Príprava dát na analýzu a personalizovanú prezentáciu na webe v doméne vedeckých publikácií

Oto Vozár, Mária Bieliková

Ústav informatiky a softvérového inžinierstva,
Fakulta informatiky a informačných technológií,
Slovenská Technická Univerzita, Ilkovičova 3, SK 842 16 Bratislava
{vozar,bielik}@fiit.stuba.sk

Abstrakt. Príspevok sa zameriava na proces prípravy dát pre analýzu a prezentáciu, ktorý je súčasťou viacerých metód práce s informáciami s cieľom efektívneho vyhľadávania a personalizovaného odporúčania používateľovi. Cieľom je vytvorenie sady údajov využiteľných pre experimentovanie s metódami a technikami práce s informáciami, ktoré sú postavené na webe so sémantikou. Opisujeme doménovo špecifické (v doméne vedeckých publikácií) a aj všeobecne použiteľné metódy pre odhaľovanie duplicít ako aj „čistenie" jednotlivých inštancií. Taktiež sa venujeme optimalizácii týchto metód a to najmä z časového hľadiska, keďže pre rozsiahle súbory dát je vykonanie v základnom tvare navrhnutých metód veľmi náročné. Overenie sme realizovali pomocou softvérového nástroja, ktorý umožnil sadu experimentov poskytujúcich prehľad o úspešnosti metód v umelých aj reálnych podmienkach.

Kľúčové slova: ontológia, príprava dát, duplicita, publikácie

1 Úvod

Rýchlo napredujúci rozvoj v oblastiach webu so sémantikou prináša so sebou množstvo nových prístupov pre navigáciu, filtrovanie, či vizualizáciu čoraz väčšieho množstva údajov. V praxi však treba na ich overenie vytvoriť dostatočne početnú experimentálnu vzorku dát (inštancií v doménovej ontológii), čo sa dá dosiahnuť vytvorením umelých dát alebo ich získaním z existujúcich zdrojov. Práca s reálnymi údajmi, čo je atraktívnejšie a pre experimentovanie vhodnejšie riešenie, však so sebou otvára novú skupinu problémov, ktoré treba riešiť.

Tento príspevok opisuje metódy pre prípravu získaných reálnych údajov pre ďalšiu ich analýzu, či experimenty v doméne vedeckých publikácií. Napriek niektorým špecifikám sú však dostatočne všeobecné, aby mohli byť nasadené aj v iných oblastiach.. Proces v sebe zahŕňa predovšetkým odstraňovanie duplicít ako aj metódy, pre čistenie jednotlivých inštancií.

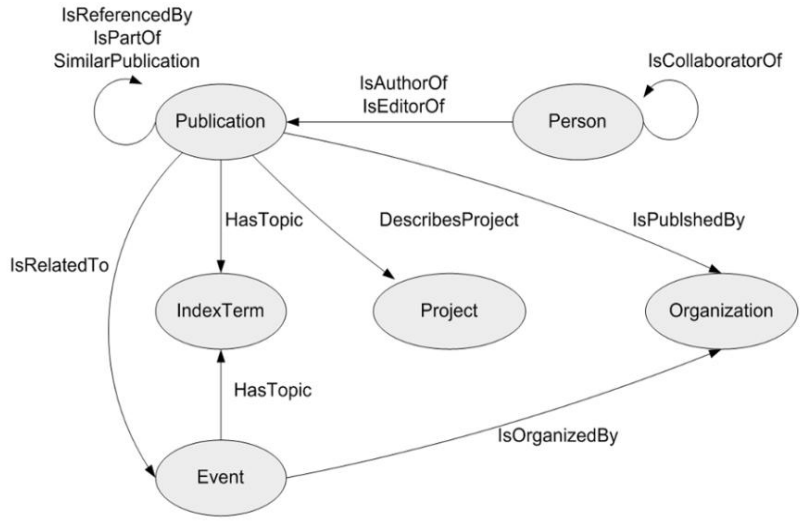
1.1 Ontológia publikácií

Vzťahy v rámci domény sú konceptualizované pomocou ontológie publikácií reprezentovanej jazykom OWL. Ontológiu sme vytvorili na základe analýzy domény ako aj analýzy potenciálnych zdrojov inštancií.

Pre ilustráciu vymenúvame významnejšie triedy:

- *Publication* – predstavuje všeobecnú triedu pre vedecké publikácie, jej podtriedami sú jednotlivé druhy publikácií, napr. kniha, článok alebo zborník.
- *Person* – zahŕňa autorov aj editorov publikácií.
- *Event* – predstavuje konferencie a iné stretnutia, ku ktorým sa publikácie viažu.
- *Organization* – zahŕňa univerzity, vedecké pracoviská a vydavateľov.
- *IndexTerm* – reprezentuje tému článkov, pričom témy sú vytvorené hierarchicky s použitím kategorizácie v digitálnej knižnici ACM.

