ИНФОТЕКА бр.1-2/2002

СКРИПТОР - програм за рашчлањивање библиографских информација

Дејан Пајић, Перо Шипка,
Филозофски факултет - Катедра за психологију, Нови Сад,
Биљана Косановић,
Народна библиотека Србије, Београд

Сажетак
Описан је Скриптор, програм за рашчлањивање садржаја и референци из периодичних публикација, развијен за потребе одржавања базе података СоциоФакта. С ослонцем на помоћне базе (листе ауторских имена, издавача и сл.) и једноставне алгоритме за обраду српског као природног језика, програм аутоматски препознаје елементе из садржаја часописа и референци чланака (име аутора, наслов, извор, колацију итд.) и додељује им стандардне лабеле, чиме се обезбеђује аутоматски трансфер података у одговарајућа поља базе.

Поред основног модула за рашчлањивање, програм садржи потпрограме за конверзију кодних распореда, претварање великих слова у мала у складу са правописом, промену редоследа ауторског имена и презимена, допуњавање недостајућих информација, као и интерактивну контролу и корекцију рашчлањеног материјала.

Скриптор садржи програм за инсталацију и детаљан систем помоћи, који оператора упућује у употребу програма и упознаје га са библиографским стандардима који се користе при изради СоциоФакта. Написан је у језику Visual Basic for Application као додатак програму Microsoft Word.

Кључне речи: **библиографске информације, рашчлањивање, библиографске базе података, цитатне информације, софтвер**

Израда библиографских база података захтева одговарајућу припрему текста за унос у базу. Најважнија операција у процесу припреме је рашчлањивање библиографских јединица у складу с неким од стандарда, чиме се обезбеђује смештај истоврсних елемената референци (име аутора, наслов рада, извор, колација, итд.) у засебна поља базе. Велика важност рашчлањивања долази отуда што оно омогућује ефикасно кориговање грешака неизбежних у процесу уноса, као и нормирање библиографских информација, нужно због великог броја стандарда који се равноправно користе. Кориговање и нормирање библиографских информација сматрају се неопходним условом ефикасног претраживања.

Захтев за ваљаним рашчлањивањем поставио се у изради СоциоФакта - југословенске библиографске базе за друштвене науке, у најоштријем могућем облику [7]. СоциоФакт је конципиран као цитатна база која треба да послужи вредновању већег броја научних субјеката и генерисању знатно већег броја наукометријских показатеља него што је то случај у класичним цитатним базама (тзв. цитатним индексима: SCI, SSCI и A&HCI), по узору на које је и настао.

Цитатни индекси у свом садашњем облику представљају комбиноване базе података цитата са апстрактима. Поред апстраката и других поља која служе претраживању и трасирању примарних извора (нпр. наслови, адреса аутора итд.) цитатни индекси садрже само неколико поља, односно потпоља намењених евалуацији научних субјеката, дакле оном што представља посебност и предност цитатних база. Нарочито је мали број поља намењен евалуацији на основу цитираности (практично једно, с три потпоља). Рашчлањивање цитатних информација на тако мали број елемената не представља нарочито сложену активност.

У СоциоФакту се референце дају у пуном облику, тако да је њихово рашчлањивање знатно захтевније. У својој последњој, онлајн верзији **(http://www.nbs.bg.ac.yu/sfakt.htm**), СоциоФакт подржава вредновање цитираности издавача (професионалних и академских издавачких организација), научних тимова/пројеката и научних скупова, као и вредновање уредничког доприноса појединаца. Само се за квалитет часописа генерише десетак независних показатеља које цитатне базе стандардно не нуде, а у наукометријској литератури се сматрају веома вреднима. Такви су нпр. језик цитираног издања или издавачки облик референце (монографија – зборник с научног скупа – часопис – збирка радова). Проширење намене базе имало је за последицу битно повећање броја поља на које се рашчлањује свака референца. Тај број није стандардан јер зависи од врсте цитираног извора и броја аутора цитираног рада. У многим случајевима број (пот)поља износи више од 10.

