Validation of entity relations in linked data

Aimilios Vourliotakis
Taiger
Madrid, Spain
Email: aimilios.vourliotakis@taiger.com

Michael Rovatsos
University of Edinburgh
Edinburgh, UK
Email: mrovatso@inf.ed.ac.uk

Carlos Ruiz Moreno
Taiger
Madrid, Spain
Email: carlos.ruiz@taiger.com

***Abstract*—The Web is continuously enriched with new information in the form of interconnected datasets of structured data using the *linked data* methodology. In order to develop systems that are capable of making the most of this phenomenon, it is imperative that we provide techniques capable of identifying and, ideally, resolving contradictions within it. This paper presents a formal definition of the problem, which involves detecting and resolving asserted erroneous relations between entities by determining hidden contextual features that refute them. As the type of heterogeneity in data can vary, any principled methodology has to offer different approaches to solving it. We describe existing approaches as well as our own planned future work on the subject, which will be focused on the scenario where the relation denotes that two entities are referring to the same real world entity.**

I. Introduction

The adoption of the *linked data* [1] paradigm for authoring, inserting and linking explicit and machine-readable data on the Web has lead to the creation of a "Web of Data", in which new datasets are continually created and published by a vast range of different authors. The Linked Open Data cloud[1](LOD) project has been the predominant example of this change, containing 1014 datasets from different domains, such as Social Web, Media and Government, according to a study conducted in April 2014 [2].

These datasets are mainly comprised of assertions about entities and their connections using RDF [3]. An RDF triple is of the form *subject-predicate-object*, where *subject* is an entity described by an unique identifiers IRI, *object* can either be another entity or a literal value and *predicate* denotes their connection. Ontology authors can either use existing namespaces as vocabularies of predicates (such as FOAF[2] or Dublin Core[3]) or create their own.

Basic structural hierarchies can be created using the predicate `rdf:type` to declare that an entity is of a certain type. In order to define more complex semantic relationships between classes, and properties, more expressive data-modelling languages are used, such as RDF Schema [4] and OWL [5]. Different datasets are linked through RDF triples, in which *subject* and *object* are entities of the two datasets and *predicate* is the connection between them. According to [2] the most commonly used predicate in linking datasets is `owl:sameAs`, which denotes that two entities refer to the same real-world entity.

As each dataset creator might require or have different amounts of information about an entity, the Open World Assumption (OWA) is made, allowing for incomplete representations of the same entity. Moreover, the decentralised nature of the LOD cloud makes any form of Unique Names Assumption (UNA) unusable, as each author is allowed to use his/her own naming convention. In addition to this, dataset interlinking depends on the effort and opinion of the author.

Assume an information extraction application consumes linked data through a SPARQL endpoint of a specific dataset. The application's goal is to retrieve information about an entity. The quality of its output depends on the consistency of knowledge regarding the entity that is retrieved, which in turn is comprised of the relations it has with other entities at dataset level (as RDF statements containing the entity) and between datasets (as RDF statements containing all entities connected to the entity via links such as `owl:sameAs`)[4]. Our goal is to detect contradictions created by these relations and to resolve them by determining the hidden contextual features that refute them.

Consider an information extraction application that is looking for information regarding the entity $e3$ (`Japan`) and discovers two entities, $e1$ (`Tokyo`) and $e2$ (`Kyoto`), connected to it through the predicate `capitalOf`. The combined information from this collection of RDF statements is wrong, as we know that the capital of Japan is Tokyo. The problem, in this case, arises from the definition of the predicate `capitalOf` (or its interpretation by the creator of the RDF triples). By adding time as a contextual feature that separates the two triples, we would be able to resolve the contradiction.

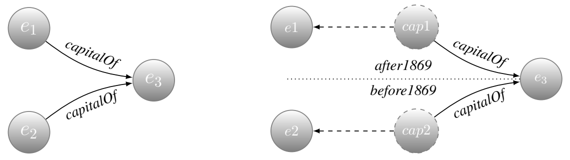

Fig. 1. Illustration of the problem and our desired goal, in which the contextual feature of time is able to resolve the contradiction as a distinguishing feature of the shared relation *capitalOf*.