Vzťahy medzi triedami sa nachádzajú na obrázku 1.



Obr. 1. Významné väzby medzi triedami.

Inštancie ontológie, teda údaje pre vybudovanie experimentálnej vzorky pochádzajú z troch väčších portálov zaoberajúcich sa vedeckými publikáciami a to ACM (www.acm.org), DBLP (www.informatik.uni-trier.de/~ley/db) a Springer (www.springer.com). Sú to ako údaje o autoroch a organizáciách, tak aj o publikáciách, kľúčových slovách či referenciách. Počty jednotlivých inštancií týchto tried pre každý zdroj sa nachádzajú v tabuľke 1.

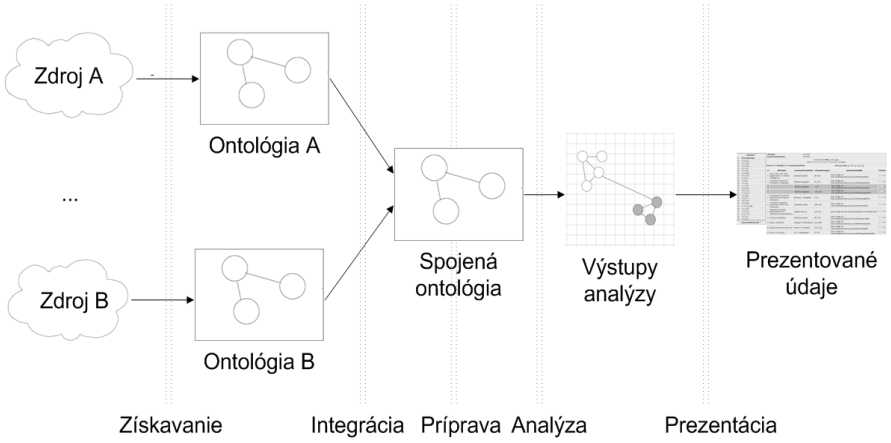
Tabuľka 1. Počty inštancií získaných z portálov ACM, DBLP a Springer.

Typ inštancie	ACM	DBLP	Springer
Autor	126 589	69 996	57 504
Organizácia	17 161	-	6 232
Publikácia	48 854	47 854	35 442
Kľúčové slovo	49 182	-	-
Referencia	454 997	-	-

1.2 Kontext procesu prípravy dát

Procesy prípravy dát boli navrhnuté v rámci výskumného projektu MAPEKUS[1] (Modeling and Acquisition, Processing and Employing Knowledge about User Activities in the Internet Hyperspace), ktorý sa zaoberá personalizovanou navigáciou vo veľkých informačných priestoroch, získavaním znalostí o používateľovi a prispôsobovaním sa jeho potrebám a záujmom [2].

Na obrázku 2 je znázornené umiestnenie prípravy dát v rámci ostatných procesov. Predchádza mu získanie údajov z jednotlivých zdrojov pomocou obaľovačov a ich integrácia. Následnými procesmi sú analýza údajov (napr. extrakcia grafov či zhlukovanie) a ich prezentácia používateľovi pomocou metód adaptívnej prezentácie a navigácie (napr. adaptívnym fazetovým prehliadačom).



Obr. 2. Následnosť procesov.

Prípravu dát predstavuje najmä odhalenie a odstránenie nekonzistencií a chýb, ktoré mohli vzniknúť najmä z týchto príčin:

- *Nekonzistencie v zdrojových údajoch* – duplicitné údaje, neúplné alebo chybné dáta spôsobené napr. preklepmi sa môžu nachádzať priamo v zdrojoch údajov.
- *Nekonzistencie, ktoré vznikli procesom obaľovania* – je možné, že niektoré duplicity vznikli pri procese obaľovania, napr. ak nie sú z časových dôvodov nasledované všetky odkazy.
- *Nekonzistencie spôsobené integráciou údajov z viacerých zdrojov* – rovnaký údaj môže mať v rôznych zdrojoch rôzny tvar (napr. iné používanie veľkých písmen v názvoch, rôzne bibliografické záznamy publikácií), čo sa môže prejaviť pri integrácii vznikom duplicitných inštancií.

[1] http://mapekus.fiit.stuba.sk/

2 Prístupy k príprave ontologických dát

Čistenie údajov (*data cleaning, cleansing, scrubbing*) je už dlhšie riešená oblasť, ktorá sa zaoberá detekciou a odstraňovaním chýb a nekonzistencií v dátach s cieľom zlepšenia ich kvality [6]. Problémy v dátach sa môžu vyskytovať v rámci jednej kolekcie, napr. súboru alebo databázy alebo pri spájaní viacerých zdrojov napr. veľkých dátových skladov, kde sa často vyskytujú redundantné dáta v rôznej reprezentácií [5].

