Implementace os XPath ve vícerozměrném přístupu pro indexování XML dat

Michal Krátký[1], michal.kratky@vsb.cz
Jaroslav Pokorný[2], pokorny@ksi.ms.mff.cuni.cz
Tomáš Skopal[1], tomas.skopal@vsb.cz
Václav Snášel[1], vaclav.snasel@vsb.cz

[1]Katedra informatiky, VŠB – Technická univerzita Ostrava
[2]Katedra softwarového inženýrství, Univerzita Karlova v Praze

Obsah

- Úvod – XML, dotazovací jazyky.
- Aktuální stav v oblasti indexování XML dat.
- Vícerozměrný přístup pro indexování XML dat.
- Dotazování, implementace os XPath, implementace podmnožin dalších jazyků.
- Datové struktury.
- Výsledky experimentů.
- Závěr.

Úvod

- E**x**tensible **M**arkup **L**anguage (**XML**) je značkovací jazyk vyvíjený W3C.
- "databázový pohled": XML je jazyk pro modelování dat.
- Dokument(y) tvoří databázi, ***DTD*** (***XML Schema)*** jsou jejími schématy.
- XML dotazovací jazyky (***XPath***, ***XQL***, ***XQuery***,…).
- Současné přístupy (relační, objektově-relační) nejsou vhodné pro indexování XML dokumentů.
- Problémem je nutnost průchodu stromem při provádění XML dotazu.

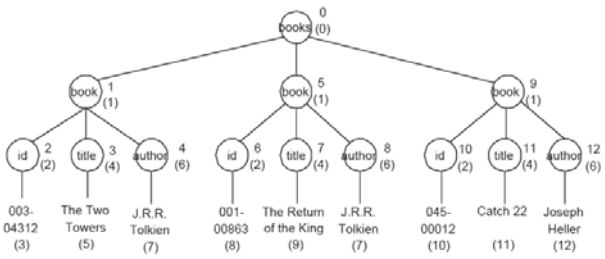
XML dotazovací jazyky – osy XPath

- XML strom.
- 13 os XPath - relace mezi uzly dokumentu.

Osa *v*/α	**Výsledek**
child	dítě uzlu *v*
descendant	potomek *v*
descendant-or-self	descendant +*v*
parent	rodič *v*
ancestor	předek *v*
ancestor-or-self	předek +*v*

following	následníci
following-sibling	následující sourozenci
preceding	předchůdci
preceding-sibling	předcházející sourozenci
attribute	atributy *v*
self	uzel v
namespace	

Osy XPath



Aktuální stav v indexování XML dat

- API: SAX, DOM - není indexování.
- Přístupy založené na relační dekompozici.
- Trie reprezentace dokumentu.
- Vícerozměrné přístupy.
- Další: signaturové přístupy apod.

Přístupy založené na relační dekompozici

- Často velmi triviální mapování XML dokumentů do relačních tabulek (Oracle apod.).
- Neposkytuje „opravdové“ indexování XML – nelze efektivně implementovat XML dotazovací jazyky.
- Sofistikovanější přístupy: ***STORED***, ***XISS***, ***Hybrid Storage Model***.

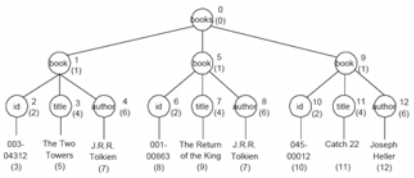
Vícerozměrné přístupy

- **prof. Bayer** – Indexování XML jako vícerozměrný problém. Velmi problematické indexování cest v jedné souřadnici.
- ***XPath Accelerator***, pro každý element je vytvořen 5-rozměrný vektor a osy XPath jsou implementovány rozsahovými dotazy.