С повећањем броја поља смањена је њихова дискриминабилност. На тај начин рашчлањивање је постало веома сложена скуп операција. За разлику од класичних цитатних индекса где је оно део процеса уноса и операторска активност, у припреми СоциоФакта рашчлањивање је засебна аналитичка делатност високог професионалног нивоа, која обједињава експертизу библиографског и научног (дисциплинарног) карактера. С поступним повећањем броја часописа представљених у СоциоФакту (до броја 58, колико их се реферише почев од 2000. године) рашчлањивање је постало "уско грло" у процесу одржавања базе. Упоредо с порастом базе нарастала је и потреба за аутоматизацијом процеса рашчлањивања.

АУТОМАТИЗАЦИЈА ПРОЦЕСА РАШЧЛАЊИВАЊА

У електронском научном издаваштву у току је процес преоријентације произвођача на понуду цитатних база у форми која омогућује њихову експлоатацију на даљину и у Веб окружењу, понекад у виду тзв. отворених архива. Паралелно се умножавају настојања да се аутоматизује процес њихове израде. Један од таквих покушаја је и ResearchIndex (раније CiteSeer) NEC Research Institute-a - систем Интернет апликација за аутоматско креирање дигиталних библиотека и аутономних цитатних индекса. ResearchIndex прикупља научне радове у Postskript, PDF и другим форматима доступним на World Wide Web-у, рашчлањује их, нормира и аутоматски генерише једноставне показатеље научног учинка типа цитираности [6]. Мада је за потребе креирања поузданог индекса цитираности у самом систему имплементирано неколико техника за изједначавање различитих формата истог цитата, полази се од тога да ResearchIndex као улазне податке најчешће има форматиране, кориговане документе у којима се литература наводи у складу са неким од познатих стилова, односно стандарда. Са друге стране, овде се не може говорити о

рашчлањивању у правом смислу те речи већ пре о издвајању релевантних информација (information extraction) пошто је најчешће довољно издвојити само име аутора (обично првоименованог), наслов, годину публиковања и бројеве страница цитираног рада [9].

Доследност и униформност у форматирању чланака и навођењу литературе, олакшава издвајање релевантних информација из електронског документа и омогућава релативно једноставно рашчлањивање референци у складу са препознатљивим, унапред дефинисаним моделима, односно правилима (нпр: АПА или Ванкувер стиловима цитирања). Трагање за подацима који формирају препознатљиве шаблоне и процесирање тих података у складу са упутствима везаним за одређени шаблон, познато је као template mining. Ова техника екстракције информација пронашла је своју примену и у парсирању, односно рашчлањивању библиографских информација [3]. Формирање таквих шаблона (templates) подразумева и утврђивање одређених, јединствених идентификатора или тзв. токена који се касније користе као синтаксички индикатори различитих делова референце, нпр: "journal" за назив часописа, "eds." за ознаку уредника и зборника, "и др." за крај списка аутора итд. Поред ових индикатора логичке структуре документа, сугерише се и потреба за анализом структуре везане за распоред типичних сегмената документа [1].

Обједињено коришћење токена, формирање шаблона и правила везаних за редослед навођења елемената референце, идентификовање честих интерпункцијских знакова и формата фонта и консултовање листа ауторских имена и назива часописа, показало се задовољавајућом техником екстракције и рашчлањивања цитатних информација [5]. Ипак, овако формирани модели представљају најчешће недовољно богат скуп образаца. Они не покривају у потпуности велику разноликост у навођењу цитатних информација. Свако иоле значајније одступање у формату цитата, нпр. изостављање имена аутора, доводи до тога да се сви наредни елементи референце погрешно класификују (рашчлане). Процес аутоматског парсирања би се свакако значајно олакшао, ако би се уредништва часописа и сами аутори доследније придржавали захтева везаних за правилно навођење цитиране литературе. С друге стране, постоје и предлози да се припремљени шаблони за аутоматско генерисање листа референци учине јавно доступним или да се интегришу у често коришћене програме за обраду текста [3]. Тренутно су, међутим, за ове потребе ауторима доступне само демо или пробне верзије комерцијалних програма какви су на пример ProCite, Bibliographix или Citation.

