Podpora sémantického vyhľadávania na webe

Jozef Vrana[1], Kristína Machová[1], Martin Dzbor[2]

[1] Katedra kybernetiky a umelej inteligencie, Technická Univerzita,
Letná 9, 04200 Košice, Slovakia
{jozef.vrana, kristina.machova}@tuke.sk
[2] Knowledge Media Institute, The Open University, Walton Hall,
Milton Keynes, MK7 6AA United Kingdom
M.Dzbor@open.ac.uk

Abstrakt. Článok je venovaný problému vyhľadávania informácií na webe s podporou sémantického aspektu. Zameriava sa na transformáciu ontológie na lexikón. Táto transformácia predstavuje dedukciu všeobecnejšej ontológie na špecifickejší lexikón, ktorý je vhodnejší na vysvetľovanie konceptov z webovej stránky pomocou systému Magpie. Systém Magpie slúži na interpretáciu obsahu webových stránok pomocou ontológií. Sémantická vrstva závisí od dostupnosti ontológie. V rámci príspevku bol implementovaný a testovaný Lexicon Generator, ktorý generuje Magpie lexikón z príslušnej ontológie.

Kľúčové slová: Sémantický web, Magpie, lexikón, vyhľadávanie na webe

1 Úvod

Víziou Tima Berners-Lee je urobiť súčasný web strojovo spracovateľný[1]. To je možné iba priradením metadát do súčasného webu. Nie však len na úrovni web stránok ale aj na úrovni objektov web stránky. Pre popis takýchto vzťahov slúži Resource Description Framework. To umožní presne definovať objekty a vzťahy medzi nimi pomocou Uniform Resource Identifier. Pri použití takéhoto postupu predstavujú slová web stránky nie len reťazce lexikálnych jednotiek bez definovaného zmyslu, ale stavajú sa objektmi s plne definovaným významom a explicitne vyjadrenými vzťahmi k ostatným objektom na webe. V takejto štruktúrovanej sieti sa môže webový agend pohybovať a vyhľadávať pre používateľa relevantné zdroje k jeho dotazu. To by umožnilo odbúrať množstvo neproduktívnej práce s vyhľadávačmi.

Koncepcia podpory zmysluplnosti dát je rozdelená do štyroch úrovní:

- **Text a databázy** – táto základná úroveň je prístupná už na webe, takom ako ho poznáme dnes, a predstavuje základ pre pridávanie metadát.
- **XML dáta v jednej doméne** – predstavuje úroveň pre popis dát, ktoré sa presúvajú medzi aplikáciami v rámci jednej domény. Príkladom sú XML štandardy pre popis dát v oblastiach, ako je poisťovníctvo, zdravotníctvo a podobne.

- **Taxonómie a dokumenty s rôznymi slovníkmi** – dáta v tejto úrovni, pochádzajú z rozličných domén. Taxonómia je v tomto prípade prostriedkom pre špecifikovanie vzťahov medzi týmito rozdielnymi doménami. Jednoduché vzťahy dané taxonómiou kategórií umožnia jednoduchšiu orientáciu a kombinovanie týchto dát.
- **Ontológie a pravidlá** – najvyššia úroveň organizácie dát umožní odvodzovanie nových znalostí na základe presne popísaného modelu a exaktne definovaných pravidiel. Takto dobre popísaný model sprístupňuje rôzne výpočty, ktoré je nad ním možné vykonávať. Príkladom takéhoto výpočtu je automatický preklad dokumentu zo zdrojovej domény do ekvivalentnej domény.

Pojem „sémantický web“ sa pertraktuje vo svete informatiky už niekoľko rokov. Realizovať sémantický web je vec veľmi žiaduca no neľahká. Žiaduca, pretože web ako taký sa používa veľmi intenzívne po celom svete a teda obsahuje milióny statických informačných zdrojov s rôznou štruktúrou ich uloženia. Je preto nutné motivovať používateľov štruktúrovať webové stránky jednotným, vyššie popísaným spôsobom, a tým dosiahnuť postupné presadzovanie sémantického webu do praxe.

2 Softvérové prostriedky pre podporu sémantického webu

Sémantický web a sémantické webové služby sú dnes prevažne v štádiu vývoja. Tento vývoj má už aj svoje reálne výsledky napríklad nástroj s názvom Magpie. Ide o produkt Knowleadge Media institut na Open University [4].

Magpie bol navrhnutý ako komponent internetového prehliadača, ktorý slúži na interpretáciu obsahu internetových stránok. Poskytuje doplňujúci informačný zdroj, ktorý obsahuje relevantné informácie týkajúce sa daného webového zdroja. Magpie automaticky spojí internetový zdroj s jeho sémantickým obsahom. Dostupnosť tejto vrstvy však závisí na dostupnosti ontológie pre daný zdroj. Tá musí byť v zdroji explicitne obsiahnutá. Magpie teda predstavuje nástroj na využívanie služieb sémantického webu a dokáže sa flexibilne prispôsobiť rôznym potrebám používateľov [3].

Magpie však nečíta súbory s ontológiou priamo, ale sprostredkovane. Sprostredkovateľom je Magpie lexikón uložený v súbore *.onto. Dôvod takéhoto sprostredkovaného prístupu je, že v ontológii nie je možné dostatočne obsiahnuť lexikálnu bohatosť konceptov. Lexikónom sa taktiež definuje oblasť záujmu. Lexikón a ontológia sú pritom veľmi úzko zviazané. Magpie lexikón je “snapshot” ontológie pre vybrané triedy v danom časovom okamihu. Je to teda súbor, ktorý sa nevyvíja na rozdiel od ontológie. Lexikón a ontológia sú pritom veľmi úzko zviazané [5].

