Data Conflict Resolution among Same Entities in Web of Data

Mojgan Askarizade
Isfahan University
Department of Computer
Isfahan Iran
Mojgan.askarizade@gmail.com

Mohammadali Nematbakhsh
Isfahan University
Department of Computer
Isfahan Iran
nematbakhsh@eng.ui.ac.ir

Enseih Davoodi jam
Computer Engineering Department
University Of Isfahan
Isfahan, Iran
nc_davodi.softcom@eng.ui.ac.ir

***Abstract*—With the growing amount of published RDF datasets on similar domains, data conflict between similar entities (same-as) is becoming a common problem for Web of Data applications. In this paper we propose an algorithm to detect conflict of same properties values of similar entities and select the most accurate value. The proposed algorithm contains two major steps. The first step filters out low ranked datasets using a link analysis technique. The second step calculates and evaluates the focus level of a dataset in a specific domain. Finally, the value of the top ranked dataset is considered. The proposed algorithm is implemented by Java Programming Language and is evaluated by geographical datasets containing "country" entities.**

Data mining; semantic Web; Linked Data; data conflict; ranking; same entities.

I. INTRODUCTION

Because of growing amount of data on the Web of data, number of the structured datasets has increased significantly in recent years. Since semantic Web allows machines to process data and retrieve data automatically, availability of accurate structured data is required. The information published in different data sources often overlaps. For example, WordFactbook[1] and Eurostat[2] data sets overlap deeply, as they both describe the type "country".

According to principle of Linked Data, entities are identified by URI[3] [1]. The same real-word entity can be represented by different URIs correspond to each other with owl:sameAs links. Multiple datasets may provide different values for the same property of an entity [3]. For example datasets such as: DBpedia[4] , Geonames[5] , and WordFactbook publish different information about population of United Kingdom that are mentioned below:

TABLEI. United Kingdom Different Data Sets

Geonames	WorldFactbook	DBpedia
60776238	62348447	58789194

However, these data inconsistencies are not acceptable in the cases where quality data is needed, for example in the case of a commercial applications. Generally three different strategies are suggested to deal with such data inconsistency [6]:

- Conflict ignoring: Because data correction is vital in most circumstances, ignorance of data inconsistency is not acceptable in these cases.
- Conflict avoiding: The second strategy handle conflicting data by conflict avoidance. With respect to large dimensions of web of data, this strategy is impossible.
- Conflict resolving: This strategy detects same entities and resolves data conflict between their values. This is the most practical strategy to deal with data inconsistency.

We deal with the inconsistency by resolving data conflict, for this reason, two main approaches are described here:

- Selection from available data: In this approach a value is selected from inconsistent values. For example, if the population of London is one of the different values (X, Y, Z), one of them must be chosen as the population of London.
- A new value generation: Using this approach a new value is generated using available values. For example, if the population of London is one of the different values (X, Y, Z), the result can be the average value of (X, Y, Z).

Our algorithm resolve data conflict by assigning new value from available values of properties.

In this article an algorithm is proposed to detect inconsistency among the values of similar properties of entities connected with sameAs link in order to choose the most accurate data. The proposed algorithm is divided into two steps. then some of them are removed from the list of data sets according to their ranks. In the second step the specialty of datasets are computed based on two

1 http://www4.wiwiss.fuberlin.de/factbook

2 http://www4.wiwiss.fu-berlin.de/eurostat/

3 Uniform Resource Identifiers

4 http://dbpedia.org/page/United_Kingdom

5 http://www.geonames.org/countries/GB/united kingdom.html

metrics: ontology and size of dataset. This article has shown that the data of the data set which is more specialized is more accurate. At the end of the ranking, the property values of the top ranked data set are considered as the output.

The proposed algorithm is implemented using Java Programming Language and evaluated on geographical datasets that contain "country" entity. This paper is organized as follows: In section 2, the related works are presented. The proposed algorithm is described in section 3. The evaluation of the proposed algorithm is presented in section 4.Finally the conclusion and the outlook on future works are presented in section 5.

