MASARYKOVA UNIVERZITA
FAKULTA INFORMATIKY

CiteSeer - Scientific Literature DL

esej do předmětu Digitální knihovny

Autor: Pavel Župa Podzim 2006

1 Úvod

Výzkum provází člověka už od pradávna. První výsledky bádání byly předávány ústně nebo formou rukopisů, které byly ručně opisovány a šířeny těm několika podivínům, co se zkoumání věnovali. Později bylo možné díky mecenášům některá význačná vědecká díla vydat knižně. V minulém století se stalo běžnou praxí vydávání odborných časopisů, jejichž redaktoři pečlivě vybírají ověřené vědecké novinky, kterým je pak dovoleno objevit se na jejich stránkách.

V současnosti se výzkumem v rozličných odvětvích zabývá ohromné množství lidí, ať již na akademické půdě, pod záštitou velkých společností či dokonce jako jednotlivci. Už od počátků je hlavním problémem výzkumu přístup vědců k potřebným informacím. Dochází tak ke zpomalování výzkumu a zbytečnému mrhání duševním potenciálem při „vymýšlení již vymyšleného“. Telefon je toho zářným příkladem. Ideálním stavem by bylo, kdyby každý mohl okamžitě získat výsledky veškerého výzkumu, který se týká jeho vlastního projektu.

1.1 Internet

Rozmach internetu přinesl každému možnost publikovat své poznatky a zpřístupnit je tak širšímu publiku, než bylo kdykoliv dříve představitelné. Projevil se tím ale problém jak se k relevantním informacím dostat. Na internetu je velké množství vědeckých publikací, ale jsou naprosto neorganizované a vědci, kteří se snaží najít potřebné informace, se musí probírat obrovským množstvím irelevantních dokumentů. Problémem tedy už není nedostatek informací, ale naopak jejich nezpracovatelné množství. *„Někde to tam určitě je, ale jak to jen najít?“*

1.2 Indexování citací

Seznam použité literatury je samozřejmou částí každého odborného textu. Je to věc užitečná, jen co je pravda, ale ještě přínosnější by byla možnost pohybovat se v odkazech opačným směrem, to znamená vyhledat dokumenty, které aktuální text nějakým způsobem rozšiřují. Toho je možné dosáhnout pomocí tzv. indexace citací (citation indexing), tedy zpětného propojení citovaného dokumentu s dokumenty, které se na něj odkazují. Tento princip umožňuje zlepšit přístup k informacím tím, že upozorní na důležité opravy a kritiky daného dokumentu, rozšíření samotným autorem i pokračování výzkumu ostatními a pomáhá omezit zbytečné zkoumání tam, kde již byly publikovány výsledky.

Většinou je k tomu ale potřeba úsilí mnoha lidí, kteří k daným publikacím vloží do databáze jednotlivé položky seznamu použité literatury a propojí je s originálními dokumenty. Objevil se například návrh na univerzální databázi citací a bibliografických odkazů [3], která by obsahovala odkazy na veškerou existující vědeckou literaturu a umožňovala vyhledávání a snadný přístup komukoli s přístupem na internet. Za vložení odpovídajících bibliografických záznamů o dokumentech by zodpovídali autoři, případně publikující organizace. Těžko si ale představit, že by každý autor byl natolik zodpovědný, nemluvě o tom, jak nesnadné (ne-li neproveditelné) by bylo shodnout se celosvětově na užívání jednoho standardu. Většina existujících digitálních knihoven, které ručně indexují citace, se zaměřuje na poměrně úzký obor článků, například mapují výběr odborných časopisů. Tudy cesta nevede – alespoň v globálním měřítku ne.

A zde právě vstupuje do hry CiteSeer.

2 CiteSeer

CiteSeer[1] je digitální knihovna vědecké literatury, zaměřující se především na literaturu o počítačových technologiích a informatice. Systém byl vyvinut v roce 1997 trojicí vědců Steve Lawrence, Lee Giles a Kurt Bollacker v NEC Research Institute.

