NÁSTROJE PRO LÉČBU INFORMAČNÍHO ZAHLCENÍ

Šmerda J., Dosoudil V., Winkler M.

Ústav výpočetní techniky, Masarykova univerzita & Laboratoř znalostních a informačních robotů, Fakulta informatiky, Masarykova univerzita

Abstrakt

V dnešní době se při práci s informacemi objevují dva jevy: zahlcení informacemi a absence povědomí o dostupnosti užitečných informací. Ukazuje se, že tyto jevy spolu úzce souvisí. Informační zahlcení nám často pomohou zvládnout informace, o kterých ani nevíme, že nám jsou dostupné. Klíčovými technikami jsou metody informační syntézy z různých datových zdrojů, práce s neurčitostí, s kontextovou závislostí informací, se sémantikou a pokročilé metody vizualizace. V příspěvku budou ilustrovány uvedené jevy na konkrétním případu užití a bude představena vyvíjená technologie, která si dává za pomoci uživateli zvládnout jak zahlcení informacemi, tak jejich nedostatek. Následně bude prezentováno konkrétní řešení z oblasti digitálních knihoven vyvíjené na představených technologiích na ÚVT MU ve spolupráci s Laboratoří znalostních a informačních robotů na FI MU.

Klíčová slova: zahlcení informacemi, informační syntéza, servisní systémy, multi-agentní systémy, technologie znalostních a informačních robotů, digitální knihovny

Abstract

Nowadays, two phenomenons appear during working with information: information overload and information insufficiency. We found out that these two phenomenons are closely related together. We can often manage information overload by using information which is available to us, but we are not aware of its existence. Crucial techniques are information synthesis from heterogeneous data sources, working with uncertainty, with context dependent information, with semantics and advanced visualization methods. In this paper, a specific developed technology will be introduced. A goal of the technology is to help user with management of information overload and information insufficiency. A specific use of the technology will be presented then – the solution in digital libraries, which is being developed at Institute of Computer Science, Masaryk University in collaboration with Knowledge and Information Robots Laboratory at Faculty of Informatics, Masaryk University.

Keywords: information overload, information synthesis, service systems, multi-agent systems, knowledge and information robots technology, digital libraries

Motivační příklady z oblasti digitálních knihoven

Na úvod do řešené problematiky uvedeme dva motivační příklady z oblasti digitálních knihoven. Tyto příklady čtenáři umožní si lépe představit cíl našeho výzkumu a vývoje na konkrétním příkladu.

Záměrně byly zvoleny příklady z lékařského prostředí. Digitální knihovny zaměřené na medicínu totiž obsahují obrovské množství cenných článků zejména pro lékaře. Ti ale patří do skupiny lidí, kteří nutně potřebují získávat informace, bohužel k jejich získávání mají velice málo času. Setkávají se s překážkami, které jim nedovolují možnosti digitálních knihoven využívat na maximum. Při práci s informacemi nastávají dva jevy: informační zahlcení a naopak absence povědomí o dostupnosti užitečných informací. V následujících odstavcích na konkrétním příkladu získávání informací z digitálních knihoven tyto dva jevy demonstrujeme a ukážeme náš přístup k jejich řešení. Příspěvek si nedává za cíl jít do hloubky a problematiku podrobně analyzovat, slouží jako stručné představení našeho přístupu a ilustrace směru, kterým v projektu jdeme.

Uznávaní autoři a renomované časopisy

První motivační příklad z oblasti získávání informací z elektronických informačních zdrojů [1]:

Jsem lékař a zajímám se o infarkt mozku. Největší problém vidím ve faktu, že existuje v různých informačních zdrojích příliš mnoho článků o tomto tématu a nemám čas je všechny procházet a pročítat.

Chci vědět, co publikovali uznávané vědecké kapacity o infarktu mozku v renomovaných časopisech. Zajímají mě relevantní články jak v elektronickém, tak v papírovém vydání.