Корак даље у аутоматизацији процеса парсирања библиографских информација, представљају покушаји да се креирају апликације које ће бити способне да с ослонцем на пробабилистичке моделе, какви су нпр. Марковљеви ланци, "уче" правила по којима су подаци уређени [10]. Правила се, затим, у виду неког софтверског решења примењују на електронске документе. Програм најпре препознаје стил навођења литературе, а потом референце рашчлањује применом правила везаних за тај стил, чак и без коришћења речника или база имена. Апликације ове врсте, међутим, захтевају да се пре самог рашчлањивања обави "увежбавање" на већ рашчлањеном материјалу структурираном у складу са одређеним стилом [4]. Број грешака у рашчлањивању значајно се повећава уколико се у материјалу јављају ауторска имена, називи часописа или термини који нису постојали у примерима на којима је обављено увежбавање, односно учење [9].

СКРИПТОР

Окружење

Скриптор је пројектован као шаблон (template) за Microsoft Word, односно скуп тзв. макроа који сукцесивно извршавају већи број операција. Написан је у програмском језику Visual Basic for Applications. Програм се након инсталације интегрише у Word 97/2000 и његове опције постају доступне делом преко новог менија који се појављује на палети менија, а делом позивањем приручног менија притиском на десни тастер миша. Овим је с једне стране обезбеђен интуитиван интерфејс за мање искусне кориснике, а с друге могућност да се користе стандардне функције програма Word за измену, брисање и копирање докумената и њихово снимање у различитим форматима. Створени су и услови за ефикаснију интеграцију с осталим програмима пакета Microsoft Office, првенствено програмом Access, пошто се речници и листе ауторских имена, назива часописа, издавача и градова који се консултују у току рашчлањивања, налазе у mdb формату.

Структура

Основу програма Скриптор чине два модула: модул за рашчлањивање садржаја периодичних публикација и модул за рашчлањивање референци у оквиру сваког чланка. Након рашчлањивања, подаци се контролишу, коригују, допуњавају и припремају за чување у формату који омогућава аутоматско смештање података у одговарајућа поља базе. Ова припрема је олакшана великим бројем помоћних рутина за: 1) конверзију различитих кодних распореда укључујући конверзију ћирилице у латиницу и обратно, уз могућност транслитерације руског алфабета, 2) претварање великих слова у мала у складу са правописом, односно уз консултовање пратећих речника, 3) замену места ауторског презимена и имена, односно иницијала ради свођења сваке референце на јединствен формат 4) допуну недостајућих података и развијање скраћених имена часописа у пун облик на основу листе, 5) интерактивну контролу рашчлањеног материјала у складу са захтевима и 6) упоређивање сваке рашчлањене референце са њеним оригиналним обликом ради ефикаснијег отклањања евентуалних грешака. Скриптор садржи детаљан систем помоћи у форми хипертекста, тачније типичну Windows датотеку помоћи која омогућава једноставно и ефикасно претраживање упутстава у којима се оператеру описује начин коришћења самог програма, нуде информације потребне за правилно рачлањивање и припрему материјала и описују библиографски стандарди коришћени при изради СоциоФакта и библиографских база уопште.

Рашчлањивање

Модул за рашчлањивање садржаја користи се за формирање тзв. "костура" - почетне датотеке која садржи основни библиографски опис сваког рада у оквиру једног броја часописа: име аутора, наслов, назив часописа, годину публиковања и колацију. Пошто корисник унесе назив часописа, колацију, годину публиковања, као и неколико основних смерница везаних за карактеристике фонта и сепараторе коришћене при раздвајању различитих елемената садржаја, рутине у оквиру овог модула, анализирајући првенствено изглед, карактеристике и распоред делова текста, а по потреби консултујући и листу имена аутора, структуришу садржај периодичне публикације додељујући

стандардне лабеле одговарајућим информацијама: АУ - аутор рада, НА - наслов рада итд. Овако формиран костур се након тога попуњава пратећим апстрактима, афилијацијама аутора и цитатним информацијама које се рашчлањују у оквиру другог модула.

