Приказ летње школе о претраживању великих повезаних скупова података заснованом на кључним речима „Keyword search in Big Linked Data“

Марија Радојичић
Ранка Станковић
Универзитет у Београду
Рударско–геолошки факултет

Себастијан Каплар
Универзитет у Новом Саду
Факултет техничких наука

РАД ПРИМЉЕН: 9. април 2017.
РАД ПРИХВАЋЕН: 8. мај 2017.

У оквиру COST акције[1] IC1302 под називом KEYSTONE (semantic KEYword-based Search on sTructured data sOurcEs)[2] одржана је друга истраживачка летња школа под називом „Keyword search in Big Linked Data“ у периоду од 18. до 22. јула 2016. године у граду Сантјаго де Компостела у Шпанији. Циљ школе био је да се учесницима представе актуелне теме у области која се брзо развија, а односи се на претрагу великих количина повезаних података (Big Linked Data) заснованих на кључним речима. Школа је намењена углавном дипломираним студентима и постдипломцима на почетку њихове академске каријере, а организатор је био Центар за истраживања информационих технологија[3] Универзитета Сантјаго де Компостела.[4]

Основна идеја летње школе била је анализа и управљање великим повезаним подацима што је појединачно укључило теме: велики подаци (Big Data), повезани подаци (Linked Data), обрада природног језика (NLP, Natural Language Processing), семантички веб (Semantic Web) и проналажење информација (IR, Information Retrieval). Осам истакнутих

[1] European Coopeartion in Science and Technology. www.cost.eu/COST_Actions
[2] http://www.cost.eu/COST_Actions/ict/IC1302
[3] Centro de Investigación en Tecnoloxías da Información (CiTIUS), https://citius.usc.es/
[4] Universidade de Santiago de Compostela, http://www.usc.es/

предавача из ових области припремили су интересантан материјал, који је јавно доступан[5]. Летњу школу похађало је 38 учесника из тринаест земаља са три континента – Европе, Америке и Африке. Летња школа је организована тако да су учесници током преподнева имали прилику да чују истакнуте предаваче са различитих европских универзитета, док су у оквиру поподневних радионица на практичним примерима могли да испробају научено. Такође, учесници су имали прилику да чују и искуства представника Hewlett Packard-a као и фирме E˜Xenio о теми „Keyword search in Big Linked Data".

Лаура По, предавач на одсеку за инжењерство Универзитета Модена и Ређо Емилија[6] из Италије, говорила је о истраживању, визуелизацији и постављању упита над великим скуповима повезаних података. Пре свега учесницима је објашњено шта су то повезани подаци, а шта отворени подаци, како се они публикују, каква је њихова улога у семантичком вебу, након чега се говорило о истраживању и постављању упита над великим скуповима повезаних података. Посебна пажња посвећена је SPARQL[7] језику који се користи за постављање упита у циљу претраживања отворених повезаних података. Представљен је концепт чворишта за податке (DataHub), претраживање метаподатака о расположивим скуповима података, начини и формати преузимања скупова података, као и коришћење SPARQL приступних тачака (SPARQL endpoint) уколико су наведени уз конкретан скуп. Истакнута је важност визуелизације повезаних отворених података и демонстриран алат LODeX са конкретним примерима упита и резултујућих интерактивних графова.

У наставку летње школе ученици су се упознали са основним појмовима и техникама везаним за проналажење информација (IR). Кроз свеобухватно и интересантно предавање Михаила Лупуа са Техничког универзитета у Бечу[8] учесницима су предочени различити начини евалуације, као и кључни критеријуми за оцену IR. Посебна пажња посвећена је примени статистичких метода за проверу добијених резултата. О тој теми говорио је и Сергеј Зер са Универзитета Саутхемптон[9] који је својим предавањем „Collective Intelligence: Crowd-

[5] https://eventos.citius.usc.es/keystone.school/slides.html

[6] Università degli Studi di Modena e Reggio Emilia (University of Modena and Reggio Emilia), http://www.unimore.it/

[7] SPARQL Protocol and RDF Query Language

[8] Technische Universität Wien (TUW), www.tuwien.ac.at

[9] University of Southampton, www.southampton.ac.uk

sourcing Groundtruth Data for Large Scale Evaluation in Information Retrieval“, држао пажњу присутних од првог до последњег минута, покренувши их притом да се укључе у дискусију тако што је приказао занимљиве примере.

