# Techniky a nástroje sémantického vyhledávání

**Jiří JELÍNEK, Tomáš KINCL**
Vysoká škola ekonomická v Praze, Fakulta managementu Jindřichův Hradec
jelinek@fm.vse.cz
kincl@fm.vse.cz

INFORUM 2005: 11. konference o profesionálních informačních zdrojích
Praha, 24. - 26.5. 2005

***Abstrakt.*** *Informační prostor přístupný prostřednictvím služby WWW je bezpochyby nejrozsáhlejším informačním zdrojem dneška. Jeho obsáhlost je však současně jeho největší nevýhodou přinášející nutnost použití efektivního vyhledávacího nástroje. Vyhledávací systémy současnosti jsou především dvou základních druhů. Do první kategorie patří indexové systémy založené na práci s klíčovými slovy a hodnocení stránek na základě jejich vzájemných vazeb. Představitelem druhé jsou pak katalogové a adresářové systémy vytvářené převážně experty. S postupným rozvojem a rostoucím nasazením technik sémantického popisu WWW stránek je však stále aktuálnější otázka existence vyhledávacích nástrojů založených na principech sémantického webu.*

*Základem většiny sémantických vyhledávačů je existence sémantické sítě zachycující vzájemné vztahy mezi vloženými pojmy. Její struktura, naplnění a dostupnost jsou hlavními oblastmi výzkumu a vývoje v současnosti. Problematika sémantického vyhledávání se ale týká rovněž uživatelů, kteří musí být připraveni a schopni takové vyhledávače používat. Pozornost musí být věnována jak otázkám konstrukce vyhledávacího dotazu, tak i formě prezentace výstupů vyhledávání, např. jejich vizualizaci.*

*Tento příspěvek si klade za cíl prezentovat techniky využívané v sémantických vyhledávačích a s tím související způsoby konstrukce dotazů a vizualizace vyhledaných dat. Snahou bude též prezentovat konkrétní systémy dostupné v současnosti a praktické zkušenosti s nimi.*

## Úvod

Informační prostor přístupný prostřednictvím služby WWW je bezpochyby nejrozsáhlejším informačním zdrojem dneška. Jeho obsáhlost je však současně jeho největší nevýhodou přinášející nutnost použití efektivního vyhledávacího nástroje.

Vyhledávací systémy současnosti jsou především dvou základních druhů. Do první kategorie patří indexové systémy založené na práci s klíčovými slovy a hodnocení stránek na základě jejich vzájemných vazeb. Tyto systémy jsou většinou automaticky aktualizovány pomocí webových robotů. Představiteli druhé kategorie jsou pak katalogové a adresářové systémy vytvářené převážně experty a používající hierarchické uspořádání URL odkazů podle jejich věcného zaměření. S postupným rozvojem a rostoucím nasazením technik sémantického popisu WWW stránek je však stále aktuálnější otázka existence vyhledávacích nástrojů založených na principech sémantického webu.

Co si lze vlastně představit pod pojmem sémantické vyhledávání? Jedná se o vyhledávání s využitím sémantické, tj. významové informace. Ta má nejčastěji podobu trojice *subjekt-vztah-objekt* a vychází z principů definovaných v jazyce RDF [RDF]. Výsledná struktura je pak vlastně (zjednodušeně řečeno) orientovaným grafem, v jehož uzlech jsou objekty a hrany označují vzájemné vztahy těchto objektů. Pokud uzly charakterizují určité pojmy či koncepty vytvářející (obvykle) hierarchickou strukturu popisující určitou doménu či problém, můžeme hovořit o ontologiích. Ontologií existuje celá řada od těch nejobecnějších (pokrývajících běžné pojmy a oblasti) ke specifickým (doménovým).

Jak už samotný název říká, musí významnou úlohu ve vyhledávacím procesu hrát sémantická informace. Kde se však má objevit? Je to cíl vyhledávání nebo prostředek v něm použitý? O obou uvedených interpretacích má smysl uvažovat.

