Doktorandské dny ’08

Ústav informatiky
Akademie věd České republiky
v.v.i.

Malá Úpa

29. září – 1. října 2008

vydavatelství Matematicko-fyzikální fakulty
University Karlovy v Praze

Ústav Informatiky AV ČR v.v.i., Pod Vodárenskou věží 2, 182 07 Praha 8

ISBN *– not yet –*

Obsah

Nevyužité možnosti sémantického webu

doktorand:
ING. MARTIN ŘIMNÁČ
Ústav informatiky AV ČR, v. v. i.
Pod Vodárenskou věží 2
182 07 Praha 8
rimnacm@cs.cas.cz

školitel:
ING. JÚLIUS ŠTULLER, CSC.
Ústav informatiky AV ČR, v. v. i.
Pod Vodárenskou věží 2
182 07 Praha 8
stuller@cs.cas.cz

obor studia:
Databázové systémy

Práce byla podpořena projektem 1ET100300419 programu Informační společnost (Tématického programu II Národního programu výzkumu v ČR: Inteligentní modely, algoritmy, metody a nástroje pro vytváření sémantického webu), projektem 1M0554 Ministerstva školství, mládeže a tělovýchovy ČR "Pokročilé sanační technologie a procesy" a záměrem AV0Z10300504 "Computer Science for the Information Society: Models, Algorithms, Applications".

Abstrakt

Vize sémantického webu byla představena před skoro již 10 lety, avšak žádná z její aplikací prozatím nedokázala oslovit takové množství lidí, jaké dnes používá web v současné podobě. Příspěvek se věnuje možnostem sémantického webu a přínosům, které může přinést pro koncové uživatele. Nejprve podává přehled o současných technologiích i jejich použití a následně diskutuje možnosti plynoucí z použití odkazů v prostředí sémantického webu tak, jak je známe z webu současného, tedy rozšiřující, zpřesňující či udávající kontext prezentované informace.

1. Vyhledávání a vize sémantického webu

Současný web čelí mnoha problémům. Mezi ty nejstěžejnější patří problematika vyhledávání relevantních informací na webu. Ta je dnes většinou řešena pomocí tzv. *information retrieval* nástrojů [1], které pracují s inverzními indexy uchovávající (četnost) výskytu jednotlivých slov v (webových) dokumentech. Relevance dokumentu je pak stanovena pomocí kosinové míry reflektující podobnost mezi zadanými klíčovými slovy a slovy obsaženými v daném dokumentu.

Tato relevance však nic neříká o kvalitě poskytovaných dat. Proto bývá rozšířena o další nepřímou míru udávající odhadnutou kvalitu dat prezentovaných v dokumentu. Jednou z takových měr je PageRank [2], který je založen na předpokladu, že dokumenty prezentující kvalitní data jsou častěji odkazovány z jiných (kvalitních) dokumentů. Zavedením této míry se podařilo uspořádat (vůči klíčovým slovům relevantní) dokumenty i podle jejich kvality.

Díky značné redundanci dat na současném internetu však ani takové uspořádání nemusí vést ke zlepšení vypovídací schopnosti výsledku hledání. Na většinu dotazů dnešní vyhledávače vrátí desetitisíce odkazů; koncový uživatel mnohdy stěží analyzuje první dvacítku odkazů a ostatní, i z hlediska úspory času, zcela ignoruje. To vede k faktu, že získání *kompletní informace* pomocí současných vyhledávacích nástrojů je velmi obtížné, ne-li nemožné.

Nejen tento problém se snaží vyřešit vize sémantického webu [3, 4], která umožňuje definovat vedle samotných dat i metadata k jejich popisu. Jinými slovy nedefinuje pouze objekty jako takové, ale vymezuje popis objektu pomocí ostatních (stejným způsobem popsaných) objektů. Například popis třídy *dítě* je možné vztáhnout k popisu třídy *osoba*.

Dokumenty sémantického webu se skládají z RDF[1] trojic

$$(\text{object}, \text{predicate}, \text{subject}) \in (\mathbb{R} \cup \mathbb{B}) \times \mathbb{R} \times (\mathbb{R} \cup \mathbb{B} \cup \mathbb{L})$$

kde [5]

- $\mathbb{R}$ značí množinu tzv. *resources* identifikující popisované objekty;
- $\mathbb{B}$ značí množinu tzv. *blank nodes*, které sami o sobě nemají žádný význam, sloužících k identifikaci složitějších (vícearitních) struktur;
- $\mathbb{L}$ značí množinu literálů. Ta může být dále rozšířena o informaci o použitém přirozeném jazyku či terminologii.