Рашчлањивање референци чланака је далеко сложенија и захтевнија операција која, да би се успешно програмерски имплементирала, подразумева коришћење већег броја критеријума, система процена и правила чијим се кумулативним примењивањем долази до коначне логичке одлуке о смештању информације у одговарајућу категорију, тј. поље базе. Алгоритми овог модула се такође ослањају на различите карактеристике делова текста (курзив, подвучен, задебљан), њихов положај у оквиру реченице/референце и постојање типичних знакова интерпункције. Ово је, међутим, веома ретко поуздан и довољно флексибилан механизам идентификовања различитих елемената референце. Стога су развијени додатни алгоритми за обраду природног језика чији је обавезан корак у процесу доношења одлуке консултовање база ауторских имена, назива часописа, издавача и градова. Истовремено се испитује постојање честих синтаксичких показатеља и знакова, тј. идентификатора одређене врсте информација. Неки од тих индикатора или токена дати су у Табели 1.

Табела 1.**Чести индикатори присутности одређених библиографских информација**

ВРСТА ИНФОРМАЦИЈЕ	**ИНДИКАТОР**
назив часописа	ЧАСОПИС, ANALES, АНАЛИ, ARCHIVES, BULLETIN, ГЛАСНИК, JOURNAL, QUARTLERLY, REVIEW, РЕВИЈА, REVUE, ZEITCHRIFT, SERIES, BULL...
име уредника	RED, EDITOR, EDS, ED, REDACTOR, REDS, EDITORS, HRSG, УР...
назив издавача	ЦЕНТАР ЗА, PRESS, ЗАВОД, PUBLISHERS, CO, INC, LTD, VERLAG, ФАКУЛТЕТ, УНИВЕРЗИТЕТ, ИНСТИТУТ, САВЕЗ, ИЗДАТЕЛСТВО, НИП...
колација	ВОЛ, ВОЛУМЕН, ГОДИНА, ГОД, БРОЈ, БР, СТР, С, СС, Р, РР, NUMBER, NR, NO, VOLUME, СТРАНА/Е, СТРАНИЦА/Е...
наслов зборника	ЗБОРНИК, ЗБОРНИК РАДОВА, У МОНОГРАФИЈИ, У, IN, У КЊИЗИ, ОБЈАВЉЕНО У, ДЕО У...

Први корак у рашчлањивању референци представља издвајање управо оних поља, односно делова референце који имају релативно стандардизовану структуру, сталан и препознатљив положај у тексту или јединствене синтаксичке идентификаторе (invariants first). Обично се прво проналази податак о години из које потиче цитат и то као низ од четири цифре од којих су прве две 18, 19 или 20 [2]. Уобичајено је да се година у оквиру референце наводи или непосредно након имена аутора или на самом крају референце и у зависности од тога, алгоритам се грана. Уколико је година наведена након имена аутора, преостаје да се утврди садржај оног дела референце који следи након године, као и да се,

уколико је наведено више аутора, појединачни аутори издвоје у засебна поља. У другом случају, када је година наведена на крају референце, поступак је нешто сложенији: издваја(ју) се аутор(и), првенствено уз консултовање листе ауторских имена и презимена, а потом се рашчлањивање наставља као у претходном случају. Због велике разноврсности текстуалног материјала који је потребно рашчланити, као и због нестандардности у навођењу референци и неретких грешака у штампању или оптичком препознавању текста, практично на сваком кораку у поступку рашчлањивања могућа су већа или мања одступања и усложњавања процеса доношења одлуке. Већ у првом кораку, када се из референце издваја поље Г (година), механизам одлучивања се усложњава ако је, на пример, у референци наведено више различитих година - ниски цифара од којих су неке заправо део наслова или чак представљају бројчану ознаку странице.

Поступак рашчлањивања референци је у највећим делу аутоматизован и то у зависности од степена слободе који корисник дефинише пре почетка парсирања. Што је степен слободе већи то ће програм чешће самостално, без консултације са корисником, доносити одлуке о томе у које поље се смешта одговарајућа информација. Са друге стране, у ситуацијама када је текст неуредно форматиран, нетипичан или садржи доста грешака, пожељно је поставити степен слободе на нижу вредност, чиме ће се кориснику омогућити да узме више учешћа у процесу интерактивног парсирања, прихватајући или преиначујући предлоге које програм сугерише. Овакав начин рада је неопходан и због постојања одређених рутина у програму Скриптор које измењују, употпуњују или бришу делове текста и тако стварају ризик да се релевантни подаци изгубе, а грешке које произведе сам програм прикрију.

