Towards Linked Data Fact Validation through Measuring Consensus

Shuangyan Liu, Mathieu d'Aquin, and Enrico Motta

The Open University, United Kingdom
{shuangyan.liu, mathieu.daquin, enrico.motta}@open.ac.uk

Abstract. In the context of linked open data, different datasets can be interlinked together, thereby providing rich background knowledge for a dataset under examination. We believe that knowledge from interlinked datasets can be used to validate the accuracy of a linked data fact. In this paper, we present a novel approach for linked data fact validation using linked open data published on the web. This approach utilises owl:sameAs links for retrieving evidence triples, and a novel predicate similarity matching method. It computes the confidence score of an input fact based on weighted average of similarity of the evidence triples retrieved. We also demonstrate the feasibility of our approach using a sample of facts extracted from DBpedia.

Keywords: Linked Open Data, Data Quality, Fact Validation, Semantic Similarity, DBpedia

1 Introduction

Linked datasets created from unstructured sources are likely to contain factual errors [5] (e.g. a wrong population number for a country). Measuring the semantic accuracy of linked sources is viewed as one of the challenging dimensions for data quality assessment [8]. Zaveri et al. defined semantic accuracy as "the degree to which data values correctly represent the real world facts."[8] A simple example to illustrate this would be: when our search engine returns the state where New York City is located as CA, this is viewed as semantically inaccurate since the state CA does not represent the real world state of NYC, i.e. NY.

Different approaches were discussed in previous studies [3,5] for linked data semantic accuracy measurement. The DeFacto approach [3] validated facts by retrieving webpages that contain the actual statement phrased in natural language using search engines and fact confirmation method. Paulheim and Bizer presented in [5] an algorithm for detecting type incompletion based on the statistical distributions of properties and types, and an algorithm for identifying wrong statements by finding large deviation between actual types of the subject and/or objects and apriori probabilities given by the distribution.

However, no studies have investigated how to validate linked data facts leveraging the very nature of linked data (via collecting matched evidence triples from other linked sources). This paper presents an approach for RDF facts validation

by collecting consensus from other linked datasets. Owl:sameAs links are followed to collect triples describing same real-world entities in other datasets. A predicate matching method is described to collect "equivalent" facts and a consensus measure is presented to quantify the agreement among the sources.

The rest of the paper is structured as follows. Section 2 presents the details of our approach. The method and results of an experiment with sample facts from DBpedia are described in Section 3. Finally, we conclude in Section 4 and provide an outlook for future work.

2 Approach

Subject Links Crawling and Cleaning. The first task addressed in this subsection deals with the process of automatically collecting the resource or subject links equivalent to the subject of the input fact(s). We approach the problem in two steps. Firstly, the values of the property `owl:sameAs`[1] of the subject of a fact are retrieved. It can be achieved by querying the underlying dataset of the input fact. Secondly, we fetch the equivalent subject links via querying the `http://sameas.org` service.

There may be duplicated and non-resolvable subject links in the results obtained via `owl:sameAs` and the `http://sameas.org` service. The duplication cases can happen since two separate services are used and the resources that they provide may overlap. It can also be due to the fact that the underlying dataset contains multilingual versions of the same resources and link them together via `owl:sameAs`. In addition, there are several reasons for non-resolvable subject links. The resources may have been deleted from the underlying dataset while the value of the relevant `owl:sameAs` property not being updated coordinately. The services of publishing the datasets may be down or have retired.

The erroneous subject links need to be cleaned before the next task can be performed effectively and efficiently. We follow the following steps for cleaning the errors. First, all subject links are verified by "pinging" the corresponding URIs. If a valid response is received within a given timeout, the subject links are considered as resolvable. Second, duplicated subject links are removed if they have the identical URIs. Finally, multilingual versions of the same resource are removed from the result set.

In our approach the reliability of the subject links are determined according to the provenance of the subject links, i.e., the methods or services used to retrieve the links, for example, the DBpedia `owl:sameAs` property and the `http://sameas.org` service. Details of how to determine the reliability of the subject links are addressed later. The provenance information of the subject links are retained for calculating the confidence score of an input fact.

[1] The following namespace conventions are used in this document: `owl=http://www.w3.org/2002/07/owl`, `dbpedia=http://dbpedia.org/resource/`, `dbpedia-owl=http://dbpedia.org/ontology/`, `dbpprop=http://dbpedia.org/property/`, `yago=http://yago-knowledge.org/resource/`

Predicate Links and Objects Retrieving. The next task of fact validation is collecting all triples that use the collected resources as the subject links. This problem cannot be tackled by simply dereferencing the URIs of the collected subject links.[2] There are three reasons. First, not all of the corresponding URIs can be dereferenced such as the URI of the mosquito Aedes vexans.[3] Second, some dereferenceable URIs may not return the real data of the resources since they were redirected to somewhere else, e.g. `yago:Borough_of_Buckingham`.[4] Finally, the content types of the representation of the information resources obtained via dereferencing can be different.

