Synthesizing a Knowledge Graph of Data Scientist Job Offers with MINTE$^+$

Mikhail Galkin[1,2,4], Diego Collarana[1,2], Mayesha Tasnim[2], Maria-Esther Vidal[3,5]

[1] University of Bonn, Germany
[2] Fraunhofer Institute for Intelligent Analysis and Information Systems, Germany
[3] TIB Leibniz Information Centre for Science and Technology, Germany
[4] ITMO University, Russia
[5] L3S Institute at University of Hannover, Germany
{collaran|galkin}@cs.uni-bonn.de
{mayesha.tasnim}@iais.fraunhofer.de
{maria.vidal}@tib.eu

Abstract. Data Scientist is one of the most sought-after jobs of this decade. In order to analyze the job market in this domain, interested institutions have to integrate numerous job advertising coming from heterogeneous Web sources e.g., job portals, company websites, professional community platforms such as StackOverflow, GitHub, etc. In this demo, we show the application of the *RDF Molecule-Based Integration Framework MINTE$^+$* in the domain-specific application of job market analysis. The use of RDF molecules for knowledge representation is a core element of the framework gives MINTE$^+$ enough flexibility to integrate job advertising from different web resources and countries. Attendees will observe how exploration and analysis of the data science job market in Europe can be facilitated by synthesizing at query time a consolidated knowledge graph of job advertising. The demo is available at: https://github.com/RDF-Molecules/MINTE/blob/master/README.md#live-demo

Keywords: Data Integration · RDF · Knowledge Graphs · RDF Molecules.

1 Introduction

According to the latest research, e.g., by PwC[1] and LinkedIn[2], the demand for data science professionals is still growing pushing data science and machine learning to the first places of various ratings of top emerging and sought-after jobs. Attempting to cover a broader audience of possible candidates employers disseminate jobs offers and hiring news at numerous Web sources including corporate websites, public job portals, community forums, social networks, and many more. Those Web sources exhibit a high degree of heterogeneity as there is no one agreed format of publishing job adds and vacancies. Therefore, in order

[1] https://www.pwc.com/us/en/library/data-science-and-analytics.html
[2] https://economicgraph.linkedin.com/research/LinkedIns-2017-US-Emerging-Jobs-Report

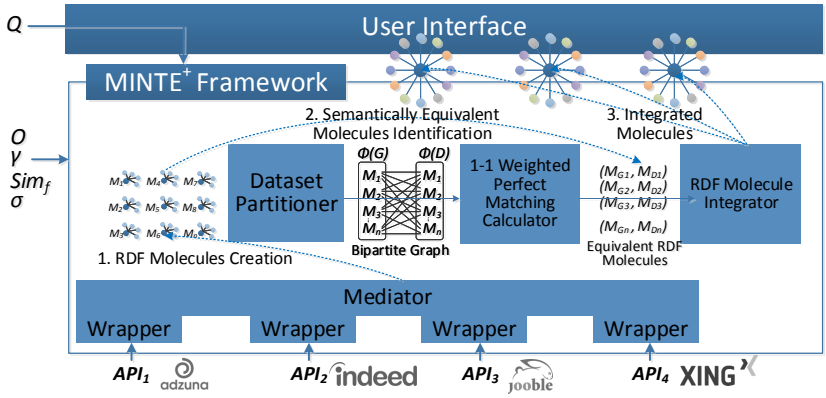


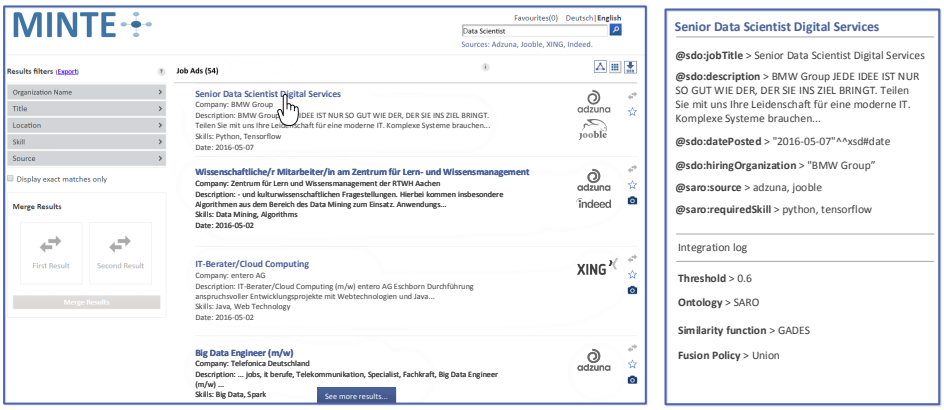
Fig. 1: **MINTE$^+$ Architecture.** Data from web sources, e.g., Adzuna, Indeed, Jooble, and XING is collected to answer a keyword query Q. A mediator creates RDF molecules from these sources; they are expressed the SARO ontology O. Dataset Partitioner builds a bipartite graph using a similarity function Sim_f, e.g., GADES and configurable threshold γ; a 1-1 perfect matching to identify equivalent RDF molecules is produced. The RDF Molecule Integrator employs a union fusion policy σ to synthesize RDF molecules into a knowledge graph.

to provide a holistic view on the job market and enable market analysis job descriptions have to be integrated using universal knowledge representation mechanisms. A flexible semantic data model provided by RDF and ontologies solves the knowledge representation task. Interoperability is tackled by a manifold of data integration frameworks [1, 2, 5]. In this demo, we demonstrate MINTE$^+$, an RDF Molecule-Based Integration Framework able to perform semantic data integration techniques in order to synthesize a knowledge graph of job adds collected from heterogeneous Web sources. Main features and applications of MINTE$^+$ are reported in Collarana et al. [1], while in this paper, the application of the MINTE$^+$ data integration techniques are illustrated in a Job Market Analysis application. Attendees of this demo will be able to examine different MINTE$^+$ components, i.e., source description, integration process configuration by tweaking semantic similarity functions, and a unified knowledge graph building.