Co ale odlišuje tuto knihovnu od dlouhé řady jiných, je použití Autonomous Citation Indexing, tedy automatického indexování citací. Knihovna se totiž nesnaží být knihovnou v pravém slova smyslu, její hlavní přínos je v tom, že prohledává web a nalezené vědecké články analyzuje a ukládá informace o nich do své databáze. Dokumenty je pak možné vyhledávat, procházet mezi jejich bibliografickými odkazy, sledovat statistiky citací a podobně.

Po několika letech úspěšného provozu získal CiteSeer postupně financování od National Science Foundation, Microsoft Research a NASA, což umožnilo další vývoj systému.

V současné době je CiteSeer hostován na Pennsylvanské státní univerzitě[2] a zrcadlen na Kansaské univerzitě[3], Massachusettském ústavu technologií[4], Zürichské univerzitě[5] a národní univerzitě v Singapuru[6]. Celkový počet indexovaných dokumentů se blíží 770 tisícům.

1 http://citeseer.ist.psu.edu/
2 Pennsylvania State University's College of Information Sciences and Technology
3 The University Of Kansas
4 Massachusetts Institute Of Technology
5 University Of Zurich
6 National University Of Singapore

2.1 Autonomous citation indexing

Jak již bylo zmíněno, srdcem systému je automatické indexování bibliografických údajů článků. CiteSeer používá prohledávání internetu a crawling (procházení odkazů) pro nalezení nových dokumentů a je také možné upozornit systém na nový článek ručně – tím dojde prakticky okamžitě k jeho zpřístupnění přes CiteSeer.

Každý nalezený dokument je z původního formátu (PostScript, PDF, html) převeden do textové podoby a následně je podroben důkladné analýze. Prvním krokem je nalezení sekce s bibliografickými odkazy, což se děje buď podle nadpisu sekce nebo rozpoznáním samotných položek odkazů. V každém odkazu na použitou literaturu je pak nutné vyhledat jednotlivá pole jako jsou identifikátor odkazu, jméno autora, název článku či knihy, rok vydání, u článků z časopisů pak název periodika, ročník a číslo, rozsah stránek a podobně. Není to jednoduchý úkol, neboť kromě existence různých typů odkazů (kniha, článek v časopise, webová stránka, atd.) může každý autor používat jinou syntaxi a navíc zde zde často vyskytují i překlepy či faktické chyby. Kupříkladu čárka se používá k oddělení polí i jednotlivých autorů a může se vyskytnout i v názvu publikace. Tečka může být použita k oddělení jednotlivých polí ale i u zkratek a iniciál autorů. Někdy dokonce nejsou pole oddělena vůbec.

K rozeznání jednotlivých polí využívá CiteSeer metod umělé inteligence založených například na pravděpodobném pořadí položek (jména autorů bývají před názvem díla), Levenshteinově vzdálenosti slov a statistikách výskytu určitých slov nebo čísel v určitých polích. V současné době se podařilo vyladit použité algoritmy tak, že pouze v pěti procentech citací se objeví v některém poli chyba.

Nakonec jsou v textu ještě vyhledána místa, která se na danou položku odkazují. To pak uživateli umožňuje pro vybraný dokument zobrazit nejen seznam citujících článků, ale i kontext v jakém je citace použita.

2.2 Metadata a fulltext

U každého dokumentu CiteSeer vyhledává a ukládá i metadata a pokud je to možné přidává odkaz na metadata z jiných zdrojů, jako například DBLP (Digital Bibliography & Library Project) nebo ACM Digital Library. Kromě zpracování citací je kompletní text každého dokumentu i fulltextově oindexován a je tedy možné vyhledávat články podle obsahu. CiteSeer navíc kombinuje fulltextové vyhledávání s metadaty a citacemi, což vylepšuje hledání například o možnost zadat iniciály autora a podobně.

CiteSeer také podporuje standard pro sklízení metadat Open Archives Initiative[7], což umožňuje jeho pohodlné propojení s jinými knihovnami.

