Semantic-based Linked Data Management Platform

Yongju Lee, Hyunseung Seok, Seonghyeon Nam
School of Computer Science and Engineering
Kyungpook National University, Daegu, Korea
e-mail: yongju@knu.ac.kr, seokhyunseung@hanmail.net, connernam@gmail.com

***Abstract*—The growing number of available Linked Datasets raises a challenging data management problem, namely, how should the big data be stored and the desired resources be located. Although many platforms, architectures, and mechanisms are proposed for Linked Data, there are still a number of limitations in the area of Linked Data. We propose a novel semantic-based Linked Data management platform, which is composed of a storage-indexing-retrieval system, an ontology learning system, and an automatic composition system. This platform serves as a basis for implementing other more sophisticated applications required in the area of Linked Data.**

***Keywords*—Linked Data; platform; storage and retrieval system; ontology learning; automatic composition.**

I. INTRODUCTION

A very pragmatic approach towards achieving the Semantic Web has gained some traction with Linked Data. Linked Data refers to a set of best practices for publishing and interlinking structured data on the Web [1]. The basic idea of Linked Data is to apply the general architecture of the Web to the task of sharing structured data on a global scale. Technically, Linked Data employ Uniform Resource Identifications (URIs), Resource Description Frameworks (RDFs), and the Hypertext Transfer Protocol (HTTP) to publish structured data and connect related data that are distributed across multiple data resources.

RDF [2] is the data model for Linked Data, and SPARQL [3] is the standard query language for this model. All data items in RDF are represented in triples of the form (*subject, predicate, object*). Spurred by efforts like the Linked Open Data (LOD) project [4], a large amount of semantic data are available in the RDF format in many fields such as science, business, bioinformatics, and social networks. These large volumes of RDF data motivate the need for scalable RDF data management solutions capable of efficiently storing, searching, and integrating RDF data. This paper introduces our current project entitled "semantic-based Linked Data management platform." Work in our project focuses on the development of a storage-indexing-retrieval system, and an ontology learning system, and an automatic composition system. These topics are described in detail in Sections 2, 3, and 4.

II. STORAGE-INDEXING-RETRIEVAL SYSTEM

This section starts by providing a software architecture for the storage-indexing-retrieval system (see Figure 1). Our system consists of four subsystems: data acquisition, RDF storage, ontology construction, and analysis subsystems.

A. Data Acquisition Subsystem

Information represented in unstructured or structured form must be mapped to the RDF data model. We crawl Web sites and extract unstructured data. This procedure is based on a crawler such as Scrapy [5]. The extracted properties are then transformed into RDF triples. Structure data (e.g., relational databases) are transformed into RDF triples using the D2R Server [6]. The D2R Server is an open source tool for publishing relational databases on the Linked Data.

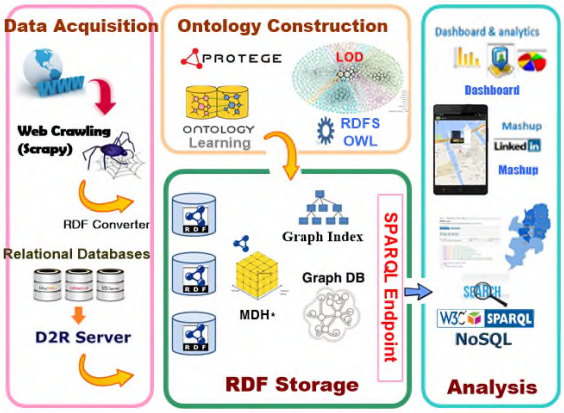

Figure 1. Architecture of storage-indexing-retrieval system.

B. RDF Storage Subsystem

Once there is a critical mass of RDF data, mechanisms have to be in place to store and index this data efficiently. Our system uses the graph-based RDF store. With this store, the platform provides a SPARQL endpoint that allows any user to access the stored RDF triples. Our system indexes the triples stored in the RDF store. These triples are mapped into multi-dimensional histograms stored in an MDH* (Multi-Dimensional Histograms for Linked Data) index structure [7].

C. Ontology Construction Subsystem

RDF Schema (RDFS) and Web Ontology Language (OWL) are key semantic Web technologies that provide a way to write down rich descriptions of your RDF data. Protégé [8] is a leading ontological engineering tool. In spite of using this ontological engineering tool, the construction of ontologies is a very expensive task which hinges on the availability of domain experts. In Section 3, we investigate an ontology learning method to generate ontologies automatically.

D. Analysis Subsystem

A number of applications provided the browsing Linked Data (e.g., Tabulator [9], Marbles [10], Magpie [11]), advanced searching facilities (e.g., Sindice [12], Sig.ma [13], Watson [14]), and meshup Linked Data (e.g., DbpediaMobile [10], LinkedGeoData [15], ActiveHiring [16]). Nevertheless, these applications hardly go beyond presenting together data gathered from different sources. Our system provides search capabilities, such as SPARQL and NoSQL (Not Only SQL) over the RDF triples. In addition, we can advance to build various applications based on Linked Data. For instance, we

Copyright (c) IARIA, 2019. ISBN: 978-1-61208-728-3

implement an aggregation of dashboards that presents various business analytics computed on the Linked Data [17].

