot. Výraz může mít nějaký výsledek. Příkaz se od výrazu liší tak, že končí středníkem a obvykle obsahuje volání nějaké funkce.

Zdrojový kód 1: Výraz a příkaz (Tutorial D)

```
/* výraz */
666 = 6661215
/* příkaz */
WRITELN(10 = 20);
```

4.2 N-tice jako hodnoty

V Tutorial D lze přímočaře vytváře řádky relací a považovat je za hodnoty, viz. kód 2.

Zdrojový kód 2: Základy n-tic (Tutorial D)

```
/* prázdná n-tice */
TUPLE {}
/* n-tice včetně dat */
TUPLE {id 666, name "Vilík", lab 5077}
/* rovnost n-tic */
WRITELN(TUPLE {id 666, name "Vilík", lab 5077} = TUPLE {id 007, name
    "James␣Bond", lab 1111});
/* vnořené n-tice */
TUPLE {id 666, TUPLE {jmeno "Vilík" prijmeni "Devil"}, lab 5077}
```

Z n-tic lze pochopitelně získávat hodnoty nebo n-tice sjednocovat, zúžovat a tak podobně.

Zdrojový kód 3: Operace s n-ticemi (Tutorial D)

```
/* vypíše hodnoty atributů id a lab (zúžení n-tice) */
TUPLE {id 666, TUPLE {jmeno "Vilík" prijmeni "Devil"}, lab 5077} {id
    , lab}
/* vypíše vše kromě lab (zúžení n-tice) */
TUPLE {id 666, TUPLE {jmeno "Vilík" prijmeni "Devil"}, lab 5077} {
    ALL BUT lab}

/* sjednocení n-tic */
TUPLE {id 666, jmeno "Vilík"} UNION TUPLE {jmeno "Vilík", lab 5077}

/* přejmenování atributů n-tice */
TUPLE {id 666, jmeno "Vilík"} RENAME (id AS cislo, jmeno AS name)
```

Sjednocení n-tic bere dvě n-tice, které se shodují na nějakém atributu a ve výsledku se objeví konkatenace n-tic s jedním výskytem společného atributu. Kompozice n-tic funguje obdobně jako sjednocení, avšak ve výsledku se společné atributy vůbec neobjeví.

Zdrojový kód 4: Další operace s n-ticemi (Tutorial D)

```
/* rozšíření n-tice */
EXTEND TUPLE {id 666, jmeno "Vilík"} ADD (5077 AS lab)
/* aktualizace hodnot v n-tici */
UPDATE TUPLE {id 666, jmeno "Vilík"} (jmeno := "God")
```

4.3 Relace v Tutorial D

Tutorial D (potažmo Rel) samozřejmě podpruje relace.

Zdrojový kód 5: Operace s relacemi (Tutorial D)

```
/* relace s několika n-ticemi*/
RELATION {
    TUPLE {id 666, jmeno "Vilík"},
    TUPLE {id 007, jmeno "Bond"}
}
/* vytvoření relace jakožto typu */
VAR seznam_lidi BASE RELATION {id INTEGER, jmeno STRING} KEY {id};
```

5 Modifikace dat

Již známe pojmy jako relace nebo relační proměnná. Modifikací dat v databázi se myslí přiřazení nové hodnoty k nějaké relační proměnné. V SQL se k tomuto používají příkazy `INSERT`, `UPDATE` nebo `DELETE`. V Tutorial D k obdobným operacím slouží operátor „:=“, tak jak jej známe například z jazyka Pascal.

Zdrojový kód 6: Modifikace relace (SQL)

```
UPDATE  zakaznici
SET     plat = 0
WHERE   (plat > 50000);
```

Zdrojový kód 7: Modifikace relace (Tutorial D)

```
INSERT zakaznici RELATION {TUPLE {id 666, jmeno "Vilík", }};
```

K získávání dat z databáze slouží tzv. *relační dotazování*. Formátování dotazů pak definuje *dotazovací jazyk*. Ten říká, jak se budou dotazy vyhodnocovat. Tyto jazyky jsou obvykle deklarativní. Známe například SQL a Tutorial D.

