Ústav informačních studií a knihovnictví
Filozofická fakulta Univerzity Karlovy

Ontologie jako součást sémantického webu

Seminární práce na předmět
Matematické a informatické modely v ontologii

ZS 2003/2004

Linda Skolková
2. ročník bak. studia
kontakt: skolkova@chello.cz

OBSAH:

Úvod

O sémantickém webu se v poslední době velmi mnoho hovoří, podobně je tomu s ontologiemi. V obou případech dochází k jistému posunu významu, který nemusí být na první pohled zcela zřetelný. S webem jakožto jednou ze služeb internetu máme povětšinou relativně dost vlastních zkušeností, sémantický web se ovšem teprve začíná klubat, a proto naše představa o tom, k čemu nám nejspíše v budoucnu bude sloužit, může být značně mlhavá. Ontologie se zase používají nikoli s dalším adjektivem přizpůsobujícím se novému významu, jako je tomu u sémantického webu, ale jejich definice se přizpůsobuje novému využití. Ontologie se také již používají v plurálu (filozofové, kteří sami sebe nejraději nazývají *filosofy*, se drží singuláru).

Proč vlastně sémantický web potřebujeme? Ing. Sklenák ve svých pracích[1] často uvádí příklad povídky *Babylónská knihovna*, kterou J.L. Borges napsal už v roce 1941, ovšem která velmi přesně vystihuje současnou situaci v prostředí www. V povídce se jedná o to, že knihovna obsahuje prakticky všechno, co kdy bylo napsáno nebo co teprve napsáno bude, ovšem nenabízí ke konkrétním informacím (resp. dokumentům) žádnou přístupovou cestu, takže nakonec počáteční pocit jakéhosi štěstí vystřídá beznaděj. A k podobnému stavu se začíná blížit současný stav webu...

Dosavadní vývoj webu lze rozdělit na dvě generace. První představuje obsah stránek vytvářený ručně za přímého použití HTML. Nabízí jednoduchý přístup s jednotným rozhraním, klade ovšem velké nároky na autory a na správu a působí obtíže při častých změnách obsahu. Za druhou generaci je možno pokládat obsah generovaný na vyžádání (*on-the-fly*). V tomto případě se využívá šablon, které jsou naplňovány z obsahu databáze. Třetí generaci bude představovat právě sémantický web, který bude podporovat nejenom vyhledávání, ale zasáhne i do dalších aplikačních oblastí.

Možnosti sémantického webu

Tvůrce webu Tim Berners-Lee podtrhuje, že sémantický web není separátním webem, nýbrž je rozšířením webu současného.**Sémantický web přiřazuje datům přesný význam umožňující spolupráci lidí a softwaru.** Myslím, že není úplně od věci připomenout si dnes již klasickou definici informačního systému od B.C. Vickeryho. Podle Vickeryho totiž informační systém představuje organizaci lidí, materiálů a strojů, která má usnadnit transfer informace od jedné osoby k osobě druhé.[2] Budeme-li se na sémantický web dívat jako na jistý druh informačního systému, pak můžeme vidět cestu od zdůrazňování *organizační* a *transferové* složky až k podtrhování *spolupráce.*

Dnes se web dynamicky vyvíjí zejména jako zprostředkovatel dokumentů pro lidského uživatele. Sémantický web se snaží naopak vyzdvihnout automatické zpracování dat a informací.

Podobně jako internet bude sémantický web co možná nejvíce decentralizovaný, což na druhou stranu bude vyžadovat určité kompromisy. Ostatně exponenciální nárůst počtu webových stránek má také jisté nedostatky. Za všechny můžeme jmenovat například zprávu o chybě 404 (nenalezení stránky).

Aby mohl sémantický web vůbec fungovat, je třeba, aby počítače měly přístup ke strukturovaným souborům dat a odvozovací pravidla k provádění automatické dedukce. Touto problematikou, která se často označuje jako *reprezentace znalostí*, se již dlouho před vznikem myšlenky sémantického

[1] Např. [6] a [7].
[2] *An information system is an organization of people, materials and machines that serves to faciliate the transfer of information from one person to another.* Viz Vickery, B.C. *Information Systems.* London: Butterworths 1973, s. 1.

webu zabývali odborníci v oblasti umělé inteligence, přesto můžeme říci, že dnes se tato technologie nachází zhruba na podobném stupni vývoje jako hypertext před vznikem webu.