III. ONTOLOGY LEARNING SYSTEM

The successful employment of Linked Data is dependent on the availability of high quality ontologies. Building such ontologies is difficult and costly, thus hampering Linked Data deployment. This research automatically generates ontologies from RDF datasets and their underlying semantics. We focus on adapting Linked Data mining techniques to the syntactic descriptions of RDF triples. Since RDF was not designed for the ontology, it does not provide placeholders for high level syntaxes of the resources. We propose an ontology learning method to semantically describe Linked Data (see Figure 2).

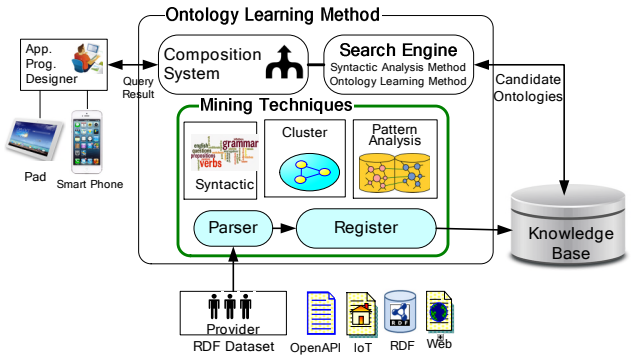

Figure 2. Ontology learning system.

We have developed a clustering technique [18] to derive several semantically meaningful concepts from RDF triples. We consider the syntactic information that resides in triples and apply a mining algorithm to obtain their underlying semantics. The main idea is to measure the co-occurrence of terms and then cluster the terms into a set of concepts. The pattern analysis technique [18] also captures relationships between terms contained in triples and matches items if both terms are similar and the relationships are equivalent. The ontology is generated from the set of triples created in accordance with the pattern analysis rules.

IV. AUTOMATIC COMPOSITION SYSTEM

With the growing popularity of Linked Data, the number of RDF datasets has increased significantly. As a result, finding and composing the right resources has become an increasingly complex task. Recent approaches have stressed the importance of Linked Data composition, data mediation, and the semantic annotations of Linked Data. Although they try to overcome the limitations of traditional mashup solutions, there are several challenging issues. First, because LOD cloud may have a large number of datasets, manually searching and composing RDF triples can be a tedious and time-consuming task. Therefore, developers wish to quickly find the desired items and easily integrate them. Second, portal sites typically only support keyword or category search. The keyword search is insufficient because of bad recall and precision. Returned lists from the category search are generally based on criteria that have no relevance to the developer's desired goals. To create mashups more efficiently, a semantic-based approach is needed such that agents can reason about the capabilities of the items that permit their discovery and composition. Third, most mashup developers want to figure out all the intermediate steps needed to generate the desired mashup automatically. An infrastructure that allows users to provide interesting or relevant composition candidates is needed.

Our research investigates algorithms for automatic Linked Data discovery and composition using the ontology learning method [19]. A common issue is how to locate the desired items. Efficient discovery can play a critical role in conducting further RDF composition. Our discovery algorithm adopts strategies that rapidly filter out items that are guaranteed not to match the query. The composition algorithm consists of constructing a Compatible Similarity Graph (CSG) and searching composition candidates. The composition process can be described as generating Directed Acyclic Graphs (DAGs) that can produce the output satisfying the desired goal. The DAGs are gradually generated by forward and backward searching over the graph (see Figure 3).

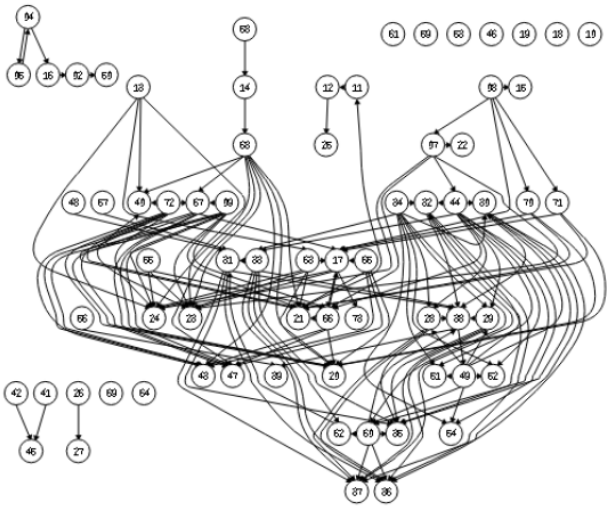

Figure 3. Compatible similarity graph for Linked Data composition.