Vícerozměrný přístup pro indexování XML dat - model

```
<!DOCTYPE books [
  <!ELEMENT books(book)>
  <!ELEMENT book(title,author)>
  <!ATTLIST book id CDATA #REQUIRED>
  <!ELEMENT title(#PCDATA)>
  <!ELEMENT author(#PCDATA)>
]>
```

```
<?xml version="1.0" ?>
<books>
 <book id="003-04312">
  <title>The Two Towers</title>
  <author>J.R.R. Tolkien</author>
 </book>
 <book id="001-00863">
  <title>The Two Towers</title>
  <author>J.R.R. Tolkien</author>
 </book>
 <book id="045-00012">
  <title>Catch 22</title>
  <author>Joseph Heller</author>
 </book>
</books>
```



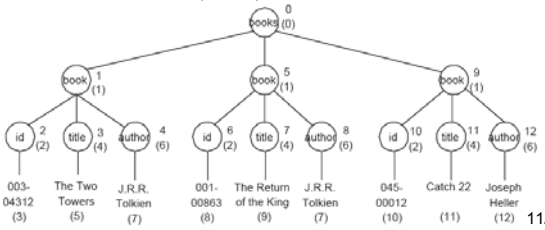
Graf je množina cest.

Cesty, značkované cesty

- $id_u(u_i)$ – jedinečné číslo uzlu u_i
- ***cesta***: p= $id_u(u_0)$, $id_u(u_1)$,..., $id_u(u_{lpn})$, s$\in X_p$,
- ***značkovaná cesta***: p=$s_0,s_1,...,s_{lpn} \in X_{pn}$,
- délka značkované cesty: l_{pn}
- délka cesty: l_{pn} nebo l_{pn+1}

Cesty, značkované cesty

- Cesty `0,1,2,'003-04212'; 0,5,6,'001-00863'` a `0,9,10,'045-00012'` náležící ke značkované cestě `books,book,id`,
- …
- Cesty `0,1,4,'J.R.R. Tolkien'; 0,5,8,'J.R.R. Tolkien'` a `0,9,12,'Joseph Heller'` náležící ke značkované cestě `books,book,author`.



Body reprezentující cesty a značkované cesty

Definice 1 (bod n-rozměrného prostoru reprezentující značkovanou cestu).

*Mějme n-rozměrný diskrétní prostor značkovaných cest $\Omega_{pn} = D^n, |D| = 2^{l_D}$. Mějme značkovanou cestu $pn = s_0, s_1, \ldots, s_{l_{pn}} \in \mathcal{X}_{PN}$, kde l_{pn} je délka značkované cesty. **Bod n-rozměrného prostoru reprezentující značkovanou cestu** je definován $t_{pn} = (id_t(s_0), id_t(s_1), \ldots, id_t(s_{l_{pn}})) \in \Omega_{pn}$, kde $id_t(s_i)$ je jedinečné číslo řetězce s_i, $id_t(s_i) \in D$. $n = max(l_{pn_i}+1, pn_i \in \mathcal{X}_{pn}, 1 \leq i \leq |\mathcal{X}_{pn}|)$. Každé značkované cestě pn_i přiřadíme jedinečné číslo $id_{pn}(pn_i)$.* ■

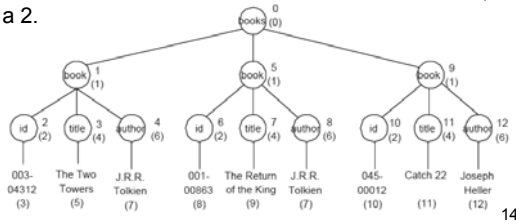
Definice 2 (bod n-rozměrného prostoru reprezentující cestu).
*Mějme n-rozměrný diskrétní prostor cest $\Omega_p = D^n, |D| = 2^{l_D}$. Mějme cestu $p = id(u_0), id(u_1), \ldots, id(u_{l_{pn}}), s \in \mathcal{X}_p$ a příslušnou značkovanou cestu pn s jedinečným číslem $id_{pn}(pn)$. **Bod n-rozměrného prostoru reprezentující cestu** je definován $t_p = (id_{pn}(pn), id_u(u_0), \ldots, id_u(u_{l_{pn}}), id_t(s)) \in \Omega_p$. $n = max(l_{p_i} + 2, p_i \in \mathcal{X}_p, 1 \leq i \leq |\mathcal{X}_p|)$.* ■

Indexy

- ***Index termů*** – uložení veškerých řetězců XML dokumentu s_i a jejich id_t (s_i).
- ***Index značkovaných cest*** – uložení bodů reprezentující značkované cesty.
- ***Index cest*** – uložení bodů reprezentující cesty.

