KATEDRA INFORMATIKY
UNIVERZITA PALACKÉHO V OLOMOUCI
www.inf.upol.cz

Jan Laštovička
https://www.inf.upol.cz/lide/jan-lastovicka
jan.lastovicka@upol.cz
17. listopadu 12, 771 46 Olomouc

Databáze ◇ poznámky k přednášce

7. Dokumentový model databáze

verze z 4. listopadu 2024

Začněme motivací. Představme si, že vlastníme jediný film *Dracula* z roku 1992 od režiséra *Francis Ford Coppola* a zajímáme se o herce *Gary Oldman*, *Anthony Hopkins* a *Tom Cruise*. Vydělujeme část reálného světa určenou vlastněnými filmy a herci, o které se zajímáme.

1 Schémata

Vhodné řetězce písmen anglické abecedy a podtržítka nazýváme **atributy**. Například `title`, `year`, `actor`, `name`.

Atributy rozdělíme na **atomické** a **kolekční**. Například můžeme určit `title`, `year` a `name` za atomické atributy a `actor` za kolekční atribut.

Každému atomickému atributu y je přiřazena spočetná množina D_y nazývaná **doména** atributu y. Například shodnou doménou $D_{\texttt{title}}$ a $D_{\texttt{name}}$ atributů `title` a `name` může být množina všech řetězců nad vhodnou abecedou a doménou $D_{\texttt{year}}$ atributu `year` množina všech celých čísel.

Zavedeme pojem (**dokumentového**) **schématu hloubky** k, kde k je nezáporné celé číslo. Prázdná množina je schéma hloubky nula. Vezměme schémata $S_1, \ldots, S_n$ hloubky ostře menší než k a po dvou různé atributy $y_1, \ldots, y_n$ takové, že pokud je y_i atomický atribut, tak S_i je prázdná množina. Pak je množina

$$\{\langle y_1, S_1\rangle, \ldots, \langle y_n, S_n\rangle\}$$

schéma hloubky nejvýše k. Například $\{\langle \texttt{title}, \emptyset\rangle, \langle \texttt{year}, \emptyset\rangle\}$ je schéma hloubky jedna a $\{\langle \texttt{title}, \emptyset\rangle, \langle \texttt{actor}, \{\langle \texttt{name}, \emptyset\rangle\}\rangle\}$ je schéma hloubky dva.

Množina S je (**dokumentové**) **schéma**, jestliže existuje k takové, že S je schéma hloubky k. Například $\emptyset$, $\{\langle \texttt{title}, \emptyset\rangle, \langle \texttt{year}, \emptyset\rangle\}$ a $\{\langle \texttt{title}, \emptyset\rangle, \langle \texttt{actor}, \{\langle \texttt{name}, \emptyset\rangle\}\rangle\}$ jsou schémata.

Dvojici $\langle y, T\rangle$ ve schématu budeme také zapisovat jako „$y.T$“. Dále zápis „$y.\emptyset$“ budeme zkracovat na y. Schémata z předchozího příkladu můžeme tedy zkráceně zapsat $\{\texttt{title}, \texttt{year}\}$ a $\{\texttt{title}, \texttt{actor}.\{\texttt{name}\}\}$

Pro schéma S označíme $\mathrm{Y}(S)$ množinu

$$\{y \mid \langle y, S'\rangle \in S\}$$

atributů, které se nacházejí v schématu S. Například $\mathrm{Y}(\{\texttt{title}, \texttt{actor}.\{\texttt{name}\}\}) = \{\texttt{title}, \texttt{actor}\}$.

Všimněme si, že schéma S je zobrazení přiřazující atributu z $\mathrm{Y}(S)$ schéma $S(y)$ nižší hloubky. Například pro $S = \{\texttt{title}, \texttt{actor}.\{\texttt{name}\}\}$ je $S(\texttt{actor}) = \{\texttt{name}\}$.

2 Kolekce

Jisté dvojice $\langle y, d\rangle$, kde y je atribut, nazýváme **komponenty**. Definujeme je pro atomické a kolekční atributy zvlášť.