Zamysleme se, jakým způsobem bychom mohli určit, zda je autor uznávaný a zda je časopis renomovaný.

Jedním ze způsobů, který je možné použít, je udržovat seznam autorů, které pokládáme za uznávané, a podobně seznam časopisů, které pokládáme za renomované. Toto řešení lze ovšem použít pouze v případě, kdy už z předchozí zkušenosti jsme schopni tyto seznamy vytvořit.

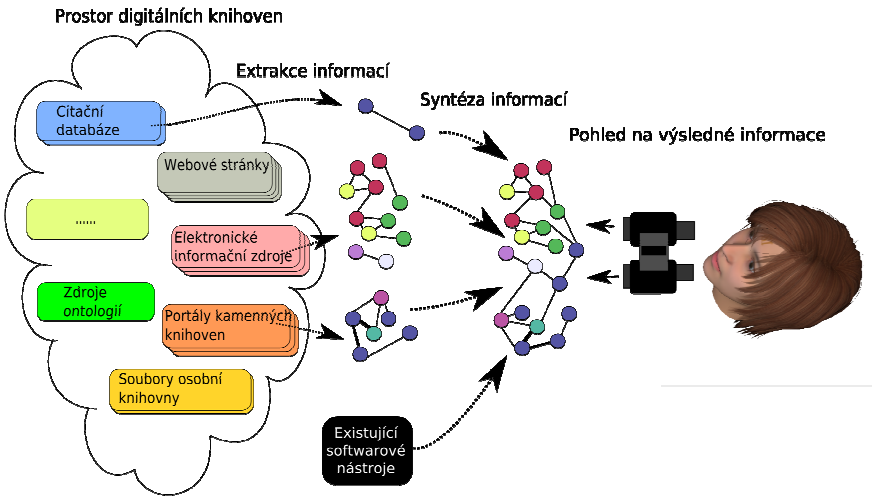
Dalším způsobem, kterým můžeme určovat míru uznávanosti autora, je na základě citačních indexů jeho publikací, tj. množství citací těchto publikací. Citační index slouží jako primární zdroj pro stanovení impaktního faktoru časopisu. Za renomovaný časopis potom považujeme časopis impaktovaný, ve kterém publikují zejména renomovaní autoři. Tento způsob nás zbavuje nutnosti si subjektivně renomovanost určovat a nabízí její objektivnější posouzení.

Navržené způsoby určování renomovanosti může být vhodné i kombinovat a získat tak další zajímavé pohledy do oblasti zájmu.

Postup řešení motivačního příkladu

K ilustraci postupu při řešení motivačního příkladu použijeme obrázek 1. Smysl prezentovaného příkladu je hlavně v demonstraci principů při práci

v digitálních knihovnách. Příklad ilustruje oblast vyhledávání a není rozhodně jediný, kterým bychom se chtěli zabývat a řešit. Prostor digitálních knihoven je široký, přidáním dalších datových zdrojů do procesu můžeme zvětšit množinu úloh, které jsme schopni řešit.



Obrázek 1: Ilustrace procesu získávání informací v oblasti digitálních knihoven

Prostor digitálních knihoven

Levá část obrázku 1 znázorňuje prostor digitálních knihoven. Prostor digitálních knihoven zahrnuje všechny kolekce a datové zdroje, které jsou považovány za digitální knihovny, a též ty, které mohou při práci s digitálními knihovnami poskytovat užitečné informace. Tento prostor je velice heterogenní, ať už z pohledu dostupnosti, způsobu přístupu nebo objemu informací.

