A Keyword Search and Citation System for RDF Graphs

Dennis Dosso[1][0000−0001−7307−4607]

Department of Information Engineering, University of Padua
dosso@dei.unipd.it

Abstract. In recent years, the Resource Description Framework (RDF) has become the *de-facto* standard to represent heterogeneous semi-structured data on the web. RDF datasets are interrogated with SPARQL, a structured query language which is often not intuitive for the non-expert users, due to its syntax and the necessity to know the structure of the underlying graph. A simpler paradigm like keyword search can help in this regard to access these databases. Moreover, nowadays datasets constitute the backbone of the scientific research, and thus they should be cited as any other scholarly publication. RDF presents a new challenge in the automatic creation of textual citation since it lacks the structure of RDB and XML databases. In this work, we discuss the design and development of a system which will perform keyword-search on RDF graphs and, given the results, will create the textual citation for the final user.

Keywords: RDF Graphs · Keyword Search · Data Citation

1 Motivation

In this paper, we describe the general structure of a system for the extraction of data through keyword search from an RDF database and the automatic creation of a human-readable citation snippet from these data.

Keyword Search In the recent years, RDF, a family of W3C specifications for the creation of directed graph databases, has become the *de-facto* standard for the publication, the access and the sharing of data on the Web. This because it allows for flexible manipulation, enrichment, discovery, and reuse of data across applications, enterprises, and community boundaries. Recently, the growth of large knowledge-sharing communities like Wikipedia and the advances in the automated information extraction from Web pages have enabled the creation of large-scale knowledge bases [10], which are represented with RDF. Among the different RDF applications, we can count Eagle-i [17], Europeana [14], Dbpedia [3], Disgenet [15] and many others.

RDF graphs can be interrogated through the SPARQL structured language. This language is difficult for non-expert users due to its complex syntax and the necessity to know the structure of the underlying dataset in order to create the correct query pattern.

One of the two directions of the research presented in this paper is to enable non-expert users to interrogate RDF databases through the easier Keyword Search paradigm, which expects the use of a bag of words as a representation of the information need.

Among the difficulties regarding keyword search systems, as described in [8] and in [9], we can count the long execution times (more than one hour on average on-line) and the memory required by most of the systems in the current literature. These systems often cannot complete their execution even on small databases (1M triples) and thus cannot scale to real-world sizes. Our aim is the development of a keyword search system able to perform keyword queries on real-world RDF datasets.

Data Citation Today data has become fundamental for building and interpreting new scientific results. Citations are one of the most significant tools used in the creation and propagation of knowledge and one of the basic means on which scholarship and scientific publishing rely. Citations permit to identify the cited material; to retrieve it; give credit to its creator; date it; assign responsibility or ownership [7]. Today most information is published in evolving databases, datasets or, in general, in structured, evolving collections of data held online. There is strong demand that these datasets are given the same scholarly status of the traditional publications [6] and nonetheless scientific datasets are ignored by large-scale citation-based systems. As a consequence, they are not considered first-class players in the science system. In this work, we will describe a possible citation system for the automatic creation of text citations for the answer graphs produced by the keyword search system.

Outline Section 2 reports the related works in the field of keyword search and data citation in the field of RDF graphs. In Section 3 we describe the general architecture of the two modules composing the keyword search and citation system we are developing. Finally, Section 4 describes some future problems and directions that will be tackled in the development of the system.

2 Related Works

Keyword Search has been extensively studied in the context of structured databases such as Relational DB and Knowledge Bases. Good reviews about these topics are [5], [18], [20].

Regarding data citation, [2] and [4] outlined four main requirement for a data citation methodology. Among them: data is a research object that should be citable; credit should be given to data creators and curators; identification and access to the cited data should be provided and the identifier and metadata should be unique; provide persistence of the cited data (*fixity*); provide completeness of the reference; enable variable *granularity* in the data citation (i.e. enable to cite whole datasets, single units or subsets of data); produce references that are both human and machine-readable.

Regarding RDF datasets in particular, among the different proposals we find a *nano-publication model* where a single RDF triple is made citable via annota-

tions [12]; another model is based on *named meta-graph* in order to cite RDF sub-graphs [16]. [1] is restricted to generating citations for single resources in an RDF dataset.

As of today, there is not a unique system that can enable users to easily extrapolate data from RDF datasets and to generate the citation for these data.

3 Model

Figure 1 reports the overall depiction of the system, composed of two blocks: a Keyword-Search System and a Citation System. We will describe these two blocks separately.