V. CONCLUSION

The evolution of Linked Data has created a strong wave of research approaches in Semantic Web community. We introduced a semantic-based Linked Data management platform, which consists of the storage-indexing-retrieval system, onto-logy learning system, and automatic composition system. In this paper, we describe briefly the overview of our ongoing project. Our platform has an integrated hybrid architecture which supports the whole life-cycle of Linked Data from data acquisitions, building ontologies, data storage and retrieval, and implementing applications. The main components of the platform are open source in order to facilitate wide usage and ease the scalability. This proposal is a first phase of our research aiming at increasing tool coverage and developing real Linked Data applications.

ACKNOWLEDGMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. 2016R1D1B02008553).

Copyright (c) IARIA, 2019. ISBN: 978-1-61208-728-3

REFERENCES

[1] S. Auer, J. Lehmann, A. N. Ngomo, and A. Zaveri, "Introduction to Linked Data and Its Lifecycle on the Web," Proc. 9th Int. Conference on Reasoning Web, Aug. 2013, pp. 1-9.

[2] W3C. *Resource Description Framework (RDF)*. Available from: https://www.w3.org/RDF/ 2014.03.15

[3] W3C. *SAPRQL 1.1 Query Language*. Available from: https://www.w3.org/TR/sparql11-query/ 2013.03.21

[4] Lod-cloud.net. *The Linked Open Data Cloud*. Available from: https://lod-cloud.net/ 2019.03.01

[5] Scrapy. *Scrapy*. Available from: https://scrapy.org/

[6] C. Bizer and R. Cyganiak, "D2R Server – Publishing Relational Databases on the Semantic Web, Poster at the 5th Int. Semantic Web Conference (ISWC), Nov. 2006.

[7] Y. Lee and S. YuXiang, "Hybrid Index Structured on MBB Approximation for Linked Data," Proc. 10th Int. Conference on Computer Modeling and Simulation, Jan. 2018, pp. 101-104.

[8] Stanford University. *A Free, Open-source Ontology Editor and Framework for Building Intelligent Systems*. Available from: https://protege.stanford.edu/

[9] T. Berners-Lee, Y. Chen, L.Chilton, D. Connolly, R. Dhanaraj, J. Hollenbach, A. Lerer, and D. Sheets, "Tabulator: Exploring and Analyzing Linked Data on the Semantic Web," Proc. 3rd Int. Semantic Web User Interaction Workshop, Nov. 2006.

[10] C. Becker and C. Bizer, "DBpedia Mobile: A Location-Enabled Linked Data Browser," Proc. 1st Workshop about Linked Data on the Web (LDOW), Apr. 2008.

[11] M. Dzbor, J. Domingue, and E. Motta, "Magpie – Towards a Semantic Web Browser," Proc. 2nd Int. Semantic Web Conference (ISWC 2003), Lecture Notes in Computer Science (LNCS), vol. 2870, pp. 690-705, 2003.

[12] G. Tummarello, R. Delbru, and E. Oren, "Sindice.com: Weaving the Open Linked Data," Proc. 6th Int. Semantic Web Conference (ISWC 2007), Lecture Notes in Computer Science (LNCS), vol. 4825, pp. 552-565, Nov. 2007.

[13] G. Rummarello, R. Cyganiak, M. Catasta, S. Danielczyk, R. Delbru, and S. Decker, "Sig.ma: Live Views on the Web of Data," Proc. 19th Int. Conference on World Wide Web (WWW 2010), Apr. 2010, pp. 1301-1304.

[14] M. d'Aquin and E. Motta, "Watson, More Than a Semantic Web Search Engine," Semantic Web Journal, vol. 2, no. 1, pp. 55-63, Jan. 2011.

[15] C. Stadler, J. Lehmann, K. Höffner, and S. Auer, "LinkedGeoData: A Core for a Web of Spatial Open Data," Semantic Web Journal, vol. 3, no. 4, pp. 333-354, Oct. 2012.

[16] A. D. Mezaour, J. Law-To, R. Isele, T. Schandl, and G. Zechmeister, "Revealing Trends and Insights in Online Hiring Market Using Linking Open Data Cloud: Active Hiring a Use Case Study," Proc. 11th Int. Semantic Web Conference: Semantic Web Challenge (SWC 2012), submission 9, Nov. 2012.

[17] E. Jung and Y. Lee, "Linked Data Based Storage/Application Platform and Implementation of Analysis System Visualization," Journal of KIIT, vol. 16, no. 9, pp. 95-102, Sep. 2018.

[18] Y. Lee, "Semantic-based Data Mashups using Hierarchical Clustering and Pattern Analysis Methods," Journal of Information Science and Engineering, vol. 30, no. 5, pp. 1601-1618, Sep. 2014.

[19] Y. Lee, "Semantic-based Web API Composition for Data Mashups," Journal of Information Science and Engineering, vol. 31, no. 4, pp. 1233-1248, Jul. 2015.

Copyright (c) IARIA, 2019. ISBN: 978-1-61208-728-3