Z nástrojov na čistenie údajov tvoria jednu skupinu doménovo zamerané. Komerčný nástroj Trillium[2] je zameraný na hľadanie zhody medzi zákazníkmi v systémoch pre manažment zákazníckych vzťahov. Obsahuje techniky pre extrakciu a transformáciu základných údajov ako mena, priezviska a adresy. Obsahuje vyše 200 000 pravidiel bežne používaných pri spracovaní dát. K všeobecným riešeniam patrí nástroj DataCleanser[3], ktorý možno použiť na čistenie duplicity v rozličných oblastiach, napr. na čistenie duplicitných e-mailových kontaktov.

Opísané nástroje aplikujú podobné princípy ako nami navrhnuté avšak sústreďujú sa skôr na reprezentácie založené na relačných databázach namiesto ontológií. Použitím ontológií vznikajú špecifické problémy, keďže treba uvažovať rôzne typy atribútov či vzťahy explicitne reprezentované v ontológii.

Prístupy pre hľadanie duplicít sú príbuzné s metódami a nástrojmi pre porovnávanie konceptov. Koncept je definovaný ako množina individuálnych objektov, ktoré obsahujú údaje označované ako atribúty. Existuje dva pohľady: extenzionálny, v ktorom koncept pozostáva z množiny svojich inštancií a intenzionálny, kde je reprezentovaný množinou svojich atribútov [3]. Intenzionálny pohľad pri porovnávaní konceptov je využitý pri formálnej analýze konceptov [4].

Metóda prezentovaná v [1] je zameraná na porovnávanie konceptov pomocou taxonómie tried a dátových ako aj objektových relácií s cieľom získania charakteristík používateľa v rámci prispôsobovania vyhľadávania v doméne pracovných ponúk. Cieľom je určenie podobnosti medzi konceptmi všeobecne na rozdiel od našich metód pre detekciu identických inštancií (z pohľadu ich sémantiky a nie reprezentácie), teda inštancií s vysokou podobnosťou.

Porovnávanie konceptov sa využíva aj pri mapovaní ontológií, kde je cieľom nájsť podobné koncepty v rámci dvoch rôznych formalizácií nejakej domény. Príkladom je prístup pre hľadanie zmien v rôznych verziách ontológie, kde je porovnávanie jedným z prístupov [7].

3 Oprava inštancií ontológie jedným prechodom

Túto metódu sme navrhli pre opravu údajov v rámci jednej inštancie danej ontológie. Sú to predovšetkým prípady:

- úprava formátu niektorých druhov údajov, napr. mien a priezvisk, ktoré majú začínať veľkým písmenom,

[2] http://www.trilliumsoftware.com/home/products/index.aspx
[3] http://www.npsa.com/edd/archfull.html

- oddelenie jednotlivých elementov údajov, napr. rozlíšenie mena a priezviska,
- filtrovanie inštancií s nedostatkom údajov, aby ich malo zmysel zaradiť do analýzy,
- filtrovanie obsahu niektorých údajov, napr. vyňatie spojok z kľúčových slov publikácií.

Čistenie inštancií je najvhodnejšie realizovať využitím konceptu dátovodov a filtrov, kde každý filter je určený pre špecifickú úlohu. Celý proces možno realizovať lineárnym prechodom cez všetky inštancie vo vzorke, preto je jeho asymptotická časová zložitosť lineárna. Pri praktickej realizácii je vhodný jeho súbežný priebeh spoločne s procesom zbierania inštancií zo zdrojov (napr. prostredníctvom obaľovača), pretože sa dá vykonať pre každú inštanciu samostatne hneď po jej získaní a počas zbierania je spravidla procesor len minimálne zaťažený.

4 Detekcia a odstraňovanie duplicitných inštancií

4.1 Základné metódy pre odhaľovanie duplicít

Na detekciu duplicít sme navrhli dve metódy, jednu na porovnávanie samotných údajov, čiže dátových relácií a ďalšiu, ktorá pracuje na základe porovnávania vzťahov medzi inštanciami, teda objektových relácií.

Základný princíp porovnávania je rovnaký pre obe metódy. Celková podobnosť inštancií sa počíta na základe podobností ich neprázdnych relácií (či už dátových alebo objektových). Pre každú reláciu sa použije váhovacia metóda s tzv. pozitívnou a negatívnou váhou. Je všeobecnejšia ako obyčajné váhovanie pomocou jednej hodnoty, ktoré je vo veľa prípadoch nepostačujúce. Uvažujme napr. v doméne vedeckých publikácií krajinu, v ktorej žijú dvaja autori. Ak je podobnosť názvu krajiny nízka, tak je vysoká pravdepodobnosť, že sa nejedná o rovnakú osobu (potreba vysokej váhy). Ak je však podobnosť vysoká, to, že autori žijú v rovnakej krajine neznamená, že ide o rovnakú osobu (potreba nízkej váhy).