Jestliže je y atomický atribut a $d \in D_y$, pak $\langle y, d\rangle$ je (**atomická**) **komponenta**. Například $\langle \texttt{title}, \texttt{"Dracula"}\rangle$ nebo $\langle \texttt{year}, 1992\rangle$ jsou atomické komponenty.

Nechť S je schéma. Pak množinu komponent

$$t = \{\langle y_1, t_1\rangle, \ldots, \langle y_n, t_n\rangle\}$$

nezveme **dokumentem** nad S, jestliže

1. $y_1, \ldots, y_n$ jsou po dvou různé,
2. $\{y_1, \ldots, y_n\} = \mathrm{Y}(S)$,
3. pro každý kolekční atribut $y \in \mathrm{Y}(S)$ je $t(y)$ kolekce nad $S(y)$.

Například

$$\{\langle \texttt{title}, \texttt{"Dracula"}\rangle, \langle \texttt{year}, 1992\rangle\}$$

je dokumentem nad $\{\texttt{title}, \texttt{year}\}$.

Konečnou množinu dokumentů nad S nazýváme **kolekcí** nad S. Například:

$$C_2 = \{\{\langle \texttt{name}, \texttt{"Gary Oldman"}\rangle\}, \{\langle \texttt{name}, \texttt{"Anthony Hopkins"}\rangle\}\}$$

je kolekce nad $\{\texttt{name}\}$. Prázdnou množinu nazýváme též **prázdnou kolekcí**.

Jestliže y je kolekční atribut a C je kolekce nad S, pak $\langle y, C\rangle$ je (**kolekční**) **komponenta**. Například:

$$\langle \texttt{actor}, \{\{\langle \texttt{name}, \texttt{"Gary Oldman"}\rangle\}, \{\langle \texttt{name}, \texttt{"Anthony Hopkins"}\rangle\}\}\rangle$$

je kolekční komponenta.

Dostáváme, že množina

$$\begin{aligned}\{&\langle \texttt{title}, \texttt{"Dracula"}\rangle,\\ &\langle \texttt{actor}, \{\{\langle \texttt{name}, \texttt{"Gary Oldman"}\rangle\},\\ &\qquad\qquad \{\langle \texttt{name}, \texttt{"Anthony Hopkins"}\rangle\}\}\rangle\}\end{aligned}$$

je dokumentem nad $\{\texttt{title}, \texttt{actor}.\{\texttt{name}\}\}$.

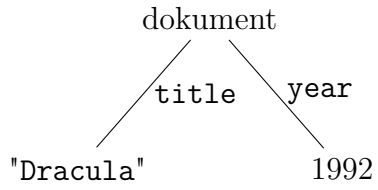
Všimněme si, že dokument nad S je zobrazení, které atributu y z $\mathrm{Y}(S)$ přiřazuje **hodnotu atributu** $t(y)$. Atributům $\mathrm{Y}(S)$ říkáme **atributy dokumentu**. Například výše uvedený dokument má atribut `title` s hodnotou `"Dracula"`.

Dokument t můžeme přirozeně zobrazit stromovým diagramem tak, že kořen zobrazuje dokument t. Následníci uzlu zobrazujícího dokument jsou v jednoznačné korespondenci s jeho komponentami. Hrana mezi uzlem a následníkem s přiřazenou komponentou je popsána atributem komponenty. Uzel s přiřazenou atomickou komponentou zobrazuje její hodnotu. Uzel s přiřazenou kolekční komponentou zobrazuje její kolekci. Následníci uzlu zobrazujícího kolekci jednoznačně odpovídají prvkům kolekce. Následník uzlu zobrazujícího kolekci zobrazuje odpovídající prvek kolekce.