Relační algebra specifikuje množinu operací s relacemmi a dotazy se skládají z postupné aplikace těchto operací. Dotazy se formulují pomocí termů a vyhodnocování dotazů odpovídá vyhodnocování termů v algebře. Vyhodnocování může být přímočaré, avšak u složitějších dotazů rovněž netriviální.

Relační kalkuly (výpočty nad relacemi) existují hned v několika variantách:

- doménový kalkul
- n-ticový kalkul

Dotaz je formule predikátové logiky, ve které relační symboly označují relační proměnné a vyhodnocování dotazu je ohodnocování formulí v dané predikátové struktuře (ta představuje instanci databáze), ve které jsou relační symboly interpretovány n-árními relacemi.

Doménový relační kalkul má zhruba stejnou sílu jako relační algebra.

6 Relační algebra

Relační algebra poskytuje formální podklad pro množinové relační operace. Ty jsou analogií ke standardním operacím na množinách.

1. *Průnik*, který obsahuje pouze společné n-tice dvou relací. Mějme tedy relace $\mathcal{D}_1$ a $\mathcal{D}_2$ nad relačním schématem $R \subseteq Y$. Následně průnik definujeme jako:

$$\mathcal{D}_1 \cap \mathcal{D}_2 = \{r \in \prod_{y \in R} dim(y) \mid r \in \mathcal{D}_1 \text{ a zároveň } r \in \mathcal{D}_2\}$$

2. *Sjednoceni*, které obsahuje ty n-tice, které se vyskytnou alespoň v jedné ze vstupních relací. Mějme tedy relace $\mathcal{D}_1$ a $\mathcal{D}_2$ nad relačním schématem $R \subseteq Y$. Následně sjednocení definujeme jako:

$$\mathcal{D}_1 \cup \mathcal{D}_2 = \{r \in \prod_{y \in R} dim(y) \mid r \in \mathcal{D}_1 \text{ nebo } r \in \mathcal{D}_2\}$$

3. *Rozdíl*, který obsahuje n-tice, které se nacházejí v první relaci, avšak nenacházejí se relaci druhé. Mějme tedy relace $\mathcal{D}_1$ a $\mathcal{D}_2$ nad relačním schématem $R \subseteq Y$. Následně rozdíl definujeme jako:

$$\mathcal{D}_1 \setminus \mathcal{D}_2 = \{r \in \prod_{y \in R} dim(y) \mid r \in \mathcal{D}_1 \text{ a zároveň } r \notin \mathcal{D}_2\}$$

Tyto operace lze provádět také v SQL či v Tutorial D.

Zdrojový kód 8: Relační operace (Tutorial D)

```
rexpr1 INTERSECT /* nebo UNION či MINUS */ rexpr2
```

Zdrojový kód 9: Relační operace (SQL)

```
SELECT * FROM table1
    INTERSECT /* nebo UNION či EXCEPT */
SELECT * FROM table2;
```

Množinové operace v SQL používají implicitně volbu `DISTINCT`, takže duplicitní n-tice jsou ignorovány. Naproti tomu u operací typu `SELECT` se jako výchozí používá `ALL`.

6.1 Efektivita množinových operací

Efektivitá závisí na implementaci fyzciké vrstvý databázového systému. Obecně platí, že pokud lze na množině potřebných n-tic zavést lineární uspořádání, tak lze množinové operace provádět pomocí slévání.

Je třeba aby dané uspořádání bylo navíc totální. Tedy každé dvě n-tice musejí být porovnatelné. Relace uspořádání musí být tranzitivní, symetrická a reflexivní. Z požadavku totality musí být uspořádání také úplné.

Pokud máme na každé doméně zavedeno toto uspořádání $\leq y$ a zavedeme uspořádání také pro dané relační schéma R, tedy $\leq R$, pak nám všechna uspořádání

$$\leq_y (y \in R) \text{ a } \leq_R$$

indukují uspořádání na n-ticích $r_1 < r_2$ právě tehdy, když existuje atribut $y \in R$ tak, že

$$r_1(z) = r_2(z) \text{ pro každé } z <_R y \text{ a navíc } r_1(y) <_y r_2(y)$$

V SQL se efektivita zajišťuje použitím tzv. indexu. Index dokáže provést ono totální uspořádání.

