Synonym Predicate Discovery for Linked Data Quality Assessment Without Requiring the Ontology Semantic Relations

Samah SALEM

Lire Laboratory

Abdelhamid Mehri – Constantine 2 University, Constantine, Algeria

Presentation at the 6th International Conference on Big Data, Small Data, Linked Data and Open Data, February 23-27, 2020 - Lisbon, Portugal

OUTLINE

01 Motivation

02 Introduction

03 Quality Assessment Approach

04 Validation

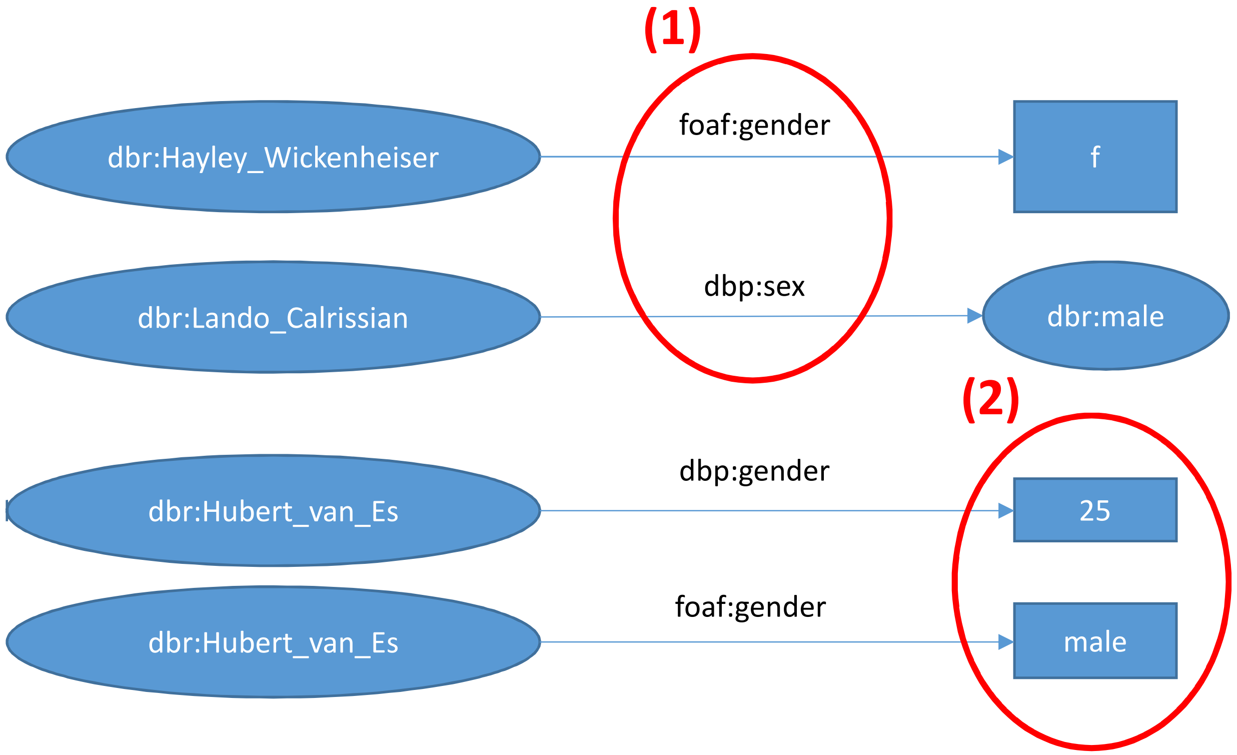
05 Limitations

06 Conclusion and Future work

Motivation

Web of Data

- **Goal:** Link and publish data using typed links to constitute a global network of information
- **Characteristics:** Evolution, heterogeneity, and usefulness
- **Challenge:** Quality problem
  - ✓ Duplicate predicates … (1)
  - ✓ Inaccurate values … (2)
  - ✓ Etc.