Our interest lies in resolving the contradiction at consumption time. As the LOD cloud keeps expanding, any efforts to try and identify interlinking errors or impose "stricter" modelling constraints in it are bound to face scalability issues. Instead, if the focus shifts to creating efficient techniques capable of

[1] http://lod-cloud.net/

[2] http://xmlns.com/foaf/spec/

[3] http://dublincore.org/documents/dcmi-terms/

[4] The two cases have subtle yet distinguishing differences. Although the word "relation" is used in both, they refer to a different type of relation. This notion is examined further in Section II

detecting and resolving contradictions at consumption level, the quality of knowledge consistency throughout the cloud will not be as sensitive to interlinking errors as it is today. In this paper, we present an outline of a system that detects and resolves erroneous equality relations by combining a number of different similarity techniques. The envisioned system would also be capable of providing supporting evidence through the estimation of granularity properties as distinguishing properties among the two entities.

The remainder of this paper is structured as follows: In Section II, we provide a formal definition of the problem. Section III contains an overview of the proposed system in terms of requirements specification, implementation and evaluation. In Section IV, we present a number of similar tasks such as entity disambiguation and ontology matching, and provide a thorough analysis of [6] as the most closely related work. In Section V, we specify our short-term goals and our overall vision for the future.

II. Problem Definition

Assume that real-world entities are uniquely catalogued in a set $O = \{o_1, \ldots, o_n\}$. These entities are connected via binary relations $R = \{r_1, \ldots, r_m\}$, such that each $r_j \subseteq O \times O$. *Linked data* can be described by a graph $G = (V, E)$ which consists of a set of nodes $V = \{v_1, \ldots, v_k\}$ and a set of edges $E\{e_1, \ldots, e_l\}$. Both nodes and edges have names. The latter are taken from a set of relation types $T = \{\tau_1, \ldots, \tau_l\}$, such that every potential edge has a set of types $\tau(e) \subseteq T$ attached to it. This formalisation allows for capturing multiple relations in a single edge.

Assume we have a mapping $\mu : O \cup R \rightarrow V \cup E$ such that $\mu(o) \in V$ and $\mu(r) \in T$, with $\forall r, o, o' \ .\mu(r) \in \tau((\mu(o), \mu(o')))$, i.e. the edge corresponding to (o, o') includes the type corresponding to each relation for which $r(o, o')$ holds.

$$\begin{tikzcd}
o \arrow[r, "r"] \arrow[d, "\mu"'] & o' \arrow[d, "\mu"] \\
v \arrow[r, "\mu(r)"'] & v'
\end{tikzcd}$$

Fig. 2. A relation r between two real world entities o and o', described in *linked data* through a mapping μ.

For simplicity, we can write the type itself as a relation (in G), by defining $t = \{e \in E | t \in \tau(e)\}$ in slight abuse of notation, so that $\mu(r(o, o')) = \mu(r)(\mu(o), \mu(o'))$.

From a *linked data* perspective, the sum of different relations of the form $\mu(r)(u, u')$ form a set of node-based assertions about the world $A = \{a_1, \ldots, a_p\}$. For instance, the entities `Nick`, `Helen` and a relation `hasSister` form the assertion `hasSister(Nick,Helen)`.

Given two entities v_1 and v_2 and a collection of assertions A about them, a relation between v_1 and v_2 can be supported by a subset of A which indicates that there exist similar assertions regarding v_1 and v_2. For example, the collection of assertions:

```
hasSister(Nick,Helen)
hasSister(George,Helen)
```

might suggest a relation between `Nick` and `George`, due to their similar `hasSister` relations with `Helen`.

Our focus is on detecting and resolving contradictions that arise from asserted relations between *linked data* entities ($V \cup E$) and the real world ($O \cup R$). The questions we want to answer are the following:

A(i) Given two entities v_1, v_2 that are members of an ontology and a asserted relation between them, is the relation valid?