Například dokument

$$\{\langle \texttt{title}, \texttt{"Dracula"}\rangle, \langle \texttt{year}, 1992\rangle\}$$

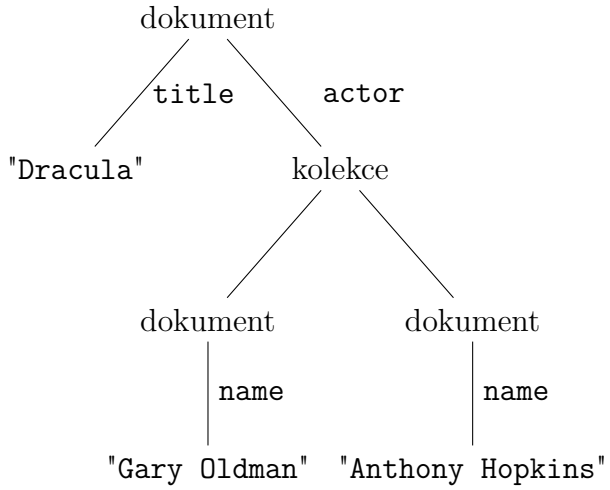
zobrazíme diagramem:



a

$$\{\langle \texttt{title}, \texttt{"Dracula"}\rangle,\\ \langle \texttt{actor}, \{\{\langle \texttt{name}, \texttt{"Gary Oldman"}\rangle\},\\ \{\langle \texttt{name}, \texttt{"Anthony Hopkins"}\rangle\}\}\rangle\}$$

diagramem:



K textovému zápisu dokumentů budeme používat formát JSON (JavaScript Object Notation). Atomické hodnoty zapíšeme přímo. Například řetězec `"Dracula"` nebo číslo `1992`. Dokument $\{\langle y_1, d_1\rangle, \ldots, \langle y_n, d_n\rangle\}$ zapíšeme řetězcem:

```
{
  "y_1": d_1,
  ⋮
  "y_n": d_n
}
```

Například dokument

$$\{\langle \texttt{title}, \texttt{"Dracula"}\rangle, \langle \texttt{year}, 1992\rangle\}$$

zapíšeme řetězcem:

```
{
  "title": "Dracula",
  "year": 1992
}
```

Kolekci $\{t_1, \ldots, t_n\}$ zapíšeme řetězcem:

```
[
  t_1,
  ⋮
  t_n
]
```

Například kolekci:

$$\{\{\langle \texttt{name}, \texttt{"Gary Oldman"}\rangle\}, \{\langle \texttt{name}, \texttt{"Anthony Hopkins"}\rangle\}\}$$

zapíšeme řetězcem:

```
[
  {
    "name": "Gary Oldman"
  },
  {
    "name": "Anthony Hopkins"
  }
]
```

a konečně dokument:

$$\{\langle \texttt{title}, \texttt{"Dracula"}\rangle,\ \langle \texttt{actor}, \{\{\langle \texttt{name}, \texttt{"Gary Oldman"}\rangle\},\ \{\langle \texttt{name}, \texttt{"Anthony Hopkins"}\rangle\}\}\rangle\}$$

zapíšeme řetězcem:

```
{
  "title": "Dracula",
  "actor": [
    {
      "name": "Gary Oldman"
    },
    {
      "name": "Anthony Hopkins"
    }
  ]
}
```

Množinu všech dokumentů nad S označíme $\mathrm{Doc}(S)$.

3 Charakteristické vlastnosti

Výroková forma $V(t)$, kde t je proměnná nabývajících hodnot z množiny $\mathrm{Doc}(S)$ se nazývá **vlastnost** nad S. Například „t(`name`) je jméno herce“ je vlastnost nad {`name`}. Proměnnou t většinou můžeme vynechat a zkráceně psát například jen „`name` je jméno herce“.

Vlastnost V nad S je **charakteristická vlastnost kolekce** C nad S, jestliže $t \in C$, právě když $V(t)$ pro každý dokument t nad S. Zkráceně budeme říkat, že C **je určeno** V.

Platí, že V je charakteristická vlastnost C, právě když

$$C = \{t \in \mathrm{Doc}(S) \mid V(t)\}.$$

Například kolekce

```
[
  {
    "title": "Dracula",
    "actor": [
      {
        "name": "Gary Oldman"
      },
      {
        "name": "Anthony Hopkins"
      }
    ]
  }
]
```

má charakteristickou vlastnost $V(t)$:

Řetězec $t(\texttt{title})$ je jméno filmu a $t(\texttt{actor})$ je kolekce s charakteristickou vlastností $V(t')$ rovnou „Herec $t'(\texttt{name})$ hrál ve filmu $t(\texttt{title})$“.