Motivation

Web of Data Quality

- Linked Data (LD) quality assessment approaches: ***with*** or ***without*** ontology

Approaches	Goal	Quality of	Quality dimensions	With/ without ontology
Lei et al., 2007	Quality assessment of semantic metadata	Metadata	Accuracy, consistency, conciseness	With ontology
Fürber and Hepp, 2011	Quality assessment of published data	Literal	Accuracy, completeness, uniqueness, timeliness	With ontology
Kontokostas et al., 2014	DBpedia quality assessment	Triple	-	With ontology
Spahiu et al., 2016	Summarize the content of a dataset and reveal data quality problems	Predicate	Accuracy, completeness, timeliness	With ontology
Jang et al., 2015	Linked data quality assessment	Triple	Accuracy and consistency	Without ontology

But,

- Most approaches, based on the ontology, such as Luzzu [1], SWIQA [2], RDFUnit [3], etc.
- Many datasets are without ontology or with an incomplete one

Motivation

- **What about the quality of datasets without schema/ ontology?**
  - Jang et al. [4] approach
  - Assess the quality of LD without requiring ontology
  - Data quality pattern [3]: DQP, RQP, and TQP
- **But,**
  - Lack of specific domain/ range setting
  - Quality assessment with only one triple
  - No quality improvement after detecting quality problems is incorporated

Introduction

- **Goal**

  - Assess the quality of triples by detecting errors and eventually measuring the error rate, without using the ontology information

Real-world data

```
dbr:Hayley_Wickenheiser foaf:gender “f”
dbr:Lando_Calrissian dbp:sex dbr:male
dbr:Hubert_van_Es foaf:gender 25
dbr:Hubert_van_Es foaf:gender “male”
```

☑

Schema

```
foaf:gender owl:sameAs dbp:sex
foaf:gender rdfs:domain Person
foaf:gender rdfs:range String
dbp:sex rdfs:domain Person
foaf:gender rdfs:range ObjectType
```