II. RELATED WORK

Due to its novelty, here are a few researches which have been done to resolve data conflict in the Web of data. One of the noticeable researches suggested by Tacchini et al is a framework aimed at resolving the inconsistency of data value [9]. The main idea of their proposed framework is based on extracting data from different Wikipedia language version and comparing these data with each other to identify the most appropriate language version. For example, despite of better quality of English version as a whole, the population of Germany (as an example of required data) could be indicated more up-to-date in German version than Italian, French or even English version. They provide several strategies to recognize the correct Wikipedia version to choose from. To do this first, data of Wikipedia are converted to RDF format (by DBpedia project[6]) and then they are published in web of data. Because all datasets use Dbpedia ontology, no adaptation is required so Dbpedia ontology is used. Therefore, it's enough to indicate the classes referring to a single entity, using a URI. In other steps, different heuristics are utilized to resolve the inconsistency of data retrieved from different resources. The main purpose of this research is to support the hypothesis that retrieving data from different language versions instead of a single one and combining them to obtain higher quality and more complete data. In this way, the French, Italian, German and English versions of Wikipedia were utilized. Due to novelty of web of data, here are a few researches which have been done to resolve data conflicts in web of data. One of the noticeable researches suggested by Bizer et al Proposes a framework aimed for resolving the inconsistency of data [10]. The main idea of their proposed framework is based on extracting data from different Wikipedia language version and comparing these data with each other to identify the most appropriate language version. For example, despite of better quality of English version as a whole, the population of Germany (as an example of required data) could be indicated more up-to-date in German version than Italian, French or even English version. They provide several strategies to recognize the correct Wikipedia version to choose from. To do this first, data of Wikipedia are converted to RDF format (by DBpedia project) and then they are published in web of data. Because all datasets use Dbpedia ontology, no adaptation is required so Dbpedia ontology is used. Therefore, it's enough to indicate the classes referring to a single entity, using a URI. In other steps, different heuristics are utilized to resolve the inconsistency of data retrieved from different resources. The main purpose of this research is to support the hypothesis that retrieving data from different language versions instead of a single one and combining them to obtain higher quality and more complete data. In this way, the French, Italian, German and English versions of Wikipedia were utilized.

6 http://dbpedia.org/About

III. THE PROPOSED ALGORITHM

We present an algorithm to resolve data conflict among the property values of similar entities. Our algorithm included separate steps. The workflow of the algorithm is displayed in figure 1. The Linked Data cloud contains several data sets that overlap to some extent; the overlapping data is extracted by downloading dump files or querying SPARQL endpoints. The input of the algorithm is a number of data sets that talk about a single subject and the output is highly accurate data obtained from input data sets.

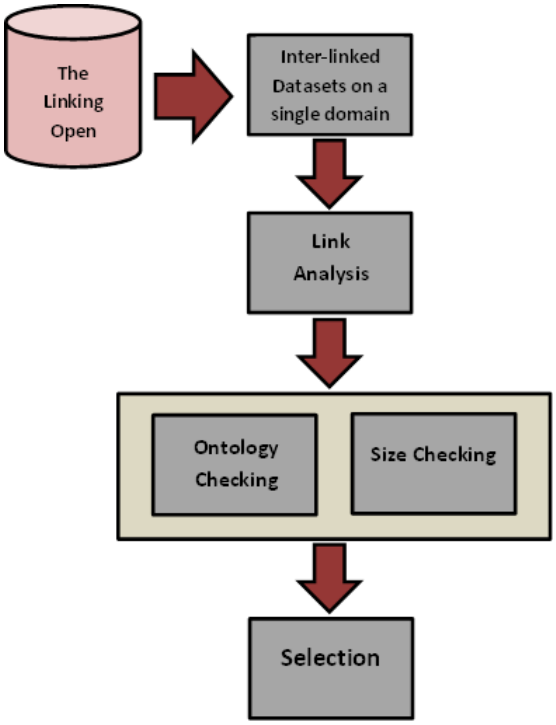

Figure 1. General Workflow of Algorithm

A. *The first step of algorithm*

First, it is necessary to begin by filtering out some minor data sets. In order to do this, data sets are ranked by performing link analysis. Afterwards, data sets that are low-ranked are taken off the data set list. The Page Rank algorithm is used to rank data sets [16].

