Portál pre odborné publikovanie ISSN 1338-0087

Přístup k sémantickému webu založený na ontologiích

Šmiraus Michal · Informačné technológie

20.05.2013

Webové stránky v současné době zahrnují velké množství informací a dokumentů, jenž jsou sice z velké části srozumitelné lidem, avšak již méně srozumitelné pro automatizované vyhledávací stroje, které v současné době nedokáží přesněji identifikovat, co obsah dané stránky vyjadřuje. Spolu s vzrůstajícím množstvím dostupných informací na webu tak vzniká potřeba efektivně označit, rozeznat a zpracovat relevantní informace nikoli jen na základě prostého full-textového vyhledávání pomocí klíčových slov, ale také na základě znalostních bází s využitím ontologie (explicitní popis určitého pojmu), jejímž předmětem je na jedné straně vývoj obecných jazyků, metodik a softwarových nástrojů a na druhé straně také konstrukce samotných ontologií popisujících různé věcné oblasti, i aplikací, které je budou využívat.

1. Úvod

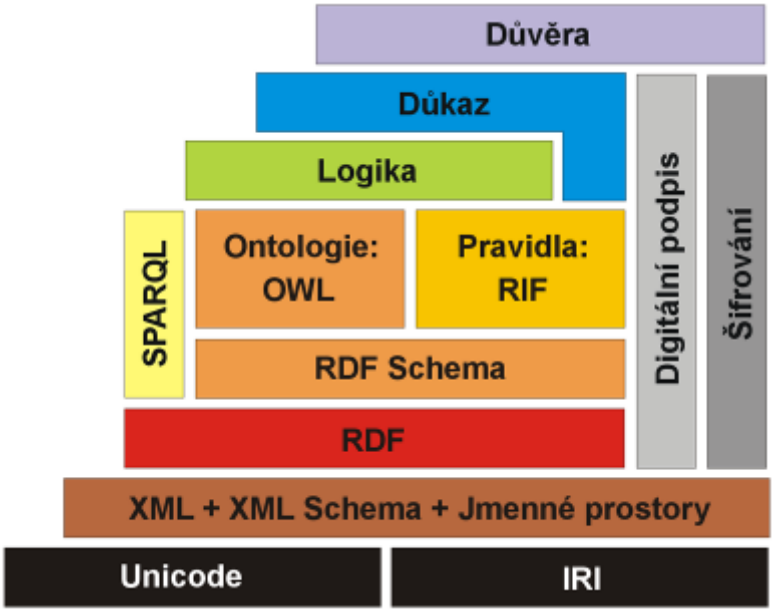
Myšlenky sémantické integrace webu ve své vizi poprvé popsal Tim Berners-Lee (2001), kde počítač vidí jako osobního asistenta, který svého majitele důvěrně zná a dokáže mu například doporučit a naplánovat celou dovolenou (včetně rezervace hotelu) v souladu s jeho časovými možnostmi a preferencemi. Takovéto možnosti byly v oblasti umělé inteligence slibovány odnepaměti, ale nikdy nedošlo k jejich naplnění. Sémantický web nespoléhá na pokročilou umělou inteligenci, která dokáže význam slov a tvrzení zpracovat sama, ale doporučuje obohacovat klasický web o značky a výroky psané ve speciálních jazycích (například RDF a OWL).

Plynulý přechod od současného WWW k sémantickému webu má být realizován prostřednictvím systematické tvorby a vkládání metadat. Pro jednoznačné vyjádření sémantiky používaných termínů je nutno použít jazyky vycházející právě z výzkumu v oblasti ontologií. Jako hlavní oblasti využití ontologií jsou v současnosti chápány: znalostní management, elektronické obchodování, zpracování přirozeného jazyka, inteligentní integrace informací z distribuovaných zdrojů, vyhledávání informací, sémantické webové portály, a inteligentní výukové systémy.

2. Architektura sémantického webu

Celý koncept sémantického webu je postaven na veřejném identifikátoru URI, který pomoci řetězce znaků dokáže identifikovat zdroj informace. Na úplném dně pomyslné pyramidy najdeme XML. Značkovací jazyk, kterým můžeme vytvořit strukturovaný dokument s vlastními značkami (tagy). Na něj navazuje vrstva RDF, která nám

dovoluje definovat vztahy mezi objekty (zdroji). Následující vrstva, která umožňuje zachycování složitějších ontologických struktur, je realizována prostřednictvím jazyka OWL. Logická vrstva nám dovoluje popsat vztahy mezi jednotlivými objekty komplexněji a díky aplikování použitelné logiky provádí odvozování implicitních informací. Poslední Trust vrstva umožňuje zajistit spolehlivost informací.