Nyní se blíže podívejme na zachycování struktury dat. Například v jazyce HTML se tak děje prostřednictvím jednotlivých tagů, i když ty ve skutečnosti slouží zejména jako pokyn pro prohlížeče, který jim říká, jak mají daný text správně zformátovat. XML[3] už umožňuje definici nových tagů podle konkrétní aplikace. V takovém případě se použitý slovník nejprve definuje prostřednictvím DTD[4] nebo složitěji a přesněji pomocí XML schématu. Jestliže se XML má používat jako výměnný formát, pak je nutná předchozí dohoda obou stran na daném slovníku a významech. Přesto i zde narážíme na problém, který si můžeme ilustrovat na následujícím příkladě: V konečném důsledku totiž počítač stejně od sebe sémanticky neodliší kupříkladu nadpis třetí úrovně předznamenaný tagem <h3> v HTML a informaci o ceně, která následuje po tagu <cena> ve vytvořené aplikaci XML.

Pro vývoj sémantického webu kromě zmiňovaného jazyka XML existuje ještě další technologie známá pod zkratkou RDF.[5] Nejedná se o jazyk, ale o model pro reprezentaci dat o zdrojích na webu. Zatímco XML umožňuje uživatelům vytvářet vlastní struktury dokumentů, ale neříká nic o jejich významu, RDF umožňuje zachycení významu, a to v podobě trojic objekt-atribut-hodnota (*subject – verb – object)*. Konkrétní *věci* (lidé, webové stránky nebo cokoliv jiného) mají *vlastnosti* (atributy, predikáty; například *býti sestrou*), které pak nabývají jistých *hodnot* (jiná osoba, jiná webová stránka). Objekt a hodnota jsou identifikovány pomocí URI.[6] RDF trojice vytvářejí pavučiny informací o souvisejících věcech. URI zajišťují, že koncepty nejsou pouhými slovy v dokumentu, ale jsou provázány na unikátní definici, kterou si každý může najít na webu.

Ovšem za těchto předpokladů je stále možné (můžeme říci, že dokonce i pravděpodobné), že například dvě rozdílné databáze budou používat různé identifikátory příslušející stejnému konceptu. Proto je nutná třetí základní složka sémantického webu, a to jsou *ontologie*.

Ontologie jako páteř sémantického webu

O ontologiích se hovoří už více než desetiletí, svědčí o tom ostatně i datování nejpoužívanější Gruberovy definice rokem 1993. S rozšířeními provedenými později dalšími autory si definici můžeme uvést v tomto znění: **Ontologie je formální, explicitní specifikace sdílené konceptualizace.** Zde asi nezbude než využít principu kompozicionality významu a shluk významu cizích slov v tomto smyslu si vysvětlit po částech. *Konceptualizací* budeme mít na mysli abstraktní model výseku reálného světa identifikující relevantní koncepty daného výseku. Adjektivum *explicitní* zdůrazňuje jednoznačnost definice typu konceptu a podmínek jeho užití, *formální* odráží možnost strojového zpracování, *sdílený* pak poukazuje na zachycení konsensuálních znalostí (širší než znalosti jedince).

Ontologie mají obrovskou výhodu v tom, že **jsou srozumitelné člověku a zároveň strojově zpracovatelné.**

Ontologie se nejčastěji rozdělují podle zdroje konceptualizace:

- *generické ontologie (*též *ontologie vyššího řádu)* – zachycování obecných zákonitostí
- *doménové ontologie* – určeny pro specifickou věcnou oblast (nejčastější)

[3] XML = Extensible Markup Language.
[4] DTD = Document Type Definition.
[5] RDF = Resource Description Framework.
[6] Nejběžnějším typem URI (URI = Universal Resource Identifier) je URL (Uniform Resource Locator).

* *úlohové ontologie* (též *reprezentační ontologie* či *metaontologie)* – zaměřeny na procesy odvozování
* *aplikační ontologie* – adaptovány na konkrétní aplikaci (nejspecifičtější, zpravidla zahrnují doménovou i úlohovou část)

Podíváme-li se pro větší názornost na ontologie z mírně knihovnického (a tedy z našeho pohledu tradičního) hlediska, pak je můžeme srovnat s tezaury. Ontologie do jisté míry vycházejí z funkcí a účelu tezaurů, v podstatných rysech jdou však mnohem dále. Přehledné porovnání ontologií s tezaury uvádí následující tabulka:

