Review of the 2nd KEYSTONE Training Summer School on Keyword Search in Big Linked Data

Marija Radojičić
University of Belgrade
Faculty of Mining and Geology

Ranka Stanković
University of Belgrade
Faculty of Mining and Geology

Sebastijan Kaplar
University of Novi Sad
Faculty of Technical Sciences

PAPER SUBMITTED: 9 April 2017.
PAPER ACCEPTED: 8 May 2017.

Within the framework of COST action[1] IC1302 with acronym KEYSTONE (semantic KEYword-based Search on sTructured data sOurcEs)[2], the 2nd KEYSTONE Training School named "Keyword search in Big Linked Data" was held from 18th to 22th of July 2016 in Santiago de Compostela in Spain. The aim of the school was to present current topics in the fast-growing area related to Keyword Search in Big Linked Data. The School was intended mainly for graduate and post-graduate students at the beginning of their academic careers, and was organized by the Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS)[3] of the Universidade de Santiago de Compostela (USC).[4]

The basic idea behind the summer school was the analysis and management of large linked data that especially included topics like: Big Data, Linked Data, natural language processing (NLP), the Semantic Web and information retrieval (IR). Eight prominent speakers from these areas have

[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/

prepared interesting materials, which are still publicly available.[5] The summer school was attended by 38 participants from thirteen countries, i.e. from three continents – Europe, America and Africa. The summer school was organized so that the participants had the opportunity to hear lectures by eminent lecturers from different European universities during the morning, while within the afternoon workshop they had the opportunity to try out what they have learned, through practical examples. Also, participants had the opportunity to hear the experiences of representatives of Hewlett Packard as well as the company E˜Xenia on "Keyword search in Big Linked Data".

Laura Po, lecturer at the Department of Engineering of the University of Modena and the Emilia Region[6] spoke about research, visualization and query formulation over large data sets of linked data. The participants were first introduced to the concepts of linked data and open data, how they are published, and their role in the Semantic Web, using very vivid examples. After that she talked about research and query formulation over large sets of linked data. Special attention was paid to the SPARQL[7] language used for query search over open linked data. The concept of the Datahub has been introduced, as well as metadata search for available data sets, methods and formats for dataset downloading (requiring), and the use of SPARQL endpoints, when they are available for a specific collection. The importance of visualization of linked open data was highlighted, with demonstration of the LODeX tool with specific query examples and the resulting interactive graphs.

During next session, the summer school students were introduced to the basic concepts and techniques in information retrieval (IR). Through a comprehensive and interesting lecture by Mikhail Lupu from the Technical University in Vienna[8], participants had the opportunity to learn about diverse evaluation techniques as well as the key criteria for evaluation in IR. Special attention was given to the application of statistical methods for verification of results. This issue was tackled by Sergei Zer from the University of Southampton[9] with his lecture "Collective Intelligence: Crowdsourcing Ground Truth Data for Large Scale Evaluation and Information Retrieval", which captured the attention of the auditorium from the first to the last

[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

minute, involving participants in the discussion and showing them interesting examples.

Substantial lecture of Genevieve Varga-Solar, researcher from the French National Centre for Scientific Research,[10] completed the picture of the achievements so far related to storage and processing of Big linked data. She paid special attention to trends in the analysis of Big data, data mining and the data science, as well as to the specifics of distributed modelling, warehousing, predictive analytics, clustering, treatment and research in stream processing & mining and declarative languages.

Elena Demidova from the University of Southampton introduced participants of the summer school to interactive search over large structured data sets based on keywords. Search of structured data using keywords, as opposed to the classic query language (SQL), requires prior preparation of data (indexing) and translating user information needs into a convenient form. The good side of this approach is the possibility of querying even when the database schema is unknown, with a drawback, namely imprecision of the interpretation of user information needs, because the initial request expressed by keywords translates into a (most probable) structured query. Query formulation and structuring for large databases, for example Freebase, with over 22 million entities, 350 million facts, 7,500 relational tables and about 100 domains, requires an effective and scalable interactive query construction, which was presented both theoretically and through examples. During the summer school professor of University of Belgrade, Faculty of Mining and Geology, Ranka Stanković, gave a lecture on language resources. She paid special attention to query expansion and semantic annotation, as the means for improvement of search results, in terms of increasing the response without loss of precision. During the lecture, participants were introduced to various tools for language processing. Within the workshop, Professor Stanković presented some of the resources developed for Serbian and demonstrated their application.

Mauro Dragoni from the Bruno Kessler Foundation,[11] on the last day scheduled for lectures, emphasized the importance of looking at a document from different perspectives, and presented several case studies that contributed to a deeper understanding of the concepts presented.