Obr.1 Architektura vrstev Sémantického webu.

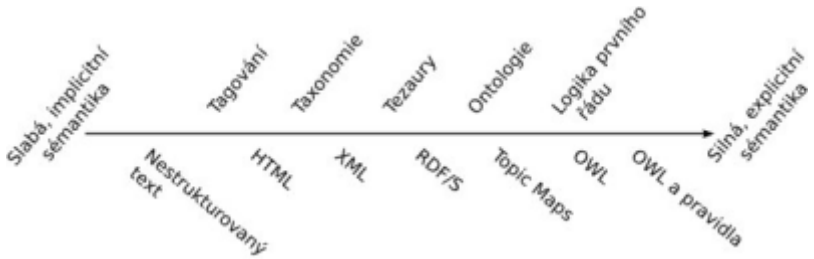
Sémantické informace vpletené do běžného webu umožňují počítači manipulovat s daty inteligentněji. Například slovo „škola“ vyskytující se na běžném webu je pro počítač pouze řetězec pěti znaků. Na sémantickém webu je však možné označit slovo „škola“ identifikátorem (URI) pojmu škola v mnohem širším popisu pojmů a jejich vzájemných vztahů, kterým se běžně říká ontologie.

Počítač pak v ontologii například zjistí, že škola je vzdělávací zařízení, které přijímá studenty a že student je člověk, který má studentský průkaz. Vyskytuje-li se pak v textu třeba informace „Pavel studuje na této škole“, je pro počítač snazší odvodit, že Pavel je student, který má studentský průkaz. Nejdříve ale musí existovat ontologie, která takové vztahy popisuje, a text musí být anotovaný (doplněný o značky). Tvorba ontologií většinou probíhá ručně. Automatické odvozování ontologií je stále předmětem aktivního výzkumu. Podobně je tomu i se značkováním – často probíhá ručně, ale existuje i spousta automatických a poloautomatických nástrojů.

3. Aplikace ontologického spektra

Samotný pojem ontologie je většinou definovaný jako explicitní popis konceptualizace. To jest, zaznamenává pojmy a vztahy mezi nimi v nějakém jazyce (Smrž, 2004). Tyto jazyky mívají velkou vyjadřovací sílu a často vyžadují expertní znalosti. Existují ale i méně silné a daleko rozšířenější prostředky pro popis konceptualizace. Používá je někdy téměř každý uživatel dnešního webu – jsou to tagování, taxonomie a tezaury. Tagování neboli přiřazování štítků (kde štítek je obyčejný řetězec znaků) má nejmenší vyjadřovací sílu – význam zprostředkovaný tagováním je jen malý. Tagování na webu pomáhá uživateli třídit informace především za účelem jejich pozdějšího vyhledání.

Taxonomie je hierarchie (strom) pojmů. Většinou popisuje pouze jeden druh vztahu (například vztah „je podtřídou"), ale může v ní implicitně existovat více druhů vztahů, jako je tomu například u adresářů v souborových systémech. Podadresář P může být v nadřazeném adresáři N, protože P (jezevčík) je druhem N (pes). Jiný podadresář Q může být v nadřazeném adresáři N, protože Q (hlava) je částí N (pes). Tezaurus také popisuje hierarchii pojmů, ale jasně říká, které vztahy mezi pojmy popisuje. Většinou jsou to pojmy „je obecnější než," „je méně obecný než", „je příbuzný".



Obr.2 Ontologické spektrum Sémantického webu.

Ontologie je nejbohatším způsobem popisu konceptualizace. V ontologických jazycích, jako je např. OWL, je možné zavádět jak pojmy, tak i nové vztahy, které jsou následně používány pro další popis pojmů. Tagování, taxonomie, tezaurus a ontologie tvoří takzvané sémantické spektrum (někdy se místo tagování uvádí obyčejný a řízený slovník).