Parametrami váhovania sú tri hodnoty a to pozitívna váha p, negatívna váha n, a prahová hodnota t, ktorá určuje, pri akej podobnosti použiť pozitívnu a pri akej negatívnu váhu. Celková podobnosť dvoch inštancií je hodnota medzi 0 (úplne rozdielne) a 1 (identické) a dá sa vypočítať z podobnosti relácií takto:

$$S = \frac{\sum_{i=1}^{N} F_i(s_i) + n_i}{\sum_{i=1}^{N} p_i - n_i} \quad (1)$$

kde N je počet neprázdnych relácií, ktoré obe inštancie obsahujú, s_i predstavuje podobnosť objektov v i-tych reláciách (hodnota medzi 0 a 1), p_i je pozitívna váha daného typu relácie, s_i negatívna a F_i funkcia daná vzťahom:

$$F_i(x) = \begin{cases} p_i x & x \geq t_i \\ n_i(x-1) & x < t_i \end{cases} \quad (2)$$

kde *i* je index relácie, p_i pozitívna váha, n_i negatívna a t_i predstavuje prahovú hodnotu (medzi 0 a 1).
Pre rozhodnutie, či budeme považovať dané inštancie identické (z pohľadu toho čo vyjadrujú) sa berú do úvahy výsledky oboch metód, t.j. porovnanie dátových a objektových relácií, z ktorých sa vypočíta priemer a následne sa uplatňuje inštančná prahová hodnota. Ak je podobnosť vyššia, inštancie prehlásime za identické.

Porovnávanie dátových relácií. Metóda je založená na porovnávaní príslušných dátových relácií patriacich dvom inštanciám rovnakej triedy. Pre porovnanie údajov sme použili pätnásť štandardných metrík na meranie podobnosti reťazcov ako napr. QGramy, Levensteinová alebo Monge-Elkanová metrika. Zároveň sme navrhli niekoľko ďalších metrík:

- *Keyboard distance* metrika – jej základným mechanizmom je výpočet vzdialeností klávesov pre nezhodujúce sa písmená dvoch reťazcov, berúc do úvahy aj klávesu shift a klávesy z rôznych rozložení kláves (napr. slovenskú a anglickú klávesnicu), je vhodná na detekciu preklepov.
- *Menná metrika* – je určená pre porovnávanie mien a priezvisk, berie do úvahy, že niektoré z nich môžu byť skrátené, napr. J. F. Smyth a John. F. Smyth určí ako zhodné mená.
- *Kompozitná metrika* – umožňuje porovnávať pomocou kombinácie všetkých ostatných metrík, umožňuje určiť váhu každej použitej metriky.

Pre každú dátovú reláciu môžeme zvoliť samostatnú metriku porovnávania. Dá sa zvoliť aj jednoduché porovnávanie zhodnosti reťazcov, čo je užitočné v niektorých prípadoch, napr. pri porovnávaní ISBN.

Porovnávanie objektových relácií. Metóda funguje na základe porovnávania objektových relácií inštancií, čiže porovnávania oboru hodnôt týchto relácií, napr. kníh, ktoré porovnávaní autori napísali. Tie sa porovnávajú pomocou ich dátových relácií. Netreba rozlišovať funkcionálne a nefunkcionálne relácie, ale treba na to pamätať pri nastavovaní pozitívnych a negatívnych váh a nefunkcionálnym reláciám nastavovať menšie hodnoty váh.

4.2 Optimalizácia porovnávania inštancií

Problematika vzájomného porovnávania inštancií ma kvadratický charakter zložitosti, čo v praxi znamená, že základný tvar metód je pre väčšie množstvá inštancií (už nad 10 000) z časového hľadiska nevyhovujúci. Z optimalizácií asymtotickú zložitosť mierne zlepšuje len jedna metóda a to zhlukovanie inštancií do skupín pred porovnaním, avšak za cenu presnosti. Ostatné spôsoby zrýchľujú vykonanie

porovnávania zmenšením konštanty, spoločne však urýchľujú proces približne o faktor dvadsať, čo je nezanedbateľné zlepšenie.

