Latif, Atif; Afzal, Muhammad Tanvir; Tochtermann, Klaus; Raza, Muhammad Asif

Conference Paper

Automating property binding into informational aspects from Linked Data


This version is available at:
http://hdl.handle.net/11108/159
Automating Property Binding into Informational Aspects from Linked Data

Atif Latif1, Muhammad Tanvir Afzal2, Klaus Tochtermann1, Muhammad Asif Raza3

1German National Library of Economics Leibniz Information Center for Economics, Kiel, Germany, {a.latif, k.tochtermann}@zbw.eu,
2Muhammad Ali Jinnah University Islamabad, Pakistan, mafzal@jinnah.edu.pk,
3National University of Sciences and Technology, Islamabad, Pakistan, niazi_gau@hotmail.com

Abstract—Semantic Web is a collaborative effort to bootstrap the Conventional Web (Document Web) content by adding meaningful data structure to it. This transition is to make conventional web's content more valuable by providing machine understandable and semantically rich structured data. Linked Open Data project is seeding the Semantic Web Vision by publishing and interlinking related structured data. This project resulted in a heap of Linked Open Data which ranges from geographic to cross-domain datasets which provide huge opportunities for knowledge discovery and mash up application development. However to make full use of these datasets, there still lies many challenges which are needed to be trailed. One of the major bottlenecks in Linked Open Data is the extraction and then presentation of semantic data to the naive users. In past many efforts has been made to develop semantic applications which can search the required information from linked data sources accurately, further on can hide the complexities of data, organize and importantly convert the data in a readable format before presenting it to the users. “Concept Aggregation Framework” is one of these applications. According to this framework, the similar type of data is grouped into different aspects and sub-aspects before displaying it. In first step, this framework was practiced in an application known as “CAFSIAL.” However, till now the grouping of the properties into aspects and sub-aspects is being done manually which need to be automated for new resource type alignment in CAFSIAL. This paper presents an automated way of grouping related properties into informational aspects using ontology structure. The evaluation has proved that the grouping of related properties can be automated with good accuracy.

Keywords—Semantic Web, Linked Open Data, Information aggregation and presentation

I. INTRODUCTION

Semantic Web is an extension of our Conventional Web (also termed as Document Web). The existing web contains huge amount of information in form of documents. Information on the web is linked by means of hyperlinks, and using these hyperlinks any information can be linked to any other information. This information is of great diversity for example, from education to news, technology to medical science and music to religion. Most of the information and documents present on the web is for humans and not for computers to process it. Semantic Web aims to convert the data or information in a properly structured format so that it can be understood by computers and programs. This new format will make the already existing data meaningful. For the realization and bootstrapping structured data movement Linked Open Data [1][2] project was proposed in 2007. Main objective of this project is to disengage practices of well-gardening data and motivating people to publish their datasets as open and structured data. In a nutshell, it’s an effort for creating a global connected data space [3] where related information is better connected. This will make information more reusable and discoverable; as well as leading towards unrestrictive data usage for better data interlinking, querying and intelligent application development. It is based on the Linked Data four principles stated by Tim Berners-Lee [4] which are:

- Use URIs as names for things
- Use HTTP URLs so that people (and machines) can look up those names (see also [5])
- When someone looks up a URI, provide useful information
- Incorporate other URI’s so that more data can be exposed

Basically, these rules provide a set of guidelines for publishing data as Linked Data. Firstly this emphasizes to identify real and abstract concepts within the datasets and then assigning identified resources with unique URIs in Resource Description Framework (RDF) [6] format which are further dereference-able to present more meaningful information. One of the beauties of this new paradigm is that linked data applications can use any number of data sets for searching and connected with them at run time. Currently, more than 925 data sets consisting of over 60 billion RDF triples which are interlinked by around 673 million RDF link in Linked Data cloud are estimated1.

The myriad of linked data is piling up by each passing day however complexity attached to this data is still retained. Interestingly, when this available linked data is intended to be used for application development it poses many challenges. One of the major challenges is the lack of user friendly interfaces which can facilitates users in exploring linked data benefits by hiding the complex structure of RDF and SPARQL querying. Concept Aggregation Framework was presented by Latif et al. [7] exactly to solve this kind of challenge. According to this framework, the similar type of data retrieved

1 http://stats.lod2.eu/ (as of March 2014)
creating new tags and connecting them to common universal concepts present in the knowledge world. A user interested in any of these tags can dereference the URI of the tag to obtain the information.