Садржајно предавање Женевјев Варга-Солар, истраживача из Француског националног центра за научна истраживања,[10] значајно је употпунило слику о досадашњим достигнућима по питању чувања и обраде великих повезаних података. Она је посебну пажњу посветила трендовима у анализи великих података, истраживању података (data mining) и науци о подацима (data science), као и о специфичностима дистрибуираног моделирања, складиштења, предиктивне аналитике, кластеровања, обраде и истраживања токова података (stream processing & mining), декларативним језицима.

Елена Демидова са Универзитета Саутхемптон упознала је полазнике летње школе са интерактивном претрагом великих структурираних скупова података заснованом на кључним речима. Претрага структурираних података путем кључних речи, наспрам класичних упитних језика (SQL), захтева претходну припрему података (индексирање) и превођење информационих потреба корисника у погодан облик. Добра страна оваквог приступа је могућност постављања упита чак и када шема базе података није позната, али је недостатак непрецизност интерпретације информационе потребе корисника јер се полазни упит изражен кључним речима преводи у (највероватнији) структурирани упит. Конструкција упита за велике базе података, на пример Freebase са преко 22 милиона ентитета, 350 милиона чињеница, 7.500 релационих табела и око 100 домена, захтева ефикасну и прошириву интерактивну конструкцију упита која је представљена теоретски и кроз примере.

Током летње школе Ранка Станковић, професорка Рударско–геолошког факултета Универзитета у Београду, одржала је предавање о језичким ресурсима. Посебну пажњу професорка Станковић посветила је проширивању упита и семантичком означавању као начинима за унапређење претраживања у смислу повећања одзива без губитка прецизности. Током предавања полазници су упознати са различитим алатима за језичку обраду. У оквиру радионице професорка Станковић

[10] Centre national de la recherche scientifique, French Council of Scientific Research (CNRS), http://www.cnrs.fr

представила је до сада развијене ресурсе за српски језик и демонстрирала њихову примену.

Мауро Драгони из Фондације Бруно Кеслер[11] последњег дана предвиђеног за предавања истакао је значај сагледавања неког документа из различитих углова и приказао неколико студија случаја које су допринеле дубљем разумевању представљених концепата.

На самом крају летње школе уследио је можда најзанимљивији део када су се учесници здружили у групе од три до четири члана како би учествовали у такмичењу Hackaton. Задатак који су учесници добили захтевао је како програмерске вештине тако и примену алата и знања представљених за време летње школе. Током такмичења коришћена су два опште позната скупа повезаних података за решавање постављених проблема: DBPedia и GeoNames. DBPedia је централни репозиторијум повезаних података, екстрахован из Википедије, који описује више од 4,5 милиона ентитета класификованих у конзистентну онтологију од којих су око 1.445.000 ентитета за особе, 735.000 за места, 411.000 за уметничка дела, 241.000 за организације. Постоје верзије за 125 различитих језика, при чему је онтологија за енглески језик највећа. Осим линкова ка сликама, спољним странама, ентитетима су придружене категорије из Википедије и YAGO онтологије, што омогућава постављање SPARQL упита над подацима изведеним из Википедије. Бројни скупови података и онтологије из репозиторијума DBPedia могу се преузети или им се може директно поступати онлајн коришћењем SPARQL приступне тачке`http://dbpedia.org/sparql`, претраживати по кључним речима коришћењем алата DBPedia Lookup[12] или се могу преузети и користити као локални повезани подаци.

