International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
Volume-3 Number-1 Issue-8 March-2013

Semi-Automatic Mapping Generation for the DBpedia Information Extraction Framework

Arup Sarkar[1], Ujjal Marjit[2], Utpal Biswas[3]
Dept. of Computer Sc. & Engineering, University of Kalyani, Kalyani, India[1,3],
CIRM University of Kalyani, Kalyani, India [2]

Abstract

DBpedia is one of the very well-known live projects from the Semantic Web. It is like a mirror version of the Wikipedia site in Semantic Web. Initially it publishes the information collected from the Wikipedia, but only that part which is relevant to the Semantic Web. Collecting information for Semantic Web from the Wikipedia is demonstrated as the extraction of structured data. DBpedia normally do this by using a specially designed framework called DBpedia Information Extraction Framework. This extraction framework do its works thorough the evaluation of the similar properties from the DBpedia Ontology and the Wikipedia template. This step is known as DBpedia mapping. At present most of the mapping jobs are done complete manually. In this paper a new framework is introduced considering the issues related to the template to ontology mapping. A semi-automatic mapping tool for the DBpedia project is proposed with the capability of automatic suggestion generation for the end users so that users can identify the similar Ontology and template properties. Proposed framework is useful since after selection of similar properties, the necessary code to maintain the mapping between Ontology and template is generated automatically.

Keywords

DBpedia, DBpedia Information Extraction Framework, DBpedia Ontology, Mapping tool.

1. Introduction

Traditional web is a place where information get stored in the form of digitized document connected with each other. There are lots of web searching tools which help the end users to track down any information from the web. The validity or the correctness of the produced results by the search tools are verified by the end users mainly. Machines do not aid much to this problem. Since machines simply don't understand those human readable data. To make them machine processable as well as understandable, it is necessary to add semantics to the available information over the web. To make this happened Semantic Web technology is developed to make the world of web more meaningful. Semantic Web is regarded as the future of the traditional document based web; Another version of web where information is not meant for human consumption only, but, for the machines also. Normally the datasets are published over the Semantic Web in a special well known format, called RDF (Resource Description Framework). When this RDF datasets get connected with each other in terms of related information we eventually meet with the more advanced characteristic of Semantic Web, i.e. the Linked Data Web. Linked Data Web is also popular as the Web of Data. No matter what standard or methodology it follows the key aim of the Semantic Web technologies are always remains to publish the connected, relevant and structured data to make them available on the web. These data must be human processable and understandable as well as machine processable and understandable. Traditional web is already holds a plethora of information. It is not always an option to ignore them and simply regenerate them for the Web of Data. In reality these information get processed by the different tools based on the Semantic Web technologies and extract structured data from the legacy datasets and publish them by making them suitable for the Web of Data. There are many such examples. But at the moment, the most important of the type is the DBpedia[1] project which plays a key role within the Web of Data. DBpedia project mainly deals with the information available over the Wikipedia site. Its aim is to analyze the pages of Wikipedia and extract and publish structured data from them and make them available throughout the Web of Data.

The extraction procedure of DBpedia is completely based on the modified DBpedia Information Extraction Framework (DIEF)[2]. This extraction job, done by DIEF, is completely based on the manually developed Ontology. Whereas the Wikipedia[3] is not based on any Ontology, instead it

holds information in key-value pairs format. While extracting information from the Wikipedia, it is required to find out the related concepts, properties from the DBpedia ontology against the different template parameters from the different Wikipedia pages. To make this working a prior mapping[4] file is needed to be generated. This mapping file states, which template and ontology properties are indicating the same information. The limitation of the DBpedia project is that it doesn't provide any tool to create the mapping files automatically. So the manual creation of the mapping files is required. Though DBpedia developers provide a semi-automatic mapping tool[5] to ease the steps of mapping a template and ontology property, but, still the main job is done by manually. That is, the selection of DBpedia Concept against the template and mapping of each template properties against the properties available under the selected ontology class. The main disadvantage of the tool is that it does not give any auto generated suggestion for the most probable candidate for mapping. The framework discussed in this paper is developed to resolve this issue to some extent. To increase the automaticity of such mapping mechanism is always a difficult task. The main difficulty arises from the absence of the use of ontology within Wikipedia itself. In general the mapping is done between two different ontologies[6], while in case of DBpedia Extraction Framework it is done between the DBpedia Ontology[7] and the Wikipedia template. Wikipedia Templates are nothing but the combination of manually edited and randomly chosen word parameters. Some times in different templates, same property represented with different terms. Naturally, it increases the heterogeneity among the Wikipedia template terms, which in turn increases the challenge for mapping between DBpedia ontology and the Wikipedia templates. In the figure 1 our proposed framework is shown which is designed to get a better result while mapping with less amount of involvement of the end users.