2 Architecture

MINTE$^+$ generates RDF molecules, i.e., a set of RDF triples that share the same subject. A MINTE$^+$ knowledge graph consists of two components, i.e., ontologies and schema definitions (TBox), and RDF molecules that contain knowledge annotated with those ontologies (ABox). The integration task, therefore, is to achieve an RDF molecule representation for data gathered from heterogeneous data sources; an ontology and configurable parameters are received as input. Fig. 1 shows the MINTE$^+$ architecture and these three integration steps.

At the **RDF Molecules Creation** step, wrappers are used to collect data from data sources; a mediator utilizes an ontology O to create RDF molecules,



(a) Exploration (b) RDF Molecule

Fig. 2: **MINTE$^+$ in the Job Market Analysis application**. (a) A faceted browser user interface allows to search and analyze the data scientist job market in Europe. (b) The integration log shows configuration of the integration process.

e.g., SARO [4] for job adds. During the **RDF Molecules Integration** step, RDF molecules are partitioned in a bipartite graph using a semantic similarity function Sim_f and a threshold γ. All similarity scores less than γ are discarded. The most similar molecules in a bipartite graph are identified via the **1-1 Weighted Perfect Matching Calculator**. Finally, the **RDF Molecule Integrator** component merges identified similar RDF molecules following the rules specified in a fusion policy σ. Details of the architecture can be found at [1].

MINTE$^+$ Configuration: In this demonstration, Adzuna, Indeed, Jooble, and XING are the sources. Wrappers implemented in Scala create RDF molecules using the Web APIs provided by each Web source. The SARO ontology [4] is used for semantically describing the job ads; to decide relatedness between RDF molecules and the semantic similarity measure GADES [3] is used; for the sake of simplicity, a *union* fusion policy is followed to merge input RDF molecules into an RDF molecule that preserves all the properties of the original molecules. The demo and all the components of MINTE$^+$ are publicly available[3].

3 Demonstration of Use Cases

In this demo, MINTE$^+$ integrates publicly available data about job ads. MINTE$^+$ provides neither a monitoring service nor persistent storage, thus avoiding data protection risks by design. MINTE$^+$ serves as a back-end of a faceted browsing user interface as illustrated in Fig. 2 which visual premises will be employed during the demo. We will present the following use cases:

Building RDF molecules from Web sources. Given a keyword query, e.g., a query for some job ad as 'data scientist', we will demonstrate how wrappers

[3] `https://github.com/RDF-Molecules/MINTE`

interact with source APIs, and how the RDF molecules are built by the mediator using the SARO ontology. Attendees will be able to add new predicates to the formal job ad description and link them to particular attributes of job postings available at original Web data sources. They will also observe how RDF molecules enable the description of data gathered from heterogeneous sources.

Effects of changing integration parameters on a knowledge graph. The attendees will be able to adjust core integration parameters, e.g., a threshold of a semantic similarity function, number of attached sources, and fusion policy rules, in order to observe how a knowledge graph evolves in the ad-hoc fashion. Fig. 2b illustrates an example of an integrated RDF molecule of the same job ad published on two websites, i.e., Azduna and Jooble, obtained after applying GADES similarity function with 0.6 threshold and the union fusion policy.

Faceted knowledge graph browser. A synthesized knowledge graph of RDF molecules for unique job ads relevant for a given keyword query as presented in Fig. 2a. A graphical user interface with a faceted browser allows attendees to configure MINTE$^+$ integration parameters, apply filters on gathered predicates and values, and inspect contents of each RDF molecule including its integration log, as well as navigate and explore the knowledge graph.

4 Conclusions

The application to job market analysis is just one out of many possible applications of MINTE$^+$. This demo of MINTE$^+$ emphasizes the flexibility and advantages of the RDF molecule-based semantic integration approach. Attendees will explore the use-cases and understand the semantic integration mechanisms employed by the framework. More importantly, evidence of the relevance of creating meaningful knowledge graphs from heterogeneous sources will be provided.

Acknowledgements: Work supported by the European Commission (project SlideWiki, grant no. 688095) and the German Ministry of Education and Research (BMBF) in the context of the project InDaSpacePlus (grant no. 01IS17031).

References

1. D. Collarana, M. Galkin, C. Lange, S. Scerri, S. Auer, and M.-E. Vidal. Synthesizing knowledge graphs from web sources with the MINTE$^+$ framework. In *Accepted for publication at ISWC 2018*.
2. R. Isele and C. Bizer. Active learning of expressive linkage rules using genetic programming. *Journal of Web Semantics*, 23:2–15, 2013.
3. I. T. Ribón, M. Vidal, B. Kämpgen, and Y. Sure-Vetter. GADES: A graph-based semantic similarity measure. In *Proceedings of SEMANTICS 2016*, pages 101–104.
4. E. M. Sibarani, S. Scerri, C. Morales, S. Auer, and D. Collarana. Ontology-guided job market demand analysis: A cross-sectional study for the data science field. In *Proceedings of SEMANTICS 2017*, pages 25–32.
5. M. Taheriyan, C. A. Knoblock, P. Szekely, and J. L. Ambite. Rapidly Integrating Services into the Linked Data Cloud. In *Proceedings of the 11th International Semantic Web Conference (ISWC 2012)*.