NTK
50°6'14.083"N, 14°23'26.365"E
Národní technická knihovna
National Technical Library

Odborný článek

Co se skrývá za vyhledáváním aneb Searching Session NTK 2011

Mgr. Kristýna Kožuchová

297 mm

říjen 2011

S = 62 370 mm² 210 mm A4

4. října 2011 proběhl v Národní technické knihovně 2. ročník semináře "Co se skrývá za vyhledáváním aneb Searching Session NTK 2011", který byl realizován v rámci projektu *Moderní informační a komunikační technologie v knihovnictví 2011* s podporou programu Ministerstva kultury *Veřejné informační služby knihoven,* konkrétně podprogramu *Mimoškolní vzdělávání knihovníků*. Seminář byl určen zejména pracovníkům knihoven, informačním specialistům, ale i odborníkům z dalších oblastí. Účastníci semináře měli možnost shlédnout celkem sedm přednášek rozdělených do dvou bloků – dopoledního a odpoledního.

Po krátkém uvítání účastníků semináře se slova v dopoledním bloku ujal Filip Hráček ze společnosti Google, který hovořil o sémantickém webu a Google vyhledávání. Přednášku uvedl větou: *„Everything is OUR problem"* (všechno je náš problém). Toto tvrzení reprezentovalo rozhodnutí společnosti Google začít se svým uživatelům přizpůsobovat. Podle Hráčka nevýhoda a zároveň výhoda Google spočívala v tom, že jej začali využívat netechničtí uživatelé, kteří nevěděli, jakým způsobem správně vyhledávat, správně pokládat dotazy, zejména s ohledem na kontext fulltextového vyhledávání. Tuto situaci Google řešil a řeší tak, že se snaží maximálně tolerovat chyby svých uživatelů při vyhledávání. V této souvislosti Hráček postupně nastolil několik "problémů" (na příklad – *uživatel nezná rozdíl mezi search bar a address bar* nebo *uživatel píše moc pomalu*), u kterých zároveň předvedl způsoby, jak se s nimi Google vyrovnal. Google ve svém vyhledávání dodává tzv. *rich snippets*, tedy "něco navíc". Může se jednat o obrázky receptů nebo krátké informace, resp. přímé odpovědi na dotazy položené vyhledávači. Na příklad zadá-li uživatel do vyhledávacího pole číslo letu, vyhledávání Google mu nabídne okamžitou informaci o tomto letu, aniž by byl uživatel nucen otevřít některý z odkazů. Podobně vyhledávání pracuje s informacemi o počasí či kurzech měn. Klíčem k tomu, aby vyhledávač rozuměl obsahu, je sémantický web, ke kterému přednášející definoval tři přístupy: 1. general data extraction – náročný proces, který lze využít jen pro malé procento informací, 2. structured data markup – užití mikrodat, mikroformátů, RDFa apod., 3. feeds – používají se přímo pro stroje, ne pro člověka např. RSS, jejich využití je vhodné pro často měnící se data. Sémantický web je kromě formátů spojen také s otázkou definování slovníku umožňujícího sémanticky popsat různé druhy entit. Příkladem takového slovníku je schema.org, který vedle společnosti Google podporují také Bing a Yahoo!.

Anti-SEO ve veřejné správě – stránky, které nechtějí být nalezeny byla druhá přednáška dopoledního bloku. Její autor, Jiří Skuhrovec, mimo jiné spolutvůrce projektu zIndex.cz, v rámci kterého jsou hodnoceni zadavatelé veřejných zakázek, se v ní věnoval problematice zveřejňování informací o veřejných zakázkách. Přednáška byla zajímavým pohledem do praxe orgánů veřejné správy, které vystupují v roli zadavatelů veřejných zakázek. Jedním z hlavních témat, které autor přednášky v této souvislosti otevřel, byl Informační systém o veřejných zakázkách. Předmětem kritiky tohoto systému bylo především špatně nastavené vyhledávání a data, která sice jsou do jisté míry strukturovaná, ale podle Skuhrovce je nelze dohledat. Jako příklad uvedl problém při vyhledávání zadavatelů podle evidenčního čísla nebo názvu instituce. Autor přednášky rovněž konstatoval, že problém v zadávání veřejných zakázek spočívá už v jejich samotném popisu. V mnoha případech jsou totiž zakázky pojmenovány a popisovány takovým způsobem, že je těžké uhodnout, co zadavatel vlastně požaduje. Příkladem mohu být zakázky definované jako *koncepce profilu aglomerace* nebo *ZMR 32* či *informační kriminalita v internetu*, pod jejímž názvem se skrýval nákup softwaru, ačkoli se slova *software* či *program* v popisu zakázky vůbec nevyskytovala. Skuhrovec upozornil také na veřejné zakázky malého rozsahu, které sice jsou "někde" na webových stránkách dané organizace umístěny, ale nevedou na ně žádné odkazy. Jiným špatným příkladem jsou zakázky, které jsou na webových stránkách vystaveny, ale pouze několik dní a