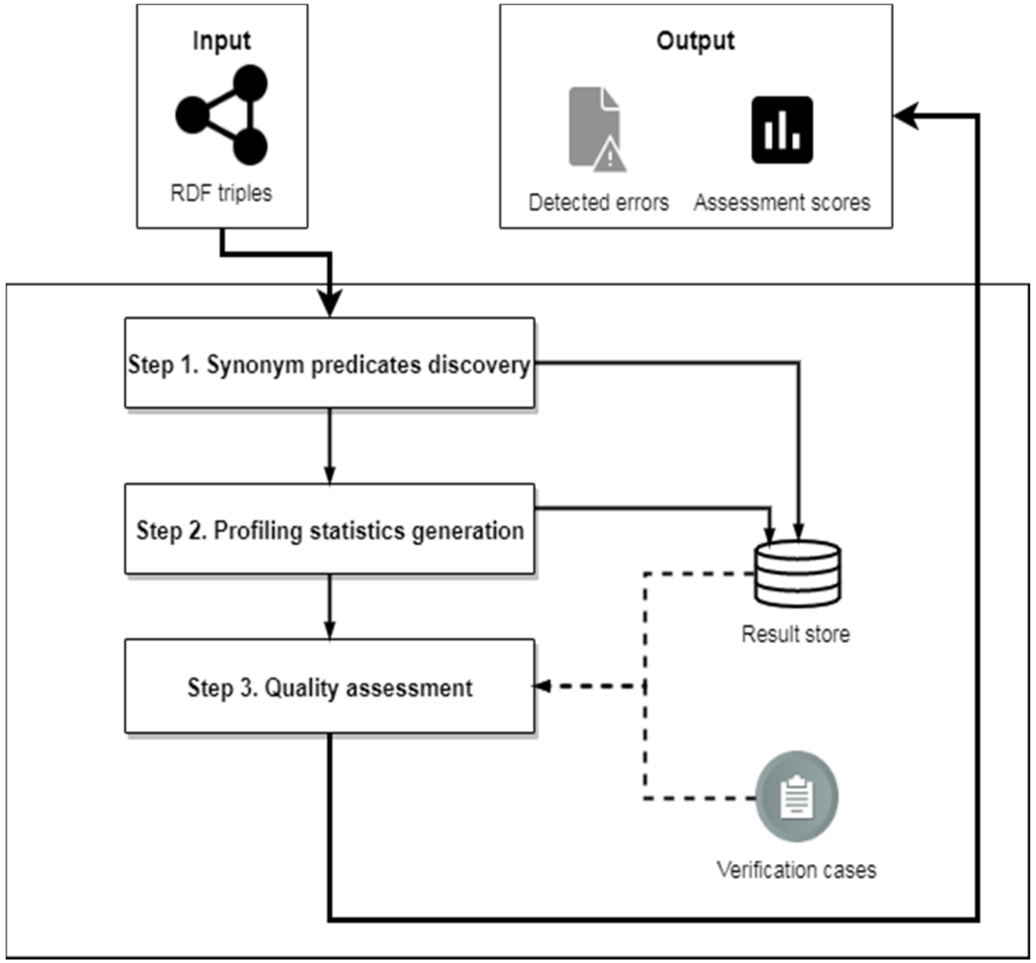
☒

  - Understand the dataset
  - Enrich the dataset with metadata

Introduction

- **Idea**
  - A large number of predicates have relationships with each other
  - The possibility of finding two or more predicates, which have the same meaning is very high
  - e.g. *foaf:nick* and *dbp:nickname*
  - Evaluate the quality based on the discovered synonyms

Quality Assessment Approach


Input
RDF triples
Output
Detected errors
Assessment scores
Step 1. Synonym predicates discovery
Step 2. Profiling statistics generation
Step 3. Quality assessment
Result store
Verification cases

Quality Assessment Approach

Step 1: Synonym predicates discovery

- Semi-automatic
- Based on natural language processing methods
  - Thesaurus-based: WordNet
  - Check spelling methods: Ispell [5], Aspell [6], and MySpell [7]
- Detect quality issues
- Semantic relationships overview

Quality Assessment Approach

Step 2: Profiling statistics generation

- Generate synonym-pattern:
  - a summary that provides a global view of the synonym predicates in the dataset and the predicate frequency $< p_i(\sum p_i) \equiv_{syn} p_j(\sum p_j) \equiv_{syn} p_n(\sum p_n) >$
  - e.g. <dbo:birthplace (13), dbp:birthCity (2)>
- Calculate simple profiling statistics, such as
  - the total number of triples in a dataset
  - the property occurrence
- Purpose: Quality score estimation

Quality Assessment Approach

Step 3: Quality assessment

- Quality problems detection
- Quality score estimation

Quality Assessment Approach

Quality problems detection

- Based on synonym predicates and Quality Verification Cases
- Quality Verification Cases
  - Verify the similarity or the difference between the subject and the object of each predicate synonyms pair
  - $t_i(s_i, p_i, o_i) \wedge t_j(s_j, p_j, o_j) \mid p_i \equiv_{syn} p_j$
  - *Case 01:* $If\ s_i = s_j \wedge o_i = o_j \Rightarrow \{p_i(o_i, s_i) \Leftrightarrow p_j(o_j, s_j)\}$: **$t_i$ or t_j is a redundant triple**
  - *Case 02:* $If\ s_i = s_j \wedge o_i \neq o_j \Rightarrow \{p_i \Leftrightarrow p_j\}$: **oi and/ or oj is an inaccurate value**
  - *Case 03:* $If\ s_i \neq s_j \wedge o_i = o_j \Rightarrow \{p_i \Leftrightarrow p_j\}$: **oi and/ or oj is an inaccurate value**
  - *Case 04:* $If\ s_i \neq s_j \wedge o_i \neq o_j \Rightarrow \{p_i \Leftrightarrow p_j\}$: **duplicate information in order to define the same predicate in the dataset**

Quality Assessment Approach

Quality scores estimation

- Based on the existing quality score metrics
- Three quality scores:
  - $\mathrm{QScore} = \mathrm{A_t} \;/\; \mathrm{T_t}$
  - $\mathrm{Acc} - \mathrm{QS} = \mathrm{PA_t} \;/\; \mathrm{T_t}$
  - $\mathrm{Co} - \mathrm{QS} = \mathrm{PC_t} \;/\; \mathrm{T_t}$