The non-dereferenceable URIs are removed from the set of subject links as a result of performing the subject links cleaning task. For those dereferenceable URIs, a combination of methods are applied to extract the desired predicates and objects, and convert them to a uniform format for performing the subsequent tasks.

The first method used in our approach is HTTP GET with the resource URI and content negotiation. It allows to obtain the RDF facts of an information resource in most cases. Programming libraries such as the Jena API[5] can be used to extract the desired data from the RDF data. The second method is HTTP GET with a SPARQL query to a dataset endpoint. This method is adopted when the resource URIs cannot return the real data of that resources, and there is a SPARQL endpoint associated with that knowledge base. Last but not the least, when there are only dumps of data available from the knowledge bases, e.g. Wikidata,[6], particular toolkits can be developed to extract desired data from the dumps.

Predicate Similarity Measurement. After completing the beforementioned tasks, a large amount of triples with subjects being equivalent to the subject links of the input facts are collected. The objective of the next task is selecting the evidence triples that have predicates matching the predicates of the input facts.

We choose to measure the predicate similarity based on the semantic similarity between the predicates of the input facts and the collected triples. String similarity measures such as the Trigram similarity metric [1] are not used since they cannot effectively detect predicates which are composed of different words but actually have the same meaning. For example, the property `dbpedia-owl:populationTotal` and the property `yago:hasNumberOfPeople` should be identified as highly related.

There are a number of semantic relatedness measures available including Jiang & Conrath [2], Resnik [6], Lin [4], and Wu & Palmer [7]. They rely mas-

[2] According to the W3Cs note on dereferencing HTTP URIs, the act of retrieving a representation of a resource identified by a URI is known as dereferencing that URI, `http://www.w3.org/2001/tag/doc/httpRange-14/2007-05-31/HttpRange-14`

[3] `http://lod.geospecies.org/ses/4XSQO`

[4] `http://tinyurl.com/mxdkv4s`

[5] `https://jena.apache.org/`

[6] `http://www.wikidata.org/`

sively on the enormous store of knowledge available in WordNet.[7] The principle of our approach for detecting highly related predicates is applying a suitable semantic relatedness measure on the predicates of the evidence triples. In addition, our method is based on WS4J[8] which can generate a matrix of pairwise similarity scores for two input sentences, according to selected semantic relatedness measures. WS4J implements several semantic similarity algorithms described earlier.

Many predicates use compound words such as `dbpedia-owl:populationTotal` and `yago:hasNumberOfPeople`. Thus, our method should be able to handle predicates of compound words as well as predicates composed of single words. Our method consists of three parts. First, a compound word splitter is used to transform predicate names into space separated words (i.e. sentences). Second, a matrix of pairwise similarity scores are generated for two input sentences by the means of WS4J. Finally, formulas are defined to measure the semantic similarity of the input sentences (i.e. the predicates) using the pairwise similarity matrix.

Table 2 provides an example of the pairwise similarity matrix for the sentences "population Total" and "has Number Of People" (as generated by WS4J).

Table 1. Pairwise semantic similarity matrix for two input sentences.

	has	Number	of	People
population	0.0	0.4286	0.0	0.9091
Total	0.0	1.0	0.0	0.3636

Let r be the number of rows of a similarity matrix and c the number of columns of the matrix. The scores in the n^{th} row or column are represented by the sets $S_{row}(n)$, $S_{column}(n)$ respectively. For each word in the shorter sentence (either $r \leq c$ or $r > c$), we choose the max score in the row or column where the word lies as the semantic similarity score of that word, noted as $W(n)$. This leads to the following formula:

$$W(n) = \begin{cases} \max\left(\, S_{row}\left(\, n \,\right)\right) & \text{if } r \leq c \\ \max\left(\, S_{column}\left(\, n \,\right)\right) & \text{if } r > c \end{cases} \tag{1}$$

Moreover, let $\Phi(W)$ be the set of similarity scores of the words in the shorter sentence of a similarity matrix, and k the number of values in the set. If any word in the shorter sentence has a value of similarity greater than the threshold θ, then the two input sentences may have similar meaning. Thus we define the average of the scores belonging to $\Phi(W)$, P, as the semantic similarity score for the two input sentences (i.e. the predicates). Thus, it leads to the following formula:

$$P = \frac{\sum_{W \in \Phi(W)} W}{k} \text{ with } \exists\ W \in \Phi(W) \text{ and } W > \theta \tag{2}$$

[7] `http://wordnet.princeton.edu/`
[8] `https://code.google.com/p/ws4j/`

If no word in the shorter sentence has a value of similarity greater than the threshold θ, then the two input sentences can not have similar meaning. In this case, the value of the similarity score for the two input sentences is assigned to zero.