[1] Resource Description Framework

Každý resource $\mathbb{R}$ je, dle definice, identifikován pomocí URI, např. ve tvaru

`http://example.com/ontologie#dite`

Vyhledávání v prostředí sémantického webu se primárně soustředí na vytváření indexu ukazující, který resource je popsán ve kterém dokumentu. Prohledávání takových indexů ale může být spojeno s odvozování, např. při hledání instancí třídy *osoba* zahrnout i instance třídy *dítě*.

Současný sémantický web se spíše orientuje na vytyčení pojmů pomocí ontologií; je známé nasazení vize sémantického webu v prostředí webových služeb, kdy jejich ontologický popis umožňuje kooperaci mezi dílčími webovými službami. Sémantický web je ale i odpovědí na otázku, jak najít na webu kompletní informaci samotnou, ne pouze odkazy na ní, tak, jak se dělají dnešní vyhledávače.

2. Formáty používané na webu

Za první formát webových dokumentů lze považovat HTML[2], který rozšířil formátovaná data o hypertextové odkazy. Tento formát je postaven na SGML, dnes se většinou používá jako základ striktnější XML[3]. Fragment takového HTML dokumentu může být ilustrován například pomocí:

```
<div class='item'>
   <img src='disk.samsung.spinpoint-F1-500GB.jpg'
        alt='Disk Samsung Spin Point F1 500GB'/>
   <div>Disk Samsung Spin Point F1 500GB</div>
   <ul>
     <li><b>Product No</b>:
         HD202IJ</li>
     <li><b>Interface:</b>
         SATA-II</li>
     <li><b>Space:</b>
         500GB</li>
     <li><b>RPM:</b>
         7200</li>
     <li><b>Warranty</b>:
         36 months</li>
     <li><b>Price</b>:
         1273 CZK</li>
     <li><b>Price incl. VAT</b>:
         1557 CZK</li>
     <li><b>Produced by:</b>
         <a href='http://www.samsung.com/global/business/hdd/productmodel.do?
            group=72&type=61&subtype=63&model_cd=240&ppmi=1155'>Samsung</a>
   </ul>
</div>
```

Takovýto fragment dokumentu může být zaindexován fulltextovými vyhledávači, jako relevantní je možné vybrat klíčová slova *SATA-II, HD202IJ, Samsung, Spin Point F1, 500GB*. Pakliže koncový uživatel zvolí některé z těchto klíčových slov, dříve či později by měl ve výsledku vyhledávání narazit na odkaz na dokument obsahující tento fragment. Pokud si uživatel bude chtít vybrat tento disk z nabídky všech prodejců, nezbude mu nic jiného, než projít ručně všechny tyto prodejce.

Naopak dokumenty sémantického webu jsou předurčeny pro další strojové zpracování. Vzhledem k tomu, že se prozatím nepodařilo v dostatečné míře prosadit publikování dat ve formátech sémantického webu, uchýlilo se konsorcium W3C, definující standarty v oblasti webu, v roce 2004 k návrhu rozšíření formátu HTML o další atributy RDFa[4]. Účelem rozšíření je zavést možnost sémantické anotace přímo do HTML dokumentů. Stejný fragment by pak vypadal následovně:

```
<div aboat='HD202IJ-in-my-shop' class='item'
     xmlns:disk-ont='http://example.com/disk-ont'
     xmlns:myshop='http://myshop.com'>
   <img src='disk.samsung.spinpoint-F1-500GB.jpg'
      alt='Disk Samsung Spin Point F1 500GB' rel='picture'/>
   <div property='disk-ont:Name'>Disk Samsung Spin Point F1 500GB</div>
   <ul>
      <li><b>Product No</b>:
         <span property='disk-ont:Product-ID'>HD202IJ</span></li>
      <li><b>Interface:</b>
         <span property='disk-ont:Interface'>SATA-II</span></li>
      <li><b>Capacity:</b>
         <span property='disk-ont:Capacity'500GB</span></li>
      <li><b>RPM:</b>
         <span property='disk-ont:Disk-rpm'>7200</span></li>
      <li><b>Warranty</b>:
         <span property='disk-ont:Warranty'> 36 months</span></li>
      <li><b>Price</b>:
         <span property='myshop:Price'>1273 CZK</span></li>
      <li><b>Price incl. VAT</b>:
         <span property='myshop:Price-inc-VAT'> 1557 CZK</li>
      <li><b>Produced by:</b>
      <a href='http://www.samsung.com/global/business/hdd/productmodel.do?
          group=72&type=61&subtype=63&model_cd=240&ppmi=1155'
        rel='disk-ont:Producer'>Samsung</a>
   </ul>
</div>
```