Z našeho motivačního příkladu do tohoto prostoru patří v první řadě elektronické informační zdroje (neboli kolekce) z oblasti lékařství (MEDLINE, ProQuest, Springer, Elsevier atd.), ve kterých chceme vyhledat relevantní články. Chceme ovšem hledat i mezi publikacemi, které se nacházejí v papírové podobě, takže je vhodné zahrnout i katalogy kamenných knihoven (např. katalog knihoven na MU). K řešení našeho příkladu dále potřebujeme informace o renomovanosti autorů a časopisů. Jedním ze zdrojů, ve kterém můžeme najít informace o citačních indexech a impaktních faktorech, je Web of Science nabízený v rámci projektu Web of Knowledge. Stejně tak můžeme impaktní faktor získat z dalších zdrojů, jako např. Google Scholar. Dalším důležitým

prvkem prostoru digitálních knihoven, který může pomoci k lepšímu výsledku, jsou zdroje, které poskytují ontologie, a to nejen čistě z lékařského prostředí (MESH, SNOMED atd.).

Extrakce, adaptace a syntéza informací

Abychom mohli dále s informacemi poskytovanými jednotlivými prvky prostoru digitálních knihoven pracovat, je třeba provést informační integraci, která se skládá z následujících operací [6]. Nejdříve informace extrahujeme a následně uchopíme konceptuálním systémem, který při práci s digitálními knihovnami používáme. To znamená, že extrahovaná data adaptujeme. Teprve potom můžeme se získanými informacemi dále pracovat a provádět informační syntézu, která způsobí, že informace z různých datových zdrojů jsou propojeny do souvislého celku. Syntézu je možné provádět několika způsoby, a to nejen odvozením souvislostí a faktů na základě společných vlastností objektů, ale i pomocí pravidel, která v prostoru digitálních knihoven platí.

Vraťme se k našemu příkladu. Z prostoru digitálních knihoven, konkrétně z kolekcí MEDLINE, ProQuest, Springer atd. a též z katalogů tradičních knihoven, extrahujeme články k tématu infarkt mozku. To vše za pomoci zdrojů, které poskytují informace o ontologiích. Informace o renomovanosti jednotlivých autorů a časopisů získáme z citačních databází v kolekci Web of Science a nebo z námi předem sestaveného seznamu autorů a časopisů. Cílem informační syntézy je významově propojit získané informace a nalézt v nich či vytvořit požadované i dosud neznámé souvislosti.

Výsledkem procesu extrakce, adaptace a syntézy informací je komplexní informace o každém získaném článku. Příkladem tedy může být informace, že článek pojednává o vyhledávaném tématu, jeho autor dosahuje vysokého citačního indexu, byl publikován ve vysoce impaktovaném časopise a dokonce se jméno autora nachází na mém seznamu vybraných autorů.

Vizualizace výsledků

Nyní přichází moment, kdy sice máme výsledné informace, avšak je třeba se v nich dobře orientovat a efektivně s nimi pracovat. Vizualizace výsledků je finálním a klíčovým krokem pro úspěch celého procesu získávání informací z prostoru digitálních knihoven. Jejím úkolem je pomoci uživateli zaměřovat pozornost a pomoci mu rozlišit podstatné od nepodstatného.

Příklad s dostupností plného textu článku a časopisu

Uveďme nyní druhý motivační příklad, tentokrát z oblasti dostupnosti, ptáme se [1]:

Je pro mě dostupný fulltext daného článku/časopisu?

Může se totiž stát, že článek, který nás zajímá, je obsažený v kolekci, ke které nemáme zaplacený přístup. Avšak ten samý článek se nachází i jinde, kam přístup už máme. U časopisů nás zase zajímá, kde se mohu dostat k jeho fulltextu. Například různé organizační celky Masarykovy univerzity mají předplacené různé časopisy a pro přístup k některému časopisu je třeba pouze zajít na to správné místo a tam fulltext získat.