Validation

- DBpedia released in 2019
- Properties of 449 triples
- Available at GitHub repository: https://github.com/SalemSamah/SPDiscovery
- **Synonym predicates generation**
  - The experiment revealed several cases of unknown synonymous relationships

DBpedia Person	
foaf:name	dbp:name
dbo:birthplace	dbp:birthCity
dbo:birthDate	dbp:birthdate
foaf:gender	dbo:gender
dbo:occupation	dbp:occupation

Validation

- **Quality problems detection**
  - 50 abnormal triples that present an error rate equal to 11 %
  - The abnormal triples: redundant predicates, redundant triples, and inaccurate values
  - Quality dimensions: accuracy, and conciseness

Triples pairs with synonym predicates	Error type	Quality dimension
dbr:Duduka_da_Fonseca, dbo:birthplace, dbr:Rio_de_Janeiro dbr:Duduka_da_Fonseca, dbp:birthCity, dbr:Rio_de_Janeiro	*Case 01:* The results show that the two triples are equivalent, which means that one of these two triples is redundant.	Conciseness
dbr:Paulie_Pennino, foaf:gender, "female"@en dbr:Paulie_Pennino, dbo:gender, dbr:Male	*Case 02:* The sex of the entity dbr:Paulie_Pennino is inaccurate in one of these two triples since once is defined as “female”, and once is defined as dbr:Male	Accuracy/ Conciseness
dbr:Cornelia_(wife_of_Caesar), dbp:diedPlace, dbr:Rome dbr:Aloysius_Lilius, dbo:deathPlace, dbr:Rome	*Case 03:* The predicates dbp:diedPlace and dbo:deathPlace are defined differently despite that they have the same meaning	Conciseness
dbr:Alice_Walker, foaf:gender, "female"@en dbr:Zack_Addy, dbo:gender, dbr:Male	*Case 04:* In this case, there is duplicate information in order to define the same predicate in the dataset	Conciseness

Limitations

- Lack of specific setting when the predicate values are represented with different patterns
  - e.g. dbr:Julius_Caesar, dbo:birthdate, ‘−100 - 07 - 13’

    dbo:birthdate, ‘− 100 - 7 - 13’
  - these triples are identified in ***Case 02***, however, they should be identified in ***Case 01***
- No quality improvement after detecting quality problems is incorporated

Conclusion & Future work

- Understand the semantics between properties, detect quality problems and estimate the quality scores, without requiring the existence of the ontology information
- Quality issues detected: inaccurate values, redundant predicates, and redundant triples
- Generates semi-automatically the synonym predicates

Ongoing research:

- Applying the approach on large datasets
- Defining more varied metrics
- Improving the quality of data

Thank you for attention

References

[1] J. Debattista, S. Auer, and C. Lange, “Luzzu--A Framework for Linked Data Quality Assessment,” In 2016 IEEE Tenth International Conference on Semantic Computing (ICSC), (pp. 124-131), IEEE, February 2016.

[2] C. Fürber and M. Hepp, "Swiqa-a semantic web information quality assessment framework," In ECIS, Vol. 15, pp. 19-31, 2011.

[3] D. Kontokostas et al., "Test-driven evaluation of linked data quality," In Proceedings of the 23rd international conference on World Wide Web, pp. 747-758, ACM, April 2014.

[4] S. Jang, M. Megawati, J. Choi, and M. Y. Yi, "Semiautomatic quality assessment of linked data without requiring ontology," In NLP-DBPEDIA@ ISWC, pp. 45-55, October 2015.

[5] R. E. Gorin, P. Willisson, W. Buehring, and G. Kuenning, "Ispell, a free software package for spell checking files," The UNIX community, 1971.

[6] K. Atkinson, "GNU Aspell," 2003, URL http://aspell. Net, 2011.

[7] C. Andrea, "My spell-checker’s «weigh» with words," The Christian Science Monitor, August 2002.