Z takto anotovaného dokumentu lze pomocí XSLT[5] transformace (obecně transformující jeden XML dokument na jiný dokument) získat přímo popis vlastností disku v RDF. Získaný fragment RDF dokumentu pak bude

```
<rdf:Description rdf:about='HD202IJ-in-my-shop'
                 xmlns:disk-ont='http://example.com/disk-ont'
                 xmlns:myshop='http://myshop.com'>
   <disk-ont:Picture rdf:resource='disk.samsung.spinpoint-F1-500GB.jpg'/>
   <disk-ont:Name>Disk Samsung Spin Point F1 500GB</disk-ont:Name>
   <disk-ont:Product-ID>HD202IJ</disk-ont:product-ID>
   <disk-ont:Interface>SATA-II</disk-ont:interface>
   <disk-ont:Capacity>500GB<disk-ont:capacity>
   <disk-ont:Disk-rpm>7200</disk-ont:disk-rpm>
   <disk-ont:Warranty>36 months</disk-ont:warranty>
   <myshop:Price>1273 CZK</myshop:price>
   <myshop:Price-inc-VAT> 1557 CZK</myshop:Price-inc-VAT>
   <disk-ont:Producer
       rdf:resource='http://www.samsung.com/global/business/hdd/productmodel.do?
         group=72&type=61&subtype=63&model_cd=240&ppmi=1155' />
</rdf:Description>
```

Ani toto rozšíření se prozatím nedočkalo velkého ohlasu mezi producenty dat, a tak koncoví uživatelé zůstávají bez možnosti efektivně (automaticky) zpracovávat data v současné době schovaná uprostřed formátování.

[2] HyperText Markup Language
[3] Extensible Markup Language
[4] Resource Description Framework Attributes
[5] Extensible Stylesheet Language Transformations

3. Distribuované prostředí

Web jako takový je distribuované prostředí, ve kterém kdokoliv může publikovat cokoliv. Web si koncoví uživatelé navykli používat; pakliže najdou zajímavý dokument, jisto jistě prozkoumají i odkazy vedoucí z tohoto dokumentu. I z tohoto důvodu se navigaci uživatele po webových stránkách věnuje značná pozornost a je jedním z hlavních kritérií hodnocení kvality (přístupnosti) webu.

Všimněme si, že každý resource v sémantickém webu je identifikován pomocí URI. Co by se však stalo, kdyby namísto (virtuálního) URI dokument odkazoval stejně jako je to u současného webu na jiný webový dokument obsahující detailnější informace o popisovaném objektu? Ve zvoleném případě by výrobce disků publikoval na adrese *http://example.com/sata-II-disks.rdf* dokument popisující například sérii disků. Příklad fragmentu takového dokumentu nechť je následující

```
<disk-ont:disk rdf:ID='HD202IJ'
               xmlns:disk-ont='http://example.com/disk-ont.rdf'>
   <disk-ont:picture rdf:resource='disk.samsung.spinpoint-F1-500GB.jpg'/>
   <disk-ont:Name>Disk Samsung Spin Point F1 500GB</disk-ont:Name>
   <disk-ont:product-ID>HD202IJ</disk-ont:product-ID>
   <disk-ont:interface>SATA-II</disk-ont:interface>
   <disk-ont:capacity>500GB<disk-ont:capacity>
   <disk-ont:disk-rpm>7200</disk-ont:disk-rpm>
   <disk-ont:warranty>36 months</disk-ont:warranty>
</disk-ont:disk>
```

přičemž jednotlivé vlastnosti mohou být definovány v externí ontologii *http://example.com/disk-ont.rdf*:

```
<rdfs:Property rdf:ID='disk-ont:Name'>
    <rdfs:label xml:lang='en'>Product Name</rdfs:label>
    <rdfs:label xml:lang='cs'>Označeni produktu</rdfs:label>
    ...
</rdfs:Property>
```

Jak je patrné, tato ontologie může obsahovat popisy vlastností v různých jazykových mutací. Ty mohou být následně využity pro generování HTML verze dokumentu, viz předchozí příklady.