A(ii) If the outcome of (i) is false, is there a way to determine the hidden contextual features that refute the relation?

The open nature of *linked data* introduces a more specific problem when the relation is about entity equality. In the absence of a Unique Names Assumption (UNA), a real world entity can be mapped to multiple nodes which are pairwise connected via equality relations (denoted by $\approx$) such as `owl:sameAs`.

$$\begin{tikzcd}
& o \arrow[dl, "\mu"'] \arrow[dr, "\mu"] & \\
v \arrow[rr, leftrightarrow, "\approx"'] & & v'
\end{tikzcd}$$

Fig. 3. An equality relation between two *linked data* entities entails that they refer to the same real world entity.

As a first attempt, we choose to concentrate on equality relations, which are specific types of entity relations used to denote that two nodes refer to the same real-world entity. As such, a real world entity can be mapped to multiple nodes which are pairwise connected via equality relations (denoted by $\approx$) such as `owl:sameAs`. With respect to equality relations, our problem statement is modified as follows: Given two entities v, v' and an equality relation linking them:

B(i) Is the equality relation property valid?

B(ii) If the outcome of (i) is false, what are the separating features that prove v,v' refer to two different real-world entities?

III. Methodological Issues

The system will receive as input an equality relation regarding two entities v and v' that are members of an RDF graph. After determining whether the relation is valid it will output the result along with supporting evidence. The operation will be carried out in real time, and the two crucial features in its design will be speed and accuracy.

With regard to equality validation, the system will follow a probabilistic approach, in the sense that its outcome will be an estimation of the probability that the equality assertion is valid, $P(v \approx v')$. The probability will be calculated from a probability distribution over different similarity estimators. More specifically, the following will be considered:

- Syntactic Similarity: $\sigma_{syn}(v \approx v')$, which will be a measure of similarity between names using string-based techniques, such as string distance.
- Structural Similarity: Similarly to [6], we will create fuzzy structural definitions of entities as $k-$deep subgraphs around them including nodes and edges, denoted by $N^k(v)$. A discount factor $\gamma < 1$ will be

used to emphasise decreasing "relatedness" over distance. We can then apply graph-based techniques, such as graph isomorphisms, over a sum of different k's for the two entities, calculating the measure of similarity as $\sum_k \gamma^k \sigma_{struc}(N^k(v), N^k(v'))$.

- Semantic Similarity: In Social Networks, homophily is a term used to indicate that people with similar characteristics are more connected [7]. Inspired by this notion, we can estimate a probability of similarity proportional to the number of shared and equal connections with other nodes. Assuming $\equiv_V$ and $\equiv_T$ as already known node and type equivalences in respect, we can define semantic similarity σ_{sem} as equal to the number of connections with other nodes $\hat{v}$,$\hat{v}'$ that have been matched and which they share, $\sigma_{sem} = |\{(\hat{v}, \hat{v}')|\hat{v} \equiv_V \hat{v}' \wedge t \equiv_T t' \wedge t \in \tau(v, \hat{v}) \wedge t' \in \tau(v', \hat{v}')\}|$.

With respect to our problem statement, the overall similarity probability calculated from the aforementioned similarity estimators would enable us to address B(i), but provides little to no insight about the features that distinguish the two entities if B(i) is false.

Assuming a *linked data* dataset in which the UNA holds, there exists a set of relation types $\tau_{prim} \subseteq T$ which has a different set of values for every entity v, i.e $\exists ! v : \tau_{prim}(v)$. In other words τ_{prim} functions just as keys in databases, uniquely identifying different entities. Inspired by this notion, we argue that in a system in which the UNA does not hold, properties can *hint* at potential distinction between two entities, in relation to the different unique values it has. Consider two entities which both have the properties *passportNo* and *hairColour*, but with different values. Looking at the whole population, each *hairColour* value can be assigned to a lot more people than the each *passportNo* number.