Navrhli sme takéto optimalizácie:

- *Využitie výsledkov dátového porovnávania pri porovnávaní objektových relácií.* Najprv sa vykoná dátové porovnávanie všetkých inštancií, ktorého výsledok by sa mal zapamätať pre každú dvojicu. To by však vyžadovalo pre väčšie vzorky obrovské množstvo pamäte, a preto sa pre každú inštanciu dá zapamätať konštantný počet k nej najbližších inštancií rovnakého typu (napr. k autorovi päť jemu najpodobnejších). Následne sa vykoná porovnávanie pomocou objektových relácií, kde sa tieto výsledky využijú. Ak pre jednu inštanciu nie je inštancia s ňou porovnávaná v zozname jej niekoľkých najbližších, aproximuje sa ich podobnosť nulovou hodnotou, v opačnom prípade sa využije zapamätaná hodnota.
- *Zhlukovanie autorov.* Táto optimalizácia funguje na základe zhlukovania autorov podľa ich prvého písmena v priezvisku. Je navrhnutá práve pre autorov z toho dôvodu, že ich je v ontológii publikácií vždy najväčší počet. Empiricky sme zistili, že je malá šanca, aby v prvom písmene priezviska vznikol preklep, nakoľko sa to dá veľmi rýchlo zbadať a ľudia sa zväčšia nemýlia pri prvom písmene slova. Po vytvorení zhlukov sa inštancie autorov porovnávajú len v rámci jednej skupiny. Táto optimalizácia znižuje presnosť porovnávania, pretože ak by bol rozdiel v prvom písmene priezviska, porovnávanie takýchto inštancií by vôbec neprebehlo.
- *Porovnanie počtu písmen*. Porovnanie možno vykonať napr, pred Levensteinovou metrikou, ak je rozdiel v dĺžke príliš veľký, netreba už ďalej využiť drahšie určenie podobnosti a stačí ju aproximovať nulou
- *Porovnanie slov.* Tento postup možno využiť pred drahšou metrikou, vhodné sú napr. dlhé názvy článkov. Pokiaľ sa v názvoch nachádza viac než povolený počet rozdielnych slov, podobnosť je považovaná za nulovú.
- *Porovnanie početnosti písmen.* Pre niekoľko písmen, napr. pre samohlásky a menší počet spoluhlások je vhodné zistiť ich početnosti v oboch porovnávaných reťazcoch. Ak sú príliš rozdielne, netreba využiť ďalšie metriky a podobnosť určiť ako nulovú.

4.3 Odstraňovanie duplicít

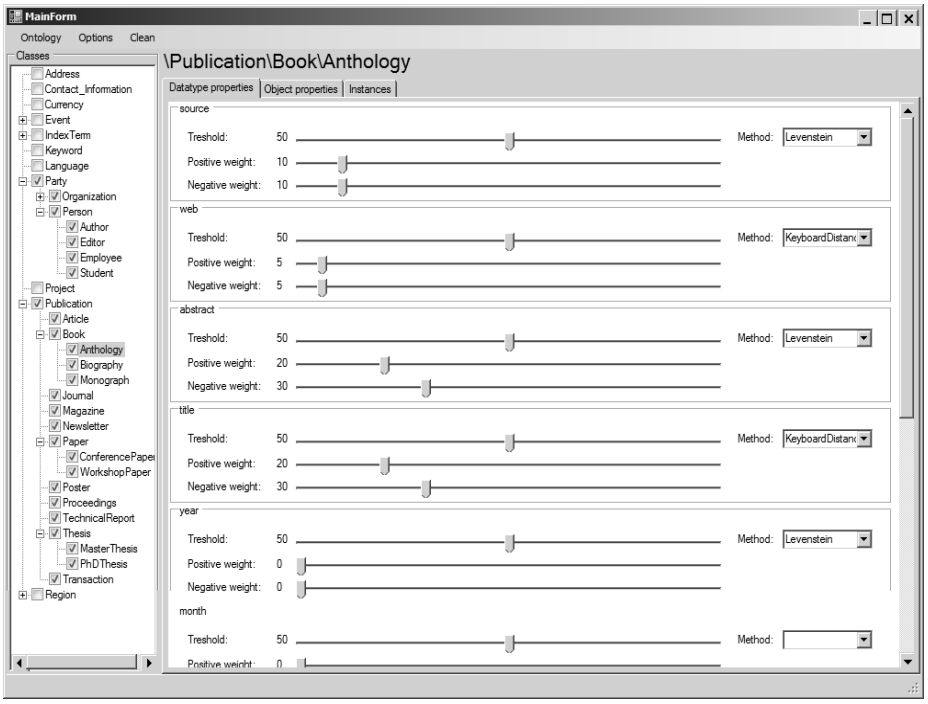
Identifikovali sme niekoľko spôsobov, ako zaobchádzať s inštanciami, ktoré boli označené za duplicity:

- označenie duplicity v ontológii pomocou špeciálnej objektovej relácie,
- manuálne vyriešenie duplicity, tento spôsob je vhodný, ak bol identifikovaný menší počet duplicít,
- vymazanie inštancie s menším množstvom informácií,
- zlúčenie inštancií, použijú sa dátové relácie, ktoré obsahujú viac údajov, nefunkcionálne objektové relácie sa zlúčia[4].

