Maticový model jako řešení pro DIS v prostředí XML

Vladimír Rejlek

Obsah

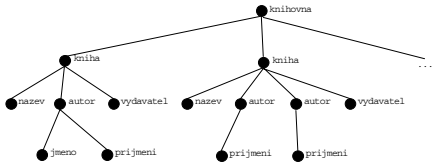
- jazyk XML
- pojem podobnosti v oblasti XML dokumentů
  - kategorizace přístupů
- přístup DIS s indexací
  - Maticový model

XML – ukázka dat

```
<knihovna>
  <kniha rok="2000">
     <nazev> XML pro každého </nazev>
     <autor>
       <jmeno> Jiří </jmeno>
       <prijmeni> Kosek </prijmeni>
     </autor>
     <vydavatel> Grada Publishing s.r.o. </vydavatel>
  </kniha>

  <kniha rok="1998">
     <nazev> Microsoft Word pro pokročilé </nazev>
     <autor>
       <prijmeni> Šimek </prijmeni>
     </autor>
     <autor>
       <prijmeni> Vacek </prijmeni>
     </autor>
     <vydavatel> Computer Press </vydavatel>
  </kniha>
...
```


XML – stromová struktura



Přístupy k podobnosti XML

1) Klasické XML dotazovací jazyky rozšířené o operátor podobnosti
2) Závislost odpovědí na dotazy nad XML daty na granularitě těchto dat
3) Podobnost mezi XML dokumenty a DTD
4) **Přístup DIS s indexací**
5) Přístup DIS bez indexace

1. XML dotazovací jazyky rozšířené o operátor podobnosti

- vychází z již navržených XML dotazovacích jazyků (XQL, XML-QL)
- přidání operátoru podobnosti (~)
- dvojí použití
  - porovnání na konstantu
  - porovnání dvou částí dat mezi sebou

1. XML dotazovací jazyky rozšířené o operátor podobnosti

- dotaz v jazyku *XXL*

```
SELECT H, S
  FROM cd01.xml, cd02.xml
  WHERE ~cd AS C
    AND C.#.interpret AS I
    AND I = "Gustav Brom se svým orchestrem"
    AND C.#.(~skladba)? AS S
    AND S.~hudebnik AS H
    AND H.# ~ "barytonsaxofon"
```

7

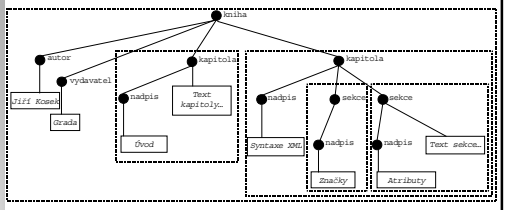
2. Závislost odpovědí na dotazy na granularitě dat

- orientace na *"text-rich"* dokumenty
- v čase konstrukce dotazu neznáme přesný tvar odpovědi
- chceme nalézt co nejrelevantnější kontext pro hledané termy
- dva způsoby řešení:
  - přídavné informace
  - speciální operátory

8

2. Závislost odpovědí na dotazy na granularitě dat

- jazyk *XIRQL* přidává *kontextové uzly*



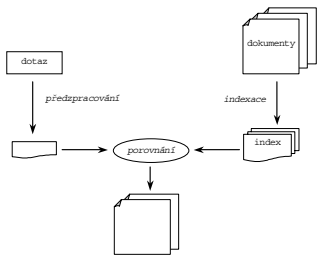
9

3. Podobnost mezi XML dokumenty a DTD

- zkoumá XML dokumenty, pro než neznáme DTD
- pro XML dokument hledáme v množině DTD to nejpodobnější
- podobnost DTD mezi sebou

10

4. Přístup DIS s indexací



11

4. Přístup DIS s indexací

- dotazy typu:

"najdi všechny dokumenty z kolekce s co největší relevancí k zadanému popisu"

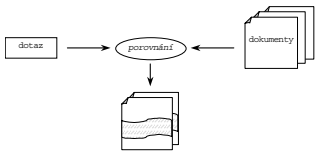
- klasické DIS vůbec nepočítají s vnitřní strukturou dokumentů
- potřeba rozšíření indexu o tyto informace

12

Podobnost XML

5. Přístup DIS bez indexace

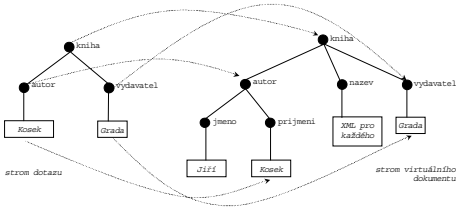
- výstupem není množina relevantních dokumentů
- ale množina relevantních *podstromů dokumentů*



Podobnost XML

5. Přístup DIS bez indexace

- přibližné vnořování stromů (jazyk *ApproXQL*)



Podobnost XML

Maticový model pro XML DIS

- přístup DIS s indexací
- přímo vychází z vektorového modelu pro DIS
- dokument je v indexu reprezentován *maticí* namísto vektorem
- přidává nový prvek: *Matice převodu cest*