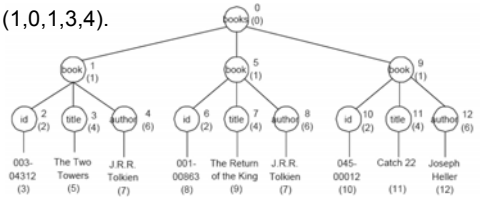
Příklad – index značkovaných cest

- `books,book,id`; `books,book,title` a `books,book,author`. Pomocí id_t názvů elementů a atributů vytvoříme vektory `(0,1,2)`; `(0,1,4)` a `(0,1,6)`, které vložíme do vícerozměrné struktury s id_{pn} 0, 1 a 2.



Příklad – index cest

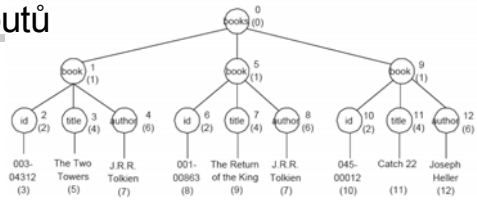
- Např. cesta k hodnotě `The Two Towers`. Jedná se o značkovanou cestu `book,book,title` s id_{pn} 1. Po vložení jedinečného čísla značkované cesty id_{pn}, jedinečných čísel elementů id_u a id_t termu `The Two Towers` získáme vektor (1,0,1,3,4).



Dotazy na hodnoty elementů a atributů

- Fáze dotazování:
  - Nalezení id_t termů dotazu v indexu termů.
  - Nalezení id_{pn} značkovaných cest dotazu v indexu značkovaných cest.
  - Nalezení vektorů v indexu cest.
- Problém vykonávání dotazů je převeden na problém definování a realizace bodových a rozsahových dotazů.

Dotazy na hodnoty elementů a atributů



- XPath dotaz: `books/book[author="Joseph Heller"]`
  - id termů z ***indexu termů***,
  - id 2 značkované cesty `books/book/author` z ***indexu značkovaných cest***: bodový dotaz `(0,1,6)`,
  - vektory z ***indexu cest***:
    rozsahový dotaz `(2,0,0,12)x(2,max,max,12)`.

Složitější dotazy

`books/book[author='J.R.R. Tolkien']`

`books//author='J.R.R. Tolkien'`

Řešeno sérií rozsahových dotazů v indexu značkovaných cest a indexu cest:

`books,author` - `books,*,author` - `books,*,…,*,author` – počet rozsahových dotazů je *n*.

Složitější (disjunktní) rozsahové dotazy je možné realizovat v datové struktuře najednou.

Implementace os XPath 1

- Z indexu cest získáme body reprezentující cesty – získáme předky uzlu *u*: $id_u(u_0)$,…, $id_u(u_{l-1})$. Osy `parent`, `ancestor` a `ancestor-or-self` uzlu *u* jsou tedy realizovány přímo.
- `descendent`: [0, $id_u(u_0)$,…, $id_u(u_{l-1})$, $id_u(u)$, 0,…, 0]×[max_D,$id_u(u_0)$,…,$id_u(u_{l-1})$, $id_u(u)$, max_D,…,max_D]
- `child` – naivní přístup - (l(u)+3). souřadnice

Implementace os XPath 2

- `preceding-siblings`: [0,$id_u(u_0)$, …,$id_u(u_{l-1})$, 0,0,…, 0] × [max_D,$id_u(u_0)$,…, $id_u(u_{l-1})$, $id_u(u)$-1, max_D,…,max_D]
- `following-siblings`: [0, $id_u(u_0)$,…, $id_u(u_{l-1})$, id(u)+1,0,…, 0] × [max_D, $id_u(u_0)$,…, $id_u(u_{l-1})$, max_D,max_D,…,max_D]
- `preceding`, `following`
- Osy `child`, `preceding-siblings` a `following-siblings` je nutné řešit složitějším způsobem.

Komplikovanější dotazy

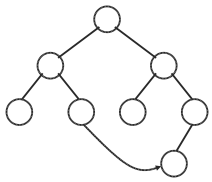
- Dotazy na hodnoty a realizace osy XPath.
- např. `books/book[author='Joseph Heller']/title`
- Kombinace dříve popsaných technik:
  - dotaz na hodnotu,
  - zjištění sourozence.

