MASARYKOVA UNIVERZITA
FAKULTA INFORMATIKY

CiteSeer (Next Generation)

Scientific Literature Digital Library and Search Engine

PV070 - Digitální knihovny

Bc. Martin Klubus, FI N-IN PSK [sem. 1., roč. 1.] 1. února 2011

Název projektu/programu, jeho nositel, URL:

CiteSeer (Next Generation), Pennsylvánská státní univerzita, http://citeseer.ist.psu.edu/

Stručná charakteristika projektu/programu:

CiteSeer^{x} je digitální knihovna a vyhledávač zaměřený na odbornou literaturu, která se týká především oblasti počítačové a informační vědy. Snaží se zefektivnit její šíření, dostupnost a použitelnost. Nejedná se však pouze o další digitální knihovnu. CiteSeer^{x} poskytuje technologie, algoritmy, metadata a další prostředky pro indexaci a zkoumání článků na webu. Zejména se jedná o Autonomous Citation Indexing (ACI) neboli automatické indexování citací.

Doba řešení, aktuální stav:

Za zrodem projektu CiteSeer stála v roce 1997 trojce vědců z NEC Research Institute, který se nachází v New Jersey. Byly to pánové Steve Lawrence, Lee Giles a Kurt Bollacker. Vytvořili digitální knihovnu se systémem automatického indexování citací. CiteSeer prohledává web a analyzuje nalezené články týkající se počítačové a informační vědy. Do své databáze je ukládá spolu s vytvořenými metadaty tak, že uživatel se k nim může dostat při vyhledávání podle rozličných kritérií. Zejména pokud daný článek někdo odcituje. Projekt je provozován na Pennsylvánské státní univerzitě a několika dalších místech, které slouží jako takzvaná zrcadla.

Cíle projektu:

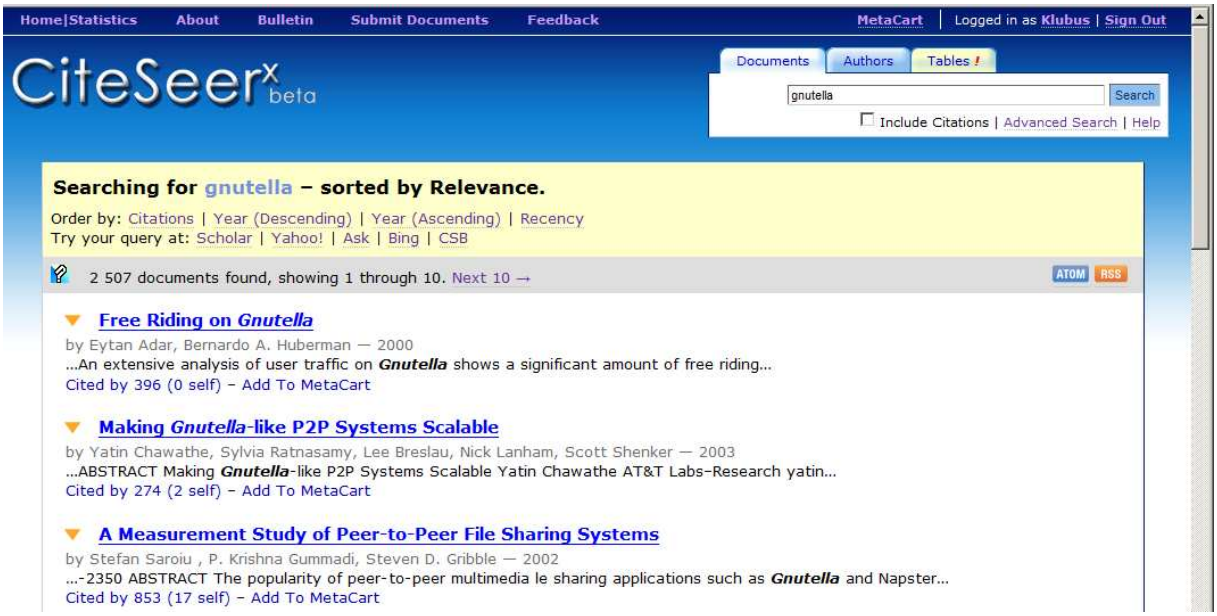
Na internetu je dnes k dispozici nepřeberné množství informací. Pokud má například někdo konkrétní problém s počítačem, může si být jistý, že spoustě lidí se přihodilo to samé a někdo z nich problém popsal a snad i vyřešil. Ale možnost cokoli snadno a v podstatě zdarma zveřejnit má dnes téměř každý. U většiny publikovaných věcí nikdo negarantuje smysluplnost a pravdivost. Informace k hledanému tématu jsou navíc roztroušeny na různých místech a valná část značně duplicitně. Klasické vyhledávače nabízejí spoustu irelevantních dokumentů a ručně indexované databáze zase prohledávají pouze svůj úzce zmapovaný okruh. CiteSeer, potažmo jeho novější verze CiteSeer (Next Generation) označovaná jako CiteSeer^{x}, se snaží uživateli co nejvíce usnadnit hledání dat, které s daným tématem přímo souvisí. Pokud tedy někdo něco vymyslel, popsal a zveřejnil, mělo by to být rozumně dohledatelné. Projekt byl financován několika významnými subjekty jako například National Science Foundation nebo NASA.