Samotný obchod pak pouze deklaruje, že prodává daný disk a tuto informaci pouze rozšíří o specifika obchodu jako jsou cena, zkušenosti nakupujících a podobně:

```
<myshop:disk rdf:ID='HD202IJ-in-my-shop'
   <myshop:ProductDetail
      rdf:resource='http://example.com/sata-II-disks.rdf#HD202IJ'/>
   <myshop:Price>1273 CZK</myshop:price>
   <myshop:Price-inc-VAT> 1557 CZK</myshop:Price-inc-VAT>
</myshop:disk>
```

Tento model distribuce dat má několik výhod. První výhodou je nižší redundance dat, v původní architektuře každý prodejce musel uvádět veškerá data. Pro poskytovatele obsahu (ať výrobce či obchodníka) pak odpadá nutnost znovu zpracovávat data - pokud obchodník bude používat značení výrobce (ontologii poskytnutou výrobcem), má výrobce jistotu, že nedochází ke klamání koncového zákazníka se strany prodejce, naopak prodejce může deklarovat (např. elektronickým podpisem výrobce), že jím zprostředkovávaná data jsou ověřena. Obecně tímto postupem může být budována důvěra mezi subjekty publikující data na webu.

Další výhoda se uplatní u vyhledávání. Pokud se zákazník rozhodne pro daný disk, hledá již pouze prodejce, kteří tento disk nabízejí. Vzhledem k tomu, že disk je vždy identifikován pomocí URL na straně výrobce, je takové vyhledávání téměř triviální.

Toto zjednodušení vyhledávání je způsobeno tím, že není potřeba (heterogenní) data od různých prodejců integrovat. Integrace dat [6], neboli hledání korespondencí mezi daty více zdrojů a jejich následné spojování, sama o sobě představuje velmi těžkou a obecně automaticky [7] neřešitelnou úlohu. Čím složitější (a expresivnější) je popis objektů, tím je složitější i integrační proces. Díky tomu, že je objekt jednoznačně identifikován cílovou URL odkazu, není potřeba data integrovat v takovém rozsahu (integrují se pouze atributy specifické pro daného prodejce).

V neposlední řadě současné prohlížeče webových dokumentů umožňují zpracovat libovolný XML dokument a zobrazit jej buďto pomocí kaskádových stylů CSS a nebo pomocí XSLT transformace. Tato funkcionalita umožňuje stáhnout XML dokument obsahující pouze prostá RDF data, v jehož hlavičce je uvedeno, jakým způsobem mají být data zformátována. V případě XSLT transformace XML dokumentu do XHTML formátu je použita následující hlavička:

```
<?xml version='1.0' encoding='utf-8'?>
<?xml-stylesheet type='text/xsl' href='rdf2html.xslt'?>
```

kde *rdf2html.xslt* je šablona popisující transformaci z RDF trojic do HTML dokumentu. Tuto transformaci provede přímo prohlížeč a zobrazí její výstup. Koncový uživatel tak vůbec nepozná, že si neprohlíží klasickou webovou stránku, ale RDF dokument. Bohužel, tato technologie, byť je již dlouhodobě podporována všemi předními webovými prohlížeči, nebývá užívána, neboť současné vyhledávače nejsou schopni takto publikovaná data zpracovat. Tento způsob značně minimalizuje objem nutných datových přenosů, což je vhodné například u mobilních zařízení.

Další výhodou distribuované architektury a potažmo celého sémantického webu je fakt, že k takovýmto dokumentům mohou velmi jednoduše přistupovat aplikace označované jako *Web X.0*. Tyto aplikace postupně načítají/modifikují zobrazovanou stránku pomocí

[6] Asynchronous JavaScript and XML

AJAX[6] technologie, na straně prohlížeče spouštěných *javascriptových* programů umožňujících interakci mezi uživatelem a poskytovanými daty. Na jednotlivé RDF dokumenty lze pohlížet jako na tzv. *REST*[7] *webové služby* [8] volané AJAX programy. Zásadní nevýhodou této technologie je nemožnost indexace obsahu (neb aktuálně zobrazená data neodpovídají žádné URL, na kterou by se mohl uživatel později odkázat).

Tuto potencionální nevýhodu lze obejít publikováním jak RDF dokumentu formátovaného pomocí XML, tak statické HTML stránky, která vznikla identickou transformací na straně serveru. Tedy uživatel má možnost získat odkaz na (přibližně) stejný obsah reprezentovaný statickou HTML verzí, u které je uvedena korespondence s původním RDF dokumentem (například i pomocí RDFa rozšíření) a další navigace (hledání podobných produktů, více detailů, konkurenční prodejci) je zprostředkována již v rámci aktivní složky obsahu stránky.