The rest of the paper is structured as follows; in section 2 the necessary background regarding the related work are discussed. Section 3 represents the proposed framework SMASG(Semi-automatic Mapping tool and Automatic Suggestion Generator) for Semi-automatic mapping of the Ontology properties and the template properties. Section 4 depicts the implementation issues. Finally section 5 concludes this paper.

2. Background

It is well known that the Wikipedia is an online encyclopedia, where structured and non-structured both types of data get published. Different types of templates are used here to store the structured data. Among them a large amount of structured data is stored using the different Infobox templates. DBpedia plays a key role here to make those structured data available over the Web of Data. It extracts those structured data from the Wikipedia into RDF format as Linked Data. Now a day, the DBpedia project got so much recognition and importance that it has become the central data set of Linked Open Data (LOD)[12] cloud. Since it extracts structured data directly from the Wikipedia, its coverage of different subject matter is vast. That's why it has become an easy target for the other projects to get connected with, to become a part of LOD cloud. The quality of the extracted data has increased by the use of the DBpedia Ontology at the backend. But to make the extraction job easier and successful, mapping between the Ontology terms and the Wikipedia template terms is an absolute requirement. Automaticity at any stage of this mapping job is always preferable since the success of the extraction framework is dependent on the quality and availability of the mappings between Wikipedia templates and the DBpedia Ontology. DBpedia project itself provides a mapping tool, but it lack of automaticity. Mapping is done manually with this tool. It only generates the mapping code automatically. Considering this need of a better tool for this mapping job, the SMASG framework is proposed in this paper. At present all programming and experimentation is done with the English version of Wikipedia only, but in future a multilingual version of the tool is in the plan.

A brief introduction to the DBpedia, Wikipedia projects as well as Linked Data technology and RDF (Resource Description Framework) is as follows.

DBpedia

Today, DBpedia has become one of the most common and well known projects among the Linked Data enthusiasts. Most of its popularity comes due to its efficiency to extract a large amount of structured data from the normal looking Wikipedia pages. It not only extracts structured data from the Wikipedia in RDF (Resource Description Framework) format but it also publishes them over the web of data as Linked Data. After certain time of interval (in terms of months)

DBpedia publishes its updated structured dataset for download. The current release of DBpedia is 3.8. The core of the DBpedia project is its information extraction framework known as DBpedia Information Extraction Framework (DIEF) as already mentioned in section 1 of this paper. DIEF consists of two sub modules, **Core Module** and the **Dump extraction Module**.

Wikipedia

Wikipedia is one of the most accessed and largest websites. It is an online encyclopedia continuously growing up by its users. It is available in several non-English editions, besides its mother edition, i.e., the English version. Wikipedia itself describe it as *"... collaboratively edited, multilingual, free Internet encyclopedia ..."*. One important aspect of the Wikipedia is that, the English version does not put any burden or any restriction over the non-English editions, which means non-English editions, may grow independently by its users. These editions may get richer with its own language specific content. Most of the pages in Wikipedia are rich with structured data. This data mainly stored as key-value form. Wikipedia provides a large collection of templates to make it possible. Among them most well known ones are the Infobox templates. Since all the content in Wikipedia is developed by human users so a huge existence of redundancy and heterogeneity is observed. This is the main reason that makes the structured data extraction from Wikipedia so much of a challenge.