Рутине, односно опције које омогућавају интерактивно учешће корисника програма у процесу рашчлањивања груписане су у модул назван "Корекције и допун. Постоје три основна типа корекција, односно допуна. Кориснику је омогућено да прихвати, "овери" почетак одређеног поља, а тиме и границе дела текста који садржи неку информацију, при чему је дозвољено да се промени не само место прелома референце (почетак поља), већ и назив (лабела) поља. Ово је посебно корисно када програм, на пример, погрешно препозна назив зборника као назив часописа. Након одлуке корисника, рашчлањивање се наставља у складу са њом. Друга врста корекција тиче се брисања непотребних делова текста, нпр. ознака "у зборнику:" или "in:", бројева страница код монографија итд. Трећу групу чине напреднији алгоритими помоћу којих се информације допуњују или усклађују са форматом дефинисаним као стандард у цитирању и посебно као стандард у нормирању података у бази СоциоФакт. Овде, на пример, спадају рутине којима се скраћени називи часописа, коришћењем техника парцијалног поређења и израчунавања међусобне удаљености ниски, развијају у пуне, стандардне називе.

Нормирање

Основне замерке које се упућују класичним цитатним базама, односно цитатним индексима тичу се, поред очигледне пристрасности у избору часописа, недопустиво великог броја броја грешака у уносу. Број грешака је толики да значајно умањује ефикасност претраживања и доводи у питање прикладност база за вредновање учинка различитих научних субјеката, које је практично немогуће без опсежног допунског кориговања информација (data cleaning) и/или

високих допунских трошкова за употребу тзв. аналитичке датотеке произвођача база, Института за научне информације из Филаделфије.

Том проблему приступило се у изради СоциоФакта као могућој озбиљној препреци за увођење СоциоФакта у практичну употребу. Имајући на уму налазе о дистрибуцији библиометријских података из базе, нарочито ниску учесталост цитата немалог броја домаћих аутора, институција и часописа, велики утицај грешака у навођењу референци на цитатне показатеље, као и потребу да СоциоФакт послужи вредновању што већем броју домаћих научних субјеката, закључено је да се подаци у бази морају кориговати и нормирати. У те сврхе развијен је посебан програм ПроКос којим се истоврсне, а различито уобличене библиографске информације изједначавају (своде), а грешке настале, било у фази уноса у базу, било у припреми и штампању оригиналних докумената - реферисаних чланака - коригују и уклањају. Први део процеса кориговања и нормирања, онај који захтева увид у оригиналне докумете, рационално је обављати већ у фази рашчлањивања.

Кориговање и почетно делимично нормирање информација обавља се у Скрипторовом Контролном моду. Позивањем тог потпрограма, корисник се обавештава о евентуалним типичним грешкама у рашчлањеном материјалу. Грешке се односе на неслагање формата текстуалног материјала са нормама коришћеним при изради базе СоциоФакт, нпр. постојање поља Ц (часопис) и И (издавач) у оквиру исте референце, дуплираност поља, изостанак одређених обавезних поља, непотпуно или неједнако навођење истих ауторских имена, итд. Подаци којима је додељена погрешна лабела, могу се лакше уочити применом технике "вертикалног поређења". Она се обезбеђује на тај начин што се информације истог типа (нпр. називи часописа) издвајају у засебан (привремени) документ и абецедно сортирају у оквиру јединствене листе. У току контроле корисник се, захваљујући обезбеђеној "просторној" кореспонденцији основног и привременог документа, може у сваком тренутку "вратити" на рашчлањену референцу, како би уочену грешку и отклонио.