[4] Chapman, S.: SimMetrics, www.dcs.shef.ac.uk/~sam/stringmetrics.html,

5 Overenie riešenia

Pre overenie metód a za účelom experimentovania sme vytvorili softvérový nástroj pre opravu inštancií a vyhľadávanie duplicít v ontológií publikácií. Nástroj realizuje jednotlivé navrhnuté metódy a zároveň umožňuje konfiguráciu a experimentovanie s parametrami jednotlivých metód na vybraných vzorkách ontológie reprezentovanej v jazyku OWL. Hlavné okno nástroja je zobrazené na obrázku 3. V ľavej časti sa nachádza hierarchia tried v stromovej štruktúre, pričom pre jednotlivé triedy je možné zvoliť, či sa ich inštancie budú alebo nebudú porovnávať. Pre zvolenú triedu môžeme v pravej časti vidieť jej dátové a objektové relácie a nastaviť metódu ich porovnávania spoločne s váhami a prahovou hodnotou. Nastavovanie má pritom tú vlastnosť, že ak majú dve rôzne triedy rovnaký typ relácie, nastavuje sa pre obe rovnako, teda napr. nie je nutné meno autora a editora nastavovať dvakrát. Takisto môžeme prezerať inštancie jednotlivých tried.



Obr. 3. Hlavné okno nástroja pre nastavovanie parametrov porovnávania.

Nástroj umožňuje jednoprechodové čistenie inštancií, napr. rozdelenie mena na krstné a stredné mená a priezvisko. Jej hlavnou činnosťou je však detekcia duplicít. Najprv zhlukuje autorov, porovná dátové relácie a následne objektové, pričom využíva nastavené, zväčša experimentálne zistené parametre. Po vykonaní procesu vypíše štatistiky pre jednotlivé triedy, čo je potrebné pre vykonanie meraní.

Za účelom zistenia úspešnosti metódy pre odstraňovanie duplicít sme použili vzorku reálnych údajov získaných z databázy DBLP, v ktorých bol umelo vytvorený

známy počet duplicít. Vytváranie duplikátov predstavuje postupný náhodný výber inštancií, pričom každá prejde procesom, ktorý zahŕňa tieto kroky:

1. Mutácia náhodne zvolených dátových relácií jednou z týchto operácií:
   - zámena slov
   - vymazanie slova
   - zámena písmen
   - vymazanie písmena
   - duplikácia písmena
2. Vymazanie náhodných objektových relácií (pre každú reláciu je 5% šanca, že bude vymazaná, táto hodnota bola určená pokusmi s cieľom priblížiť sa reálnym podmienkam).
3. Duplikácia inštancie nachádzajúcej sa v definičnom obore náhodnej objektovej relácie pomocou krokov 1 a 2 a následnú výmenu tejto inštancie v definičnom obore relácie za jej duplikát. Pre každú reláciu je pravdepodobnosť pre duplikáciu 5%, táto hodnota, bola určená rovnako ako v druhom kroku.

V rámci experimentu sme vytvorili vzorky o veľkosti 1 000, 2 000, 5 000, 10 000 a 20 000 inštancií so sto vloženými duplicitami (experimenty s väčšími vzorkami by boli časovo veľmi náročné). Pre každú z nich sme vykonali desať meraní. Kvôli časovej náročnosti sme použili rozdelenie inštancií autorov do skupín podľa priezviska (pozri časť 4.2). Pre porovnávanie názvov publikácií sme použili Levensteinovu metriku, pre mená a priezviská autorov špeciálnu mennú metriku a pre ISBN jednoduché porovnávanie reťazcov. Všetky metriky pre každú reláciu boli váhované s použitím empirických poznatkov zameraných na získanie čo najlepších výsledkov. Určili sme vždy dve váhy, jednu pozitívnu a ďalšiu negatívnu (pozri časť 4), napr. pre referenciu medzi publikáciami bola pozitívna váha 1 a negatívna 10, teda ak majú publikácie rovnaké referencie, je to menší znak ich identity, ako rôzne referencie ich rozdielnosti. Ako hraničnú hodnotu sme zvolili hodnotu 0.8, čiže všetky entity, ktoré boli vyhodnotené s touto a vyššou hodnotou podobnosti na stupnici 0 až 1 sa považujú za duplicity.