Zdrojový kód 10: Tvorba indexu (SQL)

```sql
CREATE INDEX muj_index ON moje_tabulka (sloupec_1, sloupec_2);
```

6.2 Další pojmy

Je třeba vysvětlit několik dalších pojmů:

Nadklíč (superkey) Mějme relaci $\mathcal{D}$ nad relačním schématem R, která obsahuje několik n-tic. Jelikož je $\mathcal{D}$ množina, tak nemůže obsahovat dvě stejné n-tice. Tedy dvě n-tice, které jsou shodné na všech atributech z R. Navíc mohou přirozeně existovat další podmnožiny R (i $R \subseteq R$), pro které platí, že žádné dvě n-tice by se na těchto podmnožinách neměli shodovat. Každou takovou podmnožinu R nazýváme *superklíč*. Každá relace má alespoň jeden superklíč – ten, který je tvořen všemi atributy.

Klíč (key) Mějme libovolný nadklíč a relaci $\mathcal{D}$. Takový nadklíč může obsahovat „nadbytečné“ prvky resp. atributy. Jako *klíč* K označíme superklíč takový, že odstraněním jakéhokoliv dalšího atributu z klíče K by vedlo k tomu, že K již nebude nadklíč. Tedy:

1. Pro každé různé n-tice musí platit, že nemohou mít shodné hodnoty na (všech) atributech z klíče.
2. Klíč je minimálním nadklíčem.

6.3 Relační operace

6.3.1 Projekce

Mějme relací R na schématu T. Projekce $\mathcal{D}$ na $R \subseteq T$ je relace $\pi_R(\mathcal{D})$, definovaná jako:

$$\pi_R(\mathcal{D}) = \{r \in \prod_{y \in R} dim(y) \mid \text{existuje } s \in \prod_{y \in \pi_R} \text{ tak, že } rs \in \mathcal{D}\}$$

kde rs znázorňuje zřetězení n-tic, kde část $r \in R$ a $s \in T \setminus R$.

V Tutorial D lze n-tice řetězit prostřednictvým klíčového slova `COMPOSE`.

Zdrojový kód 11: Projekce (SQL)

```
/* část id, jmeno, plat je projekcí */
SELECT DISTINCT id, jmeno, plat FROM table_1;
```

Zdrojový kód 12: Projekce (Tutorial D)

```
/* část {id} je projekcí */
TUPLE {id 666, jmeno "Vilík"} {id}
/* vše kromě id */
TUPLE {id 666, jmeno "Vilík"} {ALL BUT id}
```

Mějme na vědomí, že bez `DISTINCT` se ze zřejmých důvodů nejedná o relační operaci.

6.3.1.1 Aspekty fyzické vrstvy

Projekce v SQL přes `DISTINCT` je rychlá operace. V případě užití `DISTINCT` se už o rychlou operaci jednat nemusí. To ale neplatí v případě, že $R \subseteq T$ obsahuje klíč. Projekcí umí zrychlit:

1. Hashování, kde se potencíální n-tice, které se objeví ve výsledku, hashují.
2. Použití totálního uspořádání n-tic.

Věta 6.1 (Užití totálního uspořádání pro zrychlení projekce)
Pokud máme $S \subseteq R \subseteq T$, pak

$$\begin{aligned} \pi_S(\pi_R(\mathcal{D})) &= \pi_S(\mathcal{D}) \\ \pi_T(\mathcal{D}) &= \mathcal{D} \\ \pi_R(\mathcal{D}) &= \begin{cases} \emptyset & \text{pokud } \mathcal{D} = \emptyset \\ \{\emptyset\} & \text{pokud } \mathcal{D} \neq \emptyset \end{cases} \end{aligned}$$