Řešení motivačních příkladů dnešními nástroji

Podívejme se nyní na to, jak bychom ilustrativní příklady řešili s dnešními dostupnými nástroji. Docházíme ke zjištění, že je velice obtížné současnými nástroji uvedené příklady řešit. Současné nástroje nám pomáhají řešit dílčí úlohy, celkový pohled ovšem chybí. Informace jsou dostupné v izolovaných celcích a přicházíme tak o cenné souvislosti mezi jednotlivými částmi. Informační schopnost sjednocení všech informačních zdrojů je totiž větší než sjednocení informačních schopností jednotlivých informačních zdrojů [11]. Informační syntézu z jednotlivých zdrojů si nyní musí každý udělat svépomocí, což vyžaduje netriviální úsilí. A tak lze konstatovat, že informace sice dostupné jsou, ale nejsou snadno dostupné, což v důsledku vede k tomu, že nejsou využívané.

Informační nedostatek a informační zahlcení

Nejprve musíme čelit již dříve zmíněné absenci povědomí o dostupnosti informací. Odkud máme vlastně chtěné informace získat? Co vlastně prostor digitálních knihoven obsahuje? Postupně si najdeme několik zdrojů, ve kterých hledáme, a další již neuvažujeme. Je to přirozené, nikdo nemá čas stále hledat nové zdroje a i kdyby to někdo dělal, už nemá čas je všechny při každém hledání procházet. Každopádně i při práci s několika vybranými zdroji postupně docházíme k tomu, že obsahují veliké množství informací a „oddělit zrno od plev“ je náročné. Dochází k informačnímu zahlcení.

Informační zahlcení lze ovšem mírnit za pomocí dalších informací, které jsou nám dostupné. Z prvního příkladu jsou to informace o renomovanosti autorů a časopisů, které nám pomohou si z nalezených článků vybrat ty nejlepší. Nemusí to být ovšem pouze informace získané od externích subjektů. Dalším velice dobrým vodítkem k určení preferencí uživatele je historie jeho práce. To znamená, že při vyhledávání uvažujeme i dříve provedené uživatelovy dotazy a jím vybrané výsledky. Takto průběžně tvoříme uživatelův profil a případně s uživatelem spolupracujeme tak, že sám může poskytnout informace o tom, co se mu líbí a co se mu nelíbí.

Servisní systémy

Teoretickým východiskem architektury servisních systémů je inovativní směr SSME [8], mezioborový přístup ke studiu, návrhu a implementaci servisních systémů, který je iniciativou společnosti IBM [9]. Architektura servisních systémů je založena na agentech, kteří jsou především [6]

- autonomní,
- kooperativní,
- skladební,
- distribuovatelní,
- interaktivní,
- adaptabilní.

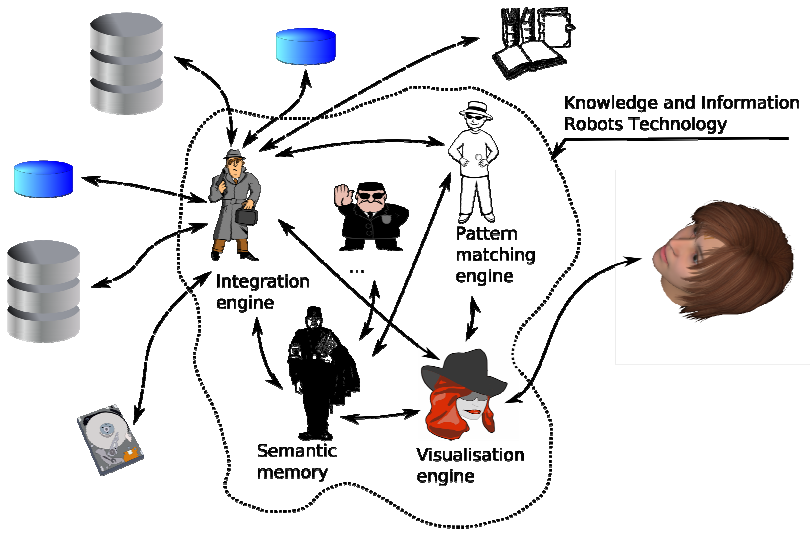
Agent servisního systému poskytuje služby ostatním agentům servisního systému (viz obrázek 2). Servisní systém je společenství navzájem spolupracujících agentů, jejichž činností se vytváří hodnota. Je to komplexní adaptivní systém, kde lidé a technologie navzájem spolupracují k tomu, aby vytvořili přidanou hodnotu [5].

