Systém pro zjišťování a evidenci publikací a citací

(návrh SW projektu)

Vedoucí: Jiří Dokulil (jiri.dokulil@mff.cuni.cz)
Irena Mlýnková (irena.mlynkova@mff.cuni.cz)
Počet členů týmu: 5
Jazyk, OS: libovolný

Motivace:
Ve světě vědy je klíčovou metrikou v posuzování úrovně vědecké činnosti daného autora počet jeho publikací a citací, tj. publikací jiných autorů, které se na jeho publikace odkazují. Zjistit takovéto informace ovšem není snadné, rozhodně je není možné triviálně získávat „ručně“ pomocí běžných vyhledávačů. Aktuální stav se navíc pochopitelně téměř každý den mění a přestože existuje několik systémů, které publikace i citace evidují, jejich praktická použitelnost není velká. Můžeme najít obecné systémy, které evidují určité množství publikací a jsou schopny z nich zjistit vzájemné citace [1,2,3], způsob jejich „plnění“ je ovšem buďto neznámý nebo převážně ruční a data v nich tedy nejsou ani úplná, ani aktuální. Na druhou stranu většina vydavatelů sborníků a časopisů má své digitální knihovny [4,5,6,7], ale v tomto případě obsahují buď pouze data daného vydavatele nebo je situace stejná jako v předchozím případě.

Cíle projektu:
Cílem projektu je implementace rozšiřitelného systému, který bude schopen automaticky zjišťovat, ukládat a v uživatelsky příjemné formě zpřístupňovat data o publikační činnosti a citacích různých autorů.
První částí systému bude vhodně upravený crawler, který bude schopen stahovat dokumenty, které pravděpodobně odpovídají publikacím, tj. vědeckým článkům ve sbornících či časopisech, technickým zprávám, kapitolám v knihách, monografiím bakalářským / diplomovým / dizertačním / habilitačním pracem apod. To, že se jedná o publikaci, bude určeno např. z textu „blízko“ odkazu na text publikace, na její stránku v elektronickém sborníku apod. Protože ne každá publikace obsahuje kompletní informace o svém původu (např. pro článek název sborníku / časopisu, v němž byl vydán, jeho vydavatele, rok a místo vydání apod.), bude nutné tyto informace dohledat. Lze ovšem předpokládat, že jsou typicky obsaženy na www stránce, z níž byla stažena. Pro evidenci nejen „citovanosti“, ale i „odkazovanosti“ budou dále evidovány linky vedoucí ke publikacím, resp. jejich stránkám v elektronických sbornících.
V další fázi bude analyzován obsah stažených dokumentů a budou odfiltrovány ty, které články ve skutečnosti nejsou. Např. součástí elektronických sborníků jsou nejen články ale i různé předmluvy, seznamy organizátorů apod. Vzhledem k tomu že články mohou být uloženy v téměř libovolném formátu (např. pdf, ps, doc, html, …) bude jejich obsah pro další zpracování transformován do vhodného jednotného meziformátu (např. XML). V rámci implementace projektu se předpokládá, že budou implementovány pouze transformace nejtypičtějších formátů (např. pdf), ale modulárně, aby bylo možné snadno přidávat další. Nad transformovanými daty bude dále provedena analýza, která z textu vyextrahuje potřebná data (název článku, autory a jejich kontaktní informace, klíčová slova, název sborníku, je-li uveden, seznam citací atd.). Získané údaje a relevantní údaje z crawleru budou uloženy v

databázi, kde bude také identifikován jejich zdroj a další doplňující údaje, například míra jejich důvěryhodnosti.
Dále budou uložená data „vyčištěna“, tj. z dostupných dat (tedy jak z dat nalezených na www stránkách, tak data vyextrahovaných z obsahu publikací), která obsahují nepřesnosti a nejsou úplná, bude naplněna druhá databáze, která již nebude obsahovat duplicity a kde budou data co nejúplnější. Například pokud existuje více autorů stejného jména, tak v této databázi by již tito autoři měli být vedeni samostatně a ke každému by měly být vedeny jen jeho publikace. Navíc i ta samá publikace získaná z více zdrojů by měla být uvedena pouze jednou s maximem dostupných informací, takže se v systému nebudou objevovat duplicity způsobené různými formáty (např. uvedení plných jmen autorů vs. uvedení pouze iniciál křestních jmen vs. uvedení prvního autora a příznaku „a kol.“ nebo stejnojmenných článků vydaných v různých formách jako např. článek ve sborníku / článek v časopise / technická zpráva apod.) a neúplnými informacemi (např. často chybějící vydavatel sborníku, ISBN knih, ISSN časopisů apod.).
Poslední částí systému bude rozhraní, které umožní takto získané informace vhodně dotazovat a výsledky v uživatelsky příjemné formě zobrazovat. Prostřednictvím rozhraní bude možné zobrazovat seznamy publikací vybraných autorů, seznamy jejich citací a referencí, seznamy publikací a referencí celých sborníků apod. Důležitou součástí bude možnost upřesnění konkrétní osoby v případě duplicit jmen. Např. bude možné specifikovat osobu jmény jeho spoluautorů, přibližnými roky publikační činnosti, tématy, kterými se zabývá apod. Škála dotazování by měla být co nejširší. Zobrazené citace bude dále možné filtrovat podle různých kritérií jako např. ignorování self-citací (tj. citací v nichž autor cituje svůj vlastní článek), citací od spoluautorů autora, referencí z jeho vlastních stránek apod.

Další požadavky na program:

- Aplikace bude stabilní a veřejně přístupná. Cílem je zajistit, aby aplikaci využívalo co nejvíce uživatelů.
- Veškerá dokumentace bude v angličtině.

Předpokládaný průběh práce:

1. Analýza existujících implementací a přístupů v jednotlivých oblastech
2. Podrobná specifikace konkrétních funkcí systému, architektury a rozhraní mezi jednotlivými moduly
3. Implementace projektu
4. Testy, ladění
5. Ověření funkčnosti na netriviálních datech
6. Dokumentace (programátorská, uživatelská, instalační)

Poznámka:
Problematiku řešenou v rámci implementace projektu je možné rozšířit do diplomových prací.

Reference:
[1] CiteSeer (http://citeseer.ist.psu.edu/)

[2] DBLP (http://www.informatik.uni-trier.de/~ley/db/)

[3] Google Scholar (http://scholar.google.cz/schhp?sourceid=navclient&hl=cs)

[4] ACM Digital Library (http://portal.acm.org/dl.cfm)

[5] IEEE Digital Library (http://www.computer.org/portal/site/csdl/index.jsp)

[6] IEEE Xplore (http://ieeexplore.ieee.org/Xplore/dynhome.jsp)

[7] Springer Link (http://www.springerlink.com/home/main.mpx)