6.3.2 Restrikce (selekce)

Mějme danou relaci $\mathcal{D}$ a formulí popisující podmínku $\varphi : t_1 \approx t_2$. t_1 je atribut a t_2 je hodnota z domény atributu. Následně restrikce:

$$\sigma_\varphi(\mathcal{D}) = \{t \in \mathcal{D} \mid t \text{ splňuje } \varphi\}$$

Zdrojový kód 13: Restrikce (SQL)

```
/* část id = 666 je restrikcí */
SELECT * FROM table_1 WHERE id = 666;
```

Zdrojový kód 14: Restrikce (Tutorial D)

```
/* část id = 666 je restrikcí */
relexpr WHERE id = 666;
```

Věta 6.2 (O komutativitě restrikcí)
Platí, že:

$$\sigma_{\varphi}(\sigma_{\psi}(\mathcal{D})) = \sigma_{\psi}(\sigma_{\varphi}(\mathcal{D})) = \sigma_{\varphi \times \psi}(\mathcal{D})$$

6.3.2.1 Záměna pořadí projekce a restrikce

Prohlašme, že $\pi_R(\sigma_{\varphi}(\mathcal{D})) \backsim \sigma_{\varphi}(\pi_R(\mathcal{D}))$. Pokud platí, že φ závisí na některém atributu z $T \setminus R$, pak pravá strana nedává smysl.

Předchozí poznatek má použití, a sice duální operaci k projekci, tedy přidání atributů.

Zdrojový kód 15: Přidání atributů (SQL)

```sql
SELECT table_1.*, expr AS additional_column FROM table_1;
```

Zdrojový kód 16: Přidání atributů (Tutorial D)

```
EXTEND relexpr ADD (expr AS additional_column)
```

Zdrojový kód 17: Kombinace restrikce a duality projekce (SQL)

```sql
SELECT  money * 2 AS doubled_money, money, id
FROM    employees
WHERE   money * 2 > 5000;
```

Zdrojový kód 18: Kombinace restrikce a duality projekce (Tutorial D)

```
((EXTEND relexpr ADD (money * 2 AS doubled_money))
WHERE doubled_money > 5000) {doubled_money, money, id}
```

6.3.3 Sumarizace (agregace)

Zdrojový kód 19: Počet n-tic v relaci (SQL)

```sql
SELECT COUNT(*) FROM my_table;
```

Výsledkem obdobného dotazu v SQL bude vždy relace. Například v tomto případě relace o jedné jednoprvkové n-tici, která obsahuje počet řádků původní relace.

Zdrojový kód 20: Počet n-tic v relaci (Tutorial D)

```
COUNT(relexpr)
```

Naopak v Tutorial D nebude výsledkem relace, avšak skalární hodnota. Tedy jednoduše číslo. Chování jazyka SQL lze v Tutorial D umitovat.

Zdrojový kód 21: Imitace chování SQL (Tutorial D)

```
SUMMARIZE relexpr ADD (COUNT() AS count_of_tuples)
```

Zdrojový kód 22: Pokročilá sumarizace (Tutorial D)

```
SUMMARIZE relexpr1
PER (relexpr2)
ADD (summary AS name)

/* místo PER lze použít také BY */
/* platí, že BY {seznam atributů} odpovídá PER (relexpr1 {seznam
    atributů}) */
SUMMARIZE relexpr1
BY (seznam atributů)
ADD (summary AS name)
```

Navíc by mělo platit, že `relexpr1` $\subseteq$ `relexpr2`. Význam zdrojového kódu 22 je takový, že jdeme přes všechny n-tice z `relexpr2` a počítáme sumarizaci ze všech n-tic z `relexpr1`, které se shodují na atributech z relačního schématu z `relexpr2`.

6.3.4 Seskupování

Uplatňuje se v modelech, které umožňují atributy s doménami, které jsou množiny relací.

6.3.4.1 Metoda UNGROUP

Zdrojový kód 23: Seskupování UNGROUP (Tutorial D)

```
/* atributy jsou typu relace */
relexpr UNGROUP (atribut, dalsi_atribut)
```

Výsledkem je relace nad schématem, které vznikne tak, že místo atributů ze seznamu budeme mít atributty vnořených tabulek a výsledek obsahuje n-tice tak, že pro každou n-tici je ve výsledku (obecně) několik n-tic, jejichž hodnoty jsou brány z tabulek ve výchozí n-tici metodou každý s každým.

