CSSO: Ontology-Based Framework for Semantic Interoperatability Between Heterogeneous Web Sources

Emergent Research Forum (ERF)

Samreen Kazi
Department of Computer Science,
Institute of Business Administration
sakazi@iba.edu.pk

Quratulain Rajput
Faculty of Computer Science,
Institute of Business Administration
qraajput@iba.edu.pk

Shakeel khoja
Faculty of Computer Science,
Institute of Business Administration
skhoja@iba.edu.pk

Abstract

The majority of data on the web is of semi-structured or unstructured form but due to syntactic and semantic differences, attempts to combine semi-structured data that stems from diverse Internet sources are often unsuccessful. The use of ontologies can significantly improve semantic richness of texts presented on the Web but taxonomy building is a tedious task requiring extra human resource. This paper explores methods of automatizing the process of creating an ontology-based approach as a solution for scholarships retrieval, using heterogeneous data sources and resolving scheme-level semantic conflicts among data. To achieve this, a framework called CSSO (Computer Science Scholarship Ontology) is created to perform semantic annotation by employing ontology to build knowledge, generate rules and obtain information using reasoning with help of SWRL rules. In the CSSO framework a semantic framework suggests a general methodology of transforming existing resources into “semantically enhanced” ones to obtain generic information.

Keywords

Semantic web, Ontology, RDF, SWRL rules, OWL

Introduction

The World Wide Web has over the years been the best platform and the biggest source of useful data and information. Businesses and individuals are now relying on the data found on the web for their critical decision making. The information found on the web can only be useful if it is; precise, accurate and also published on time when the need for it arises. Moreover, the web is densely packed with unstructured content, not to mention the restrictions that using typical keyword-based search engines sets. Due to the difference in the requirements by different application domains, the information system is therefore implemented differently. For instance, if a web user is searching for information around the scholarship for Crest Oran-B dental hygiene, the result for the exact query in the HTML pages may fail because there are conflicts on issues such as use of that may differ. E.g. conflict between the words “sponsor” and “awarding institute.” Also, there may be conflict in the currency units that may be in use. E.g. EUR and USD.

The best solution to the problems highlighted above is to ensure that you have a system specifically for integrating the web data. Therefore, there is need to resolve all semantic heterogeneities in web applications. The purpose of this work is to give a model that can resolve this problem. The model proposed is semantically annotating all documents relevant to a search to allow easy retrieval through structured query language scripts without manual browsing.
To achieve this goal, we suggest an architecture for a web data integration system with support of ontologies. This proposal employs ontology as a way of reconciling the data and the schema model through explicitly mapping of terms in the ontology. From the perspective of a user querying the Internet for useful information on Computer Science scholarships; it is understood as significant area that needs better information management. This paper is arranged as follows: Section 2 is an elaboration of related work, Section 3 discusses semantic annotation framework, Section 4 is a discussion of results and the last Section 5, is the conclusion with a discussion on challenges and recommendation for future research.

**Literature Review**

Several classifications suffer from integration problem due to heterogeneous naming conflicts. The literature in the community of semantic extraction of sources (Rajput et al, 2011a; Beneventano et al. 2000) is abundant with taxonomies that revolve around heterogeneity issues and integration conflicts. In broader context, semantic interoperability at application and knowledge level create problems because querying data for aggregated results from heterogeneous sources is not possible due to taxonomy conflicts. Researchers asserted that in order to deduce conflicts developing from various data sources, more elaborate data semantics are required. Thus recently, researchers have suggested employing ontologies in data integration (Rajput et a, 2011a).

Moreover, ontologies’ role in interoperability became evident and several ontology-based methods evolved towards attaining information interoperability (Rajput et al, 2011a). MOMIS (Mediator Environment for Multiple Information Sources) is a method proposed in (Beneventano et al. 2000) to integrate and query semi-structured and structured heterogeneous data sources alike. MOMIS aims to outline a universal schema that enables symmetric and transparent access semantically heterogeneous data sources. This is done by generating a global virtual view (GVV) of data sources regardless of these sources’ location or the heterogeneous nature of their data. KRAFT (Preece et al. 2000) is intended to assist the merging of knowledge from multiple, heterogeneous databases and knowledge bases. It fuses heterogeneous information while utilizing ontologies to reconcile semantics issues. The BNOSA (Bayesian Network and Ontology-based Semantic Annotation) (Rajput et al, 2011a) is proposed to dynamically gather information from unstructured and ungrammatical resources given pre-defined ontology and resolving data conflicts and missing information using Bayesian network. Throughout the process of extraction, Bayesian networks are employed to predict missing values or resolve conflict in cases where a value is assessed missing or too many values were extracted for a given attribute. BNOSA is thoroughly extensible due to its ability to dynamically associate an ontology with the conforming Bayesian network in order to harvest data from a designated domain.