Podobnost XML

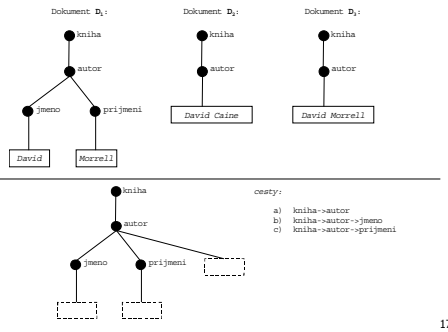
Reprezentace dokumentu

- *Reprezentací dokumentu d_i* v rámci kolekce *c* v maticovém modelu rozumíme matici $\mathbf{D_i}$ o rozměrech **m×k**, kde **m** je počet měřených termů a **k** je počet cest v XML struktuře kolekce *c*. Hodnota $\mathbf{d_{i,j,s} \in <0,1>}$ udává váhu termu $\mathbf{t_j}$ na cestě **s** v dokumentu d_i.

$$D_i = \begin{bmatrix} d_{i,1,1} & d_{i,1,2} & \dots & d_{i,1,k} \\ d_{i,2,1} & d_{i,2,2} & \dots & d_{i,2,k} \\ \dots & \dots & \dots & \dots \\ d_{i,m,1} & d_{i,m,2} & \dots & d_{i,m,k} \end{bmatrix} \in <0,1>^{m,k}$$

Podobnost XML

Příklad – strom kolekce



Podobnost XML

Příklad – matice dokumentů

```
        a  b  c      a  b  c      a  b  c
D1: [(  0, 1, 0), (  0, 0, 1), (  0, 0, 0)]

D2: [(0.5, 0, 0), (  0, 0, 0), (0.5, 0, 0)]

D3: [(0.5, 0, 0), (0.5, 0, 0), (  0, 0, 0)]

      "david"      "morrell"     "caine"
```

cesty:

a) kniha->autor

b) kniha->autor->jmeno

c) kniha->autor->prijmeni

Definice podobnosti

- podobnost

$$Sim_1(D_i, Q) = \sum_{l=1}^{m} \sum_{j=1}^{k} d_{i,l,j} * q_{l,j}$$

$$Sim_2(D_i, Q) = \sum_{l=1}^{m} \frac{\sum_{j=1}^{k} d_{i,l,j} * q_{l,j}}{\sqrt{\sum_{j=1}^{k} (d_{i,l,j})^2 * \sum_{j=1}^{k} (q_{l,j})^2}}$$

19

Matice převodu cest

- matice převodu cest
  - čtvercová reálná matice *A* o rozměrech *k* x *k*, kde *k* je počet cest v kolekci; $a_{i,j} \in$ <0,1> a $a_{i,i} = 1$
  - pro každou cestu vektor, který vyjádří vztah této cesty ke všem ostatním

20

Matice převodu cest

- Jednokrokový převod
  - Mějme matici dokumentu D o rozměrech *mxk* a matici převodu cest A o rozměrech *kxk*, kde $a_{i,j} \in$ <0,1> a $a_{i,i}=1$.
    Pak *jednokrokovým převodem* rozumíme funkci *JP(D,A)=UD*, kde *UD* je opět matice o rozměrech *mxk* a platí, že:

$$UD = \left(\max\left(d_{l,i}, \max_{j=1}^{k} \left(a_{j,i} * d_{l,j} \right) \right) \right)_{l,i}$$

21

Matice převodu cest

- Převod
  - převodem matice dokumentu *D* podle matice převodu cest *A* rozumíme *tranzitivní uzávěr* funkce *JP(D,A)*
- zjednodušeně:
  - váha termu se distribuuje po cestách podle matice převodu cest pomocí funkce maximum

22

Příklad – matice převodu cest

matice převodu cest

	a	*b*	*c*
a	1	0.2	0.2
b	0.5	1	0
c	0.5	0	1

a) kniha→autor
b) kniha→autor→jmeno
c) kniha→autor→prijmeni

matice dokumentů po převodu

	a	*b*	*c*	*a*	*b*	*c*	*a*	*b*	*c*
UD_1:	[(0.5,	1,	0.1),	(0.5,	0.1,	1),	(0,	0,	0)]
UD_2:	[(0.5,	0.1,	0.1),	(0,	0,	0),	(0.5,	0.1,	0.1)]
UD_3:	[(0.5,	0.1,	0.1),	(0.5,	0.1,	0.1),	(0,	0,	0)]

23

Maticový model pro XML DIS

- využití matice převodu cest
  - každá matice dokumentu je před uložením do indexu upravena převodní maticí
  - jednotlivé cesty (elementy) se tak dostávají do vztahů
  - dva dokumenty se stejným termem na různých cestách si budou (mohou) více či méně podobné