V tabuľke 2 sú uvedené počty správne nájdených duplicít (stĺpec *Správne*), zle identifikovaných duplicít (stĺpec *Nesprávne*) a neodhalených duplicít (stĺpec *Neodhalené*). Z týchto hodnôt sme vypočítali hodnoty *presnosť*, *pokrytie* a F_1-štatistika, čo sú štandardné charakteristiky pre vyhodnocovanie výsledkov pri práci s informáciami. Každé pole tabuľky obsahuje priemernú hodnotu a smerodajnú odchýlku. V meraniach uvažujeme len tie inštancie, ktoré možno identifikovať po zhlukovaní autorov, pretože táto optimalizácia nie je časťou samotnej identifikácie. Bez nej by boli výsledky rovnaké pre všetky duplikáty, ale proces by sa predĺžil na viac ako desaťnásobok.

Vo výsledkoch možno pozorovať, že hodnota charakteristiky *presnosť* je konštantná v každej vzorke. Počet chybne identifikovaných duplicít vzhľadom na počet inštancií lineárne raste, ale počet inštancií rastie tiež lineárne a počet ich vzájomných porovnaní kvadraticky.

Tabuľka 2. Výsledky experimentu identifikácie duplicít na rôzne početných upravených DBLP vzorkách.

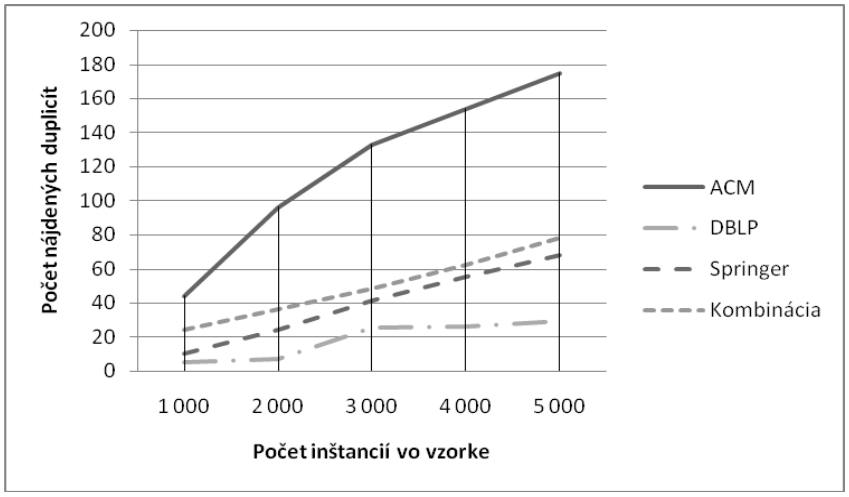
Veľkosť vzorky	Správne	Nesprávne	Neodhalené	Presnosť	Pokrytie	F1
1 000	53.2 ± 4.23	5.7 ± 7.52	14.1 ± 4.37	0.79 ± 0.06	0.90 ± 0.08	0.84 ± 0.04
5 000	70.6 ± 5.93	6.6 ± 5.95	18.8 ± 7.05	0.79 ± 0.07	0.92 ± 0.05	0.85 ± 0.05
10 000	74.5 ± 6.07	20.7 ± 5.20	17.7 ± 4.19	0.81 ± 0.05	0.78 ± 0.04	0.80 ± 0.04
20 000	81.3 ± 4.38	24.7 ± 5.58	13.9 ± 2.98	0.85 ± 0.03	0.77 ± 0.02	0.81 ± 0.02

Ďalší experiment sme vykonali s použitím reálnych údajov jednotlivo z každého zdroja ako aj vzoriek, v ktorých sa nachádzali údaje zo všetkých zdrojov (tabuľka 3). Nájdené duplicity sme ručne kontrolovali. Približne 90 % bolo správnych.

Tabuľka 3. Výsledky experimentu identifikácie duplicity na reálnych dátach.

Veľkosť vzorky	ACM	DBLP	Springer	Kombinácia
1 000	44	5	10	24
2 000	96	7	24	36
3 000	133	25	41	48
4 000	154	26	55	62
5 000	175	29	68	78

Pre lepšiu názornosť sú údaje z tabuľky 3 zobrazené v grafe na obrázku 4.



Obr. 4. Výsledky experimentovania s identifikáciou duplicity na reálnych dátach.

Najviac duplicít sa nachádza v zdroji ACM, väčšina z nich je spôsobená nejednotným používaním veľkých písmen článkov. Pri údajoch z kombinovaných zdrojov sa výraznejšie neprejavili duplicity spôsobené integráciou, čo je zapríčinené menšou

početnosťou inštancií vo vzorkách, pri ktorých je šanca pre výskyt jednej inštancie v dvoch zdrojoch nízka.

Počas experimentovania sme tiež zistili niekoľko zaujímavých skutočností v rámci domény publikácií:

- veľa duplicitných publikácií vzniká len kvôli rôznym veľkým písmenám,
- čínske mená a priezviská pôsobia pri rozpoznávaní väčšie problémy, pretože sú krátke,
- rôzne edície a vydania kníh sa ťažko rozpoznávajú, pretože sa ich názvy často líšia len jedným číslom alebo slovom,
- v jednom špecifickom prípade metóda nerozlíšila dvoch bratov píšucich rovnaké publikácie – niektoré takéto prípady sú ťažko rozlíšiteľné aj človekom.

