AN INTERLINKING APPROACH FOR LINKED GEOSPATIAL DATA

Mingda Zhang, Jie Yuan, Jianya Gong, Peng Yue*

State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan
University, 129 Luoyu Road, Wuhan, 430079, China-
pyue@whu.edu.cn

KEY WORDS: Interlinking approaches, Identity links, Geospatial data, Linking open data cloud, Tversky's contrast model

ABSTRACT:

Geospatial metadata from metadata catalogue can be published as part of Web of data used Linked Data technologies. The published data could be named as linked geospatial metadata. A key issue of Linked Data technologies is to create links among datasets. There are three important types of RDF links: relationship links, identity links, and vocabulary links. This paper proposes a matching method to construct linkages between linked geospatial metadata and geospatial datasets in the linking open data cloud (LOD). This matching method is based on semantic similarity to construct identity links. A matching algorithm using Tversky's contrast model and Jaro-Winkler distance is proposed and evaluated.

1. INTRODUCTION

The term Linked Data refers to a set of best practices for publishing and connecting structured data on the Web (Bizer, 2009). Linked data technologies use the Resource Description Framework (RDF) language and HTTP protocol to publish structured data on the Web (Bizer, 2008), which have shown great promise for effectively sharing and interlinking of Web resources (Berners-Lee, 2006). Semantic Web researchers and practitioners have started to make geospatial data available as Linked Data on the Web, which promotes sharing and interlinking of geospatial data. For example, LinkedGeoData (Sören, 2009) makes OpenStreetMap data available as RDF.

Creating links is a key issue of the Linked Data, which can connect the data to an unbounded Web in which one can find all kinds of things (Berners-Lee, 2006). There are three important types of RDF links: relationship links, identity links and vocabulary links (Heath & Bizer, 2011). Relationship links set reference from entities in one data set to entities in another, which add more descriptions to the source data set. Identity links aim at constructing interlinks between deferent URIs indentifying the same entity. Vocabulary links map the relationship between terms from different vocabularies.

This paper proposes a method based on semantic similarity to construct identity links between linked geospatial data. This method is based on Tversky's contrast model (Tversky, 1997), which determines semantic similarity by comparing properties of two different instances. Jaro-Winkler distance (Winkler, 1990) is used to compute the similarity of these properties. It measures similarity between two strings. Combining the two methods together, instances between two datasets could be linked using values of similarity. The remainder of the paper is organized as follows. Section II introduces related work. Section III describes the method to construct identity links. Conclusion is given in Section IV.

2. RELATED WORK

Semantic Web is about "making links, so that a person or machine can explore the Web of Data" (Berners-Lee, 2006).

The interlinking of Linked Data is an important factor for the success of the Semantic Web. Researchers have done a lot of work to publish data-sets in RDF on the Web according to the principles of Linked Data. These datasets are interlinked with each other. For example, World Wide Web Consortium (W3C) Linking Open Data (LOD) community project has published various open data sets as RDF on the Web and set RDF links between data items from different data sources. Figure 1 shows datasets that have been published and interlinked by the project. There are already 295 datasets consisting of over 31 billion RDF triples, which are interlinked by around 504 million RDF links by September 2011.
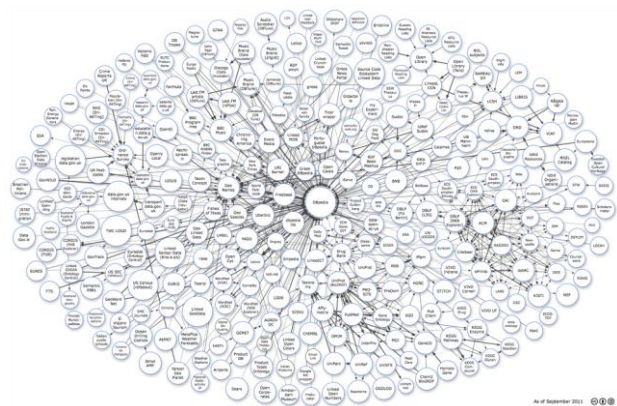