Použití distribuované architektury tak, jak je popsána výše, v praxi naráží na pomalé odezvy webových serverů (čas potřebný k navázání spojení je podstatně větší nežli čas potřebný k samotnému přenosu dat). Tento problém lze vyřešit buďto efektivním cacheováním načtených dokumentů, které navíc může být podpořeno postupným načítáním obsahu pomocí AJAX aplikace.

4. Odhad struktury dat

Sémantický web umožňuje popisovat vlastnosti objektů pomocí vztahů. Tyto vztahy jsou definovány obecně pomocí resource - každý návrhář ontologie může použít své vlastní zavedení vlastností. Tento fakt obecně velmi ztěžuje jakékoliv složitější operace, včetně integrace ontologií. Z tohoto důvodu se mnohé nástroje poohlíží po podstatně jednodušších, byť méně popisných formalismech.

Vzhledem k nedostatku dat ve formátu sémantického webu je žádoucí najít způsob, jak využít data z webových stránek a extrahovat je do formátu sémantického webu (například anotací pomocí RDFa atributů). Pro anotaci je však potřeba znát strukturu dat; ta na webových stránkách nebývá uvedena a pak nezbývá nic jiného, než se ji pokusit odhadnout.

Strukturu dat lze popsat mnohými formalismy, ilustrujme ji na příkladu formalismu inspirovaném relačními databázemi [9]. Struktura dat je odhadnuta analýzou *extensionálních funkčních závislostí* platných na dané množině dat.

Funkční závislost mezi dvěma atributy je integritní omezení zajišťující jednoznačnou odvoditelnost hodnoty atributu na pravé straně při znalosti hodnoty atributu na levé straně. Příkladem funkční závislosti je například

$$\text{Stát} \rightarrow \text{Měna}$$

Samotné záznamy jsou popsány v odpovídající relaci. Všimněme si, že unární funkční závislost[8] je možné popsat pomocí odpovídající trojice

$$(\text{Stát}, \text{implies}, \text{Měna})$$

Abychom mohli stejným způsobem zavést i vztahy mezi hodnotami atributů, je vhodné pro každou funkční závislost definovat její *instance* [10]

$$A_1 \rightarrow A_2 \in \mathscr{F} \rightsquigarrow (A_1, A_1(t)) \rightarrow (A_2, A_2(t)) \in \mathscr{I}$$

kde

- $A_1, A_2 \in \mathscr{R}$ jsou atributy relace $\mathscr{R}$
- $A_\star(t)$ je zobrazení přiřazující záznamu t hodnotu atributu $A_\star$

Nazveme-li dvojici atribut-hodnota *elementem* (A, v), pak je možné tyto instance rovněž vyjádřit jako vztahy mezi elementy, které jsou popsány pomocí trojic

$$((A_1, v_1), \text{implies}, (A_2, v_2))$$

Takováto reprezentace dat ve formátech sémantického webu je vhodná v případě, že není zajištěna korektnost odhadnuté struktury dat. Pokud je odhadnutý model označen jako korektní, je možné data transformovat do formy [11]

$$(v_1, \text{name}(A_1 \rightarrow A_2), v_2)$$

kde $name$ je funkce pojmenovávající funkční závislosti. Pokud se přidržíme zvolené funkční závislosti, příkladem výsledku trasformace instance může být trojice