24

Podobnost XML

Maticový model pro XML DIS

- nevýhody:
  - časová a prostorová složitost je oproti vektorovému modelu horší
  - potřeba přídavných informací (převodní matice)
- výhody:
  - zpracovává strukturu XML dat
  - kolekce může být z různých zdrojů
  - dotaz a dokument ztotožněny
  - velmi flexibilní (převodní matice je značně univerzální)

Podobnost XML

Závěr

- zavedení podobnosti do prostředí XML
- kategorizace přístupů k podobnosti XML
- Maticový model jako řešení pro přístup DIS s indexací

Podobnost XML

Literatura

[1] Tim Bray, Jean Paoli, C. M. Sperberg-McQueen, Eve Maler (2000): Extensible Markup Language (XML) 1.0 (Second Edition). *W3C Recommendation* (http://www.w3.org/TR/2000/REC-xml-20001006)

[2] Jiří Kosek (2000): XML pro každého. *Grada Publishing s.r.o.*

[3] Jonathan Robie, Joe Lapp, David Schach (1998): XML Query Language (XQL). (http://www.w3.org/TandS/QL/QL98/pp/xql.html)

[4] Hiroshi Ishikawa, Kazumi Kubota, Yasuhiko Kanemasa (1998): XQL: A Query Language for XML Data. *Fujitsu Laboratories Ltd.* (http://www.w3.org/TandS/QL/QL98/pp/flab.doc)

[5] Alin Deutsch, Mary Fernandez, Daniela Florescu, Alon Levy, Dan Suciu (1998): XML-QL: A Query Language for XML. *Submission to the World Wide Web Consortium* (http://www.w3.org/TR/1998/NOTE-xml-ql-19980819.html)

[6] Jaroslav Pokorný (2001): XML a databáze. *KSI MFF UK* (http://kocour.ms.mff.cuni.cz/texty/xml-dj/)

[7] Scott Boag, Don Chamberlin, Mary F. Fernandez, Daniela Florescu, Jonathan Robie, Jérôme Siméon, Mugur Stefanescu (2002): XQuery 1.0: An XML Query Language. *W3C Working Draft* (http://www.w3.org/TR/xquery/)

[8] Jonathan Robie, Don Chamberlin, Daniela Florescu (2000): Quilt: an XML Query Language. (http://www.almaden.ibm.com/cs/people/chamberlin/quilt_euro.html)

[9] Jaroslav Pokorný, Václav Snášel, Dušan Húsek (1998): Dokumentografické informační systémy. *Skripta MFF UK, Karolinum – nakladatelství UK*

[10] Michal Kopecký (2000): Dokumentografické informační systémy. *KSI MFF UK* (http://www.ms.mff.cuni.cz/~kopecky/dis/)

[11] Anja Theobald, Gerhard Weikum (2000): Adding Relevance to XML. *Department of Computer Science University of the Saarland, Germany*

Podobnost XML

Literatura

[12] Taurai Chinenyanga, Nicholas Kushmerick (2001): An Expressive and Efficient Language For XML Information Retrieval. *J. American Society for Information Science & Technology*

[13] William W. Cohen (1998): Integration of heterogeneous databases without common domains using queries based on textual similarity. *Proc. SIGMOD, stránky 201-211*

[14] Norbert Fuhr, Kai Großjohann (2000): XIRQL – An Extension of XQL for Information Retrieval. *University of Dortmund, Germany*

[15] Norbert Fuhr, Kai Großjohann (2000): XIRQL: A Query Language for Information Retrieval. *University of Dortmund, Germany*

[16] Norbert Fuhr (2000): Probabilistic Datalog – Implementing Logical Information Retrieval for Advanced Applications.

[17] Albrecht Schmidt, Martin Kersten, Menzo Windhouwer (2001): Querying XML Documents Made Easy: The Nearest Concept Queries. *17th International Conference on Data Engineering* (http://dlib.computer.org/conferen/icde/1001/pdf1001032l.pdf)

[18] Yoshihiko Hayashi, Junji Tomita, Gen'ichiro Kikui (2000): Searching Text-rich XML Documents with Relevance Ranking. *ACM SIGIR 2000 Workshop on XML and Information Retrieval* (http://www.haifa.il.ibm.com/sigir00-xml/final-papers/Hayashi/hayashi.html)

[19] Torsten Schlieder (2001): Similarity search in XML data using cost-based query transformations. *Proceedings of the Fourth International Workshop on the Web and Databases (WebDB'01)* (http://www.inf.fu-berlin.de/~schlied/publications/webdb2001.ps)

[20] Torsten Schlieder, Holger Meuss (2000): Result ranking for structured queries against XML documents. *DELOS Workshop on Information Seeking, Searching and Querying in Digital Libraries* (http://www.inf.fu-berlin.de/~schlied/publications/delos2000.ps)

[21] Elisa Bertino, Giovanna Guerrini, Marco Mesiti (2001): Measuring the Structural Similarity among XML Documents and DTDs. *Dipartimento di Informatica e Scienze dell'Informazione*

[22] Jakub Vrána (2002): Specificita slov. *MFF UK*