Географска база података GeoNames[13] садржи преко десет милиона географских назива и девет милиона јединствених географских ентитета, класификованих у девет класа, означених са 645 маркера. Ове класе и маркери су описани GeoNames онтологијом[14] и предефинисаним кодовима.[15] Сваки елемент у GeoNames има URI и одговарајући документ са RDF XML подацима. Тако је, на пример, елемент за Београд има URI `http://www.geonames.org/792680/` и одговарајући документ `http://www.geonames.org/792680/about.rdf` у ком је, осим назива на

[11] Fondazione Bruno Kessler, `www.fbk.eu`
[12] `http://wiki.dbpedia.org/projects/dbpedia-lookup`
[13] `http://www.geonames.org/`
[14] `http://www.geonames.org/ontology/documentation.html`
[15] `http://www.geonames.org/export/codes.html`

енглеском језику, Београд описан на још 67 језика. GeoNames елементи су међусобно повезани коришћењем три типа географских релација: подређени, у смислу административних потцелина, потом суседни, на пример суседне државе и просторно блиски, на пример насеља које се налазе на малој удаљености. Тако се свим регионима у Црној Гори може приступити када се на основни њен http://www.geonames.org/3194884 дода суфикс contains.rdf.

Податке из GeoNames базе могуће је преузети и користити локално, а онлајн приступ је могућ путем кореног чвора његове хијерархије http://sws.geonames.org/6295630/about.rdf и праћењем *contains* релација, или коришћењем веб-сервиса заснованог на кључним речима.[16] Коначно, може се видети да постоје линкови од GeoNames елемената ка Википедији и DBPedia-ји.

Полазници су добили задатак да користећи репозиторијум DBPedia и GeoNames направе веб-апликацију за: 1) проналажење административних области погођених задатим ураганом, 2) проналажење свих урагана који су погодили задате административне области и 3) процењивање како су одређене области биле припремљене за задати ураган.

1. У првом сценарију корисник треба да зада кључне речи којима идентификује ураган, што укључује његово име (Катрина, Емили итд.) и опционо годину (2005, 2006. итд.). Текст о урагану треба пронаћи у репозиторијуму DBPedia, потом га издвојити, при чему претрага треба да укључи апстракт и текст који описује погођене области. Из исечака текста потом треба издвојити имена административних области коришћењем GeoNames базе и Standford NER[17] алата за препознавање именованих ентитета.
2. У другом сценарију корисник кључним речима задаје административну област, а потом се GeoNames база користи за проналажење насеља у задатој области, након чега се претражује DBPedia како би се идентификовали урагани који су погодили та насеља, односно област.
3. Трећи део је подразумевао одређивање нумеричког индикатора којим се оцењује припремљеност неке административне области за ураган коришћењем података из DBPedia-је (жртве, штета, брзина ветра, трајање, тип и слично) и броја становника у угроженим

[16] http://www.geonames.org/export/geonames-search.html
[17] http://nlp.stanford.edu/software/CRF-NER.shtml

областима пронађеним у претходном упиту. Потребно је генерисати рангирање засновано на предложеном индикатору.

Једна од стратегија српског тима за решавање датог проблема била је креирање апликације која кроз програмски језик Python користи све расположиве отворене ресурсе како би приступила датим подацима и обрадила их. Идеја је била да након што корисник одабере ураган о коме жели да сазна више информација, систем шаље SPARQL упит ка отвореном API-ју DBPedia-је који је издвајао текст апстракта и текст који описује погођене области. Након тога из резултата претходног упита су извучене погођене области и слао се нови упит ка GeoNames бази података који је враћао информације о свим погођеним насељима за дату област. Из добијених ресурса издвојене су информације попут броја становника, броја жртава, материјалне штете, брзине ветра и слично, на основу чега су се одредили нумерички индикатори који су приказали спремност неке одређене административне области за ураган. Ти нумерички индикатори су предвиђени за креирање математичког модела и његов визуелни приказ и интерпретацију.

Последњег дана летње школе, који је протекао у правом такмичарском духу и преданом рада учесника, представљена су решења екипа. Најуспешније решење је направила екипа студената мастер и докторских студија из Сарагосе, тако да су као најбоља екипа освојили вредну награду у виду стипендије за боравак и истраживање на Универзитету Сантјаго де Компостела. Евалуација система је подразумевала процену прецизности и одзива система, брзине извршавања, као и ефектности решења. Након свечаног затварања учесници су имали прилику да уживају у живописном граду Сантјаго де Компостела, као и у чарима Галиције.

Сви материјали са летње школе доступни су на адреси https://eventos.citius.usc.es/keystone.school/index.html.