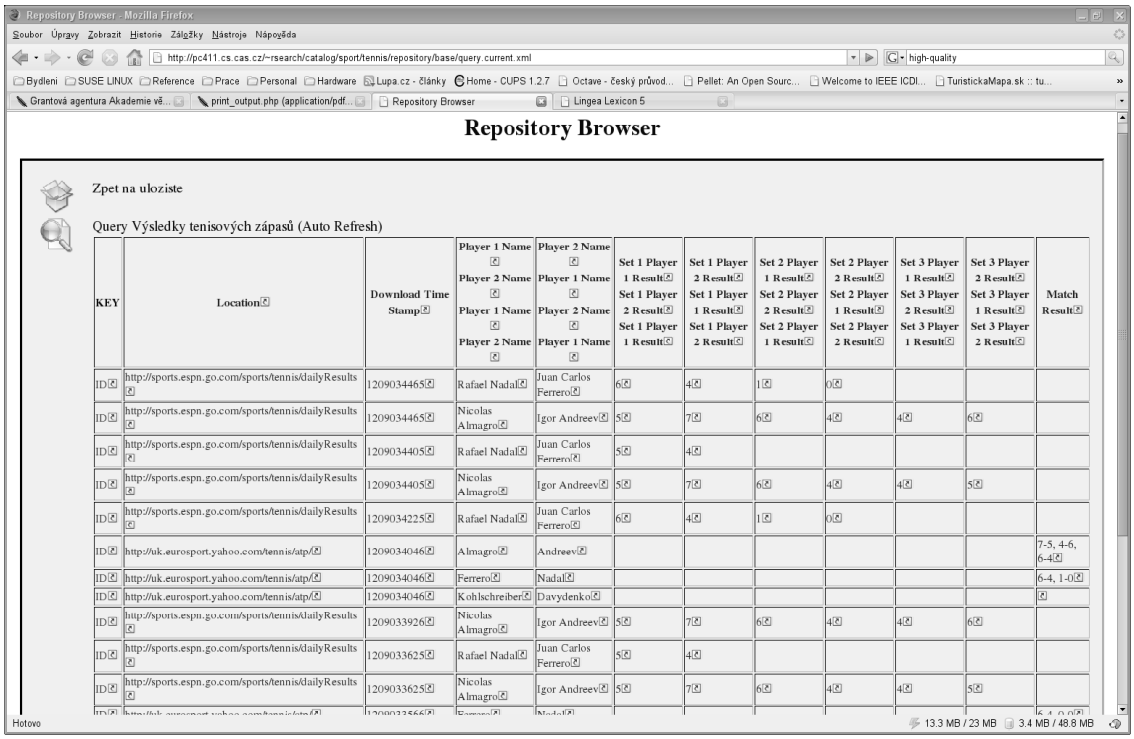
$$(\text{Česká Republika}, \text{has-a-Měna}, \text{Česká koruna})$$

Tyto trojice mnohou být uloženy do XML formátu. Například

```
<state rdf:ID='CeskaRepublika'>
   <has-a-Mena rdf:resource=
     'CeskaKoruna.rdf#CeskaKoruna'/>
</state>
```