Figure 1. Datasets in LOD by September 2011
(http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData)

DBpedia is part of the Linking Open Data community project, which is a crowd-sourced community effort to extract structured information from Wikipedia and make this information available on the Web. In recent years, an increasing number of data publishers link their datasets to DBpedia resources, making DBpedia a central interlinking hub for the emerging Web of Data (Auer et al., 2007). GeoNames is another part of the project, which is a linked geospatial data set about place names. In GeoNames database, latitude, longitude, elevation,

population, and other information about places are stored. In this paper, we use the DBpedia as a target dataset to build links.

There are two kinds of approaches to build links: manual and automatic methods. Manual methods are suitable for small and static datasets. They are effective but require skilled human data publishers (Araújo et al., 2011). When comes to big datasets, automatic or semi-automatic methods are needed. There are two main types of automatic approaches: key-based and similarity-based approaches (Heath & Bizer, 2011). In several domains, there are domain-accepted identification codes, for example, ISBN numbers in the publication domain. The code may be used as a property value of the resource or part of the URIs. By using the common codes as keys, links between resources could be established. Where there are no common identifiers among different datasets, similarity-based approaches are needed. Multiple properties of resources are often selected to be compared and similarity scores are calculated. These similarity scores are aggregated, and if the aggregated value is above a given threshold, resources could be linked.

Researchers have proposed some linking approaches for linked geospatial data. Barnaghi et al. (2012) propose a platform, called Sense2Web, to publish Semantic Sensor Network data as linked data and link them to resources on the Web of Data manually. For example, when a user publishes a new sensor, Sense2Web uses Jena API to query DBpedia and GeoNames to obtain descriptive information such as location and sensor types, which then can be selected manually by users to link with sensor data. Pschorr et al. (2010) present an automatic approach to publish sensors as linked data. Longitude/latitude pairs are extracted from both semantic sensor data and GeoNames respectively using SPARQL queries. Using longitude/latitude, links between semantic sensor data and GeoNames can be established automatically. The links help discovery of sensors using two basic operations (Pschorr et al., 2010): Find the named location closest to a given sensor; Find all sensors near a given named location. Yuan et al. (2013) propose an approach to publish geospatial data provenance in a catalog service into the Web of Data using the Linked-Data approach. They compare the boundingbox of data items from the linked geospatial data provenance with the spatial region of data items from LOD datasets. The topological relation is calculated between two geometries. Once the relation is determined, the dataset will be linked to the data item using one type of geometric relations, for example 'within'. In this paper, we will use the linked geospatial data mentioned above as our source dataset to construct more links to the LOD cloud.

3. CONSTRUCTING LINKS TO THE LOD CLOUD

There could be many different providers publishing same entities as linked data with different URIs. It is a common practice to use the link http://www.w3.org/2002/07/owl#sameAs to state that two URI references refer to the same thing (Bizer, 2007). This section describes a method based on semantic similarity to construct such links between different datasets. Linked geospatial data from a catalogue service (Yuan, 2013) and DBpedia are used as source and target datasets respectively.

Figure 2 describes the properties of agent that stands for the provider of geospatial data or service in linked geospatial dataset. The property, dc:title, is the agent's name; pro:city is the city that the agent is located in; pro:province is the province that the agent is located in; pro:country is the country; pro:tel is

the telephone number of the agent. In the DBpedia, more detailed information is provided for the entity, Wuhan University. Figure 3 is part of properties of Wuhan University in DBpedia (http://dbpedia.org/page/Wuhan_University). If links between the two datasets are constructed, we could get more details about the agent in linked geospatial dataset using links to DBpedia.
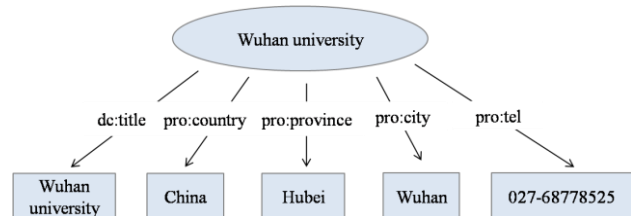