Popis projektu a jeho výsledku

Na úvodní stránce projektu je nejdůležitější pole, ze kterého můžeme vyhledávat v dokumentech, případně se přepnout na prohledávání podle jména autora nebo prohledávání v tabulkách. Po zadání hledaného hesla, např. „gnutella“, dostaneme výpis nalezených dokumentů se zadaným heslem. Nalezené dokumenty jsou vypsány ve tvaru název, autor/autoři, část věty obsahující heslo a počet citací, o nichž systém ví, že na daný dokument odkazují. Když na některý název klikneme, zobrazí se další informace. Především krátký abstrakt, odkazy na různé verze původního dokumentu, seznam zdrojů, které jsou v dokumentu citovány a také odkaz na zmiňované články, které na tento dokument odkazují.

AUTONOMOUS CITATION INDEXING

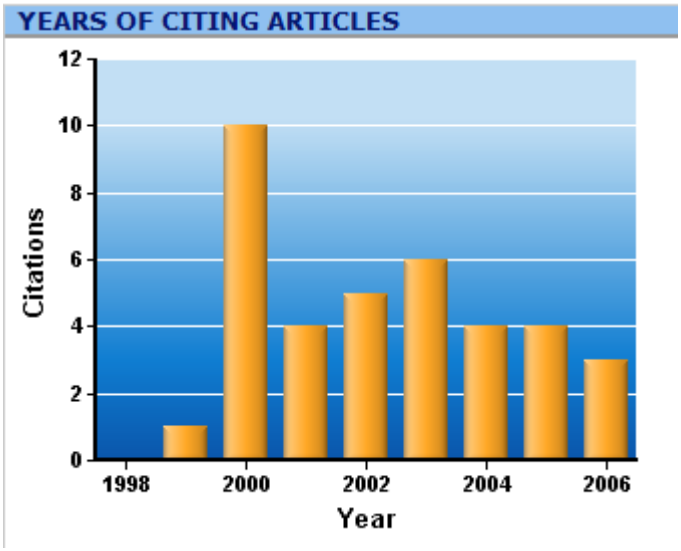
Všechny tyto výstupy jsou generovány automaticky. Systém sám prochází články na internetu a extrahuje z nich a ukládá data. Je také možné zadat konkrétní odkaz ručně. V obou případech je dokument nejprve převeden z původního formátu do textové podoby a poté jsou v něm vyhledány bibliografické údaje jako název, jméno autora, rok vydání a podobně. S velkým úspěchem se přitom využívají různé pravděpodobnostní algoritmy a metody umělé inteligence. K takto získaným metadatům se případně přidá i odkaz na metadata z jiných zdrojů. Dále je text zpracován fulltextově, aby uživatel mohl jednoduše vyhledávat tak, jak je běžně zvyklý třeba z vyhledávače Googlu. CiteSeer^{x} zároveň nabízí vyhledávání pomocí Google Scholar, Yahoo!, Ask, Bing a CSB (The Collection of Computer Science Bibliographies). Při zadání dotazu v CiteSeer^{x} je pak kombinováno jak fulltextové vyhledávání, tak vyhledávání pomocí metadat a citací. Díky tomu může systém nabídnout heslo i s okolním kontextem.



Obr. 1: Výsledky hledání hesla „gnutella“

Práce s vyhledávačem

Výsledky hledání lze sledovat pomocí kanálu RSS případně ATOM, což lze využít zejména v případě, že sledujeme nějaké aktuální téma, o kterém právě vychází nové a nové články. Bibliografické údaje jsou poskytovány ve formátu BibTeX. V internetových zdrojích jsem narazil na zmínky o grafech, které mají ukazovat počet článků, které daný dokument citují v průběhu času. Často se ale žádný obrázek nezobrazí a nepřišel jsem na žádnou souvislost s počtem citací, rokem napsání článku a podobně. Z úvodní stránky jsou dostupné statistiky, které ukazují například nejvíce citované články nebo autory.