6 Záver a ďalšia práca

Súčasné metódy zaoberajúce sa čistením dát sa aj v rámci objavovania znalostí sústreďujú na predspracovanie a prípravu dát v relačných databázach, takže sa priamo pre web so sémantikou nedajú použiť. Pri automatizovanom spracovaní meta-dát z reálnych zdrojov treba automatizovane zabezpečiť predspracovanie dát, v rámci zlepšovania ich kvality, čo môže výrazne zlepšiť výsledky iných metód a nástrojov, ktoré dané údaje využívajú (napríklad zhlukovače alebo fazetové prehliadače).

V príspevku sme opísali metódy, ktoré sme navrhlo na prípravu údajov z domény vedeckých publikácií pre analýzu s cieľom ich efektívnej prezentácie a navigácie v nich s využitím meta-dát reprezentovaných ontológiou. Sú to najmä metódy pre opravu inštancií a vyhľadávanie duplicít. V porovnaní s existujúcimi riešeniami sme sa zamerali na problémy v doméne vedeckých publikácií, ktorej analýza nie je ešte na tak úrovni ako pri iných doménach. Navrhli sme koncept kladných a záporných váh, ktorý prispel k zlepšeniu vyjadrovania v rámci parametrov a tým aj zvýšil presnosť riešenia. Navrhli sme niekoľko zaujímavých metrík, napr. na určenie vzdialenosti na klávesnici. Taktiež sme vytvorili niekoľko optimalizácií a hlavne dokázali využiť výsledky analýzy dát pri analýze vzťahov medzi dátami.

Vyvinuli sme softvérový nástroj a metódy overili na základe experimentov s použitím umelých aj reálnych vzoriek údajov rôznych veľkostí pochádzajúcich z rôznych zdrojov. Výsledky experimentov ukazujú reálnu využiteľnosť pri príprave doménových ontológií pre využitie ďalšími procesmi a softvérovými nástrojmi, ktoré tieto procesy realizujú. Naša práca sa v súčasnosti zameriava najmä na ďalšie zlepšovanie presnosti a efektívnosti vyhľadávania duplicít a overenia vplyvu kvality ontológie na následnú analýzu a prezentáciu.

Poďakovanie. Tento príspevok vznikol súvislosti s riešením projektu Agentúry na podporu vedy a techniky na základe Zmluvy č. APVT-20-007104.

Literatúra

1. Andrejko A., Bieliková M. Comparing Instances of the Ontological Concepts, In Proc. of Second Workshop on *Tools for Acquisition, Organisation and Presenting Information and Knowledge,* P. Návrat et al. (Eds.), Poľana, Slovakia, 2007, pp. 26-35.
2. Bieliková M., Návrat P. Modeling and acquisition, processing and employment of knowledge on user behavior in hyperspace of the Internet. In *Proc. of Znalosti 2007*, Ostrava, pp. 368-371 (in Slovak).
3. Burek P. Adoption of the Classical Theory of Definition to Ontology Modeling, *Artificial Intelligence: Methodology, Systems, and Applications*, Lecture Notes In Computer Science Ch. Bussler, D. Fensel (Eds.) 3192, Springer, Berlín, 2004, pp. 1-10.
4. Formica A. Ontology-based Concept Similarity in Formal Concept Analysis, In Information Sciences 176(18), 2006, pp. 2624-2641.
5. Hernándes M. A., Stolfo S. J., Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem, *Data Mining and Knowledge Discovery* 2(1), 1998, pp. 9-37.
6. Rahm E., Do H. H. Data Cleaning: Problems and Current Approaches, *IEEE Techn. Bulletin on Data Engineering*, 23(4), 2000, pp. 3-13.
7. Tury M., Bieliková M. An Approach to Detection Ontology Changes In *ICWE'06: Workshop Proceedings of the Sixth International Conference on Web Engineering*, Palo Alto, California, ACM Press, 2006.

Annotation:

Preprocessing of data for analysis and personalized web presentation in scientific publications domain

In this paper we describe methods for data preprocessing, which constitute part of methods working with information processing that are aimed at effective searching and personalized recommendation information to a user. Our goal is creation of a data set, which can be used in experiments with methods and techniques for data analysis, which are based on the Semantic Web. We describe domain specific (in domain of publications) and also domain independent methods for duplicity detection and instance correction. We also focus on optimalization of these methods, because they are in their basic form rather time expensive for large information spaces. Proposed methods have been evaluated using developed software tool, which enabled execution of set of experiments on artificially created and also real data.