Další XML dotazovací jazyky

- XML-QL

```
where
 <list><person>
   <name>Radim</name>
   <surname>Horáček</surname>
   <email>$e<email>
  </person></list> in …
construct $e
```

ID, IDREF



- XQuery

```
document("data.xml")//article/[title='XML Book']/
   first_author/@author_id->author//surname
```

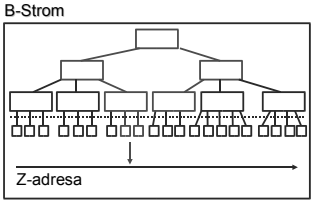

Indexující datové struktury

- Perzistentní vícerozměrné datové struktury – (B)UB-stromy, varianty R-stromů.
- Poskytují bodové a rozsahové dotazy.
- Problémy:
  - indexování vektorů různé dimenze,
  - úzké rozsahové dotazy.

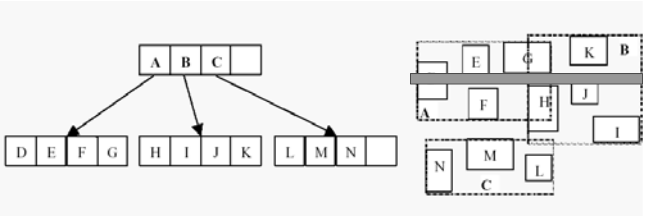
(B)UB-strom

UB-Strom

B-Strom



R-strom



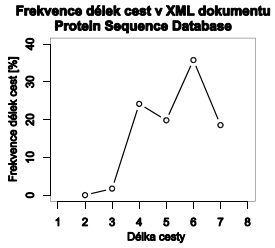
Indexování bodů různé dimenze

- Body reprezentující cesty a značkované cesty mají různou dimenzi.

	Maximální hloubka	Průměrná hloubka
DocBook	16	8.56
Shakespeare	6	4.77
XHTML	15	5.47

Vícerozměrné lesy

- BUB-les, R-les
- Každý strom lesa indexuje prostor různé dimenze.
- Např. XML dokument Protein Sequence Database: *n*=7 a *n*=9.

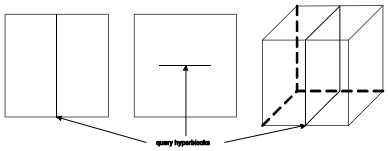


Úzké rozsahové dotazy

Definice 2 (Úzký rozsahový dotaz).

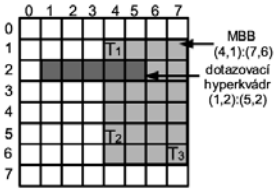
Mějme n-rozměrný diskrétní prostor $\Omega = D^n$. Dotazovací hyperkvádr je definován dvěma body $QL = (ql_1, ql_2, \ldots, ql_n)$ a $QH = (qh_1, qh_2, \ldots, qh_n)$, kde $\forall i : ql_i \leq qh_i$. Nechť konstanty ψ a ϕ jsou mohutnosti dosti malého ($\psi \rightarrow min(D)$) resp. velkého ($\phi \rightarrow |D|$) intervalu $\subseteq D$. Rozsahový dotaz nazveme úzký pokud:

1. *$\forall i : qh_i - ql_i \leq \psi \vee qh_i - ql_i \geq \phi$.*
2. *Označme počet dimenzí, pro které platí $qh_i - ql_i \leq \psi$ n_ψ a n_ϕ pro které platí $qh_i - ql_i \geq \phi$. Pro úzký rozsahový dotaz musí platit $1 < n_\psi < n \wedge 1 < n_\phi < n$.*

- 

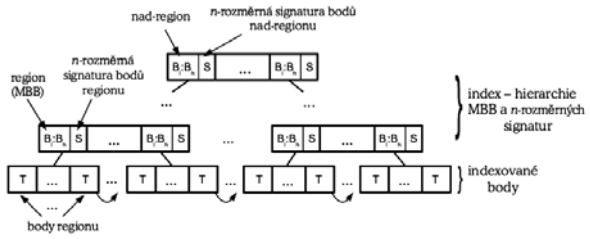
Signaturové vícerozměrné stromy

- Regiony aproximují přibližnou distribuci dat.