The Page Rank algorithm which is used in most search engines such as Google is an easy means of ranking linked data. The probability of finding any given page is evaluated by this algorithm by starting from a point and random surfing. The algorithm assumes a link between a page i to a page j demonstrates the importance of page j. In addition, the importance of page j is associated to the importance of page i itself and inversely proportional to the number of pages i point to. To adapt this algorithm to the web of data, any page considered as a data set and links between pages considered as

links between data sets. According to Page Rank Algorithm, the rank of data set j in k level is equals to:

$$R^k(j) = \sum_{i \in B(j)} \frac{R^{k-1}(i)}{|L(i)|} \qquad (1)$$

Let $B(j) = \{source(l) \mid \forall l \in L, target(l) = j\}$ be the set of data sets point to to j and L(i) = {target(l)| $\forall l \in L$ } be the set of data sets linked by a data set i. the operation is repeated until the algorithm reach to a specific threshold.

B. *The second step of algorithm*

There are several datasets have been published on a specific domain. Some datasets are general so they cover various domains. On the other hand, some datasets cover only specified domain. For example DBpedia is a general dataset covering different entities such as: persons, places, music albums, films etc. while LMDB is a specialized domain-specific dataset on film domain. The idea behind the algorithm is that the domain specific datasets have more accurate data than general datasets. For evaluation of specialty degree of a dataset two criteria is considered: ontology and size of dataset.

1) *Ontology*

In this step, we assess the ontology of datasets semantically to measure how much an ontology is specialized. First, all properties that are included in the ontology are extracted by SPARQL querying. Then a word which describes the domain of our data sets is selected. This word which can be for example 'music', 'film', 'book', 'drug' or some other domain word is given to WordNet[7] dictionary. All words that are synonyms with the mentioned word are extracted from WordNet and kept in a list; we name the list the synonyms list. Afterwards, the synonyms list is compared with the properties list to find any words naming properties which are in the synonyms list. The result is a list of words which are semantically related to the domain of the data set. These words are defined as the specialized entities.

2) *Size*

Another criterion is the size of the data set. It is possible that a data set be specialized but be small. It is assumed that the data set is which larger is more important and more reliable.

In the Web of data each resource is represented as a number of triples. Each triple consist of three parts that are subject, predicate and object. The subject and object of a triple each can be a simple literal value, such as a string, number, or date; or the URI of another resource. In this step, we count the number of instances in which the subject or object is contained in the specialized list. This number is considered as |ST|. And |TT| is the number of instances of all entities.

Using equation 2 we score each of the data sets. The first fraction shows the impact of ontology, where |SE| is the number of specialized entities and |TE| is the number of total entities. The second fraction shows the impact of data set size where |ST| is the number of specialized instances and |TT| is the number of total instances. As a result each data set which gains a higher score is more specialized.

$$Score = 0.8 * \frac{|SE|}{|TE|} + 0.2 * \frac{|ST|}{|TT|} \qquad (1)$$

Afterwards, the values of the data set which obtains the higher score are selected as a high-accurate data values.

IV. EVALUATION AND IMPLEMENTATION

The proposed algorithm has been implemented using Java Programming Language and evaluated via geographical domain data sets. Five major data sets of the geographical domain are used as the input of the proposed algorithm; they are DBpedia, Geonames, WordFactbook, Eurostat, GeoLinked Data[8] and the output is the values of the most accurate data selected from these five data sets. In order to evaluate the proposed algorithm, 35 countries were selected and the population property of each country entity examined.

First, the countries of the five data sets are extracted by SPARQL querying using Jena. For analyzing links between data sets, we find all links between the data sets as shown in figure 4. Then data set ranks are computed by the Page Rank algorithm according to link analysis. The result is shown in figure 5.