Problematika sémantického vyhledávání se ale týká rovněž uživatelů, kteří musí být připraveni a schopni jej používat. Pozornost musí být věnována jak otázkám konstrukce vyhledávacího dotazu, tak i formě prezentace výstupů vyhledávání.

# Sémantické vyhledávání

Tato kapitola objasňuje dva možné pohledy na pojem *sémantické vyhledávání*. V prvním jde o pojetí, ve kterém je sémantická informace cílem vyhledávání. Druhý přístup zdůrazňuje užití sémantiky v procesu vyhledávání.

## *Sémantická informace jako cíl vyhledávání*

V této interpretaci je možné sémantické vyhledávání chápat jako snahu objasnit určitý pojem a ještě častěji zařadit ho do ontologické struktury. Základním předpokladem aplikace výše naznačeného přístupu je dostupnost širokého spektra ontologií zahrnujících dotazované pojmy a domény. Ideálním stavem by byla jediná všezahrnující ontologie popisující všechny pojmy související se všemi zájmy uživatelů, která by kromě toho byla též široce akceptována a přijímána uživateli. Je zřejmé, že tento úkol je při dnešní šíři poznání neuskutečnitelný a řešení musí být podstatně jednodušší. Mohlo by jít např. o definování ontologií pro hlavní oblasti zájmu nebo řešené problémy. I zde však (stejně jako ve všech oblastech, kde je nutné pracovat s lidským faktorem) narazíme na rozdílnost v názorech na ontologické uspořádání pojmů a tím i existenci většího množství různě obsáhlých ontologií. Standardy pro jednotlivé domény prozatím nejsou kodifikovány, vše záleží na podpoře a akceptaci jednotlivých kandidátů širokou veřejností.

Příkladem systémů vyhledávajících sémantické informace jsou např. [Swoogle] nebo [SWS]. Nástroje této kategorie jsou si dosti podobné. Vzhledem k větší propracovanosti bude podrobněji popsán systém Swoogle.

Swoogle je klasický indexový vyhledávací systém určený však pro práci se sémanticky popsanými dokumenty (např. pomocí jazyka RDF nebo OWL). Sémantický popis charakterizuje pojmy a vazby ve specifické oblasti, kterou se dokument zabývá. Popisuje tedy jeho vnitřní ontologii. Takové dokumenty mohou samozřejmě kromě strojově srozumitelné sémantické informace obsahovat i další data, např. v jazyce XML, či hlavně HTML, určená uživateli. V praxi se však jedná téměř výhradně o ontologické definice z nejrůznějších oblastí. Swoogle takovéto dokumenty indexuje a ukládá do vlastní databáze. Ta pak prakticky realizuje mapování jednotlivých stránek (URL) k ontologickým pojmům. Při dotazu na konkrétní pojem pak vrací seznam dokumentů, ve kterých je tento pojem použit (definován). Uživatel může též blíže specifikovat, zda se mají prohledávat pouze názvy tříd nebo vztahů. Výsledek vyhledávání je možné zobrazit několika způsoby a lze použít i některé externí nástroje. Systém také disponuje rozhraním pro webové služby, takže ho je možné využít v rozsáhlejších automatizovaných nástrojích. V současné době Swoogle indexuje cca 337 tisíc dokumentů a cca 47 mil. RDF trojic *subjekt-vztah-objekt.*

Dosud nevyjasněná zůstává u systémů pro vyhledávání sémantických informací otázka vhodné reprezentace výstupu. Podoba převzatá z klasických vyhledávačů totiž není pro prezentaci ontologických struktur příliš vhodná. Pokud k systému přistupujeme přes API, předpokládá se další strojové zpracování a strohý, jasně strukturovaný formát výstupu je na místě. Případná výsledná reprezentace leží na navazujících systémech. Pokud však volíme přístup přes webovou stránku, měla by být výstupní informace prezentována přehledně pro uživatele, nejlépe graficky. Zde zmíněné systémy se však v tomto bodě spoléhají na externí nástroje. Např. Swoogle umožňuje přímo volat nástroj pro práci s ontologiemi Swoop. Pro další práci s výstupními dokumenty (nejčastěji ontologiemi) lze použít i další externí nástroje (např. [Protege]).