Představovaná technologie znalostních a informačních robotů je právě servisním systémem. Je založena na principu multi-agentního systému, v němž každý agent něco umí a v rámci služeb své schopnosti nabízí ostatním agentům [10]. Agenti v systému komunikují pomocí zpráv, které si navzájem předávají. Servisní systém založen na technologii znalostních a informačních robotů je takový multi-agentní systém, který disponuje speciální službou nazývanou sémantická paměť, přesněji heterogenní selfreferenční sémantická síť. Východiska pro servisní systémy založené na technologii znalostních a informačních robotů jsou principy zmiňování-užívání a univerzálního modelování [12]. Výzkum v oblasti znalostních a informačních robotů je prováděn na Fakultě informatiky na Masarykově univerzitě v Laboratoři znalostních a informačních robotů [4], konkrétní implementace se děje v rámci spin-offu Masarykovy univerzity, firmy Mycroft Mind, a.s.

V rámci vývoje se soustředíme zejména na následující technologické prvky [2]:

- vizualizační engine – kombinuje různé *vizualizační metody* v závislosti na povaze dat a preferencích uživatele
- sémantická paměť – umožňuje pracovat s *kontextově závislou*, *neurčitou* a *pozornostně ohodnocenou informací*
- pattern engine – umožňuje pracovat se vzory struktur a chování

- integrační engine – umožňuje *přiblížení dat* z místa jejich vzniku či původního uložení k následnému zpracování
- organizační platforma – propojuje předešlé technologie a umožňuje *škálovatelnost výpočetního výkonu* systému pomocí distribuovatelnosti



Obrázek 2: Ilustrace architektury technologie znalostních a informačních robotů

Schéma užitečnosti

Užitečnost aplikace postavené na výše uvedených technologiích spatřujeme v následujících čtyřech fázích [2]:

1. ukaž
   - zprostředkuje vhled a interaktivní procházení daty z různých pohledů
   - umožňuje interakci pomocí dvou základních pokynů: *obšírněji*, *stručněji*
2. poznej
   - rozeznává a upozorňuje na definované vzory struktur v medicínských datech
3. poraď

- navrhuje opatření pro případ, kdy se daný vzor objeví

4. udělej
   - pomáhá realizovat navržená opatření (např. podpora administrativní práce generováním reportů a jejich odesílání e-mailem)

Důležitým aspektem uvedeného schématu je fakt, že průchod fázemi je sice postupný (od 1. ke 4.) avšak nikoliv striktně lineární, nýbrž cyklický. K cyklení zpravidla dochází např. mezi fázemi 1 a 2, kdy na základě toho, co uživatel vidí (fáze 1), sám rozpoznává nějaký opakující se vzor. Pokud se tento vzor naučí rozpoznávat a vizualizovat i aplikace (fáze 2), je vcelku přirozené a dá se očekávat, že nad novou vizualizací se uživatel rozpozná nový vzor (fáze 1), kterému může být aplikace opět naučena (fáze 2), atd.

Aplikace postavené na technologiích Laboratoře znalostních a informačních robotů tak mají inherentně zabudovánu podporu pro svůj kontinuální rozvoj na základě aktuálních zkušeností a potřeb uživatelů.

Aplikační oblasti

Výše představené technologie mají široké spektrum použití. V současné době se soustřeďujeme zejména na dvě konkrétní oblasti jejich užití: oblast bezpečnosti počítačových sítí a oblast digitálních knihoven (tato bude stručně představena dále). Začínáme též výzkum v oblasti jejich použití pro informace získané z inteligentních senzorů, dále pak pro project management a pro analýzu zdravotnických dat.