V rámci tejto práce bol vytvorený program Lexicon Generator, pre generovanie Magpie lexikónu z ontológie. Ide o aplikáciu, ktorá na základe zadaných vstupov generuje odpovedajúci *.onto súbor. Po zadaní vstupov je po niekoľkých sekundách k dispozícii súbor s Magpie lexikónom. Lexicon Generator je aplikácia napísaná

v Jave a teda je platformovo nezávislá. Keďže ide o konzolovú aplikáciu, je možné ju spúšťať aj na vzdialenom serveri, bez nutnosti exportovania grafického výstupu.

3 Sémantické webové služby a ich miesto v sémantickom webe

Vo všeobecnosti je možné za webové služby softwarový systém, podporujúci interakcie medzi strojmi cez sieť. Najznámejšími štandardmi sú WSDL (Web Services Description Language) a SOAP(Simple Object Access Protocol). Je podstatné že webové služby sú zväčša postavené na štandarde XML [1].
Popisovaný nastroj Magpie poskytuje dva druhy sémantických webových služieb a to služby „On Demmand“ a „Triger semantic service“. Zatiaľ čo „Triger semantic servis“ je len pokusnou službou a nebola úplne implementovaná do spomínaného komponentu, „On Demmand“ je plne funkčná, čo dokazujú prevádzané testy. Magpie je aplikáciou typu „client-server“ a teda požiadavky používateľa komprimuje do špecifickej formy požiadavky a odosiela serveru, ktorý túto žiadosť následne obslúži [5].
Komunikácia nepoužíva žiadnu zo spomínaných služieb (WSDL; SOAP) ale pracuje s vlastným formátom, ktorý nevyužíva XML. Tento prístup je veľmi jednoduchý a umožňuje aby na strane servera bola jednoduchá aplikácia, ktorá žiadosť obslúži a výsledok odošle a prezentuje používateľovi ako nové okno prehliadača.

4 Testy

Ako ontológia bol k dispozícii súbor „kmi-basic-portal-kb.owl“ získaný z Knowledge Media institute. Táto ontológia je zameraná na projekt Advanced Knowledge Technologies. Ontológia teda obsahuje mená výskumných pracovníkov, názvy konferencií, mená študentov podieľajúcich sa na projektoch a podobne.

Tabulka 1. Hodnoty návratnosti (Recall) entít na zvolených testovacích stránkach pre jednotlivé verzie lexikónov

	KMi-people	Magpie	AKT-people	AKT-publications
kmi-basic-portal-kb_ver.1	1	0,25	0,933	0,224
kmi-basic-portal-kb_ver.2	1	0,3	0,933	0,483
kmi-basic-portal-kb_ver.3	1	1	1	1

Z tejto ontológie boli generované pomocou programu Lexikon Generator lexikony so zameraním na triedy: „*kmi-research-staff-member, kmi-phd-student a kmi-academic-staff-member*“. Výber týchto tried bol podmienený skutočnosťou, že heuristické

programy, ktoré sú súčasťou programu Lexikon Generator, sú zamerané na prácu s menami. Tieto triedy boli zastrešené top triedou „*Community*“.
Teda, testovanie sa uskutočnilo nad tromi lexikónmi, ktoré sa líšia iba stupňom lexikálnej bohatosti. Pre testovanie boli zvolené stránky KMi a AKT, pre ktoré bola ontológia „kmi-basic-portal-kb.owl“ vytvorená. V Tab.1 sú uvedené hodnoty návratnosti entít na každej z testovaných stránok a pre každý z testovaných lexikónov.

5 Záver

Z hodnôt uvedených v Tab. 1 vyplýva význam použitia heuristík (kmi-basic-portal-kb_ver.2, kmi-basic-portal-kb_ver.3) pre zvýšenie lexikálnej bohatosti. Väčší význam má ich použitie na tých stránkach, kde sú mená súčasťou väčšieho textu (napr. zoznam literatúry, odstavec textu). To je prípad stránok Magpie a AKT-publications.

Tento príspevok vznikol s podporou VEGA grantu MŠ SR č. 1/4074/07 „Metódy anotovania, vyhľadávania, tvorby a sprístupňovania znalostí s využitím metadát pre sémantický popis znalostí“.

References

1. Studer, R., Grimm, S., Abecker, A.: *Semantic web services*. Springer-Verlag Berlin Heidelberg, 2007, New York, ISBN 978-3-540-70893-3.
2. Bielikova, M., Návrat. P. a kol.: *Štúdie vybraných tém softvérového inžinierstva*, STU, Bratislava, Vazovova 5, 2006.
3. Dzbor, M., Motta, E. - Domingue, J. B.: *Opening Up Magpie via Semantic Services*, In Proc. of the 3rd Intl. Semantic Web Conference, November 2004, Japan.
4. Domingue, J.B., Dzbor, M., Motta, E.: *Collaborative Semantic Web Browsing with Magpie*, In Proc. of the 1st European Semantic Web Symposium (ESWS), May 2004, Greece.
5. Domingue, J., Dzbor, M., and Motta, E., *Semantic Layering with Magpie, In Handbook on Ontologies in Information Systems*, Staab, S. and Studer, R. (Eds.) 2003, Springer Verlag.

Annotation:
Podpora sémantického vyhľadávania na webe
This paper is devoted to the problem of the Internet searching with the support of semantic aspect. It is focused on ontology transformation into a lexicon. This transformation represents a deduction of a more general ontology into a more specific lexicon, which is more suitable for concept explanation from web pages with the aid of system Magpie. This system serves for web page concept interpretation with the aid of ontology. The semantic layer depends on ontology available. Lexicon generator was implemented and tested.