## *Sémantická informace jako prostředek vyhledávání*

Běžného uživatele WWW bude patrně více zajímat druhá interpretace sémantického vyhledávání. Tou je užití sémantické informace *v procesu vyhledávání* umožňující uživateli vyhledat vysoce relevantní informace k tématu, které ho zajímá.

Prvním krokem při konkrétní realizaci je zařazení vyhledávaného pojmu do ontologické struktury, tedy postup popsaný v předchozím oddílu. Následný postup spočívající v nalezení WWW stránek zabývajících se danou tématikou také není zcela bezproblémový. Klíčovou otázkou je zde existence vazeb či přiřazení (mapování) mezi ontologickými pojmy a WWW stránkami, které se jimi zabývají. Pro tento úkol nelze využít stávající indexové vyhledávací systémy postavené na existenci klíčových slov v dokumentech či stránkách. Výskyt klíčových slov však téměř nic neříká o vztahu stránky k tématu vyhledávání. Přiřazení stránek musí nastat pouze u pojmů, kterými se stránka významně zabývá či se jich dotýká, a které lze ontologicky zařadit. Nejde tedy o prostou indexaci na základě slov vyskytujících

se na stránce, ale o definici zaměření stránky vyjádřenou hlavními pojmy. I přes tento nedostatek je však možné klasické vyhledávače použít s rizikem určité chyby ve výstupu. Důvodem pro tento krok je především záběr současných indexových systémů [Google]. Typickým řešením je pak vytvořit nad těmito systémy metavyhledávač.

Efektivnějším postupem by jistě bylo použít pro vyhledávání sémantickou informaci z příslušné stránky a mít tak jistotu, že nalezené dokumenty se skutečně příslušným tématem zabývají. Počet takto anotovaných stránek s informačním obsahem je však stále malý. Přesto lze s rozvojem sémantického webu očekávat nárůst jejich počtu. Sémantický popis informačních stránek by se mohl týkat pouze základních pojmů, kterými se stránky zabývají, případně příslušnosti pojmů k ontologiím.

Jaká zlepšení lze očekávat při využití nástrojů sémantického vyhledávání? Zde je nutné si ujasnit, jaké faktory ovlivňují kvalitu vyhledávání. Na jedné straně jsou to schopnosti vyhledávacího systému, na druhé pak chování a schopnosti uživatele. Mezi ty patří např. znalosti uživatele, jeho vztah k procesu vyhledávání, schopnost pracovat s jazykem vyhledávacího systému či používaný slovník [Alb04]. Přínos sémantického vyhledávání lze vidět především v oblasti používaného (především odborného) slovníku. Uživateli mohou být nabídnuty též ontologicky blízké pojmy.

# Současnost sémantického vyhledávání

Současné problémy sémantického vyhledávání lze vidět především v následujících bodech:

- Dostupnost sémantické informace
- Využití sémantické informace v procesu vyhledávání na WWW
- Vizualizace procesu a výstupu vyhledávání, případně použitých ontologických struktur

## *Dostupnost sémantické informace*

Pro sémantické vyhledávání je podstatná nejen dostupnost ontologií popisujících danou doménu, ale i dostupnost přiřazení konkrétních WWW stránek k pojmům z těchto ontologií [Hyv].

**Dostupnost ontologií**

Tvorba obecně použitelných ontologií, zahrnujících s dostatečnou mírou detailu hlavní oblasti prohledávané uživateli, je velmi náročná. Téměř nikdy se nelze spolehnout pouze na automatické nástroje a je nutné využít služeb expertů. Hlavním trendem současnosti je definování tzv. doménových ontologií, zabývajících se pouze relativně úzkou pojmovou oblastí, které však mohou být kombinovány např. díky jednotnému jazyku nebo rozhraní využívajícímu stejné formy a protokoly komunikace.