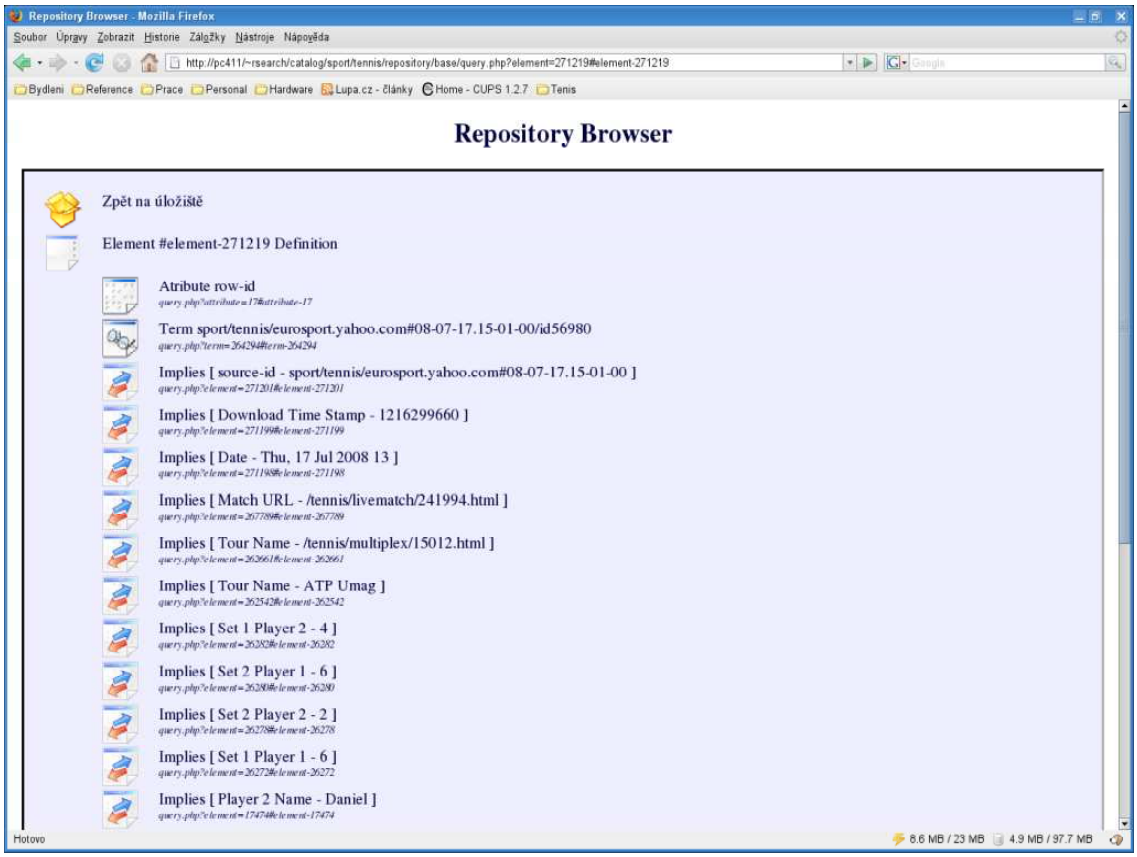
[7] Representational State Transfer

[8] Unární funkční závislost je funkční závislost mezi jednoduchými atributy (t.j. s aritou 1)



KEY	Location	Download Time Stamp	Player 1 Name / Player 2 Name / Player 1 Name / Player 2 Name	Player 2 Name / Player 1 Name / Player 2 Name / Player 1 Name	Set 1 Player 1 Result / Set 1 Player 2 Result / Set 1 Player 1 Result	Set 1 Player 2 Result / Set 1 Player 1 Result / Set 1 Player 2 Result	Set 2 Player 1 Result / Set 2 Player 2 Result / Set 2 Player 1 Result	Set 2 Player 2 Result / Set 2 Player 1 Result / Set 2 Player 2 Result	Set 3 Player 1 Result / Set 3 Player 2 Result / Set 3 Player 1 Result	Set 3 Player 2 Result / Set 3 Player 1 Result / Set 3 Player 2 Result	Match Result
ID	http://sports.espn.go.com/sports/tennis/dailyResults	1209034465	Rafael Nadal	Juan Carlos Ferrero	6	4	1	0			
ID	http://sports.espn.go.com/sports/tennis/dailyResults	1209034465	Nicolas Almagro	Igor Andreev	5	7	6	4	4	6	
ID	http://sports.espn.go.com/sports/tennis/dailyResults	1209034405	Rafael Nadal	Juan Carlos Ferrero	5	4					
ID	http://sports.espn.go.com/sports/tennis/dailyResults	1209034405	Nicolas Almagro	Igor Andreev	5	7	6	4	4	5	
ID	http://sports.espn.go.com/sports/tennis/dailyResults	1209034225	Rafael Nadal	Juan Carlos Ferrero	6	4	1	0			
ID	http://uk.eurosport.yahoo.com/tennis/atp/	1209034046	Almagro	Andreev							7-5, 4-6, 6-4
ID	http://uk.eurosport.yahoo.com/tennis/atp/	1209034046	Ferrero	Nadal							6-4, 1-0
ID	http://uk.eurosport.yahoo.com/tennis/atp/	1209034046	Kohlschreiber	Davydenko							
ID	http://sports.espn.go.com/sports/tennis/dailyResults	1209033926	Nicolas Almagro	Igor Andreev	5	7	6	4	4	6	
ID	http://sports.espn.go.com/sports/tennis/dailyResults	1209033625	Rafael Nadal	Juan Carlos Ferrero	5	4					
ID	http://sports.espn.go.com/sports/tennis/dailyResults	1209033625	Nicolas Almagro	Igor Andreev	5	7	6	4	4	5	

Obrázek 1: Ukázka stránky experimentálního portálu



Obrázek 2: Rekonstrukce záznamu

Jistě popis dat získaný odhadem jejich struktury z množiny vstupních dat nebude dosahovat expresivity známé z lidmi tvořených ontologií, avšak poskytuje za lehce splnitelných podmínek RDF dokumenty jistým, pro technická data postačujícím, způsobem. I takto jednoduchý popis dat může být použit pro učení extrakčních metod, které získávají anotovaná data z webových stránek [12, 13, 14]

V současné době je experimentálně provozován portál shromažďující informace o sportovních utkáních, kdy struktura dat byla odhadnuta z dat několika heterogenních zdrojů a data uložena na základě této struktury. Ilustrace portálu je na obrázcích 1 a 2.

5. Závěr

Příspěvek se snaží shrnout aktuální trendy, problémy a technologie jak na současném webu, tak v prostředí webu sémantického. Zvláště se pak věnuje problematice vyhledávání dat, diskutuje související problémy a navrhuje jejich řešení.

V sekci 2 ukazuje na příkladu fragmentu HTML dokumentu, jak může být zaindexován pro fulltextové vyhledávání. Ukazuje použití rozšíření RDFa, které umožňuje anotovat části HTML dokumentu. Pokud jsou hodnoty anotovány, je možné automaticky převést takový HTML dokument do RDF dokumentu a ten dále zpracovat další nástroji.