2.3 Acknowledgement indexing

V současné době probíhá zkušební provoz další velmi zajímavé funkce, kterou je automatická indexace poděkování (acknowledgements). Jelikož seznamy použité literatury se staly samozřejmou částí vědeckých textů, jsou často užívány k poměřování vědeckého přínosu prací. Poděkování však mohou být obdobným měřítkem, možná dokonce lepším. U seznamu použité literatury se očekává, že zde autor uvede všechny zdroje, ze kterých čerpal. Naproti tomu poděkování je více osobní a pokud zde autor někoho zmíní, pak se dá očekávat, že pro něj byl prospěšný opravdu významnou měrou.

Na rozdíl od seznamu použité literatury, který má alespoň nějakou formální strukturu, je poděkování psáno civilním jazykem. Přesto se vývojářům CiteSeeru podařilo implementovat mechanismy, které jsou v dostatečné míře schopné poděkování porozumět a získat z něj požadované údaje.

Nejdříve je nutné identifikovat části textu, kde se pravděpodobně poděkování objeví. To je většinou část pod nadpisem „Acknowledgements“ a dále se poděkování hledají na první straně po titulu a poslední před seznamem citací nebo přílohami. V rámci tohoto textu jsou s použitím algoritmů Support Vector Machines (SVM) identifikována jednotlivá poděkování, SVM jsou poměrně složité algoritmy umělé inteligence, které jsou obecně schopné kategorizace dat. Dalším problémem je, že se v textu může objevit například název společnosti a později jen zkratka a je žádoucí, aby oba tyto případy byly rozpoznány jen jako jeden objekt. Nalezená poděkování je se dělí do šesti základních významových skupin:

- morální podpora
- finanční podpora
- vydavatelská podpora
- presentační podpora
- technická podpora (poskytnutí technického vybavení)
- koncepční podpora, nebo také vzájemná podpora vědců

7 http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm

Přičemž poslední zmíněná se zdá být nejvýznamnější při poměřování vědeckého přínosu autorů. Finanční podpora zase umožňuje vyhodnotit například přínos jednotlivých organizací pro určitá odvětví výzkumu.

Postupy indexace poděkování se podařilo vyladit tak, že CiteSeer v aktuální beta verzi dosahuje úspěšnosti kolem 80 procent při identifikování jednotlivých odkazů na lidi či instituce kterým autoři děkují.

2.4 Uživatelský pohled

Jak již bylo zmíněno, CiteSeer umožňuje hledat dokumenty podle názvu, autorů i fulltextově, o každém dokumentu nabízí přehledně veškeré dostupné informace na samostatné stránce. V první části jsou zde uvedeny základní informace jako název práce, autoři, rok publikování, prvních několik řádků abstraktu. Dále odkaz na samotný dokument včetně jeho verze uložené CiteSeerem při indexování a to hned v několika verzích (např. pro PostScript je zde i PDF verze a obrázky jednotlivých stran ve formátu PNG). Je zde i odkaz na původní stránku, která odkazuje na daný dokument, a dokonce někdy i odkaz na domácí stránku autora. Tam, kde je to možné jsou odkazy na stejný dokument v jiných knihovnách.

Následuje seznam dokumentů, které tento článek citují a naopak zdrojů citovaných aktuálním dokumentem. Je zde i seznam podobných dokumentů včetně percentuální míry podobnosti, dále seznam příbuzných dokumentů na základě toho, že citují stejné zdroje a nakonec i seznam dokumentů, které pochází ze stejné webové stránky. Jsou zde zobrazena metadata ve formátu BibTeX a seznam takových formátů citací, které uvedeny v seznamu použité literatury, budou při indexování spolehlivě rozeznány jako odkaz na aktuální dokument. Nakonec je zde zobrazen ještě graf, ukazující počet dokumentů, které citovaly aktuální článek v jednotlivých letech.