Výzvou současnosti je pak sdílení ontologických informací. Proto je nutné zvolit stejný popisný nástroj (RDF, OWL) a postupy vedoucí k široké akceptaci výsledků. Mezi ty mohou patřit např. anotační systémy jako je Annotea [Annotea]. Jedná se o nástroj umožňující doplňovat stávající dokumenty nebo webové stránky dalšími metainformacemi bez zásahu do těchto dokumentů. Kromě komentářů, poznámek či vysvětlení mohou být těmito informacemi také např. ontologické pojmy charakterizující danou stránku, protože Annotea používá jako základní nástroj jazyk RDF. Kromě vkládání a zobrazování metainformací lokálně umožňuje Annotea jejich sdílení prostřednictvím dedikovaných anotačních serverů, které pak mohou být využity jako zdroj informací o přiřazení dokumentů k ontologickým pojmům. Přístup k serverům zajišťují specializovaní klienti, kteří však mohou být součástí webových prohlížečů (např. W3C Amaya).

Pro tvorbu obecně akceptovatelných ontologií je možné též vyjít ze struktur adresářových a předmětových katalogů, které mají za sebou dostatečně dlouhý vývoj a konsolidaci.

**Dostupnost sémantického popisu stránek**

Lze rozlišit dvě hlavní cesty, jak získat sémantický popis stránek. První z nich je přímá extrakce tohoto popisu z WWW stránek, kterou lze použít, pokud jsou stránky anotovány v některém z podporovaných formátů pro sémantický popis (RDF či OWL). Sémantický popis WWW prostoru však zatím není dostatečně rozšířen, proto je nutné získávat informace i jinými způsoby, např. použitím nástrojů, které se o extrakci sémantické informace pokusí i z neanotované stránky (např. postupy Web Content Miningu). V takovém případě je nutné detekovat na stránce významné termy, ty porovnat s ontologickými pojmy a na základě nalezené shody přiřadit stránku k příslušnému pojmu. Kvalita výstupu může však být u obou uvedených přístupů značně odlišná. Pojmy charakterizující obsah dané

stránky pak mohou být zařazeny do ontologických struktur používaných vyhledávačem společně s přimapováním příslušné WWW stránky k těmto pojmům.

Jestliže výše uvedené mapování není k dispozici, je hlavním užitím ontologie zpřesnění a rozšíření dotazů na stávající vyhledávací služby. Uživateli může být nabídnut seznam blízkých pojmů, ke kterým má ten, jež zadal, nějaký vztah. Pro „blízkost“ pojmů musí být k dispozici jednoznačně definovaná míra. Tento interaktivní proces je velmi úzce svázán s provedením a možnostmi (především vizualizačními) použitého rozhraní.

### *Využití sémantické informace v procesu vyhledávání*

Pouhé nalezení sémantické informace však obvykle není finálním cílem sémantického vyhledávání. Je nutné se zamyslet též nad možnostmi jejího užití. Ty se odvíjejí od podoby a rozsahu dostupné ontologie a existence mapování jednotlivých stránek k pojmům zachyceným v ontologii. Vyhledávací systémy mohou pracovat s externími ontologiemi nebo si postupně vytvářet své vlastní, např. na základě sémantických informací získaných indexačními roboty. *Ontologické struktury vyhledávačů* pak mohou být použity při podpoře vyhledávání.

Jedním z klíčových problémů je též vhodná volba mechanizmu konstrukce vyhledávacího dotazu a dotazovacího jazyka samotného. Je ověřenou skutečností, že zkušení uživatelé a odborníci v dané doméně dávají přednost tvorbě dotazu s použitím pojmů a logických operátorů daného jazyka, zatímco neodborníci preferují vizualizační nástroje.

### *Vizualizace procesu a výstupu vyhledávání*

Praktické pokusy ukázaly, že stejně důležitá jako schopnost vyhledat požadovanou informací je také schopnost tuto informací vhodně prezentovat (vizualizovat). Vizualizační schopnosti jsou též podstatné v případě interaktivního vyhledávání a zpřesňování dotazů.