To obtain the set of matched predicates for the predicate of the input facts, a threshold is applied, e.g., all predicates with $P \geq 0.5$ are considered as matched predicates.

Confidence Calculation. As mentioned in the first task above, the reliability of the subject links collected are determined according to the provenance of the subject links (i.e., `owl:sameAs` and `http://sameas.org service`). A weighting factor is assigned to the subject links of the evidence triples to represent their reliability. The value of a weighting factor ranges from 1 to 5. The greater the value, the more reliable the subject link is.

We define a confidence score for the input fact to represent the degree to which the evidence triples agree with the input fact (or triple). The confidence of the input fact is based on the weighted average of the values of the objects of the evidence triples, represented as γ.

The values of the objects, defined as ν, are considered to be literal values (either numerical or string). If the type of the objects is string, string similarity scores of the objects for the input facts and the evidence tripes are applied as the values of ν. If the type of the objects is numerical, the numerical values of the objects are directly used. The weight ω is the product of the reliability of the subject link and the similarity of the predicate link of an evidence triple. Additionally, let m be the number of evidence triples collected through the abovementioned tasks. Thus, γ is represented as:

$$\gamma = \frac{\sum_{i=1}^{m} \omega_i \cdot \nu_i}{\sum_{j=1}^{m} \omega_j} \tag{3}$$

Formula (3) is applied to represent the confidence score of an input fact where the value of the objects of the evidence triples are the type of string.

Furthermore, the following formula is applied to represent the confidence score of the input fact, denoted as Γ when the values of the objects are numerical. In Formula (4) x represents the numerical value of the object of the input fact while γ is the weighted average number calculated via formula (3).

$$\Gamma = 1 - \frac{\sqrt{(x-\gamma)^2}}{\gamma} \tag{4}$$

Based on Formula (4), a smaller difference in the numerical values of the objects between the input fact and the weighted average value will lead to a higher confidence score.

3 Experiment

In order to test the feasibility of the approach described in the previous section, we conducted an experiment with a property from DBpedia (`dbpedia:populationTotal`) and a sample of facts using this property as the predicate. This property was selected since the type of its values are numerical.

We made a query to the DBpedia SPARQL endpoint for obtaining all towns in Milton Keynes that have a population of more than 10,000. The resulting 18 triples were utilised as the input facts. The subjects of these facts were used as seeds to crawl equivalent subject links from other knowledge bases.

The number of subject links retrieved for a single fact ranges from dozens to several hundred. For example, `dbpedia:Stantonbury` has 23 subject links found while `dbpedia:Buckingham` has 232 subject links retrieved. The number of the cleaned subject links is reduced greatly, ranging from a few to several tens.

We selected a representative resource `dbpedia:Buckingham` to examine the correctness of the subject links cleaning process. A total of 207 noise subject links were found for the resource `dbpedia:Buckingham`. It consisted of 172 non-resolvable links, and 35 duplicate links. We manually examined the causes of the non-resolvable links, and corrected 56 out of 172 as valid links (Figure 1). Initially the 56 links were identified as invalid links due to a small value of the read timeout field set for the tool used for the subject links cleaning process. It allowed us to adjust the timeout field for a suitable value.

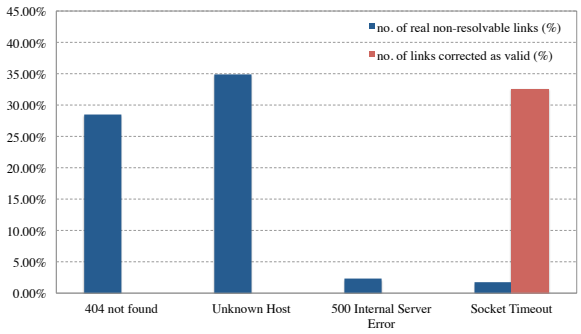

Fig. 1. Correctness of non-resolvable subject links cleaning for `dbpedia:Buckingham` with analysis of causes (Total=172)

We also found that different data access services were provided by the knowledge bases where the subject links originated from. Accordingly, we needed to adopt different methods to deal with this diversity in terms of retrieving the predicate links and objects from these knowledge bases.