TEZAURUS	**ONTOLOGIE**
důraz na termíny a vztahy mezi nimi v přirozeném jazyce	důraz na koncepty (pojmy)
vztahy: BT (nadřazený deskriptor) NT (podřazený deskriptor) RT (asociovaný termín) UF (nedeskriptor, nepreferovaný termín)	rozšířené a přesnější možnosti vyjádření vztahů mezi jednotlivými koncepty, např.:[7] hypernyma a hyponyma (vztah mezi třídami a specifickými instancemi)[8] meronyma a holonyma (vztah mezi částí a celkem)[9] kdokoliv může přidat nový termín nebo druh vztahu
terminologické pokrytí určité předmětné oblasti vymezení vztahů mezi jednotlivými termíny	
uspořádání terminologie za použití stromové struktury (u ontologií však může struktura tvořit i síť)	

Právě v tom, že ontologie se podobají tezaurům či například křížovým odkazům, se přímo knihovnické komunitě nabízí otázka, zda se raději nestát správci ontologií než správci sbírek...

Ontologie určené pro web se typicky skládají z taxonomie a ze souboru odvozovacích pravidel. Taxonomie definuje třídy objektů a jejich vzájemné vztahy. Třídy, podtřídy a vztahy jsou velmi mocným nástrojem, protože díky nim můžeme vyjádřit velké množství vztahů mezi entitami. Vychází se i z toho, že podtřídy dědí vlastnosti tříd.

Tvorba ontologie sestává z následujících kroků:

1. Stanovení rozsahu a cíle ontologie
2. Identifikace entit specifických v dané doméně
3. Uspořádání entit do hierarchie
4. Definice entit
5. Vlastnosti entit
6. Identifikace vztahů

[7] Následující příklady vztahů jsou použity např. ve WordNetu (terminologická ontologie dostupná z http://www.cogsci.princeton.edu/~wn/). Zde se setkáváme mj. i se zvláštním termínem pro množiny synonym – *synsety*. Ještě upřesněme, že WordNet není zcela čistou ontologií, nicméně velmi dobře slouží jako prostředek zkoumání přirozeného jazyka.
[8] Y je hypernymem X , jestliže X je druhem Y. X je hyponymem Y, jestližeX je druhem Y. Př.: X je slon, Y je savec.
[9] Y je holonymem X, jestliže X je částí Y. X je meronymem Y, jestliže X je částí Y. Př.: X je obývací pokoj, Y je obytný dům.

7. Upřesnění a rozšíření

Příklad ontologie zachycující typicky anglickou činnost přípravy čaje je popsán (i s příslušným grafickým doprovodem) v [3].

Ontologie mohou v mnohém vylepšit fungování webu. V nejjednoduším případě se může jednat např. o přesnost vyhledávání[10] – vyhledávač se může zaměřit je na ty stránky odpovídající danému konceptu (a nikoli dvojznačným nebo dokonce víceznačným klíčovým slovům).

Flexibilita sémantického webu mimo jiné umožní i zjednodušení využívání služeb, které pouze částečně splňují uživatelem zadané požadavky. Praktické aplikace flexibility sémantického webu se přímo nabízejí – například elektronický obchod. Zákazník a producent (resp. spíše prodejce) si mohou lépe porozumět, vymění-li si ontologie, které oběma poskytnou slovník nutný k diskusi.

Vyhledávání informací

Tradiční systémy reprezentace znalostí se vyznačují centralizovaností a požadují po všech zúčastněných sdílení přesných definic společných konceptů. Dále je pro ně typické značné omezování škály kladených otázek pouze na ty, které je počítač schopen spolehlivě zodpovědět. Tradiční systémy také většinou mají vlastní soubor odvozovacích pravidel použitelných pouze pro data v konkrétním systému.

Sémantický web by neměl být tímto směrem omezován, počítá se i s nezodpověditelnými otázkami, které jsou daní za víceúčelovost (*versatility*). Tato filozofie je ostatně podobná filozofii konvenčního webu, který se nikdy nestane dobře uspořádanou knihovnou. Bez centrální databáze a stromové struktury je téměř zaručeno, že ne vše je zpětně dohledatelné.

Současné služby pro vyhledávání informací na webu lze rozdělit na dva druhy:

* vyhledávací stroje s roboty (charakteristickým rysem je vyšší úplnost a nižší přesnost)
* vyhledávací služby s asistencí člověka (typická je naopak vyšší přesnost a nižší úplnost)

Sémantický web přidá smysluplnému obsahu webových stránek strukturu, čímž dojde k vytvoření prostředí, v němž se budou moci tzv. softwaroví agenti pohybovat ze stránky na stránku a přitom vykonávat sofistikované úkoly zadané uživatelem.