Konkrétní použití vyvíjených technologií v oblasti digitálních knihoven

Následující části příspěvku představují směr, kterým postupuje řešení vyvíjené v rámci projektu digitálních knihoven na MU. Na tomto řešení pracujeme na Ústavu výpočetní techniky MU [3] spolu s Laboratoří znalostních a informačních robotů na Fakultě informatiky MU [4].

Naším cílem je vyvinout řešení, které poskytne – metaforicky řečeno – dalekohled do prostoru digitálních knihoven. Výsledná aplikace podpoří všechny části procesu získávání informací, které byly demonstrovány na výše uvedeném motivačním příkladu. Toto řešení primárně vychází z potřeb Masarykovy univerzity.

Technologicky je řešení založeno na použití servisních systémů [5], a to konkrétně na technologii znalostních a informačních robotů [6] [7]. Tato technologie je připravena podpořit proces získávání informací z následujících důvodů [1]:

- umí extrahovat data z heterogenních datových zdrojů: má připraveny adaptéry pro připojení k různým druhům přístupu k datovým zdrojům a poskytuje možnosti tyto adaptéry dynamicky přidávat a kombinovat
- umí provádět inteligentní informační syntézu: dávat dohromady různé objekty a vytvářet mezi nimi nové souvislosti na základě společných vlastností nebo na základě definovaných pravidel
- umí pracovat s neurčitou a kontextově závislou informací: díky tomu je například možné pracovat s informacemi v kontextech jednotlivých datových zdrojů a následně je možné definovat, kterému zdroji věříme více a kterému méně, a případně pracovat i s informacemi, které si navzájem odporují
- použití metod „pattern matching“: vyhledávání a klasifikace informací dle předem daných vzorů a pravidel
- pokročilé vizualizační metody: nejenom klasické lineární seznamy, ale též vizualizace pomocí dynamických myšlenkových map, které jsou vhodné pro přehlednou vizualizaci objektů, především však souvislostí mezi nimi

Kombinací těchto metod lze uživateli umožnit soustředit se na důležité věci a odfiltrovat věci nedůležité.

Současný stav vývoje

Je jasné, že dosáhnout vytyčeného cíle není věc jednoduchá, a nelze se domnívat, že vše bude vyřešeno rychle a bez problémů. Náš přístup se snaží budovat aplikaci postupně tak, že rozšiřujeme spektrum datových zdrojů z prostoru digitálních knihoven, z nichž jsme schopni data extrahovat. Též se postupně budou rozvíjet možnosti informační syntézy – od jednoduché syntézy na základě jasných identifikátorů po pokročilejší metody založené na hledání společných vlastností, souvislostí a pravidel. Navíc se budou rozvíjet způsoby interakce uživatele se systémem, a to zejména na základě uživatelských potřeb a požadavků.

V současné době se nejvíce soustředíme na práci s metadaty publikací a článků, tj. hlavně s bibliografickými záznamy. Obsah plných textů tedy neprohledáváme. První konkrétní výsledky se očekávají v průběhu roku 2008.

Další témata zájmu v bližší i vzdálenější budoucnosti

Mezi další oblasti, kterými bychom se chtěli zabývat a o kterých víme, že jsou pro uživatele zajímavé, patří oblast osobních knihoven. Problematiku opět nejlépe představí příklad [1]:

Ve své práci přečtu velké množství článků a chci v nich mít pořádek. Chci si k přečteným článkům zapisovat poznámky, abych věděl, který článek jsem četl a který mi připadá dobrý. Též bych chtěl mít možnost svoje poznámky ke článkům sdílet s ostatními kolegy. Často též vyhledávám podobné věci. Chci mít možnost dotazy ukládat, vytvářet si tak knihovnu dotazů a následně dotazy z knihovny jednoduše vyvolávat.

Druhou významnou oblastí je monitorování prostoru digitálních knihoven a upozorňování uživatele, že se v něm objevilo něco, co by ho mohlo zajímat. Opět uvedeme ilustrující příklad:

Jsem neurolog, mám úzce vyhraněný obor a zajímají mě relevantní novinky a články z tohoto oboru, ale nemám čas procházet všechny příslušné časopisy. Chci, abych byl informován o nových článcích a knihách, které mne zajímají.