Figure 2. The agent, Wuhan University, in linked geospatial data

Figure 3. The entity, Wuhan University, in DBpedia

The links can be constructed by a match between instances from the two distinct datasets using Tversky's contrast model (Tversky, 1997). In the contrast model, the similarity between two entities, A and B, is expressed as a linear combination of the measures of theirs common and distinctive properties, as shown in equation below.

$$\text{Tversky}(\,A, B\,) = \lambda f\,(\,A \cap B\,) - \alpha f\,(\,A{-}B\,) - \beta f\,(B{-}A\,) \qquad (1)$$

where α, β, and $\lambda \geqslant 0$.

The contrast model is composed by three disjoint set functions. The scale function $f\,(A \cap B)$ represents the set of common properties between A and B. The function $f\,(A - B)$ and $f\,(B - A)$ represents the set of distinct properties between A and B, and B and A, respectively. The constants α, β, and λ represent weights of the communalities and differences in the equation. In the case taken by this paper, the agent in the linked geospatial dataset has a limited set of properties. Most of them share similarities with properties in DBpedia. Therefore, common properties are experimented in this paper, and α and β are set 0. In the source dataset, the three properties, dc:tile, pro:province and pro:city are selected. Correspondingly, in the target dataset, i.e. DBpedia, dbpprop:name, dbpprop:province, and dbpprop:city are selected. The equation based on Tversky's contrast model for computing similarity is shown below.

$$sim(a,b) = 0.6 * n\,(a,b) \; + \; 0.2 * p\,(a,b) \; + \; 0.2 * c\,(a,b) \qquad (2)$$

where $n(a,b)$, $p(a,b)$, $c(a,b)$ denote the similarity function between dc:title and dbpprop:name, pro:province and dbpprop:province, and pro:city and dbpprop:city respectively. The computation of similarity for each function is based on a string-similarity function, i.e. Jaro-Winkler distance. The higher the Jaro–Winkler distance for two strings is, the more similar the strings are. The score is normalized such that 0 equates to no similarity and 1 is an exact match. When property matching is done, the measurement of similarity between two instances will be calculated using Equation 2. If the result is greater than 0.9, identity links will be constructed automatically and a RDF triple is created. If the result is between 0.8 and 0.9, the results can be sent to users for manual decision. If the result is lower than 0.8, it means that they are not matched. A matching workflow is shown in Figure 4. Detailed descriptions are provided as follows.
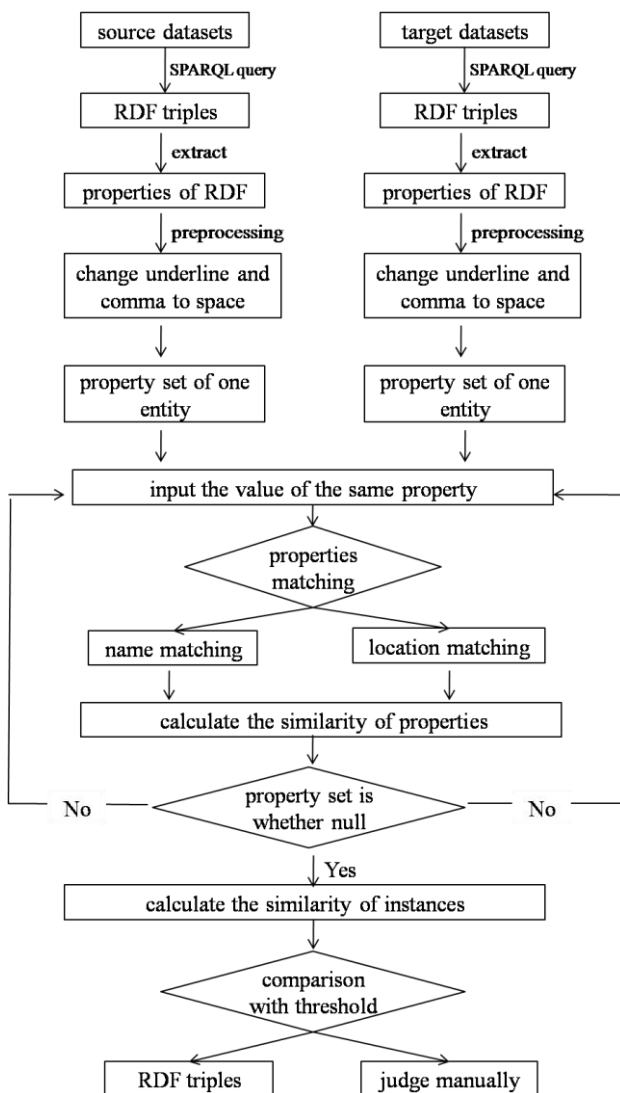
Figure 4. Matching workflow