Sekce 3 pak inovativně diskutuje výhody distribuce dat dokumentů sémantického webu, kdy resource není reprezentován pouze URI, ale URL obsahující detailnější informace o odkazovaném objektu. Zásadní výhodou tohoto přístupu je, že odpadá nutnost jinak velmi obtížné, automaticky téměř neřešitelné, integrace dat jednotlivých zdrojů. Celý problém je ilustrován na příkladě.

Jelikož v současné době nejsou k dispozici taková data požadovaného rozsahu a zaměření, sekce 4 navrhuje problém řešit pomocí metod odhadu struktury dat a tyto metody využít pro základní definici popisu dat prostřednictvím formátů sémantického webu.

Pokud by se podařilo myšlenky prezentované v článku naplnit, celá vize by našla uplatnění pro širokou veřejnost dnes používající internet.

Literatura

[1] P. Raghavan, “Information retrieval algorithms: a survey,” in *SODA ’97: Proceedings of the eighth annual ACM-SIAM symposium on Discrete algorithms*, (Philadelphia, PA, USA), pp. 11–18, Society for Industrial and Applied Mathematics, 1997.

[2] A. N. Langville and C. D. Meyer, *Google’s PageRank and Beyond: The Science of Search Engine Rankings*. Princeton University Press, July 3 2006.

[3] G. Antoniou and F. van Harmelen, *A Semantic Web Primer (Cooperative Information Systems)*. The MIT Press, April 2004.

[4] T. Lee, “Relational databases on the semantic web,” *¡http://www.w3.org/DesignIssues/RDB-RDF.html¿ [on-line]*, 1998.

[5] L. Baolin and H. Bo, “Network and parallel computing, ifip international conference, npc 2007, dalian, china, september 18-21, 2007, proceedings,” in *NPC* (K. Li, C. R. Jesshope, H. Jin, and J.-L. Gaudiot, eds.), vol. 4672 of *LNCS*, pp. 364–374, Springer, 2007.

[6] M. Lenzerini, “Data integration: a theoretical perspective,” in *PODS ’02: Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of Database Systems*, (New York, NY, USA), pp. 233–246, ACM Press, 2002.

[7] E. Rahm and P. A. Bernstein, “A survey of approaches to automatic schema matching,” *VLDB Journal: Very Large Data Bases*, vol. 10, no. 4, pp. 334–350, 2001.

[8] R. Battle and E. Benson, “Bridging the semantic web and web 2.0 with representational state transfer (rest),” *Web Semant.*, vol. 6, no. 1, pp. 61–69, 2008.

[9] C. J. Date, *An Introduction to Database Systems*. Addison Wesley Longman, October 1999.

[10] M. Řimnáč, “Data structure estimation for rdf oriented repository building,” in *Proceedings of the CISIS 2007*, (Los Alamitos, CA, USA), pp. 147–154, IEEE Computer Society, 2007.

[11] M. Řimnáč, “Transforming current web sources for semantic web usage,” *Proc. of SOFSEM 2006*, vol. 2, pp. 155–165, 2006.

[12] Z. Li and W. K. Ng, “Wdee: Web data extraction by example,” in *DASFAA* (L. Zhou, B. C. Ooi, and X. Meng, eds.), vol. 3453 of *LNCS*, pp. 347–358, Springer, 2005.

[13] W. Holzinger, B. Krüpl, and M. Herzog, “Using ontologies for extracting product features from web pages,” in *International Semantic Web Conference* (I. F. Cruz, S. Decker, D. Allemang, C. Preist, D. Schwabe, P. Mika, M. Uschold, and L. Aroyo, eds.), vol. 4273 of *LNCS*, pp. 286–299, Springer, 2006.

[14] M. Nekvasil, “Využití ontologií při indukci wrapperů,” *Proc. of Znalosti 2007*, pp. 336–339, 2007.

Ústav Informatiky AV ČR v.v.i.
DOKTORANDSKÉ DNY '08

Vydal
MATFYZPRESS
vydavatelství
Matematicko-fyzikální fakulty
University Karlovy
Sokolovská 83, 186 75 Praha 8
jako svou *– not yet –* . publikaci

Obálku navrhl František Hakl

Z předloh připravených v systému LaTeX
vytisklo Reprostředisko MFF UK
Sokolovská 83, 186 75 Praha 8

Vydání první

Praha 2008

ISBN *– not yet –*