Závěr

V příspěvku bylo ukázáno několik motivačních příkladů z oblasti práce v digitálních knihovnách v medicíně a na základě příkladů byl ilustrován proces získávání informací z prostoru digitálních knihoven. Následně bylo poukázáno na problémy a omezení, se kterými se uživatelé digitálních knihoven setkávají a bylo zdůvodněno, proč čelí informačnímu nedostatku nebo naopak informačnímu zahlcení.

V druhé části příspěvku byly představeny obecné principy technologie znalostních a informačních robotů, a to zejména architektura založená na servisních systémech a její jednotlivé prvky. Bylo nastíněno schéma užitečnosti aplikací založených na zmíněných technologií a stručně popsány oblasti, ve kterých jsou aplikace vyvíjeny.

Třetí část příspěvku byla věnována konkrétní aplikaci v oblasti digitálních knihoven. Bylo představeno vyvíjené řešení, které si dává za cíl práci s digitálními knihovnami zjednodušit, zefektivnit a získat z prostoru digitálních knihoven další informace a souvislosti, které současnými nástroji lze získat jen velice obtížně. Ke konci příspěvek naznačuje další směry rozvoje v oblasti knihoven.

Literatura

[1] Šmerda, J., Dosoudil, V., Staníček, Z., Procházka, F., *Digitální knihovny ve zdravotnictví: Jak léčit informační zahlcení pomocí moderních metod umělé inteligence,* In *MEFANET 2007 (bude publikováno).*

[2] Oškera, M., Procházka, F., Staníček, Z., *Využití sémantické paměti a pokročilých metod pattern matchingu pro analýzu medicínských dat ,* In *MEFANET 2007 (bude publikováno).*

[3] *Masarykova univerzita – Ústav výpočetní techniky.* [Online]. 2007. Available: http://www.ics.muni.cz/

[4] *Laboratoř znalostních a informačních robotů.* [Online]. 2007. Available: http://kirlab.fi.muni.cz/

[5] Spohrer J. Et al., "Steps Toward a Science of Service Systems", *Computer*, vol. 40, no. 1, pp. 71–77, 2007. ISSN 0018-9162.

[6] Dosoudil V., *Získávání informací v heterogenním prostředí digitálních knihoven pomocí technologie znalostních a informačních robotů*, Masarykova univerzita, Fakulta informatiky. Diplomová práce, 2008.

[7] Procházka F., *Universal Information Robots : a way to the effective utilisation of cyberspace*, Masaryk University, Faculty of Informatics. PhD thesis, 2006.

[8] Maglio, Paul P., et al. Service systems, service scientists, SSME, and innovation. Communications of the ACM. 2006, vol. 49, no. 7, s. 81-85. Dostupný z WWW: <http://doi.acm.org/10.1145/1139922.1139955>. ISSN 0001-0782.

[9] IBM. IBM Research : Services Sciences, Management and Engineering [online]. 2004 [cit. 2007-12-30]. Dostupný z WWW: <http://www.research.ibm.com/ssme/>.

[10] Maes, Pattie. Agents that reduce work and information overload. Communications of the ACM. 1994, vol. 37, no. 7, s. 30-40. Dostupný z WWW: <http://doi.acm.org/10.1145/176789.176792>. ISSN 0001-0782.

[11] Staníček, Z.: Doprovodné slidy k přednáškám datového modelování II. 2006, fakulta informatiky, Masarykova univerzita, Brno. Dostupné z: <http://www.fi.muni.cz/~stanicek/PA116/DM_IInew2006_L01.ppt>.

[12] Staníček, Z.: Universální modelování a konstrukce IS. [s.l.], 2003. iii, 159 s. Masarykova univerzita. Fakulta informatiky. Dizertační práce.