Let G_τ the *granularity level* of a property τ, N_{total} the total number of values it has and N_{unique} the number of its unique values. We define the granularity level of a property as being proportional to the number of unique values it has over its total values, N_{total}, $G_\tau \propto N_{unique}/N_{total}$. This technique can be used as a probability estimator to solve B(ii).

In truth, an entity can have a combination of different properties as "distinguishing keys". However, the open nature of *linked data* again contains a number of pitfalls that require a concise analysis of the manner in which these combinations of properties can be used. As dataset authors can describe entities using incomplete representations, individual properties may have to be assigned different weights in determining the over granularity level of the set. The exact manned in which property sets could be used, as well as the preference of using property sets instead of individual properties, are among our short-term goals.

Evaluation of the system's performance will be conducted with regard to both desired outputs, as specified by the problem definition. In particular, given a number of results using different inputs, the performance of our proposed solution will be measured as follows:

- Equality validation - Quantitative: Similarly to [6], the performance of the system can be quantified by using $precision$ and $recall$. $Precision$ refers to the number of cases of correctly detected erroneous equality relations over the total number of detected erroneous equality relations. $Recall$ is calculated as the percentage of correctly detected erroneous equality relations over the total number of cases with an erroneous equality relation.
- Argumentation - Qualitative: The system's performance in this task will be evaluated using an argument-based measurement. For example, results can be graded with respect to validity by different participants, with the final performance score being the average of the individual results.

IV. Related Work

A. Related fields

Our problem has strong ties with ontology matching [8], a common challenge in the field of ontologies. Ontology matching is concerned with the discovery of connections between ontological entities. Given that the field has produced a large number of techniques, we are interested particularly in those that focus on the entity (also called instance) level, such as [9]. In addition to this, numerous ontology matching applications have emerged from the Ontology Alignment Evaluation Initiative[5] (OAEI) campaigns, which have been taking place annualy since 2005. For an up-to-date survey on ontology matching, see [8], [10].

In the fields of Linked Data and Semantic Web research, Entity Linking has been used to describe the task of retrieving possible links between two datasets. Two important efforts in the field can be found in [11], [12]. It should be noted that entity linkage (also called record linkage or data linkage) predates the aforementioned fields, as it has been used before in the context of databases [13], [14]. The term has also been used in data cleaning and duplicate record detection tasks [15]. A comprehensive survey on Entity Linkage can be found in [16], [17]. Entity Linking has also being used as a task in the field of Named Entity Disambiguation, along with slightly different tasks such as linking named entities to their Wikipedia entries and document clustering in relation to their named entities [18]. Although such work has a slightly different focus, all of it has involved extracting named entities from textual documents. Conversely, our focus is in entities found in structured data.

The task of discovering erroneous equality relations is quite novel. In our view, this can be attributed to the relatively recent development of the LOD cloud and to the existence of a number of related fields. However, a small number of attempts does exist. In [19], erroneous equality relations are discovered in large datasets using constraint violation detection. In [20], a number of network theory inspired measures are used in combination with others that exploit the nature of `owl:sameAs`. In contrast to both cases, we are interested in real-time validation of the relation, whereas these previous contributions have been more concerned with large-scale analysis.

B. The Identity Crisis

The definition of `owl:sameAs` is that it "indicates that two URI references actually refer to the same person"[6]. Its use,

[5] http://oaei.ontologymatching.org/

[6] http://www.w3.org/TR/owl-ref/#sameAs-def

however, is usually conducted in a more lenient fashion than its intended definition, as ontology authors often disregard any semantic incompatibilities between real world and ontological entity connections. This phenomenon, also referred to as “the identity crisis” in Linked Data, has sparked a debate about the use and definition of `owl:sameAs`. Moreover, the problem occurs even among the definitions of the different versions of identity links contained in other vocabularies [21]

In [22], the different uses of `owl:sameAs` are categorised into situations where two things *a*) are the same real-world entity, but the properties used to describe one of them can be inappropriate for the other; *b*) are claimed to refer to the same real-world entity; *c*) are different but share enough properties to be declared as matched; *d*) are similar; and *e*) are different but related. Their proposed solution was the use of an Identity Ontology that was created based on the aforementioned taxonomy. In a similar approach [23] the proposed solution was another Identity Ontology which was fashioned to suit the needs of a specific domain. Finally, in [24] the use of indiscernibility in place of equality is proposed, to denote that two entities are equal with respect to their share and equal predicates.