Linked Data

Publishing structured data in a globally accepted standard over the web is not the only aim the Semantic Web is meant for. If Semantic Web publishes structured data in machine processable and understandable format for the machines and just left it there for consume, then, only half of the goal will achieve. To make them completely useful these data must be discoverable. To discover related information among the datasets they need to be linked up. The Linked Data[8] concept first proposed by Tim Berners-Lee[9] is just about to handle this issue. Linked Data is based on four simple standards mentioned by Berners-Lee are as follows,

1. Every entity over the web of data must be represented with standard URIs.
2. All the URIs should be HTTP URIs. It makes easy to go through these URIs to find out more related information.
3. While users are invoking any URI as URL legitimate information should be returned.
4. Most important of all is that in Linked Data, entities and their properties should be connected through URIs to make them discoverable.

Resource Description Framework (RDF)

RDF[10] is a standard model to publish structured data over the web of data. It was mainly developed to make the information about any resources on the web available, such that the machines can process them and understands what they are representing. Currently RDF becomes a leading standard for the Linked Data Web. RDF data is mainly a combination of several statements. Each statement consists of three parts, subject, predicate and object. All the concepts in a particular domain of work are represented as subjects. Every concept has certain number of properties. The predicate part in the RDF statement represents these properties. The object in a statement basically holds the concrete value or literal value for the properties against each subject. Sometimes objects may refer to another concept within the RDF dataset. Every RDF dataset is represented as RDF graph which makes them easy to exploit by the machines.

3. SMASG(Semi-automatic Mapping tool and Automatic Suggestion Generator) framework

In this section our proposed framework for DBpedia Ontology to Wikipedia Template mapping is discussed in details. The whole framework is divided into five separate sections as shown in the figure 1. Section 1 and 5 comprise of the input and output section of the framework. Here input is the DBpedia ontology and the template properties. Output is actual mapping code generated by the framework as per the DBpedia specification. The Section 2 consist of three modules named as **String/Syntax based matching, WN based matching,** and **Aggregation of match result.** The section 2 consists of two modules named as **Initial mapping** and **Auto- suggestion generator**. Section 4 represents the actual **Mapping Generator** with its two sub-modules **Selector** and **Mapper**. The role of each component is discussed below,

a. **String/Syntax based matching:** it simply applies different string based matching techniques to find out the different measurement against each combination of Ontology properties and template properties.

More than one measurement techniques are applied here to get more than one measurement, so that we can analyze the different measures and combine them to get the best candidates for mapping. The matching methods may include, Jaro measurement[6], N-Gram matching[6], Levenshtein distance[6] etc.

b. **WN based matching:** WN stands for WordNet[11], one of the well known lexical databases of English lexical words and their meanings. This module may use one or more different WordNet based matching techniques to find out the semantic based similarity. The choice of different measures may consist of the Leacock & Chodrow, Jiang & Conrath, Resnik, Lin, Hirst & St-Onge, Wu & Palmer etc.

c. **Aggregation of match results:** this component takes inputs from the previous two components, to aggregate them into a single similarity (confidence) value. So that the decision can be taken into the favour of the best candidate of the mapping.