Pro klasický web se v poslední době objevilo několik inovativních přístupů k vizualizaci výsledků vyhledávání ([Grokker], [Aduna], [Kartoo]). Jejich nejednotnost je však největší překážkou jejich použití, uživatel se musí adaptovat vždy na nový model zobrazení. Rovněž je nutné poznamenat, že zmíněné systémy se výrazně liší svými schopnostmi (např. tím, zda umí pracovat se stávajícími vyhledávacími systémy, či pouze s předem definovanými a indexovanými servery).

V oblasti vizualizace ontologií je situace poněkud lepší a existují známé vizualizační nástroje a postupy, a to buď jako samostatné produkty nebo jako vestavné moduly do větších systémů [Protege].

I vzhledem k výše uvedenému je v oblasti vizualizace výstupů sémantického vyhledávání situace poněkud nejasná. Podobně jako pro klasický web není definován a podporován jednotný standard viyualizace.

## Závěr

Z popisu celkového stavu v oblasti sémantického vyhledávání je zřejmé, že hlavní rozmach tohoto odvětví ještě nenastal, ale velmi rychle se blíží. Hlavními problémy současnosti jsou absence ontologických standardů, sémantického popisu běžných WWW stránek a nejednotnost přístupů (např k vizualizaci výsledků). Po jejich překonání najdou vyhledávací nástroje založené na principu sémantického webu jistě široké uplatnění.

## Literatura


[Aduna] Aduna AutoFocus 2005.1. In: http://aduna.biz/products/autofocus/index.html

[Alb04] Albertoni R., Bertone A., De Martino M.: Semantic Web and Information Visualization. Semantic Web Applications and Perspectives (SWAP), 1st Italian Semantic Web Workshop, 10th December 2004, Ancona, Italy, In: http://semanticweb.deit.univpm.it/swap2004/cameraready/albertoni.pdf

[Annotea] Annotea project. In: http://www.w3.org/2001/Annotea/

[Ben] Benschop A.: The future of the semantic web. In: http://www2.fmg.uva.nl/sociosite/websoc/semantic.html

[Google] Google. In: http://www.google.com/

[Grokker] Knowledge management, data mining and information mapping with Grokker. In: http://www.grokker.com/

[Guh03] Guha R., McCool R., Miller E.: Semantic Search. In: Proceedings of the twelfth international conference on World Wide Web, Budapest, Hungary, 2003, pp. 700 – 709, ISBN1-58113-680-3

[Har99] Van Harmelen F., Fensel D.: Practical Knowledge Representation for the Web. In: http://www.cs.vu.nl/~frankh/postscript/IJCAI99-III.pdf

[Hef03] Heflin J., Hendler J., Luke S.: SHOE: A Blueprint for the Semantic Web. In: http://www.cse.lehigh.edu/~heflin/pubs/swbook03.pdf

[Hyv] Hyvönen E., Saarela S., Viljanen K.: Application of Ontology Techniques to View-Based Semantic Search and Browsing. In C. Bussler, J. Davies, D. Fensel, R. Studer (eds.): The Semantic Web: Research and Applications. Proceedings of the First European Semantic Web Symposium (ESWS 2004), Springer-Verlag, LNCS 3053, 2004.

[Kartoo] KartOO visual meta search engine. In: http://www.kartoo.com/

[OWL] Web Ontology Language OWL / W3C Semantic Web Aktivity. In: http://www.w3.org/2004/OWL/

[Protege] The Protégé; Ontology Editor and Knowledge Acquisition Systém. In: http://protege.stanford.edu/

[RDF] Resource Description Framework (RDF) / W3C Semantic Web Activity. In: http://www.w3.org/RDF/

[Swoogle] Swoogle. In: http://swoogle.umbc.edu/index.php

[SWS] Intellidimension Semantic Web Search. In: http://www.semanticwebsearch.com/

[Teoma] Teoma - Search with Authority. In: http://www.teoma.com/