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

[11] Fondazione Bruno Kessler, www.fbk.eu

The summer school ended with perhaps the most interesting session, where participants organized in groups of 3 to 4 participated in the Hackathon competition. The task given to the participants required both programming skills and application of tools and knowledge presented during the summer school. Within the competition, two well-known data sets of linked data were used for solving the task: DBPedia and GeoNames. DBPedia is the central repository of linked data extracted from Wikipedia, which describes more than 4.5 million entities classified in a consistent ontology, of which approximately 1,445,000 entities about people, 735,000 related to cities, 411,000 for artworks, and 241,000 about organizations. There are versions for 125 different languages, but the ontology for English is the most used and the largest. In addition to links to images and external sites, categories from Wikipedia and YAGO ontology are associated to entities, which allows users to create and submit SPARQL queries over data derived from Wikipedia. Many data sets and DBPedia ontologies can be downloaded or directly accessed online using SPARQL `http://dbpedia.org/sparql` endpoint, as well as searched by keywords using DBPedia Lookup[12] tool, or downloaded and used as local linked data.

GeoNames,[13] the geographical database, contains over 10 million geographical names and 9 million unique geographical entities, classified into nine classes, and annotated with 645 markers. These classes and markers are described by GeoNames ontology[14] and predefined codes.[15] Each element has a GeoNames URI and a corresponding RDF document with XML data. For example, the element for Belgrade (Beograd) has URI `http://www.geonames.org/792680/` and the corresponding document is `http://www.geonames.org/792680/about.rdf`, in which, besides its name in English, Belgrade is described in additional 67 languages. GeoNames elements are linked to each other using the three types of geographical relations: subordinate, in terms of administrative sub-unities, adjacent, as for example a neighbouring country, and spatially close, as for example settlements that are located at a small distance. Thus, all regions in Montenegro can be accessed when the suffix contains.rdf is added to its base URI `http://www.geonames.org/3194884`.

The data from the GeoNames database can be downloaded and used locally, but they can also be accessed online via its root hierarchy node `http:`

[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`

//sws.geonames.org/6295630/about.rdf, and then by following *contains* relations, or by using web services based on keywords[16]. Finally, there are links from GeoNames elements to Wikipedia and DBPedia.

Participants were given the task to use DBPedia and GeoNames to create a web application for: 1) finding the administrative areas affected by a specified hurricane, 2) finding all the hurricanes that hit specified administrative areas, and 3) assessing how certain authorities were prepared to cope with a hurricane.

1. In the first scenario, the user has to specify keywords that identify the hurricane, which include its name (Katrina, Emily, etc.) and optionally year (2005, 2006, etc.). The text on the hurricane needs to be located in DBPedia, and then extracted, while the search should include the abstract and the text describing the affected area. From snippets of text, the system should extract the names of administrative areas using GeoNames database and Stanford NER[17] tool for Named Entity Recognition.
2. In the second scenario, the user defines the administrative area by keywords, after which the GeoNames database is used to find settlements in the given area, and finally DBPedia searched in order to identify the hurricanes that hit those settlements, i.e. area.
3. The third part comprised of determination of a numerical indicator that assesses the preparation level of the administrative area for a hurricane using data from DBPedia (casualties, damage, wind velocity, duration, type, and the like) and the number of inhabitants in the affected area found in the previous query. The task also included ranking based on the proposed indicator.

One of the strategies of the Serbian team for solving the given problem was the creation of an application in *Python* programming language, which uses all available open resources in order to access the given data and process them. The idea was that once the user selects a hurricane of interest, the system sends a SPARQL query to the open DBPedia API to extract texts from the abstracts and from the description of the affected area. After that, the affected areas were extracted from the results of the previous query, and a new query was sent to the GeoNames database, which retrieved information on all affected settlements for the given area. Information such as population, number of casualties, material damage, wind speed,

[16] http://www.geonames.org/export/geonames-search.html

[17] http://nlp.stanford.edu/software/CRF\begingroup\let\relax\relax\endgroup[Pleaseinsert\PrerenderUnicode{вЃθ}intopreamble]NER.shtml

and the like, were extracted from resources obtained, and formed the basis for determination of numerical indicators that showed the preparedness of a particular administrative area for the hurricane. These numerical indicators are meant for creating a mathematical model and its visual representation and interpretation.

On the last day of summer school, marked by a genuinely competitive atmosphere and devoted work of all participants, the teams presented the developed solutions. The most successful solution was developed by students of master and doctoral studies in Zaragoza, and as the best team, they won a valuable prize: the scholarships for stay and research at the University of Santiago de Compostela. The system evaluation implied the assessment of the precision and recall of the system, speed of execution, as well as the effectiveness of the solution. After the closing ceremony, participants had the opportunity to enjoy the picturesque city of Santiago de Compostela and the charms of Galicia.

All materials from the summer school are available at https://eventos.citius.usc.es/keystone.school/index.html.