3.1 Extracting instances

The first step is querying datasets with SPARQL queries. Table 1 shows an example to get agent instances with the three properties. Results are shown as RDF triples in Figure 5.

```
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
Prefix
pro:<http://geopw.whu.edu.cn:8099/provb/datas/provenance.owl
#>
prefix dc: <http://purl.org/dc/terms/>
select ?resource ?name ?city ?province ?country
where { ?resource rdf:type pro:agent .
        ?resource dc:title ?name.
        ?resource pro:city ?city.
        ?resource pro:province ?province.
        }
```

Table 1. SPARQL query of source datasets

Figure 5. RDF triples of Wuhan University

Users can access the DBpedia datasets via a SPARQL endpoint. With the SPARQL query in Table 2, 187 instances of dbResource:National_university are stored in XML file, arranged as Figure 6.

```
PREFIX dbResource:<http://dbpedia.org/resource/>
PREFIX dbProperty:<http://dbpedia.org/property/>
SELECT ?resource ?name ?city ?province ?country
WHERE{
?resource dbProperty:type dbResource:National_university .
?resource dbProperty:name ?name .
OPTIONAL { ?resource dbProperty:city ?city .}.
OPTIONAL { ?resource dbProperty:province ?province .}.
        }
```

Table 2 SPARQL query of target datasets

Figure 6. Snippet of results queried from DBpedia

3.2 Extracting and preprocessing properties

In the results, some triples describe the resource by literal, while others provide URIs as links to other resource. For the latter case, the URI is processed and the last part is extracted as the value of the property. This is reasonable because that URIs are the names for resources on the Web according to the principles of Linked Data (Berners-Lee, 2006). For example, the last part of the resource, http://dbpedia.org/resource/Wuhan, is Wuhan, which can be used as the name of the resource. Thus, instances from Step A are preprocessed to extract values for each property. In order to improve the accuracy of matching, underline and comma in values are replaced by space. When all above steps

are done, each instance with its properties is inserted into Postgres as a record, shown in Figure 7.

(a) Instances extracted from source datasets

(a) Instances extracted from target datasets

Figure 7. Instances in the database

3.3 Matching instances

This step calculates the similarity of every record in the source database with all records in the target database using the Equation 2. For each record in the source database, the target record with the maximum value of similarity is selected. Then, the maximum value is compared with the threshold to determine whether to construct links between the two records.

Interval	[1, 0.9]	(0.9, 0.8)	[0.8, 0]
Count	21	3	3

Table 3 Matching results

Table 3 shows the matching results. Among the results, 11% of instances need to be determined manually by users. 11% of instances cannot be matched. When further checking records in the target database, some empty strings are found. This is because some properties such as dbpprop:province are null in DBpedia. That is also why the SPARQL query in Table II includes the term 'OPTIONAL' for the condition '{?resource dbProperty:province ?province. }' . In this case, instances with the property value as null could also be returned. There is no meaning to match empty strings. Therefore, when the property, dbpprop:province or dbpprop:city, is null, its weight is set 0, and and more weight is added to dbpprop:name. For example, if the dbpprop:province is null, the weight for $n(a,b)$ is adjusted to 0.8, thus the equation will be

$$sim(a,b) = 0.8 * n\ (a,b)\ +\ 0.2 * c\ (a,b) \tag{3}$$

Once matching instances are found, RDF triples are generated automatically as shown in Figure 8, which will be merged to source datasets using Apache Jena API (http://jena.apache.org/). Once we query the updated source dataset, a new property, http://www.w3.org/2002/07/owl#sameAs, is returned, shown in Figure 9.