K veškerým dokumentům, které se zde objevují ať již jako citované, citující nebo podobné jsou přidány odkazy a je tedy možné procházet stromem (přesněji asi grafem) citací dokumentů. U každého údaje, který vznikl pomocí automatického indexování je také odkaz na formulář k opravení případných chyb údajů (název, rok vydání, autor, ...). Veškeré takto opravené údaje jsou samozřejmě před zveřejněním kontrolovány.

Při vyhledávání dokumentů je možné zvolit hledání pomocí CiteSeer prohledávače, případně je automaticky nabízeno i hledání pomocí dalších běžných vyhledávačů jako Google (s možností

prohledat CiteSeer nebo web), Yahoo!, případně v knihovně DBLP[8] nebo v CSB[9]. Při běžném hledání je možné omezit hledání pouze na název článku a řadit výsledky podle počtu citací, frekvence použití nebo data vytvoření dokumentu.

3 Závěr

Co dodat? CiteSeer je velmi zajímavým projektem, který první přišel s myšlenkou automatického zpracování citací a později i poděkování. Na jeho architektuře je postavena i další knihovna Pennsylvánské univerzity SMEALSearch[10], která se zaměřuje na vědeckou literaturu v oblasti ekonomie a obchodu.

Autoři tohoto projektu si dali za cíl zlepšit přístup k akademické a vědecké literatuře. Přístup do knihovny může využívat kdokoliv a CiteSeer je proto považován za přínos hnutí za otevřený přístup, které snaží zlepšovat přístup k odborné literatuře.

V současnosti již existují i další digitální knihovny, které implementují podobnou funkcionalitu automatického zpracování citací, jako Google Scholar, Scirus a další. Ovšem fakt, že kompletní zdrojový kód CiteSeer je k dispozici zdarma pro nekomerční využití, jej posouvá o stupínek výše. Zázemí silných finančních partnerů, otevřený zdrojový kód a zjevný energický vývoj zajímavých nových technologií věstí CiteSeeru slibnou budoucnost.

8 http://dblp.uni-trier.de/
9 http://liinwww.ira.uka.de/bibliography/
10 http://smealsearch2.psu.edu/index.html/

Použitá literatura

1. Steve Lawrence, C. Lee Giles, and Kurt Bollacker. *Digital libraries and autonomous citation indexing.* IEEE Computer, 32(6):67--71, 1999. http://mack.ittc.ku.edu/article/lawrence99digital.html

2. C.L. Giles & I.G. Councill. *Who gets acknowledged: Measuring scientific contributions through automatic acknowledgement indexing.* PNAS 101(51): 1759917604, 2004. http://citeseer.ist.psu.edu/giles04who.html

3. Robert D. Cameron. A universal citation database as a catalyst for reform in scholarly communication. First Monday, 2(4), 1997. http://www.firstmonday.dk/issues/issue2_4/cameron/

4. *Wikipedia: CiteSeer* [online]. 2005, 17 November 2006 [cit. 2007-01-21]. http://en.wikipedia.org/wiki/Citeseer

5. *CiteSeer: About CiteSeer* [online]. 1997 [cit. 2007-01-21]. http://citeseer.ist.psu.edu/citeseer.html

Metadata


```
<link rel="schema.DC" href="http://purl.org/dc/elements/1.1/" />
<meta name="DC.Title" content="CiteSeer - Scientific Literature DL" />
<meta name="DC.Creator" content="Pavel Župa" />
<meta name="DC.Subject" content="CiteSeer" />
<meta name="DC.Description" content="Esej do předmětu Digitální knihovny, popisující projekt CiteSeer." />
<meta name="DC.Date" content="21.1.2007" />
<meta name="DC.Type" content="Text" />
<meta name="DC.Format" content="application/pdf" />
<meta name="DC.Format" content="computerFile" />
<meta name="DC.Identifier" content="(SCHEME=URL) http://www.fi.muni.cz/~xzupa/pv070/CiteSeer.pdf" />
<meta name="DC.Language" content="cs" />
```