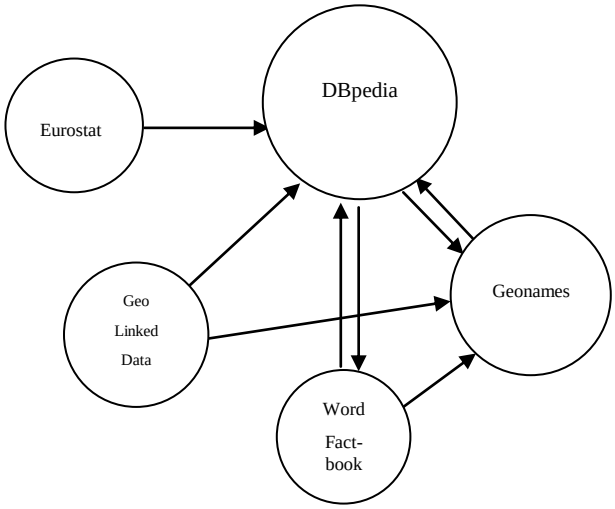

Figure 2. Links Among Data Sets

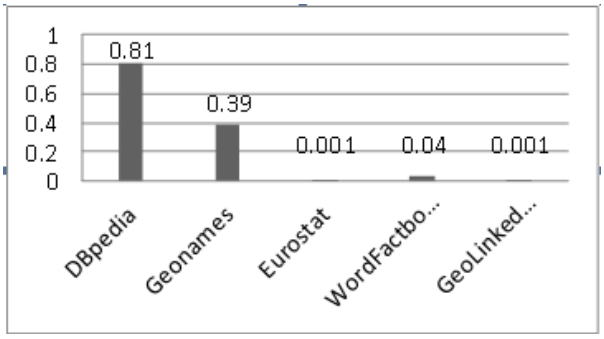

Figure 3. Ranking Data Sets by Page Rank

7 http://wordnet.princeton.edu/

8 http://geo.linkeddata.es

As depicted in figure 3 DBpedia is the top ranked data set while GeoLinked Data and Eurostat are the low ranked data sets. In this stage the data sets whose rank scores are less than half of the rank scores belonging to the top ranked data set are removed from the assessment.

For gathering data we track the links between the remaining two data sets; DBpedia and Geonames. The population of the countries is considered as an attribute for evaluation. Different vocabularies are used in the Web of data. In DBpedia dbpprop:populationCensus and in Geonames gn:population is representative of the population attribute.

Data sets are evaluated based on their ontology and their size. Since the domain that we focused on is geographical data, we chose "country" as the input of WordNet dictionary. Synonyms of "country" are extracted as the output of WordNet. Some of them are land, state, city, fatherland, nation etc. which were named the synonyms list.