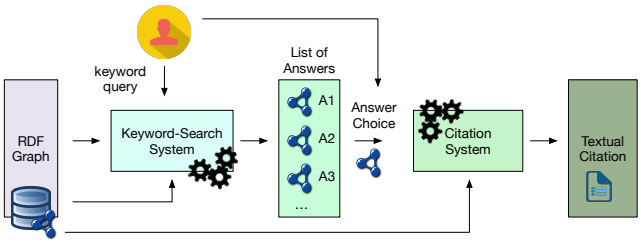

Fig. 1. Overview of our pipeline.

3.1 Keyword Search System

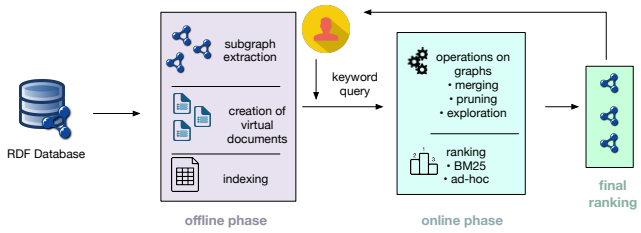

Fig. 2. General structure of a Keyword Search System for RDF.

In general, as depicted in Figure 2, keyword search systems present two modules: an off-line module, performing operations on the dataset before the user's query arrives, and the on-line module.

In our implementation, the offline module extracts subgraphs from the RDF datasets with an algorithm called TSA [11]; based on a variation of a greedy

adaptation of BFS (Breadth First Search). In this way, we obtain a collection of graphs which covers the whole dataset. We then produce textual documents (bag of words) obtained by extrapolating words from the Literals and the IRIs composing each of these subgraphs. These are called the associated *virtual documents*. These documents are also indexed at the end of the process.

Once the user's keyword query arrives, we can leverage on the index to perform a first fast ranking of the graphs using BM25 and then taking the top k element of the ranking (e.g $k = 1000$). In this way, we obtain in the first rankings the subgraphs that are more relevant to the information need of the users. Taking only the top k graphs of the ranking limits the space of potential answers and limits the required computation time. Subsequent operations on the top-k graphs include merging of graphs (when significant overlappings are present), pruning (when the presence of triples without keywords in the outer part of the graphs are detected) and a re-ranking using an MRF function [13] which takes into consideration also the structure of the answer subgraphs. This final ranking is then returned as the final answer to the user.

Thanks to this division of operations in an on-line and off-line phase we performed queries over databases of ten and even hundreds of thousands of triples (LinkedMDB, LUBM, BSBM, IMDB, Dbpedia).

3.2 Citation System

Once that the user has obtained the results for her keyword query, she may want to choose one or more of the answers that she perceives closer to her information need, and create a citation for that piece of information.

The role of the citation system is to automatically create the data citation associated with the graph chosen by the user without her direct intervention in the process. Since RDF datasets lack the structure of RDB and XML databases, it is necessary to redefine some concepts and problems of data citation, as for example the identification of the *citable unit* (the minimal element inside the graph that can be cited), the granularity of the citation (the kind of structures inside the graph that can be identified as citable).

Considering the nature of the system we are developing, the citation system should be able to cite subgraphs of different dimensions, composed of one or more triples. One possible approach can be derived from the work proposed in [12] with nano-publication. There a single RDF statement is made citable in its own right. It is enriched via annotations adding context information such as time, authority and provenance. The statement in this way becomes a publication itself, carrying all the information to be understood, validated and re-used.

The general structure of a system using nano-publication is depicted in figure 3. Ideally, in the system, every triple in the graph presents an associated nano-publication. The different nano-publications can be combined together to create the final human-readable and machine-readable citation of the answer graph. However, it is still to be studied how to aggregate the annotations over more than one triple, or how to create a citation over a set of triples only partially provided with nano-publications. An approach similar to the one presented in [19] can

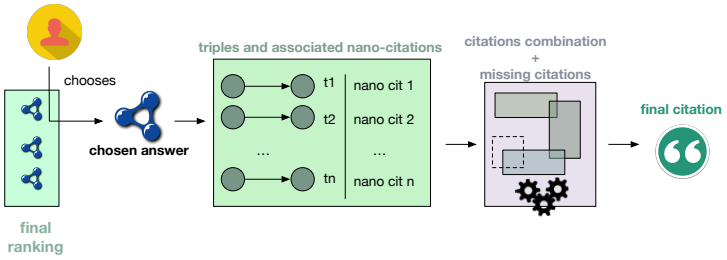

Fig. 3. Citation system for the exploitation of nano-publications.

be used, where citation semirings were deployed to deal with the combination of different citations coming from different views in order to build one unique citation, following specifications given by the DBA.

4 Conclusions & Future Directions

In this paper, we presented the two modules of a keyword search and citation system for RDF datasets that will enable users to automatically find and cite data inside an RDF dataset. Our future endeavors will be directed toward the completion and betterment of the two models.