Aby bylo možné zjednodušit vlastnost. Zavedeme operaci přejmenování.

Operace přejmenování. Atomické atributy y_1 a y_2 jsou **stejného typu**, jestliže $D_{y_1} = D_{y_2}$.

Vezměme dvě schémata S_1 a S_2 a bijekci $h\colon \mathrm{Y}(S_1) \to \mathrm{Y}(S_2)$, která splňuje, že

1. $S_1(y) = S_2(h(y))$ pro každý $y \in \mathrm{Y}(S_1)$,
2. y a $h(y)$ jsou stejného typu pro každý atomický atribut $y \in \mathrm{Y}(S_1)$.

Pak bijekce h je **přejmenování atributů** mezi S_1 a S_2. Například pro $S_1 = \{\texttt{name}\}$ a $S_2 = \{\texttt{move.name}\}$ je $h = \{\langle \texttt{name}, \texttt{movie.name}\rangle\}$ přejmenování atributů.

Nechť h je přejmenování atributů mezi S_1 a S_2 a t dokument nad S_1.

Pak **přejmenování dokumentu** t podle h je dokument

$$\rho_h(t) = \{\langle h(y), d\rangle \mid \langle y, d\rangle \in t\}$$

nad S_2. Například pro výše uvedené h a dokument $t = \{\langle \texttt{name}, \texttt{"Gary Oldman"}\rangle\}$ je $\rho_h(t) = \{\langle \texttt{movies.name}, \texttt{"Gary Oldman"}\rangle\}$.

Přejmenování kolekce C nad S_1 podle h je kolekce

$$\rho_h(C) = \{\rho_h(t) \mid t \in C\}$$

nad S_2.

Předpokládejme, že $\mathrm{Y}(S_1) = \{y_1, \ldots, y_n, x_1, \ldots, x_m\}$, kde $h(x_i) = x_i$ pro každé $1 \leq i \leq m$. Pak ρ_h můžeme značit $\rho_{h(y_1)\leftarrow y_1,\ldots,h(y_n)\leftarrow y_n}$. Přejmenování podle výše uvedeného h můžeme značit $\rho_{\texttt{actors.name}\leftarrow\texttt{name}}$.

Příklad vlastnosti. Vrátíme se k příkladu charakteristické vlastnosti. Přejmenujeme atributy kolekce $t(\texttt{actor})$:

Řetězec $t(\texttt{title})$ je jméno filmu a $\rho_{\texttt{actor.name}\leftarrow\texttt{name}}(t(\texttt{actors}))$ je kolekce s charakteristickou vlastností $V(t')$ rovnou „Herec $t'(\texttt{actor.name})$ hrál ve filmu $t(\texttt{title})$“.

Nyní můžeme proměnné t, t' odstranit:

Řetězec `title` je jméno filmu a `actor` je kolekce s charakteristickou vlastností „Herec `actor.name` hrál ve filmu `title`“.

Vlastnost zjednodušíme:

Řetězec `title` je jméno filmu a `actor` je určeno vlastností: herec `actor.name` hrál ve filmu `title`.

Běžně budeme vlastnost kolekce a v ní vložené vlastnosti uvádět zvlášť. Tedy charakteristická vlastnost kolekce je:

Řetězec `title` je jméno filmu.

a charakteristická vlastnost kolekčního atributu `actor` je:

Herec `actor.name` hrál ve filmu `title`.

Obecně při formulaci charakteristické vlastnosti nad S s proměnnou t provedeme implicitně přejmenování $\rho_h(t(y))$ pro každý kolekční atribut $y \in \mathrm{Y}(S)$, kde h přejmenuje atribut $y' \in S(y)$ na $y.y'$.

4 Kolekční proměnné

Pokud má dokument atribut `_id`, říkáme, že **má identifikátor**. Hodnota atributu `_id` se nazývá **identifikátor dokumentu**.

Uvažujme schéma S takové, že $\texttt{\_id} \in \mathrm{Y}(S)$. **Kolekční proměnná** nad schématem S je proměnná, jejíž hodnotou je kolekce nad S. Dokumenty v S musí mít po dvou různé identifikátory. Schéma S nazýváme **typem proměnné**. Můžeme mít například kolekční proměnnou `movie` nad $\{\texttt{\_id}, \texttt{title}, \texttt{actor}.\{\texttt{name}\}\}$.