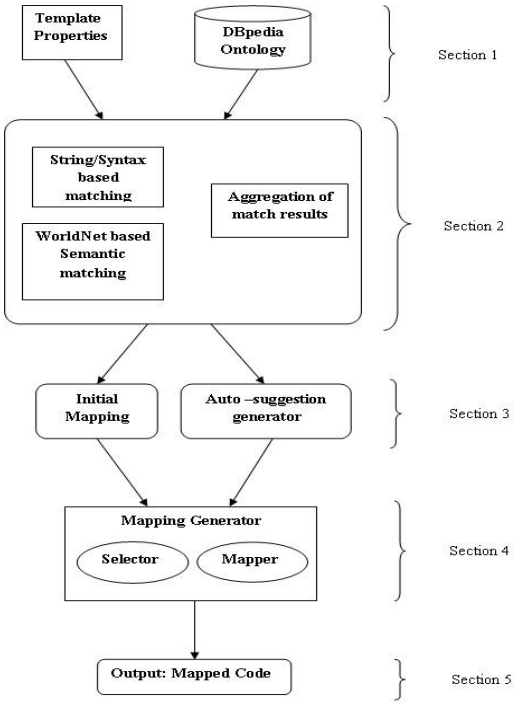

Figure 1: Schematic diagram of SMASG frame work

d. **Initial mapping:** Initial Mapping is done in a complete automatic way. It gives the initially mapped candidates based on the aggregated result received from the previous component. Normally the candidates with higher similarity confidence measure are selected for the initial mapping result.

e. **Auto-suggestion generator:** Initial mapping selects the template name and ontology class pair and after that the template parameter and ontology property pair considering the highest aggregated similarity confidence value. But still there is a possibility that the selected pairs do not reflect the correct pairs for mapping. The correct pairs may come with lower confidence value. In that case the Auto-suggestion generator will generate an automatic suggestion for the end user listing the other credible candidate pairs to help the user to select the best one. Auto suggestion may consider the lexical database like WordNet to generate the suggestion. This phase is at experimental stage.

f. **Mapping Generator:** Section 4 of the framework represents this module. Its jobs are divided into two sub-modules, ***Selector*** and ***Mapper***. Selector finally selects the candidate pairs for mapping after considering the initial mapping and the user input based on the suggestion generated by the Auto-suggestion generator. Mapper takes the final list of candidate pairs and generate the mapping code as per the DBpedia specification for Mapping codes.

4. Implementation

The full implementation of the proposed framework is not done yet and is undergoing. Java platform is chosen to perform the implementation of the framework. Jena package is used to manipulate the DBpedia ontology files. Jena framework is one of the most efficient tools to manipulate the ontology files on purpose. Besides this there are many other WordNet supported tools available with API to perform different operation with WordNet database which is required for this project. We are also using few of them. At present the String based matching is done successfully but the WordNet based matching is still at the Experimental stage. Though enough information is available about the other possible candidate for mapping after the string based

matching, but, still we didn't implement the auto-suggestion generation part, since it requires more evaluation. To get the aggregated measure/confidence value a simple aggregation formula is used here using the Jaro measure, N-Gram matching and Levenshtein Distance measure. The formula is as follows,

$$Val_{sim} = [(J + NG - LD) + W_t] / N \quad (1)$$

In equation (1) Val_{sim} holds the aggregated confidence measure based on the string based matching. In the equation five different symbols are used. The first one is "J" denotes the value of the Jaro measure, "NG" represents the NGram based matching score and "LD" represents the measure of Levenshtein distance. In (1) a weight value "W_t" is added. If the matching is done after the tokenization of the terms then the weight value is set at 0.5. But if the original property is non-tokenizable like the terms "title", "date" etc. Then a weight value of 1.0 will be added instead of 0.5. A threshold value is also used internally during the implementation to discard the irrelevant pairs from the suggestion list. Our experiment shows us that the values above 0.578250322662 give the most considerable candidate pairs for mapping. For our prototype the threshold value is kept at the 0.5. The variable "N" in the equation represents the number of matching techniques applied during the process. Here, the "N" value will be 3, since, only three of the measure is applied (i.e., Jaro, N-Gram and Levenshtein).

String based matching between the properties of Infobox template "book" and Ontology properties of the class Book is shown in Table 1. The given table reflects only those aggregated confidence values which are greater than 0.5 for the sake of experiment.

5. Conclusion

In this article, a new framework has been developed for the DBpedia mapping initiative, which works in a semiautomatic manner. At the very first stage it will generate an initial mapping for the classes as well as for their properties. In addition to that a suggestion is generated for the end users to select another candidate pair for mapping if they are not satisfied with the initial mapping. After all these steps it generates the mapping code for publication over the DBpedia mapping in the form, as prescribed by the DBpedia mapping site[4]. There is a possibility that the formulas or the methods applied for the matching of terms may change in future with better substitute. But the main framework may remain same. At present the multilingual capability is not included in the framework. Future development may include this one too.

