Automatic Alignment between Wikipedia Attributes and DBpedia Properties

Thi-Nhu Nguyen[1,2][0000-0002-7997-2229], Tuan-Dung Cao[2][0000-0003-4805-8132]

[1] Haiphong University, Vietnam
nhunt@dhhp.edu.vn

[2] Hanoi University of Science and Technology, Vietnam
dungct@soict.hust.edu.vn

Abstract. DBpedia plays a central role in Linked Open Data (LOD), due to the large and growing number of resources linked to it. Currently, this project extracts information from Wikipedia to represent in RDF triples. The extraction procedure required to manually map Wikipedia infobox attributes into the DBpedia properties. However, the number attributes are so large for all Wikipedia editions in different languages. This task therefore is time-consuming and labor intensive. We propose a novel method to mapping automatically basing on instance-based approach enhanced by using label translation. Experiments on Vietnamese Wikipedia confirm the significant improvement when applying our method.

Keywords: DBpedia, Ontology, Mappings, Infobox Attributes, Properties.

1 Introduction

DBpedia is built upon the community effort to extract the knowledge from Wikipedia [1]. Currently, this project is maintaining an extraction framework and shared ontology to retrieve knowledge based on the most frequent infoboxes within Wikipedia editions [5]. Each infobox is a set of attribute-value pairs that represents a summary of a Wikipedia article. In detail, contributors from many countries have joined the DBpedia mapping project, whose target is to map the Wikipedia infoboxes into the classes and their attributes into corresponding properties in DBpedia ontology [2]. The DBpedia Ontology 2016-04 version encompasses 754 classes which has a form in a subsumption hierarchy and are described by more than 3000 different properties[1]. Thanks to crowdsourcing, a large number of infoboxes has been mapped. However, the number of accomplished mappings is still small and limited. Thus, the alignment among multilingual DBpedia is currently incomplete. Although DBpedia extracts information from 128 languages, there are only 32 languages that have mappings and

[1] http://wiki.dbpedia.org/dbpedia-version-2016-04

19 own chapters in different languages[2]. It is clear that mapping community is immature. Meanwhile, increasing the number of DBpedia versions helps to improve the association and richness of LOD. Therefore, mapping automatically attributes into corresponding properties is useful solution to deployment DBpedia chapters fast as well as is highly prone to changes in Wikipedia, a noticeable drawback considering how fast edits are made [3].

The main idea is built on an instance-based approach. In detail, it is assumed that attribute and property are the same if theirs values are equivalent. In this paper, we propose a new method that improved the mentioned approach with using label translation; specifically Vietnamese Wikipedia is our case study.

2 Mapping extraction

In this section, we describe how to determine whether an attribute a_I contained in the Wikipedia infobox I can be mapped to a given property r in DBpedia. Given two language RDF datasets, the alignment is to harvest similar pairs in term of value between them. Given a target language and a set of source languages, after the data processing step, we will classify them into individual sets such as date, number, string and object. After that, we compute the value-based similarity between attributes and properties. For each candidate pair, the similarity sim_{l_i} of an alignment $[a_I, r]$ is measured as follow:

$$sim_{l_i}\left(a_I, r_{l_i}\right) = \delta(f(a_I, P_l), g(r_{l_i}, P_{l_i})) \qquad (1)$$

where δ is inner function that has value in [0,1], it is used to calculate the similarity between the values of and for each pivot language. And, $f(a, P_l)$ is used to address the value of Wikipedia attribute a_I in target language l, $g(r_{l_i}, P_{l_i})$ is used to extracts the value of property r_{l_i} of DBpedia in l_i. We denote their definite values by $eval_W, eval_D$ respectively. As mentioned above, we distinguish two kinds of property to compute the similarity. Thus, we apply these functions for each property type.