nikdo tak nemůže mít přehled o tom, jaké zakázky již proběhly, jaké probíhají, popř. jaké se chystají. Další ukázkou "preciznosti" některých úředníků jsou zakázky, ke kterým se lze dostat až po vyplnění formuláře opatřeného povinností přihlásit se. V mnoha případech navíc tyto formuláře vyžadují nové přihlašovací údaje pro každou zakázku. Taková praxe odrazuje nejen potenciální dodavatele, ale brání vyhledávačům takový obsah vůbec nalézt. Přednášející v současné době nevidí řešení této situace a neprůhledných praktik pouze v soustavném novelizování zákonů, ale především v důsledném postihování jasných případů porušování zákona.

Podobnostnímu hledání v netextových datech se věnoval Pavel Zezula z Fakulty informatiky Masarykovy univerzity. V úvodu konstatoval, že zatímco data byla dříve pouze v číselné podobě, dnes může být v digitální podobě téměř vše. Datová základna se přibližuje tomu, jak přemýšlí člověk. A protože člověk uvažuje na základě podobnosti, měl by počítač umět to samé. Je to totiž právě podobnost, která umožňuje lidem rozeznávat, vyhledávat, třídit apod. V této souvislosti seznámil Zezula účastníky semináře s projektem MUFIN (**MU**lti-**F**eature **I**ndexing **N**etwork), který představuje unikátní systém umožňující vyhledávání v multimediálních datech. Zatímco běžně dostupné vyhledávací služby vyhledávají obrázky podle jejich textového popisku, MUFIN vyhledá obrázky na základě jejich podobnosti, tedy na základě jejich skutečného obsahu. Princip této metody spočívá v extrakci určitých vlastností z obrázku. K těmto vlastnostem patří na příklad barevné spektrum, textura nebo rozložení tvarů. V rámci analýzy těchto vlastností určí program pro každý obrázek bod ve vícedimenzionálním prostoru a porovnává jeho vzdálenost s ostatními. Čím blíže jsou tyto body, tím podobnější si budou i k nim přiřazené obrázky.

Ve čtvrté přednášce se Jindřich Mynarz z Národní technické knihovny věnoval otázce, jak automaticky přiřadit dokumentům hesla Polytematického strukturovaného hesláře (PSH). Účastníky nejprve provedl procesem indexace, který charakterizoval jako sumarizaci dokumentu na zástupné znaky, které daný dokument nejvíce vystihují, a které lze lépe a rychleji vyhodnocovat na příklad z hlediska relevance. Podle slov přednášejícího indexace rovněž usnadňuje precizní vyhledávání a prohlížení dokumentů. Specifickým příkladem indexace je automatická indexace. Ta čerpá z analýzy plného textu dokumentu, analýzy použitého řízeného slovníku a analýzy způsobu použití daného slovníku nad korpusem dokumentů. Komponentami automatické indexace jsou tak vlastní indexátor, korpus plných textů a řízený slovník. V Národní technické knihovně (NTK) je systém automatické indexace postavený na indexátoru Maui Indexer s hesly Polytematického strukturovaného hesláře a je nasazený v Národním úložišti šedé literatury. Součástí procesu automatické indexace je také odstranění nevýznamových slov, jejichž nejčastější frekvence byla v NTK založena na Českém národním korpusu FF UK. Po procesu vygenerování kandidátů PSH dochází ještě k jejich dalšímu filtrování. Dle výzkumu Oleny Medelyan, tvůrkyně zmiňovaného indexátoru Maui-Indexer, jsou výsledky automatické indexace podobné průměrným výsledkům intelektuální indexace.

V odpoledním bloku vystoupil Lukáš Slánský z Univerzity Pardubice, který představil nově spuštěnou aplikaci pro indexaci a vyhledávání osob podle vědecko-výzkumného zaměření. Výchozím bodem nového aplikace byl původní systém pro vyhledávání zaměstnanců univerzity. Postupně však vznikla potřeba efektivně prezentovat vědecko-výzkumnou činnost akademických pracovníků univerzity. Požadavky na novou aplikaci se skládaly z potřeby vyhledávání standardně podle příjmení, pracoviště, klíčových slov publikací a vědecko-

výzkumného zaměření. Po zvážení indexace vědecko-výzkumné činnosti podle Mezinárodního desetinného třídění a možnosti tvorby vlastních klíčových slov došli na Univerzitě Pardubice k rozhodnutí využívat Polytematický strukturovaný heslář (PSH). Důvody výběru PSH byly do určité míry dány vlastnostmi PSH jako je hierarchičnost, polytematičnost, dvojjazyčnost a dostatečná podrobnost (ani příliš podrobný, ani příliš obecný – podle slov přednášejícího: „*zkrátka tak akorát*"). Indexaci vědecko-výzkumné činnosti provádí sami zaměstnanci prostřednictvím editace vlastního profilu na webových stránkách univerzity. V rámci editace si mohou zaměstnanci vybírat příslušná hesla reprezentující jejich vědecko-výzkumné zaměření. U vybraných hesel pak mohou stanovovat pořadí preferencí. Nová aplikace, v současnosti přístupná pouze v rámci intranetu univerzity, tak umožňuje vyhledávání jednotlivých akademických pracovníků podle vědecko-výzkumného zaměření.