Table 1: Aggregated Confidence measure between Infobox template and DBpedia ontology properties

Template (Infobox): book DBpedia Ontology class: Book		
Template property	**Ontology property**	**Aggregated Similarity (Confidence > 0.5)**
name	formerName	0.655
title_orig	title	0.670
cover_artist	coverArtist	1.333
subject	nonFictionSubject	0.626
release_date	releaseDate	1.333
,,	accessDate	0.593
,,	deliveryDate	0.574
english_release_date	releaseDate	0.537
media_type	mediaType	1.333
pages	numberOfPages	0.603

References

[1] L. Yu, "DBpedia". In A Developer's Guide to the Semantic Web, Springer, Berlin, Heidelberg, pp. 379-408, 2011.
[2] The DBpedia Information Extraction Framework documentation, Electronic resource, Available at: http://dbpedia.org/documentation, Accessed on November, 2012.
[3] Wikipedia, Electronic resource, Available at: http://en.wikipedia.org/wiki/Wikipedia, Accessed on January, 2013.
[4] DBpedia Mappings Wiki, Electronic resource, Available at: http://mappings.dbpedia.org/index.php/Main_Page, Accessed on December, 2012.
[5] MappingTool, Electronic resource, Available at: http://mappings.dbpedia.org/index.php/MappingTool, Accessed on September, 2012.
[6] J. Euzenat and P. Shvaiko, "Ontology Matching", Springer, Berlin, Heidelberg, 2007.
[7] The DBpedia Ontology, Electronic resource, Available at: http://wiki.dbpedia.org/Ontology, Accessed on December 2012.
[8] C. Bizer, T. Heath and T. Berners-Lee, "Linked

Data The Story So Far," Int. Journal On Semantic Web and Information Syatem, vol. 5, no. 3, pp. 1-22, 2009.

[9] T. Berners-Lee, "Linked Data Design Issues", June, 2009, Electronic resource, Available at: http://www.w3.org/DesignIssues/LinkedData.html, Accessed on January, 2013.

[10] Resource Description Framework (RDF), Electronic resource, Available at: http://www.w3.org/RDF/, Accessed on November, 2012.

[11] G. A. Miller, "WordNet: A Lexical Database for English", Communications of the ACM 38(11): 39-41, 1995.

[12] State of the LOD Cloud, Electronic resource, Available at: http://wifo5-03.informatik.uni-mannheim.de/lodcloud/state/, Accessed on January, 2013.

Arup Sarkar is a Research Scholar of the Department of Computer Science & Engineering, University of Kalyani, India. He obtained his Diploma, B.Tech. and M.Tech. degree in the years of 2005, 2008 and 2010 from The Calcutta Technical School, Govt. College of Engg. & Ceramic Technology and University of Kalyani respectively. His research interests include Web Service, Semantic Web, Semantic Web Service, Ontology, Knowledge Management, Agent Oriented Programming and Web technology.

Ujjal Marjit is the System-in-Charge at the C.I.R.M.(Centre for Information Resource Management), University of Kalyani. He obtained his M.C.A. degree from Jadavpur University, India in 2000. Currently he is pursuing his Ph.D. from University of Kalyani, India. His vast areas of research interest reside in Web Service, Semantic Web, Semantic Web Service, Ontology, Knowledge Management, e-Governance as well as Software Agents etc. More than thirty eight of his papers have been published in the several reputed national and international conferences and journals.

Dr. Utpal Biswas received his B.E, M.E and PhD degrees in Computer Science and Engineering from Jadavpur University, India in 1993, 2001 and 2008 respectively. He served as a faculty member in NIT, Durgapur, India in the department of Computer Science and Engineering from 1994 to 2001. Currently, he is working as an Associate Professor in the department of Computer Science and Engineering, University of Kalyani, West Bengal, India. He is a co-author of about 90 research articles in different journals, book chapters and conferences. His research interests include Optical Communications, Ad-hoc and Mobile Communications, Sensor Networks, Semantic Web Services, E-governance etc.