- Regiony protnuté dotazovacím hyperkvádrem jsou prohledány.
- S rostoucí dimenzí poměr relevantních (N_R) a protnutých (N_I) regionů $c_R \ll 1$.
- Vložením signatury dochází k přesnější aproximaci distribuce dat.

Signaturový R-strom



Výsledky experimentů

- Databáze bílkovin z XML UW projektu:
  - velikost souboru: 683MB,
  - počet elementů: 21 305 818,
  - počet atributů:1 290 647.
  - maximální délka cesty 7.
- BUB-les, R*-les, Signaturový BUB-les a R*-les. Index struktury: stromy indexující prostory dimenze 7 a 9.

Výsledky experimentů

Tabulka 1. Charakteristika indexujících datových struktur.

Dimenze prostoru n	Počet bodů	Velikost indexu [MB]			
		BUB-strom	Signaturový BUB-strom	R*-strom	Signaturový R*-strom
7	8 268 357	440.9	471 [+7%]	478.6	512.2 [+7%]
9	8 739 522	562.1	635.2 [+13%]	603.1	680.7 [+13%]

Dimenze prostoru n	Počet vnitřních uzlů				Počet listových uzlů			
	BUB strom	Signaturový BUB-strom	R* strom	Signaturový R*-strom	BUB strom	Signaturový BUB-strom	R* strom	Signaturový R*-strom
7	10 917	22 432	15 731	33 186	214 842	234 412	256 520	258 187
9	17 751	36 412	24 750	55 750	270 065	298 142	318 370	331 474

Výsledky experimentů

```
ProteinDatabase/ProteinEntry/[reference/refinfo/authors/
author='Smith, E.L.']
```

Tabulka 2. Charakteristika testovacích úzkých rozsahových dotazů.

Množina dotazů	Prostor dimenze n	Velikost výsledku	N_I		N_R		c_R	
			BUB-strom	R*-strom	BUB-strom	R*-strom	BUB-strom	R*-strom
1	7	5	116	828	1	1	0.009	0.0012
2	7	3 397	139 927	1 542	4 251	717	0.030	0.47
3	9	8	243	641	15	7	0.060	0.01
4	9	2 794	112 927	136	3 745	136	0.033	1

Výsledky experimentů

Tabulka 3. Statistika realizace rozsahových dotazů v indexu cest.

Množina dotazů	Procento prohledávaných listových uzlů				DAC				Čas vykonání dotazu [s]			
	BUB strom	Sign. BUB strom	R* strom	Sign. R* strom	BUB strom	Sign. BUB strom	R* strom	Sign. R* strom	BUB strom	Sign. BUB strom	R* strom	Sign. R* strom
1	0.05	0.04	0.32	0.061	324	319	960	74	0.13	0.1	0.08	0.015
2	2.3	0.4	0.60	0.45	15 275	1 124	1 671	1 043	5.8	0.9	0.19	0.094
3	0.04	0.03	0.20	0.013	513	489	817	477	0.2	0.2	0.20	0.140
4	2.2	0.4	0.05	0.03	14 451	1 245	220	184	5.6	0.8	0.017	0.015
Průměr	1.15	0.22	0.29	**0.14**	7 566	794	917	**445**	2.9	0.5	0.12	**0.07**

Závěr

http://www.cs.vsb.cz/arg

- Dotazování na částečnou shodu hodnot elementů a atributů.
- Komprese datové struktury.
- Implementace složitějších rozsahových dotazů.
- Kombinace s přístupy indexování nestrukturovaných dokumentů v IR.
- Implementace podmnožiny dotazovacích jazyků jako např. XQuery.

Reference

- M. Krátký, J. Pokorný, T. Skopal, V. Snášel: *The Geometric Framework for Exact and Similarity Querying XML data*. ***In Proceedings of EurAsia-ICT 2002***. Shiraz, Iran, Springer Verlag, LNCS 2510.
- M. Krátký, T. Skopal, and V. Snášel: ***Multidimensional Term Indexing for Efficient Processing of Complex Queries***. Kybernetika, Journal of the Academy of Sciences of the Czech Republic, 2003, accepted.