In addition, the compound word splitter[9] was utilised in the predicate similarity measurement process. It could split compound predicate names into sen-

[9] `http://www.lina.univ-nantes.fr/?Compound-Splitting-Tool.html`

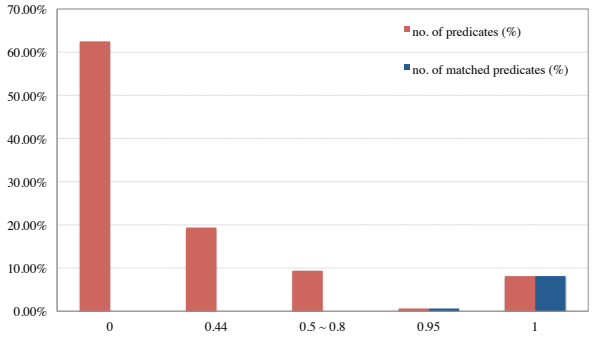

Fig. 2. Distribution of predicate similarity by applying the WUP semantic similarity measure and Formulas (1) and (2)

tences. The Wu & Palmer [7] semantic similarity measure (WUP) was selected since the result similarity scores are normalised from 0 to 1. We also tested other measures such as Lin [4]. The WUP measure demonstrated the highest rate of correctness (threshold $\theta \geq 0.8$). The distribution of the predicate similarity scores generated is provided in Figure 2.

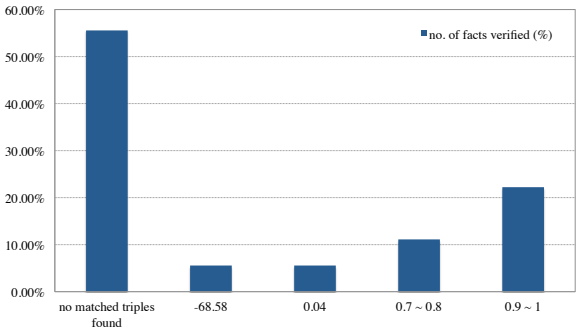

Fig. 3. Confidence of the sample of facts collected from DBpedia

Furthermore, 45% of the sample facts (i.e. statements about the population of the 18 subjects) were assigned to a confidence score and 55% were not (as no evidence triples were found). Figure 3 demonstrates the distribution of the confidence scores generated for the sample facts. 22% of the facts were identified as highly reliable ($\Gamma \geq 0.9$). Two facts were assigned to very low confidence scores (0.04 and -68.58). We manually examined the causes of the low confidence values, and discovered that a matched triple for each fact had a very large or small population number. It caused the weight average of the object values of the evidence triples to be too large or small. It was due to the fact that the subject links of the erroneous triples (retrieved from `sameas.org` service) were pointed to resources not identical to the subjects of the facts (wrong subject links). We corrected the errors by removing the erroneous triples from the set of evidence triples. It leaded to the fact (initially with 0.04 confidence) to get

a much higher confidence (0.94), and no confidence score produced for the fact (initially with -68.58 confidence) because no evidence triples are found. Based on this experiment, we plan to extend our approach to verify abnormal evidence triples with "fake" subject links in future work.

4 Conclusion and Future Work

In this paper, we presented an approach for validating linked data facts using RDF triples retrieved from open knowledge bases. Our approach enables the assessment of the accuracy of facts using the vast interlinked RDF resources on the Web. This would become increasingly important due to the fast growth of LOD on the Web.

The presented work is still at its early stage, the experiment discussed in this paper focused on testing the feasibility of each component of the presented approach. This can help refine our approach before an evaluation of the approach as a whole is carried out. We are planning to demonstrate that the proposed approach can be applied proficiently to arbitrary predicates, and evaluate the predicate similarity matching method with standard evaluation measures (Precision/Recall) on well-known datasets. Moreover, we are also going to define a gold standard and apply the standard for evaluating our method for validating RDF facts.

References

1. Angell, R.C., Freund, G.E., Willett, P.: Automatic spelling correction using a trigram similarity measure. Information Processing & Management 19(4), 255–261 (1983)
2. Jiang, J.J., Conrath, D.W.: Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of International Conference on Research in Computational Linguistics (1997)
3. Lehmann, J., Gerber, D., Morsey, M., Ngomo, A.C.N.: Defacto-deep fact validation. In: The Semantic Web–ISWC 2012, pp. 312–327. Springer (2012)
4. Lin, D.: An information-theoretic definition of similarity. In: ICML. vol. 98, pp. 296–304 (1998)
5. Paulheim, H., Bizer, C.: Improving the quality of linked data using statistical distributions. International Journal on Semantic Web and Information Systems (IJSWIS) 10(2), 63–86 (2014)
6. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence,. pp. 448–453 (1995)
7. Wu, Z., Palmer, M.: Verbs semantics and lexical selection. In: Proceedings of the 32nd annual meeting on Association for Computational Linguistics. pp. 133–138. Association for Computational Linguistics (1994)
8. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S.: Quality assessment for linked data: A survey. Semantic Web journal (to appear)