Authors of (Rajput et al, 2011b) presented a comparison between two frameworks BNOSA (Rajput et al, 2011a) and OntoX et a, 2007); both of which are intended for unstructured and ungrammatical resources. Results of comparison were in favor of BNOSA in terms of performance over a designated data set.

However, CSSO (Computer Science Scholarship Ontology) have an edge in terms of reasoning to generate new set of knowledge which is not available, it can also be extended using different machine learning approaches to predict missing values. The KRAFT’s (Kusumaningtyas and Mustofa 2017) work can significantly influence the way mapping between local and shared ontologies is done to improve performance of CSSO. However, these systems are lacking in their abilities to reason and be extended as well as their semantic data modelling. CSSO offers an App-based inclusive interface to query scholarship Ads. Of all the aforesaid tools. The main distinction of CSSO is in how data was extracted from unstructured and ungrammatical resources and mapped dynamically to ontology; afterwards more knowledge was inferred using extensive usage of SWRL rules.

**Methodology**

Our system utilizes an ontology-based database access (OBDA) for data extraction and processing from multiple data sources on the web. The system starts by reading the scholarship URLs in A and sets the list of heterogeneous URLs for extracting CS scholarships and the properties of relevant concepts in ontology to U and C respectively. The algorithm checks if U’s XPath is unstructured. If so, the algorithm extracts HTML elements, uses regex rules to extract properties of ontology and maps HTML tags to C. Otherwise,
if no more matches of the required properties are found, the algorithm moves to the next set of URLs. Upon extraction of all data, it is kept in RDF format.

System’s algorithm is depicted in Algorithm 1.

```
begin
| Read URLs of scholarships in A
| U ← List of heterogeneous URLs for extracting CS scholarships
| C ← Properties of relevant concepts in ontology
while (C≠NULL) do
| if (XPath(U) is structured) then
| | Extract elements of HTML;
| | Use regex rules to extract properties of ontology;
| | Map HTML tags to C
| else
| | if (no C in A) then
| | | Next(A);
| | end
| | D ← Scrape(U);
| | Write D in RDF format
| end
end
```

Algorithm 1. Ontology based information extraction for semantic annotation

- In the first layer of this system, domain Ontology is constructed.
- Second layer comprises of a data scrapping module.
- Third layer in combination with second layer maps terms for conflict resolution using ontology concepts
- UI layer consists of a user interface module for query construction and a resultant data to user.

In follow-up of above points; details will be given in sequence now:

**CSSO: Ontology construction**

In our case, Portege was chosen as the ontology editor to build the CS Scholarship Ontology (CSSO). OWL or Web Ontology Language and HermiT was used as the reasoner. We will start closely researching the semantics and syntax in the information related to our specific domain along these lines:

A) Identification of key concepts or classes and identification of relationships between ontology classes.

B) Creation of precise texts and definitions for the concepts and the relationships.

C) Add object properties and data properties to the classes.

It has two main classes, Scholarship and ScholarshipAgent. The Object Properties for Class Scholarship are derived based on the fact: Scholarships has award amount, location and contact details associated with it. The Object Properties for Class ScholarshipAgent are derived based on the fact: Scholarship can be awarded by university, company, or any individual. There are 4 Object properties and 15 Data properties the domain and range of which are listed in Table 1.

<table>
<thead>
<tr>
<th>Object Properties</th>
<th>Subproperties</th>
<th>Options</th>
<th>Domain</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>awardedBy</td>
<td></td>
<td>inversOf giveScholarship</td>
<td></td>
<td></td>
</tr>
<tr>
<td>giveScholarship</td>
<td></td>
<td>inversOf awardedBy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>has Address</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>has Award</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
To evade uncertainty in ontology we have selected the most often used terms in Scholarship Ads in the conceptualization. The conceptual model presented to structure Scholarship Ads helps in searching Ads from heterogeneous sources and aggregating information available for particular Ad. The Scholarships can be categorized into Ph.D., Graduate and Undergraduate level etc. The entire Organization instance of ScholarshipAgent class is distinguished by means of a distinctive property given as awardedBy property. The data properties additionalInformation, amount, awardProperty, deadline, hasCity, hasCountry, hasEligibility, hasEmail, hasPhone, hasStreet, hasZIP, name, recipients, sourceUrl capture the description related to scholarship awards collected from numerous Scholarship Ads.

### Rule Acquisition

In this stage SWRL rules are separated in two sets; one to determine the associating award amount to scholarship and the other to utilize this information to distinguish Importance of Scholarship (most prestigious, more prestigious). The proposed system does not identify type of scholarship, but takes hasDegree information along with award amount associated with each degree. Then the resultant such as Most prestigious, Research-based scholarship etc. is inferred. This is done by an SWRL rule as shown in rule 1 of Table 2.