Figure 8. RDF triples generated automatically

Figure 9. The result of SPARQL query of Wuhan University

4. CONCLUSION

This paper proposes a matching method based on semantic similarity to construct the identity links. The matching algorithm is based on Tversky's contrast model, and Jaro-Winkller distance is used to match values of properties. This method is demonstrated by constructing the links between a linked geospatial dataset and DBpedia. Further work will explore how to construct identity links with spatial characteristics between geospatial linked data.

REFERENCE

Apache Jena, http://jena.apache.org/

Araújo, S., Hidders, J., Schwabe, S., Vries A. P. de, 2011. SERIMI–Resource Description Similarity, RDF Instance Matching and Interlinking. *CoRR*, vol. abs/1107.1104.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak R., Ives, Z., 2007. DBpedia: a nucleus for a web of open data. In: *Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference*, Busan, Korea, pp. 722-735.

Barnaghi, P., Ganz, F., Abangar, H., Presser, M., Moessner, K., 2012. Sense2Web: A Linked Data Platform for Semantic Sensor Networks [EB/OL]. http://semantic-web-journal.org/sites/default/files/swj189.pdf (Accessed 18 Aug. 2013).

Berners-Lee, T., 2006. "Linked Data". http://www.w3.org/DesignIssues/LinkedData.html (Accessed 18 Aug. 2013)

Bizer, C., Cyganiak, R., Heath, T., 2007. "How to publish Linked Data on the Web", http://sites.wiwiss.fu-

berlin.de/suhl/bizer/pub/LinkedDataTutorial/ (Accessed 18 Aug. 2013)

Bizer, C., Heath, T., Berners-Lee, T, 2009. Linked Data—The Story So Far. *International Journal on Semantic Web and Information Systems*, 5 (3), pp. 1–22.

Bizer, C., Heath, T., Idehen, K., and Berners-Lee T., 2008. Linked Data on the Web (LDOW2008). In: *Proceedings of the 17th international conference on World Wide Web*, New York, NY, USA, pp. 1265–1266.

Heath, T. & Bizer, C, 2011. Linked Data: Evolving the web into a global data space. In: *Synthesis Lectures on the Semantic Web: Theory and Technology*, San Rafael, CA, USA: Morgan & Claypool, 2011.

Pschorr, J., Henson, C., Patni, H. and Sheth, A., 2010. Sensor discovery on linked data. In: *Proceedings of the 7th Extended Semantic Web Conference, ESWC2010*, Heraklion, Greece, 2010.

Sören A., Jens L., and Sebastian H., 2009. LinkedGeoData - adding a spatial dimension to the web of data. In: *Proceedings of 8th International Semantic Web Conference (ISWC 2009)*, Chantilly, VA, USA, October 25-29, 2009, pp. 731-746.

Tversky, A., 1997. Features of similarity. *Psychological Review*, 84 (4), pp. 327–352.

W3C Linking Open Data community project, 2006. "Linking Open Data", http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData (Accessed 18 Aug. 2013)

Winkler, W.E., 1990. String Comparator Metrics and Enhanced Decision Rules in the Fellegi - Sunter Model of Record Linkage. In: *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 778-783.

Yuan, J., Yue, P., Gong, J., and Zhang, M., 2013. A linked data approach for geospatial data provenance. *IEEE Transactions on Geoscience and Remote Sensing*, DOI: 10.1109/TGRS.2013.2249523.