However, determining the similarity based on their value puts together a raise of noise in the returned results. Because, their values may be the same, but they are not equal in fact. In order to overcome this drawback, we have improved by generating dictionary to translate label. Here, we considered a Wikipedia article A and DBpedia that they have the same entity. Only the attributes, which their values are equivalent, are used.

We denote them by a set A $= \{a_{I_1}, a_{I_2}, \dots, a_{I_m}\}\, with\ f\left(a_{I_1}, P_l\right) = f\left(a_{I_1}, P_l\right) = f\left(a_{I_2}, P_l\right) = \cdots = f\left(a_{I_m}, P_l\right) = c_1$ and $R = \{r_1, r_2, \dots, r_n\}\, with\ g\left(r_1, P_{l_i}\right) = g\left(r_2, P_{l_i}\right) = \cdots = g\left(r_n, P_{l_i}\right) = c_2$ for DBpedia properties in language $l\ P_{l_i}$. If $c_1 = c_2$ we will have the number of mappings most, accounted at m*n matched pairs $< a_{I_i}$, $r_j >$, where $a_{I_i} \in A,\ r_j \in R$. It is desirable this number is decreased to k (k < m*n). In detail, label of attributes and properties are translated in the same language. Obvi-

[2] http://wiki.dbpedia.org/about/language-chapters

ously, it is convenient to translate to English. Then, we can use Wordnet to get synset. Finally, we use majority voting method to retrieve the best pairs.

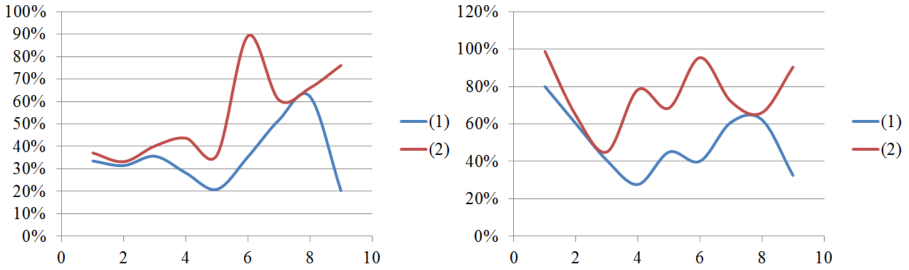
3 Experiment and evaluation

In order to evaluate our approach for automatically mapping, we have carried the experiments on Vietnamese Wikipedia with existing DBpedia editions in three pivot languages (English, German and Dutch) as training data. In currently, editors of Vietnamese Wikipedia employ infoboxes in both Vietnamese and English. For English attributes, we take advantage from existing mapping to extract mappings easily. Thus, the Vietnamese attributes are concerned mainly because they contain more information with high accuracy. Choosing an infobox for mapping is based on two following principles: the attribute number of test data and infobox are equal and all articles chosen have to link to the articles in pivot languages. However, the value of infobox attributes are often incomplete even null in Wikipedia articles. Thus, we have to create a dataset for each infobox in Vietnamese Wikipedia so that the number of attributes had value as much as possible. Most frequent infoboxes are mapped first. This guarantees a good coverage, as infoboxes are distributed according to the Zipf's law. Then, we pick up the 100 infoboxes with the most occurrences in the statistics[3] .

Table 1. The mapping results of the most occurrence infoboxes

Infoboxes /# Attributes	**D**		**E**	
	(1)	**(2)**	**(1)**	**(2)**
Bảng phân loại (Categories) / **167**	33.5%	37.1%	80%	98.8%
Thông tinh khu dân cư (Infobox settlement) /**419**	31.5%	33.2%	60.1%	64.5%
Thông tin hành tinh (Infobox planet) /**87**	35.6%	40.2%	40.5%	45.1%
Thông tin đơn vị hành chính Việt Nam / **71** (Infobox administrative divisions of Vietnam)	28.2%	43.7%	27.7%	78.4%
Thông tin nhân vật hoàng gia/**158** (Infobox royalty)	20.9%	36.1%	45%	68.5%
Thông tin tiểu sử bóng đá /**250** (Infobox football biography)	35.4%	89.2%	40.1%	95.5%
Thông tin nhạc sĩ (Infobox musical artist) / **112**	51.8%	60.7%	60.9%	72%
Thông tin phim (Infobox film) /**53**	62.3%	66%	62.3%	66%
Thông tin viên chức (Infobox officeholder) /**197**	20.3%	76.1%	32.5%	90.4%