<table>
<thead>
<tr>
<th>Rule#</th>
<th>Rules Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Scholarship that is awarded by Google is most prestigious</td>
</tr>
<tr>
<td>2</td>
<td>Scholarship that is awarded by Microsoft is highly prestigious</td>
</tr>
<tr>
<td>3</td>
<td>Scholarship that allows domestic people only is local Scholarship</td>
</tr>
<tr>
<td>4</td>
<td>Scholarship for man/male is a Man's scholarship</td>
</tr>
<tr>
<td>5</td>
<td>Scholarship for woman/female is a Woman's scholarship</td>
</tr>
<tr>
<td>6</td>
<td>Scholarship that is by organization is an Organization's scholarship</td>
</tr>
<tr>
<td>7</td>
<td>Scholarship that is for masters is graduate Scholarship</td>
</tr>
<tr>
<td>8</td>
<td>Scholarship that is for PhD is Post-graduate Scholarship</td>
</tr>
<tr>
<td>9</td>
<td>Scholarship that is not local is overseas Scholarship</td>
</tr>
<tr>
<td>10</td>
<td>Scholarship that is local is Domestic Scholarship</td>
</tr>
<tr>
<td>11</td>
<td>Scholarship that has duration six months is Research Scholarship</td>
</tr>
</tbody>
</table>

Table 2. SWRL Rules for detecting different types of scholarships

The reasoning process is carried out by Hermit 1.3.8.413 reasoner. Few of SWRL rules defined in the ontology (shown in Table 2) represent the classes’ restriction expressions as shown in Table 3.

The highlighted terms in bold from Table 3 are explained as follows:

**Most Prestigious:** Scholarship whose award money is greater than 20000 is most prestigious.

**Highly Prestigious:** Scholarship whose Award is less than 20000 and greater than or equals to 10000 is highly prestigious.

**Post-graduate Scholarship:** Scholarship that provides Ph.D. degree is Post-graduate Scholarship.

Some key issues while data extraction were related to ambiguity of terms; one of them was Semantic variability of the term scholarship with Fellowship, Grant, Paid internship, Awards. The proposed framework used a table of already defined terms (such as synonyms etc.) for resolving conflicts. The term...
scholarship appears to be used synonymously and independent extractions with these terms were mapped
to single term scholarship. Similarly, conflicts related to terms Deadline, Award Amount, Domestic
Scholarship were also resolved.

<table>
<thead>
<tr>
<th>Scholarship</th>
<th>Subclass</th>
<th>Restrictions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Address</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Award</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contact</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scholarship</td>
<td>MostPrestigiousScholarship</td>
<td>Scholarship that (hasAward some (amount some xsd:double[&gt;=20000]))</td>
</tr>
<tr>
<td></td>
<td>HighlyPrestigiousScholarship</td>
<td>Scholarship that (hasAward some (amount some xsd:double [&lt;20000,&gt;=10000]))</td>
</tr>
<tr>
<td></td>
<td>MediumPrestigiousScholarship</td>
<td>Scholarship that (hasAward some (amount some xsd:double[&lt;10000, &gt;=5000]))</td>
</tr>
<tr>
<td></td>
<td>CompanyScholarship</td>
<td>Scholarship that (hasAward some (amount some xsd:double[&lt;5000, &gt;=0]))</td>
</tr>
</tbody>
</table>

Table 3. Classes and restriction expression list

Conclusion

This paper demonstrated an automated method of conflict resolution at terms level between
heterogeneous data sources. The mapping of terms and annotation method is based on the
conceptualization of the problem which makes it adoptable to other domain ontologies as well. To inspect
the practical aspects of the proposed method, we directed several experiments on the online scholarship
ads and it gave us promising results. The presented ontology is comprised of scholarship’s knowledge base
and rules that were coded using SWRL for inferring new knowledge of interest. Future work shall
concentrate on deep learning approaches to semantically annotate data in order to address missing values
while instantiation. The conflict resolution process can be made more efficient by extracting named
terms automatically from un-structured, un-grammatical resources using core natural language
processing techniques and mapping them to concepts.

REFERENCES

of the 13th European Conference on Artificial Intelligence (ECAI-98). Brighton, UK.
information search,” in Proceedings of the 3rd International Conference on Science and Technology -
Computer (ICST). IEEE.
architecture for knowledge fusion and transformation,” Knowledge-Based Systems (13:2-3), pp. 113–
120.
Yildiz, B., and Miksch, S. 2007. “ontoX – A Method for Ontology-Driven Information Extraction,” in
pp.660–673.