In the next step, we extract all entities from the data sets. For example the following SPARQL extracts all entities from DBpedia:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
SELECT distinct ?s WHERE {?s rdf:type owl:Class}
```

Finally we count the entities that contain any words in the synonyms list. The result of first the fraction in equation 2 is shown in figure 4.

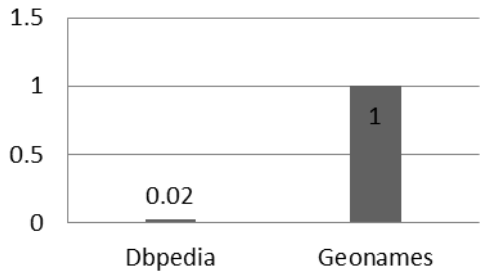

Figure 4. Result of Fist Fraction

In the last stage the size of the data set is examined. The number of instances of specialized entities and total entities are calculated. For example "London" is an instance of a specialized entity. In TABLEII one of the properties of "London" is displayed in form of a triple. As is shown, the object of the triple is dbpedia-owl:Place that contains "place" as a member of the specialized entity. And 'A Trip to the Moon' is an instance of non-specialized entity because neither the subject nor the object is in the specialized entities list. After calculating the number of specialized instances and total instances the result of the second fraction of equation 2 is shown in figure 5.

TABLEII. Type of Instances

Specialized Instance		
<http://dbpedia.org/resource/London>	rdf:type	dbpedia-owl:Place
Non Specialized Instance		
<http://dbpedia.org/page/A_Trip_to_the_Moon>	rdf:type	dbpedia-owl:Film

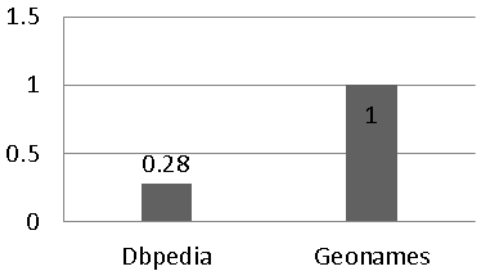

Figure 5. Result of Second Fraction

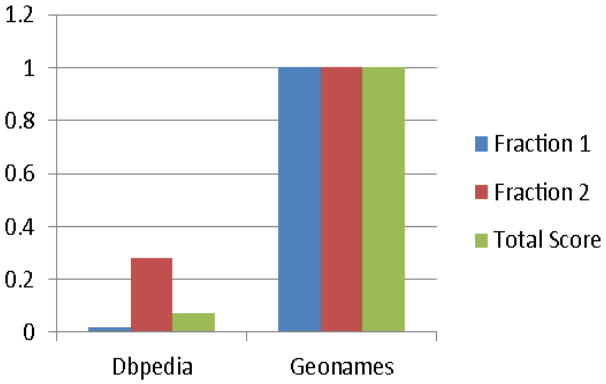

Figure 6. final ranking

Finally, the last score with regard to the equation 2 has been calculated and is shown in figure 6. As it shows Geonames gains the highest score. Considering the proposed idea, Geonames' data is the most accurate. So when we are faced with data conflicts, Geonames' data must be chosen.

In order to evaluate our algorithm we found real data on The Population Reference Bureau[9]. The Population Reference Bureau informs people around the world about population, health, and the environment, and empowers them to use that information to advance the well-being of current and future generations. 33 countries were selected for evaluation. The results of our algorithm that are the most accurate values of populations are compared with the data found on The Population Reference Bureau. The numbers of the most accurate data are found on three data sets: WordFactbook, Geonames, DBpedia are shown in following table.

TABLEIII. The number of most accurate data

WordFactbook	**Geonames**	**DBpedia**
7	22	4

The WordFactbook is also illustrated in the above table. We can see if WordFactbook is not removed from the synonyms list it would have more accurate data compared with DBpedia data.

As we know WordFactbook and Geonames are domain-specific data sets on the geographic domain while, DBpedia is a general data set. The results indicate specialized data sets have more accurate data compared with general data sets. The accuracy of the final data set compared with the real data is 67%.

9 http://www.prb.org/

V. CONCLUSION AND FUTURE WORKS

This article proposes an algorithm to resolve data conflicts among property values of similar entities in the web of data. The proposed algorithm contains three steps. In the first step, the input data sets published on a single domain are ranked by performing link analysis. In the second step, the recognition of same entities is performed by tracking the sameAs links and vocabulary matching is done manually. In the last stage, the specialty of data sets is computed based on two criteria: ontology and the size of the data set. The proposed algorithm has been evaluated using geographic data sets that contain the "country" entity. Results were compared with real data obtained from The Population Reference Bureau. The data obtained by the proposed algorithm had a 67% matching degree with real data. Future work in relation to this article will focus on:

- Using different domains to evaluate the accuracy of the algorithm.
- Using PageRank algorithm applied on internal entities and intra-links of a data set in order to rank entities.

VI. REFERENCES

[1] Tom Heath and Christian Bizer, Linked Da ta: Evolving the Web into a Global Data Space, 1st edition. Synthesis Lectures on the Semantic Web: Theory and Technology, 2011, pp.1-10.

[2] Christian Bizer, Quality-Driven Information Filtering in the Context of Web-Based Information Systems, PhD thesis, Free University of Berlin, Berlin, Germany, 2007.

[3] Cristian Bizer, Tom Heath, and T. Berners-Lee, Linked Data - The Story So Far, Journal on Semantic Web and Information Systems, Special Issue on Linked Data, Vol.5, No.3, 2009, pp.1-22.

[4] Anja Jentzsch, Cristian Bizer, and Richard Cyganiak, State of the LOD Cloud, September 2011.

[5] Felix Naumann, Alexander Bilke, Jens Bleiholder, and Melanie Weis, Data Fusion in Three Steps: Resolving Schema, Tuple, and Value Inconsistencies, J. IEEE Data Eng. Bull, 2006, pp.21-31.

[6] Xin Luna Dong, and Felix Naumann, Data fusion - Resolving Data Conflicts for Integration, J. VLDB Endowment, Vol.2, No.2, 2009, pp.1654-1655.

[7] Jens Bleiholder and Felix Naumann, Data fusion, J. ACM Computing Surveys (CSUR), Vol.41, No.1, 2008.

[8] Andriy Nikolov, Victoria Uren, Enrico Motta, Anne N. De Roeck, Integration of Semantically Annotated Data by the KnoFuss Architecture, Proc. 16th international conference on Knowledge Engineering: Practice and Patterns, 2008, pp.265-274.

[9] Olaf Hartig and Jun Zhao, Using Web Data Provenance for Quality Assessment, in Proc. of the Workshop on Semantic Web and Provenance Management at ISWC, Vol. 526, 2009.

[10] Eugenio Tacchini, Andreas Schultz, and Christian Bizer, Experiments with Wikipedia Cross-Language Data Fusion, Proc. the 5th workshop on scripting and development for the semantic web, Vol.449, 2009.

[11] Andreas Schultz, Andrea Matteini, Robert Isele, Christian Bizer, and Christian Becker, LDIF : Linked data integration framework, 2nd International Workshop on Consuming Linked Data, Bonn, Germany, 2011.

[12] Kendall Grant Clark, Lee Feigenbaum, and Elias Torres, SPARQL Protocol for RDF, W3C recommendation, 2008.

[13] Alexander Bilke , Jens Bleiholder , Christoph Böhm , Karsten Draba , Felix Naumann , Melanie Weis, Automatic Data Fusion with HumMer, Proc. International Conference on Very Large Databases (VLDB), 2005, pp.1251-1254.

[14] Jens Bleiholder, and Felix Naumann, Declarative Data Fusion - Syntax, Semantics, and Implementation, Proc. Databases and Information Systems (ADBIS), 2005, pp.58-73.

[15] Wenpu Xing, and Ali A. Ghorbani, Weighted PageRank Algorithm, Proc. CNSR, 2004, pp.305-314.

[16] Lawrence Page , Sergey Brin , Rajeev Motwani , Terry Winograd, The PageRank Citation Ranking: Bringing Order to the Web, 1998, Technical Report 1999-66 .

[17] Renaud Delbru , Nickolai Toupikov , Michele Catasta , Stefan Decker , Fondazione Bruno Kessler, Hierarchical Link Analysis for Ranking Web Data, Proc. ESWC (2), Vol.6089, 2010, pp.225-239.

[18] Renaud Delbru , Nickolai Toupikov , Michele Catasta , Stefan Decker , Fondazione Bruno Kessler, DING! Data set Ranking using Formal Descriptions, WWW 2009 Workshop: Linked Data on the Web LDOW2009, 2009.

[19] Tim Berners-Lee, Linked data. [Online]. Available: http://www.w3.org/DesignIssues/LinkedData.html, October, 2011.

[20] Cristian Bizer, Richard Cyganiak, and Tom Heath. How to publish linked data on the web. [Online]. Available: http://sites.wiwiss.fu-berlin.de/suhl/bizer/pub/LinkedDataTutorial. October , 2011.