Ačkoliv je sémantický web vyvíjen už více než deset let a naplnění jeho původní vize je stále v nedohlednu, existuje zde již nyní mnoho myšlenek a konkrétních technologií, které se podařilo realizovat v praxi. Jedním z příkladů je tzv. fasetové vyhledávání, které ve formě "našeptávání" již delší dobu používá Google. Mezi jiné patří například zobrazování tzv. "mikroformáty", což jsou kousky informací doplněné o sémantické anotace, které vyhledávači sdělí, že daná informace je popisem produktu, vizitky, události nebo struktury určitého webu či vztahy mezi lidmi v případě sociálních sítí.

Podporu mikroformátů a RDF do jádra svého vyhledávacího algoritmu přidal Google již v květnu 2009 a od té doby postupně tuto podporu rozšiřuje na celý vyhledávací index, což motivuje autory stránek, aby si na své weby začali některé informace sémanticky označovat (týká se to hlavně recenzí, osob, firem a produktů), protože Google bot s nimi již dokáže pracovat (rozpoznat je) a bude je dále nabízet uživatelům pod označením "rich snippets" (Hansson 2009).

Velmi cenným pomocníkem jsou také tzv. Linked Data Cloud. Jde o neustále se rozšiřující množinu sémantických slovníků (ontologií), které jsou veřejně dostupné, znovupoužitelné a vytvořené podle principů Linked data (Bizer, 2009). Jde o několik bodů, jak bychom měli správně propojovat informace na webu, aby jimi šlo dobře procházet, prohledávat a nacházet další zdroje.Každý objekt má mít svoje URL, na kterém zájemce nalezne informace o objektu a odkazy na další relevantní zdroje. Informace jsou poskytovány na základě toho, kdo se ptá. Jestliže se ptá počítač, tak dostane strojově čitelná data. Pokud se ptá člověk, dostane je v lidské podobě (Petrák, 2007).

4. Závěr

Vize i technické detaily sémantického webu jsou stále předmětem diskusí, především mezi softwarovými a vědomostními inženýry, jichž se realizace týká nejvíc. Krom toho se o sémantickém webu diskutuje i v řadách tvůrců webových stránek a aplikací. Většina z nich očekává výsledky vývoje se zájmem, i když postoje nejsou jednoznačně kladné. Někteří očekávají nárůst komplikací při publikování na webu a orientaci sémantického webu ve prospěch uživatelů (pro které je web zdrojem informací, vědomostí a služeb) vidí jako nebezpečnou změnu poměrů. Jsou to patrně ti uživatelé webu, pro něž je prvotní službou webu možnost co nejjednodušeji publikovat.

Je zřejmé, že jednotnosti se snáze dosáhne v určité uzavřené oblasti – doméně. Proto dnes existuje poměrně velké množství tzv. doménových ontologií. Aktuální otázkou výzkumu je, zda pro úspěšnou realizaci informačních systémů budoucnosti postačí rozšiřování počtu těchto kamínku mozaiky sémantického webu, či zda je nutné zapojit i nějakou obecnou, široce pojatou, všeobjímající“ ontologii.

Nejvýznamnější je patrně časová náročnost tvorby takového zdroje. K překonání tohoto slabého místa může vést tradiční cesta – pokusit se použít to, co je již hotovo. Cílem takového přístupu k budování ontologií je integrace existujících databází, rozsáhlých bází znalostí budovaných pro jiné účely a dalších zdrojů, jejich „vyčištění“, zpřesnění uložených informací a nakonec integrace do jednotné ontologické struktury.

Použitá literatura

1. Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. Scientific american, 284(5), 28-37.
2. Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data-the story so far. International Journal on Semantic Web and Information Systems (IJSWIS), 5(3), 1-22.
3. Hansson, O. (2009, May 12). Google engineering explains microformat support in searches. Retrieved from http://radar.oreilly.com/2009/05/google-adds-microformat-parsin.html
4. Petrák, J.: (2007, July) Vše o propojování dat (Linked Data). Červen 2007 Retrieved from http://zapisky.info/?item=vse-o-propojovani-dat-linked-data&category=vse-o-semantickem-webu
5. Smrž, P., & Pitner, T. (2004). Sémantický web a jeho technologie (3). Zpravodaj ÚVT MU. ISSN, 1212-0901.

Fakulta aplikované informatiky, Univerzita Tomáše Bati ve Zlíně, Nad Stráněmi 4511, Zlín, 760 01