[3] http://mappings.dbpedia.org/server/statistics/vi/



a) Comparsion of value D

b) Comparison of value E

Fig.1. The coverage basing on values

In fact, some attributes are occurred many times and vice versa some ones are appeared rarely. Therefore, the occurrence is one of the most criteria to evaluate the mapping results. To evaluate, we use Eq. (2) and Eq. (3) to compute the proportions of mapped attributes and the percentages based on occurrences.

$$D = \frac{|N|}{|M|} \qquad (2) \quad \text{and} \qquad E = \frac{\Sigma_N \, o_i}{\Sigma_M \, o_i} \qquad (3)$$

where M is a set of attribute in infobox I, N is a set of correct mapped attribute and o_i is an occurrence of an attribute a_i. We compare our results with two cases: (1) instance-based approach before and (2) after improving with translation.

Table 1 implies the mapping results on the top 10 infoboxes and our method gives the better result; especially for infoboxes with almost Vietnamese attributes. Figure 1 illustrates more clearly the results before and after the improvement . Conversely, the remained attributes have not mapped yet. Most of them belong attributes with low occurrences. Moreover, the relation between attribute and property does not exist or that attribute is too specific for only Vietnamese Wikipedia so that it is difficult to find out a corresponding property in DBpedia. For an instance, let's consider infobox "Thông tin đơn vị hành chính Việt Nam" (Infobox administrative divisions of Vietnam). The attributes "cỡ bản đồ" (map size), "nhãn bản đồ" (map label) and etc. have occurrences that are less than or equal 1. Besides, attribute "xã" (commune) or "phường" (ward) could not match with any exist property in DBpedia.

4 The demo

We build a tool to convert a Vietnamese Wikipedia articles into DBpedia resources basing on generated mappings. A user can input the keywords for some subjects in Vietnamese language, our system will show some suggestions about them. When the user clicks any subject, the article will be showed and the button aims to convert it to DBpedia resource. Besides, our system also allows users see and extract data in RDF triples. This demo uses the algorithm to automatically mapping attributes of Vietnamese Wikipedia infoboxes into corresponding properties in DBpedia ontology with database in any language or a URI to query the entity. We build a tool named AMA[4] as the first simulator of building Vietnamese DBpedia chapter automatically.

5 Conclusion and future work

In this work, we propose a new method that recovered basing on instance approach with translation. The experiment shows that our method has improved with better result although it remains several weaknesses. This shows that our methology is promising to evolve into the development of linked data and fast deployment localized DBpedia chapter in the context of theirs mapping communities are still weak. For future work, we will investigate this algorithm for some different languages.

References

1. Lehmann, J.; Isele, R.; Jakob, M; et al.: DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. The Journal Semantic Web – Interoperability, Usability, Applicability. vol. 6, no. 2, pp. 167-195. (2015)
2. Mendes, PN.; Jakob, M.; Bizer, C.;. DBpedia: A Multilingual Cross-Domain Knowledge Base. In Proceedings of the Eighth International Conference on Language Resources and Evaluation, pp. 1813-1817. (2012).
3. Palmero, A.; Giuliano, C.; Lavelli, A.: Automatic Mapping of Wikipedia Templates for Fast Deployment of Localised DBpedia datasets. In Proceedings of the 13th International Conference on Knowledge Management and Knowledge Technologies, pp. {1:1-1:8} (2013).

[4] https://sites.google.com/site/jist2017demo/