Každá kolekční proměnná má určenou **charakteristickou vlastnost**. Vždy platí, že hodnota proměnné má její charakteristickou vlastnost. Například `movie` má charakteristickou vlastnost:

„Film `_id` se jmenuje `title`“

a `actor` je určeno vlastností:

„Herec `actor.name` hrál ve filmu `title`.“

Hodnota proměnné `movie` je kolekce:

```
[
  {
    "_id": 1,
    "title": "Dracula",
    "actor": [
      {
        "name": "Gary Oldman"
      },
      {
        "name": "Anthony Hopkins"
      }
    ]
  }
]
```

Budeme pracovat s databázovým systémem MongoDB. Návod na jeho instalaci naleznete v dokumentu `07_mongodb.pdf`. Proměnné ani jejich schémata nemusíme v systému deklarovat. Každá kolekční proměnná *`collection`* má na začátku hodnotu rovnou prázdné kolekci.

Příkaz:

db.*collection*.insertMany(C)

změní hodnotu kolekční proměnné *collection* na $C \cup C'$, kde C' je původní hodnota proměnné *collection*. Musí platit, že množiny C a C' jsou disjunktní. Například:

```
db.movie.insertMany([
  {
    "_id": 1,
    "title": "Dracula",
    "actor": [
      {
        "name": "Gary Oldman"
      },
      {
        "name": "Anthony Hopkins"
      }
    ]
  }
])
```

Kolekční výraz je výraz, jehož hodnota je kolekce. Hodnota kolekčního výrazu:

db.*collection*.find()

je rovna hodnotě proměnné *collection*. Například:

```
> db.movie.find()
[
  {
    _id: 1,
    title: 'Dracula',
    actor: [ { name: 'Gary Oldman' }, { name: 'Anthony Hopkins' } ]
  }
]
```

Systém reprezentaci JSON zjednodušuje tak, že atributy není potřeba uzavírat do dvojitých uvozovek a řetězce mohou být uzavřeny mezi jednoduché uvozovky.

Můžeme přidat další dokumenty do kolekce:

```
db.movie.insertMany([{
  _id: 2,
  title: 'Oppenheimer',
  actor: [
```

```
    {
      name: 'Gary Oldman'
    }
  ]
}, {
  _id: 3,
  title: 'The Matrix',
  actor: []
}])
```

Nová hodnota proměnné:

```
> db.movie.find()

[
  {
    _id: 1,
    title: 'Dracula',
    actor: [ { name: 'Gary Oldman' }, { name: 'Anthony Hopkins' } ]
  },
  { _id: 2, title: 'Oppenheimer', actor: [ { name: 'Gary Oldman' } ] },
  { _id: 3, title: 'The Matrix', actor: [] }
]
```

Příkaz:

db.*collection*.drop()

nastaví proměnnou *collection* na prázdnou kolekci. Například:

```
> db.movie.drop()
```

Otázky a úkoly na cvičení

1. Vymezme si část reálného světa určenou režiséry, jejichž filmy vlastníme. U režiséra nás zajímá rok jeho narození a jeho filmy, které vlastníme. U filmu pak chceme znát jeho název a rok vydání. Navrhněte schéma S pro uložení popsaných dat.

2. Vezměte schéma S z prvního úkolu a určete množinovým zápisem kolekci C popisující následující stav vymezené reality. Od režiséra Tim Burton narozeného v roce 1958 vlastníme film *Edward Scissorhands* z roku 1990 a film *Alice in Wonderland* z roku 2010 a od režiséra Francis Ford Coppola narozeného v roce 1939 vlastníme film *Megalopolis* z roku 2024.

3. Vezměme kolekci C z druhého úkolu a zapište ji ve formátu JSON.

4. Deklarujte kolekční proměnnou nad schématem S doplněném o atribut `_id` z prvního úkolu a určete její charakteristickou vlastnost.

5. Změňte hodnotu proměnné ze čtvrtého úkolu tak, aby odpovídala stavu reality popsaném v druhém úkolu.