C. Logic-based invalidation of equality

The work presented in [6] is perhaps the one most closely related to ours. Their focus is on detecting invalid *sameAs* links between ontological entities using logical inferences. For that end, they propose the use of information extracted from property relationships and the creation of locally-complete property sets[7]. For the selection of properties to be used in the aforementioned techniques, contextual graphs are built around the entities by considering relationships and objects “around” it up to a certain degree. A contextual graph is defined by the entity in the epicentre (figuratively and literally) and the selection of properties that exist within, all of which are predefined by experts.

In the next step the contextual graphs are checked for inconsistencies in the form of different values in locally complete properties and asymmetries induced by functional and inverse functional properties hint at potential differences between them, which are also members of their contextual graphs. A functional property maps exactly one object to a given subject. For example, *hasBloodType* can be regarded as functional as each person can have exactly one blood type. Inverse functional properties map exactly one subject to a given object, e.g. *biologicalMotherOf*. In other words, for u_1 and u_2 to be the same, any shared functional, inverse functional and locally-complete properties should have the same values.

This method, although capable of providing good results, relies on the efficiency of the experts (with respect to locally-complete properties) and dataset authors (with respect to functional and inverse functional property relation declaration). Quite differently, we propose to develop an automated method for identifying (and possibly correcting) erroneous equality relations. Dataset authors are advised, not obliged to use semantically expressive statements such as the ones used here. Therefore we cannot make the assumption that the use of such properties is common practice.

V. Conclusion

Linked Data presents us with a constantly expanding network of interconnected datasets. To be able to exploit its full potential, we need to be able to navigate through its inherent heterogeneity and ambiguity. As a first step in doing so, we provided a definition for the problem of *detecting and resolving contradictions between linked data entities and the real world*, specifically with respect to relations between them. We have presented the outline of an approach which estimates the probability that an `owl:sameAs` link equating two ontological entities is valid, by defining similarity as a combination of different properties of resemblance, from syntax to graph structure and semantics.

In the short term, we aim at designing a system using the specifications laid out in this paper, focusing on detecting and resolving equality relations in real time. We expect that the results from this attempt will enable us to focus on detecting and validating contradictions without restricting ourselves to equality relations between entities. In this scenario, however, there exist a number of open research question, such as the manner in which the system will be able to discover potentially erroneous relations in a system where the relation vocabulary is unknown.

In the future we aim at exploring the notion of systems with local lightweight knowledge bases augmentable with learning from past experiences. The current scenario could also be enhanced with the addition of another contextual parameter, e.g. from a search query. With this addition, the present problem of resolving an erroneous equality relation would change, as resolution would involve selecting the entity which is specified by the query itself. Our long-term vision is the creation of a collection of algorithms and techniques capable of detecting and resolving ambiguity in structured data.

Acknowledgment

This research was supported by the People Programme (Marie Curie Actions) of the European Union’s Seventh Framework Programme under REA grant agreement number 607062. The authors would also like to thank anonymous reviewers for providing insightful input on an earlier version of this paper.

References

[1] C. Bizer *et al.*, “Linked data - the story so far,” *Int. J. Semantic Web Inf. Syst.*, vol. 5, no. 3, pp. 1–22, 2009.

[2] M. Schmachtenberg *et al.*, “Adoption of the linked data best practices in different topical domains,” in *The Semantic Web ISWC 2014*, ser. Lecture Notes in Computer Science, P. Mika *et al.*, Eds. Springer International Publishing, 2014, vol. 8796, pp. 245–260.

[3] G. Klyne and J. Carroll, “Resource description framework (RDF): Concepts and abstract syntax,” W3C, W3C Recommendation, Feb. 2004.

[4] R. Guha and D. Brickley, “RDF vocabulary description language 1.0: RDF schema,” W3C, W3C Recommendation, Feb. 2004.