Obr. 2: Citování konkrétního článku v jednotlivých letech

Všiml jsem si také, že pod některými odkazy vedoucí na nápovědu se žádné informace nenachází. Na řadu nefungujících odkazů jsem narazil i při vyhledávání dokumentů. Systém sice nabídne u některého dokumentu odkazy na různá místa v internetu nebo verze v různých formátech, ale některé z nich občas nefungují.

Vzhledem k tomu, že metadata jsou generována automaticky, může se v nich občas vyskytnout chyba, případně může některý, z článku vyplývající, údaj chybět. V takovém případě existuje možnost chyby opravit. Po úspěšné opravě systém upozorní, že může nějakou dobu trvat, než se změny projeví ve všech příslušných databázích.

Při vyhledávání a procházení dokumentů lze s úspěchem využít odkaz „Add To MetaCart“, který odkazy na vybrané dokumenty ukládá do dočasné schránky. Můžeme si tedy vyhledat a označit sadu dokumentů a poté jimi procházet skrze tuto schránku. Ale jenom dočasně, než CiteSeer^{x} opustíme.

MyCiteSeer^{x}

Aby si o nás systém něco pamatoval, je potřeba se zaregistrovat a přihlásit do autentizované části MyCiteSeer^{x}. Mělo by to přinést určité výhody při vyhledávání, kdy by měl systém brát v potaz naše preference, tak jak je při práci uplatňujeme. Toto bych však mohl zhodnotit až po dlouhodobějším a frekventovaném používání a jestli vůbec. Hmatatelnou výhodu personalizovaného vyhledávání vidím spíše v možnosti označovat články tagy, které zůstanou zachovány i po odhlášení a opětovném přihlášení. Pod tyto tagy můžeme ukládat související články.

Obr. 3: Vytvořené tagy

Dále by mělo jít skládat vlastní kolekce spolu souvisejících článků. Tato aplikace se mi nepodařila úspěšně otestovat. Odkaz na přidání do kolekce mě dostal do příslušné nabídky, ale samotné vytvoření kolekce opětovně selhávalo s tím, že došlo k chybě a mám to zkusit později.

Obr. 4: nezdařený pokus o vytvoření kolekce

Vlastní zhodnocení projektu a jeho přínosu

CiteSeer^{x} je služba, která ve své době přišla s odlišným přístupem mezi již zavedenými digitálními knihovnami hlavně způsobem automatického zpracování citací. To umožnilo pokrýt velkou oblast informační a počítačové vědy. Přínosem je především to, že vidíme, kdo daný dokument citoval. Systém umožňuje prohledávat podle různých kritérií jako je název, jméno autora, klíčové slovo a podobně, ale také fulltextově. Dokáže nabídnout i statistiky citací, i když nejspíše ne u všech výsledků. Můžeme využít i vlastní registrace a s tím spojeného personalizovaného vyhledávání. Podle mého názoru služba funguje a je ve své podobě využitelná, ale pravděpodobně se momentálně dále nerozvíjí. Narazil jsem na některé věci, které nejspíš nefungují tak, jak by měly.

Seznam literatury/zdrojů, URL

1. *Wikipedia: CiteSeer* [online]. 2005, 17 November 2006 [cit. 2010-12-09]. http://en.wikipedia.org/wiki/Citeseer
2. *CiteSeer: About CiteSeer* [online]. 1997 [cit. 2010-12-09]. http://citeseer.ist.psu.edu/citeseer.html
3. *CLGiles* [online]. 2005 [cit. 2010-12-20]. CiteSeerX. Dostupné z WWW: http://clgiles.ist.psu.edu/CiteSeerX.shtml
4. *Www2006.org* [online]. 2006 [cit. 2011-01-28]. CiteSeerX. Dostupné z WWW: http://www2006.org/programme/files/pdf/p187.pdf

Metadata v DC

dc:title	CiteSeer (Next Generation) - digitální knihovna a vyhledávač
dc:creator	Klubus, Martin
dc:subject	CiteSeer (Next Generation)
dc:date	2011-02-01
dc:description	Describes main features of project CiteSeer (Next Generation)
dc:type	text
dc:language	"cs"