Regarding the keyword search system, we plan to work on the creation of virtual documents from the subgraphs, introducing fields to better exploit the information included in the IRIs. Moreover, we will explore the possibility to introduce query expansion to help the ranking performed by BM25 and the MRF-based function. We will also work on the provided answers that by now are subgraphs. We will perform entity extraction and the extrapolation of NL descriptions of the answers in order to help non-expert users to read them. We will also face the critical aspect of scalability. While our system scales well in time on databases of tens and hundreds of millions of triples, we saw that the performances of effectiveness tend to decrease dramatically. We will study new methods to face the problem of big real-world databases.

Regarding the citation system, we will study methods to deal with the different granularity of the potential graphs to be cited. Among the major problems of the automatic citation of graphs, there is the semi-structured nature of RDF, which lacks the hierarchical nature of XML databases. New scalable techniques need to be designed in order to build human-readable and machine-readable citations, including the use of nano-publications and named-graphs.

Acknowledgments

This work is partially supported by the Computational Data Citation (CDC-STARS) project of the University of Padua and by the ExaMode project, as part of the European Union Horizon 2020 program under Grant Agreement no. 825292.

References

1. Alawini, A., Davidson, S.B., Silvello, G., Tannen, V., Wu, Y.: Data citation: A new provenance challenge. IEEE Data Eng. Bull. **41**(1), 27–38 (2018)
2. Altman, M., Crosas, M.: The evolution of data citation: From principles to implementation. IAssist quarterly **37** (2013)
3. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.G.: Dbpedia: A nucleus for a web of open data. In: The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007. pp. 722–735. Springer (2007)
4. Ball, A., Duke, M.: How to cite datasets and link to publications. Digital Curation Centre (2011)
5. Bast, H., Buchhold, B., Haussmann, H.: Semantic search on text and knowledge bases. Foundations and Trends in Information Retrieval **10**(2-3), 119–271 (2016)
6. Buneman, P.: How to cite curated databases and how to make them citable. In: 18th International Conference on Scientific and Statistical Database Management, SSDBM. pp. 195–203. IEEE Computer Society (2006)
7. Buneman, P., Davidson, S.B., Frew, J.: Why data citation is a computational problem. Commun. ACM **59**(9), 50–57 (2016)
8. Coffman, J., Weaver, A.C.: A framework for evaluating database keyword search strategies. In: Proc. of the 19th ACM International Conference on Information and knowledge management. pp. 729–738. ACM Press (2010)
9. Coffman, J., Weaver, A.C.: An Empirical Performance Evaluation of Relational Keyword Search Systems. IEEE Transactions on Knowledge and Data Engineering **26**(1), 30–42 (2014)
10. Doan, A., Ramakrishnan, R., Vaithyanathan, S.: Managing information extraction: state of the art and research directions. In: Proceedings of the 2006 ACM SIGMOD international conference on Management of data. pp. 799–800. ACM, ACM (2006)
11. Dosso, D.: Keyword search on RDF datasets. In: Advances in Information Retrieval - 41st European Conference on IR Research, ECIR 2019. pp. 332–336 (2019)
12. Groth, P., Gibson, A., Velterop, J.: The anatomy of a nanopublication. Information Services & Use **30**(1-2), 51–56 (2010)
13. Metzler, D., Croft, W.B.: A markov random field model for term dependencies. In: SIGIR 2005: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 472–479. ACM (2005)
14. Petras, V., Hill, T., Stiller, J., Gäde, M.: Europeana - a search engine for digitised cultural heritage material. Datenbank-Spektrum **17**(1), 41–46 (2017)
15. Queralt-Rosinach, N., Piñero, J., Bravo, À., Sanz, F., Furlong, L.I.: Disgenet-rdf: harnessing the innovative power of the semantic web to explore the genetic basis of diseases. Bioinformatics **32**(14), 2236–2238 (2016)
16. Silvello, G.: A methodology for citing linked open data subsets. D-Lib Magazine **21**(1/2) (2015)
17. Torniai, C., Bourges-Waldegg, D., Hoffmann, S.: eagle-i: Biomedical research resource datasets. Semantic Web **6**(2), 139–146 (2015)
18. Wang, H., Aggarwal, C.C.: A survey of algorithms for keyword search on graph data. In: Managing and Mining Graph Data, pp. 249–273. Springer (2010)
19. Wu, Y., Alawini, A., Davidson, S.B., Silvello, G.: Data citation: Giving credit where credit is due. In: Proceedings of the 2018 International Conference on Management of Data, SIGMOD. pp. 99–114 (2018)
20. Yu, J.X., Qin, L., Chang, L.: Keyword Search in Relational Databases: A Survey. IEEE Data Eng. Bull. **33**(1), 67–78 (2010)