[5] F. van Harmelen and D. McGuinness, “OWL web ontology language overview,” W3C, W3C Recommendation, Feb. 2004.

[7] local-completeness in this context refers to the local use of the Open World Assumption

[6] L. Papaleo *et al.*, "Logical detection of invalid sameas statements in RDF data," in *Knowledge Engineering and Knowledge Management - 19th Int. Conf., Linköping. Proc.*, 2014, pp. 373–384.

[7] M. McPherson *et al.*, "Birds of a feather: Homophily in social networks," *Annu. Review of Sociology*, vol. 27, no. 1, pp. 415–444, 2001.

[8] P. Shvaiko and J. Euzenat, "Ontology matching: State of the art and future challenges," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 1, pp. 158–176, 2013.

[9] A. Isaac *et al.*, "An empirical study of instance-based ontology matching," in *The Semantic Web, 6th Int. Semantic Web Conf., 2nd Asian Semantic Web Conf., Busan.*, 2007, pp. 253–266.

[10] L. Otero-Cerdeira *et al.*, "Ontology matching: A literature review," *Expert Syst. Appl.*, vol. 42, no. 2, pp. 949–971, 2015.

[11] J. Volz *et al.*, "Silk - A link discovery framework for the web of data," in *Proc. of the WWW2009 Workshop on Linked Data on the Web, Madrid*, 2009.

[12] A. N. Ngomo and S. Auer, "LIMES - A time-efficient approach for large-scale link discovery on the web of data," in *IJCAI 2011, Proc. of the 22nd Int. Joint Conf. on Artificial Intelligence, Barcelona, Catalonia*, 2011, pp. 2312–2317.

[13] I. P. Fellegi and A. B. Sunter, "A theory for record linkage," *J. of the Amer. Statistical Assoc.*, vol. 64, pp. 1183–1210, 1969.

[14] W. E. Winkler, "Overview of record linkage and current research directions," US Bureau of the Census, Tech. Rep., 2006.

[15] A. Ferrara *et al.*, "Data linking for the semantic web," *Int. J. Semantic Web Inf. Syst.*, vol. 7, no. 3, pp. 46–76, 2011.

[16] B. Tansel *et al.*, "A survey of entity resolution and record linkage methodologies," *Commun. of the IIMA*, vol. 6, pp. 41–50, 2006.

[17] N. Koudas *et al.*, "Record linkage: similarity measures and algorithms," in *Proc. of the ACM SIGMOD Int. Conf. on Management of Data, Chicago*, June 2006, pp. 802–803.

[18] B. Hachey *et al.*, "Evaluating entity linking with wikipedia," *Artif. Intell.*, vol. 194, pp. 130–150, Jan. 2013.

[19] G. de Melo, "Not quite the same: Identity constraints for the web of linked data," in *Proc. 27th AAAI Conf. Artificial Intelligence, Bellevue*, July 2013.

[20] C. Guéret *et al.*, "Assessing linked data mappings using network measures," in *The Semantic Web: Research and Applications - 9th Extended Semantic Web Conf., ESWC 2012, Heraklion. Proc.*, May 2012, pp. 87–102.

[21] A. Guy, "Roles and relationships as context-aware properties on the semantic web," in *Proc. of Workshop on Context, Interpetation and Meaning*, 2014.

[22] H. Halpin *et al.*, "When owl: sameas isn't the same: An analysis of identity in linked data," in *The Semantic Web - ISWC 2010 - 9th Int. Semantic Web Conf., Shanghai, Revised Selected Papers, Part I*, Nov. 2010, pp. 305–320.

[23] J. P. McCusker and D. L. McGuinness, "Towards identity in linked data," in *Proc. of the 7th Int. Workshop on OWL: Experiences and Directions, San Francisco, California,*, June 2010.

[24] W. Beek *et al.*, "Rough set semantics for identity on the web," in *Principles of Knowledge Representation and Reasoning: Proc. of the 14th Int. Conference*, 2014.