Díky ontologiím bude zjednodušen vývoj programů na řešení komplexních otázek, na které nelze odpovědět pouze díky informacím na jedné stránce, ale kde je třeba navštívit stránek několik. V tomto ohledu je třeba dodat, že vyvstává otázka spolehlivosti, důvěryhodnosti použitých zdrojů. Ke slovu přijdou elektronické podpisy prokazující, že danou informaci poskytl důvěryhodný zdroj.

Shrňme, že na sémantickém webu by proto mělo být možno efektivněji vyhledávat nejenom *informace* (identifikace relevantních dokumentů a jejich řazení), ale také *jednoduché i komplexní odpovědi na otázky* (např. *Kdo je britským premiérem?* a *Jaká je současná situace v Británii*?). Je zřejmé, že u komplexních odpovědích jsou navíc zapotřebí techniky extrakce a sumarizace informací.

Dodejme ještě, že se zde setkáváme s pojmem *hodnotového řetězce* – v rámci něho dochází sestavování dílčích informací (jako součástí odpovědi na otázku), které si mezi sebou vyměňují jednotliví agenti, přičemž každý z nich přidává hodnotu. Konečným výsledkem je pak odpověď uživateli. Tímto způsobem se také sníží množství dat putujících po sítích.

[10] Podrobněji k tématu viz následující kapitola.

Zamyšlení na závěr

Dalším krokem následujícím po uplatnění sémantického webu ve virtuálním prostředí bude jeho extenze do reálného, fyzického světa. URI budou pak moci označovat například fyzické entity. Díky RDF bude možné popsat nejrůznější zařízení od mobilních telefonů až třeba po televizní přijímače.

RDF schéma jako takové ovšem bylo navženo jako minimalistický jazyk. Další vrstvou je proto ještě jazyk DAML+OIL (DARPA Agent Markup Language + Ontology Inference Layer). Pod záštitou konsorcia W3C je dále vyvíjen další jazyk, který nese název OWL (Ontology Web Language).

Trochu současnější otázkou pak je, zda by se sémantický web měl prosazovat cestou evoluční nebo spíše revoluční. První možnost asi bude přijatelnější – plynulý přechod od současného k sémantickému webu má být realizován pomocí systematické tvorby a vkládání metadat.V každém případě sémantický web ve svém důsledku může velmi napomoci celkovému rozvoji lidského poznání.

Použitá literatura

[1] BERNERS-LEE, Tim; HENDLER, James; LASSILA, Ora. The Semantic Web : A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American.* May 17, 2001. [cit. 2003-11-06]. Dostupný z www: <http://www.scientificamerican.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21&catID=2>

[2] BRATKOVÁ, Eva. Metadata jako nový nástroj pro komunikaci webovských informačních zdrojů. *Národní knihovna. Knihovnická revue*. 1999, č. 4, s. 178-195. Dostupný též z www: <http://full.nkp.cz/nkkr/Nkkr9904/9904178.html>. ISSN 0862-7487.

[3] CROFTS, Nicholas; LE BŒUF, Patrick; ODILE, Artur. Ontologies. Semantic Web and Libraries (26th Library Systems Seminar) 2002. [2003-10-02]. Dostupný z www: <http://www.ifnet.it/elag2002/papers/pap9.html>.

[4] HOPPENBROUWERS, Jeroen. Semantic Modeling. ELAG 2003. [cit. 2003-10-01]. Dostupné z www: <http://www.elag2003.ch/pres/pres_hoppenbrouwers.pdf>.

[5] SVOBODA, Martin. Zpráva z cesty na seminář ELAG 2003. *Ikaros* [online]. 2003, č. 08 [cit. 2003-08-01]. Dostupný z www: <http://www.ikaros.cz/Clanek.asp?ID=200308001>. ISSN 1212-5075.

[6] SKLENÁK, Vilém. Sémantický web. *Inforum 2003*. [cit. 2003-11-01]. Dostupný z www: <http://www.inforum.cz/inforum2003/prispevky/Sklenak_Vilem.pdf>.

[7] SKLENÁK, Vilém. Vyhledávací nástroje v prostředí Internetu – co bude dál? *Automatizace knihovnických procesů 2003*. [cit. 2003-10-20]. Dostupný z www: <http://platan.vc.cvut.cz/akp2003/sbornik/03_sklenak.pdf>.

[8] SVÁTEK, Vojtěch. Ontologie a WWW. Datakon 2002. [cit. 2003-10-22]. Dostupný z www: <http://nb.vse.cz/~svatek/onto-www.pdf>.

[9] http://www.semanticweb.org

[10] http://www.w3.org/, zejména http://www.w3.org/2001/sw/ (část věnovaná přímo tématice sémantického webu)