6.3.4.2 Metoda GROUP

Zdrojový kód 24: Seskupování GROUP (Tutorial D)

```
/* atributy jsou typu relace */
relexpr GROUP (atribut, dalsi_atribut)
```

Na základě relace a seznamu atributů množin atributů, které se mají seskupit pod danými novými jmény se vytvoří relace, ve které jsou zbylé atributy výchozí relace a nové atributy, které nabývají hodnot, v nichž jsou maximální relace obsahující n.tice hodnot výchozí tabulky tak, že individuální hodnoty výsledných n-tic jsou v relaci se vzniklými relacemi právě tehdy, když zřetězení libovolných individuálních n-tic se zbylými individuálními hodnotami z každé výsledné n-tice se nachází ve výchozí tabulce.

Tabulka 2: Přirozené spojení tabulek

(a) První operand

w	x	y
1	1	2
2	2	2
3	3	2
4	3	4
5	5	5

(b) Druhý operand

x	y	z
1	1	2
2	2	1
3	2	4
4	4	5
5	5	5

(c) Výsledná tabulka

w	x	y	z
2	2	2	1
3	2	3	4
5	5	5	5

6.3.5 Přejmenování

Z relace $\mathcal{D}$ nad relačním schématem R se vytvoří nová relace

$$\rho\, y'_1 \Leftarrow y_1, \ldots, y'_n \Leftarrow y_n(\mathcal{D})$$

nad schématem R, ve kterém byly atributy $y_1, \ldots, y_n$ přejmenovány na atributy $y'_1, \ldots, y'_n$, ale za předpokladu, že každé dva atributy y_i, y'_i pro $1 \leq i \leq n$ mají stejný typ.

Zdrojový kód 25: Přejmenování (Tutorial D)

```
relexpr RENAME (old AS new, old_2 AS new_2))
```

Zdrojový kód 26: Přejmenování (SQL)

```
SELECT old AS new, old_2 AS new_2 FROM table_1;
```

6.3.6 Přirozené spojení

Definice 6.1 (Přirozené spojení)
Mějme relace $\mathcal{D}_1$ nad schématem $R \cup S$ a $\mathcal{D}_2$ nad schématem $S \cup T$ tak, že R, S, T jsou po dvou disjunktní a S je množina všech společných atributů. Pak je přírozené spojení $\mathcal{D}_1, \mathcal{D}_2$ relace nad schématem $R \cup S \cup T$ dáno předpisem

$$\mathcal{D}_1 \bowtie \mathcal{D}_2 = \{rst \mid rs \in \mathcal{D}_1, st \in \mathcal{D}_2\}$$

Zdrojový kód 27: Přirozené spojení (Tutorial D)

```
/* výsledkem je TUPLE {x 10, y 20, z 30} */
TUPLE {x 10, y 20} UNION TUPLE {y 20, z 30}
```

Příklad 6.1 (Přirozené spojení)
Přírozeným spojením získáme tabulku, který má společné záhlaví, avšak obsahuje pouze n-tice, které se shodují na společných atributech daných vstupních tabulek resp. relací. Názorně na tabulkách ve skupině tabulek 2.

Seznam zkratek

DBMS Database Management System

SQL Structured Query Language

Seznam teorémů

Literatura

[1] Thomas H. Cormen, Charles E. Leiserson, and Ronald L. Rivest. *Introduction to Algorithms*. 1st ed. Cambridge, MA: The MIT Press, June 18, 1990. ISBN: 0-262-03141-8.

[2] Carlos Coronel, Steven Morris, and Peter Rob. *Database Systems: Design, Implementation, and Management*. 9th ed. Boston, MA: Course Technology, Nov. 23, 2009. ISBN: 0-538-74884-2.

[3] Ramez A. Elmasri and Shamkant B. Navathe. *Fundamentals of Database Systems*. 6th ed. Boston, MA: Addison-Wesley, Apr. 9, 2010. ISBN: 0-136-08620-9.