Semináře se i v letošním roce zúčastnil Josef Šlerka ze Studií nových médií Filozofické fakulty Univerzity Karlovy, který po úvodních tvrzeních, že: „*Počítač bez internetu je pro účetní*" a „*Internet bez her je taková knihovna*", hovořil o gamifikaci webu. Přednášející nejprve pojem gamifikace (z anglického *gamification*) představil a zároveň naznačil, v jaké fázi zájmu se nachází. Na obrázku (Gartner-hypecycle-2011) je zobrazen graf, který každoročně (od roku 1995) vytváří společnost Gartner. Tento specifický graf sleduje perspektivy a vývoj různých technologií, jejich zralost a společenské využití s ohledem na čas a očekávání. Z grafu lze vysledovat, že gamifikace se nachází téměř před pomyslným vrcholem zájmu. Pro ty, kteří neholdují grafům postačí Šlerkovo tvrzení, které říká: „*Gamifikace je chvilku před tím, než o ní začne psát víkendová MF Dnes a stane se předmětem divokých úvah intelektuálů.*" Ačkoli gamifikace představuje neostré pole aktivit a může mít mnoho podob, vždy využívá herních prvků a mechanismů v neherním prostředí a v neherních aplikacích. K prvkům herních mechanismů patří na příklad sbírání bodů (points), získávání odznáčků (badges), systémy úrovní (levels), různé ukazatele postavení hráčů (leaderboards), podpora soutěživosti s ostatními hráči (challenges) apod. Příklady zavedení různých herních prvků a mechanismů lze nalézt na příklad v aplikacích Foursquare, Mint a Healthmonth. Šlerka dále představil typologii hráčů, do které patří tzv. achievers, killers, explorers, socializers. I když má daná typologie zůstat podle autora přednášky bez překladu, na otázku jak by publikum přeložilo skupinu *achievers,* odpověděl jeden z přítomných: „*Achievers? To jsou ti šplhouni.*". První skupina, tedy achievers (nebo také šplhouni), má ráda výzvy. Hráči této skupiny rádi sbírají body, plní úkoly a rádi ukazují své výsledky. Hráčům patřícím do skupiny socializers jde především o sociální aspekt hry, o kontakt a seznámení se s ostatními hráči. Zatímco explorers rádi objevují nové oblasti nebo píší recenze, hráči killers milují výhru, pro kterou udělají cokoli, i kdyby to "cokoli" mělo znamenat dosáhnout svého cíle hackováním a podváděním.

Na závěr semináře vystoupil Štěpán Bechynský ze společnosti Microsoft, který představil HTML5. Ten označil za značkovací jazyk a zároveň ucelené prostředí pro programování aplikací. HTML5 je vyvíjen v rámci konsorcia W3C a je ve stádiu "last call", což podle Bechynského představuje poslední možnost pro připomínkování, zároveň však reprezentuje fakt, že HTML5 je téměř hotový a připravený k užití. Základní myšlenka HTML5 spočívá v návratu k počátkům HTML, které bylo vyvinuto v laboratořích CERN. Důraz je kladen na oddělení prezentační a obsahové roviny. V HTML5 jsou definovány nové sémantické tagy jako jsou na příklad *section*, *nav*, *article*, *aside*, *footer*, *figure*. Stejně jako Filip Hráček, také Bechynský neopomněl zmínit slovník schema.org, který lze používat k rozšíření sémantického popisu dat. Přednášející rovněž doporučil zájemcům o problematiku HTML5 publikaci Marka Pilgrima Dive Into HTML5, jejíž komunitní český překlad vzniká na webových

stránkách HTML5.cz, kde jsou k dispozici také další odkazy na české, ale i zahraniční zdroje zabývající se touto tematikou.

Jednotlivé prezentace přednášejících si lze prohlédnout na stránkách NTK. O seminář byl v letošním roce velký zájem a dle vyhodnocených anonymních dotazníků účastníků se jednoznačně vydařil. Podle reakcí přítomných byla celá akce inspirativní, zajímavá, plná nových a atraktivních témat a informací. Věřím, že i třetí ročník semináře *Co se skrývá za vyhledáváním aneb Searching Session NTK 2012*, který se bude konat 2. října 2012, vzbudí v účastnících podobné reakce.