Евалуација

Скриптор се интензивно користи у одржавању базе СоциоФакт, тако да се његова ефикасност непрестано оцењује, појединачне рутине веома често усавршавају, а базе допуњавају. Повремене систематске провере његове ефикасности показују да резултати у великој мери зависе од језичке и штампарске уредности, као и врсте материјала који се рашчлањује. У случају типичних, комплетних и високо структурираних референци, односно референци које припадају научним областима покривених СоциоФактом, а наведене су у складу са неким од стандарда у цитирању, успешност рашчлањивања изражена пропорцијом правилних категоризација прелази 0,97. Тај удео, међутим, значајно опада онда када су референце нетипичне, карактеристике фонта недоследно примењиване, распоред елемената измењен у односу на већину стандарда или када су библиографски извори уграђени у напомене (опаске, коментаре), што је често случај када се референце дају у тзв. фус- или енд-нотама, уместо у засебном одељку на крају текста. Но, с обзиром на то да је Скриптор алатка која се користи у интерактивном режиму рада и да кориснику-аналитичару пружа веома комфорне услове за проверу и преправке аутоматски генерисаних решења, тачност одлука у рашчлањивању не мора бити најважнији критеријум његове ефикасности. Бар подједнако важан критеријум је брзина

рашчлањивања који аналитичар постиже у односу на рад на класичан начин ("пешке"), какав омогућују савремени програми за обраду текста својим стандардним рутинама. У вези с тим, потреба за некаквом систематском евалуацијом није се ни указивала, пошто је несумњиво да Скриптор вишекратно убрзава поступак рашчлањивања.

ЗАКЉУЧАК

Због великих и тешко предвидљивих варијација у форматима навођења литературе, као и атипичности значајног броја референци, аутоматско рашчлањивање библиографских информација домаћег порекла чинило се идејом која не обећава успешну реализацију. Међутим, показало се да је чак и без примене сложених и ресурсима захтевних статистичих метода, један ограничен, али богат и флексибилан скуп правила могућно претворити у апликацију која процес парсирања значајно убрзава и олакшава. Предност овакве апликације није само у економичности, већ и у томе што обезбеђује виши квалитет у рашчлањивању захваљујући томе што остварује одређени ниво "професионалности" (знања), консултујући знатно већу количину корисних информација од оне којом располаже просечан обучени аналитичар.

Ваља истаћи да је ефикасност Скриптора "домен-специфична". Сваки покушај његове примене изван научних области покривених СоциоФактом захтевао би не само допуну и ажурирање помоћних база, већ и прилагођавање алгоритама стандардима цитирања који се више користе у другим научним дисциплинама. То једнако важи за друге видове његове евентуалне употребе, какве су форматирање референци у току припреме за штампу (књига, часописних чланака, библиографија и сл.) или ретроспективна конверзија.

ЛИТЕРАТУРА

1. Aiello, M., Monz, C., Todoran, L. (2000) Combining linguistic and spatial information for document analysis. In: Mariani, J., Harama, D. (ed.) Proceedings of RIAO '2000 Content-Based Multimedia Information Access, CID, 266-275
2. Bergmark, D. (2000) Automatic extraction of reference linking information from online documents, Technical Report TR 2000-1821, Cornell University - Computer Science Departement, November, URL: http://www.cs.cornell.edu/cdlrg/Reference Linking/extraction.pdf, преузето 25.02.2002.
3. Chowdhury, G.G. (1999) Template mining for information extraction from digital documents, Library Trends, 48 (1), 182-208
4. Connan, J., Omlin, C.W. (2000) Bibliography extraction with hidden Markov models, URL: http://citeseer.nj.nec.com/294556.html, преузето 11.06.2001.
5. Ding, Y., Chowdhury, G.G., Foo, S. (1999) Template mining for the extraction of citation from digital documents. In: Second Asian Digital Libraries Conference, National Taiwan University, November 8-9
6. Giles, L.C., Bollacker, D., Lawrense, S. (1998) CiteSeer: An automatic citation indexing system, у: Witten I., Akscyn R., Shipman F.III (ed.) Digital Libraries - Third ACM Conference on Digital Libraries, ACM Press, New York, 89-98
7. Косановић, Б., Шипка, П. (1996) СоциоФакт - Југословенска база за друштвене чињеничке науке. У: Костић П. (ур.) Мерење у психологији, ИКСИ и Центар за примењену психологију, Београд, 2, 85-95
8. Lawrense, S., Giles, L.C., Bollacker, D. (1999): Digital libraries and autonomous citation indexing, IEEE Computer, 32 (6), 67-71

9. Seymore, K., Mccallum, A., Rosenfeld, R. (1999) Learning hidden Markov model structure for information extraction. In: AAAI '99 - Workshop on Machine Learning for Information Extraction