DBpedia Mobile\(^2\) is a client application meant for mobiles. It uses GPS signals from a mobile to get its current geographic position. After locating its position it renders its map indicating the nearby locations from the DBpedia datasets. Using this map user can further navigate into interlinked datasets and can obtain background knowledge about locations nearby. This is an interesting application of linked open data.

BBC music\(^6\) is a web application for searching music. It is built on Musicbrainz metadata and identifiers. Information like name of the artist is picked from Musicbrainz and information like introduction of the artist is extracted from DBpedia.

PowerMagpie\(^9\) is a new version of original Magpie. The primary goal of this system is that user has to make very little effort of semantic understanding of a web content. It automatically relates the major terms present in the text of the web page to semantic entities with help of dynamically discovered ontologies present on the (Semantic) Web. In order to achieve this goal, PowerMagpie has to handle four major tasks. a) Identifying relevant terms in the currently browse web page. b) Selecting online ontologies to interpret the domain terms. c) Relating the text to semantic information d) Navigating textual and semantic information together.

The Tabulator\(^10\) is an RDF browser designed for naive users to provide interaction opportunities with the entire web of RDF data as well giving incentives to developers for posting RDF data and promoting RDF linking standards. Meanwhile it also designed for data provides to see how their data interact with rest of data on the Semantic Web. This project is an attempt to demonstrate and utilize the power of linked RDF data with a user-friendly Semantic Web browser that is able to recognize and follow RDF links to other RDF resources based on the user’s exploration and analysis.

Visual Query Tools\(^11\) allow users to form their own queries using a visual interface. Users should have knowledge of RDF and SPARQL in order to use construct queries and use these tools. Explorator\(^12\) is a visual query tool. It allows user to construct queries visually. Explorator uses an operational model. A visual interface allows user to give criteria for query and the under lying operational model implements it. It allows information searching, exploration and visualization facilities. User can extract information without having domain knowledge.

Analysis of above mentioned systems shows us the major approaches and challenges which motivated us for developing CAFSIAL and its subsequent extension.

- In order to use most of the systems, user must possess knowledge about RDF, OWL and SPARQL.
- Most of the applications lack filtering mechanism.

\(^{1}\) http://cafsial.lod-mania.com
\(^{2}\) http://www.dbpedia.org
\(^{3}\) http://www.faviki.com
\(^{4}\) http://www.faviki.com
\(^{5}\) http://dbpedia.org/DBpediaMobile
\(^{6}\) http://www.bbc.co.uk/music/beta/
Making a visual SPARQL query requires background knowledge.

A novice user cannot use these systems.

Most of the systems are based on limited resource types.

Presentation of data to user in textual form.

III. TEST DATASET AND CAFISL APPLICATION

A. DBpedia

DBpedia is a semantic version of Wikipedia - a popular free Internet encyclopedia. This project is based on the extraction of structured content from Wikipedia articles which is further made available as Linked Open Data. DBpedia allows querying the properties, relationship and external links of other resources which are associated with Wikipedia pages. DBpedia is considered as a nucleus and famous cross-interlinking hub within Linked Data Cloud as also described by Tim Berners-Lee [13][14]. The English version of the DBpedia knowledge base currently describes 4.0 million things, out of which 3.22 million are classified in a consistent Ontology, including 832,000 persons, 639,000 places and 209,000 organizations etc. Resource Description Framework (RDF) is used to represent these records and is accessible in form of RDF triples. For our part of study, we made use of “Organization” records present in DBpedia and accessed these records by querying DBpedia SPARQL endpoint [15].

B. CAFISL Application

CAFSIAL stands for “Concept Aggregation Framework for Structuring Informational Aspects from Linked Open Data”. CAFSIAL contains data extracted from DBpedia known as cross-domain and nucleus of linked data. It contains 23 different types of resources like (Person, Place and Organization etc). Initial experiment of CAFSIAL was carried out with resource type “Person”. According to this framework, relevant concepts of a resource can be aggregated from a knowledge base and the most related informational aspects can be organized in to informational aspects [8]. It addresses challenge related to data presentation. The aim is to hide the process of data extraction and processing from users and allowing users to search for information in the way user they already accustomed to. The last and most important feature in this application is the presentation of the information in a structured textual form which is easy to understand.

IV. EXTENSION OF CAFISIAL

In this section strategy for automatic property binding in CAFISIAL is discussed in details.

A. Populating Resources

Initial version of CAFSIAL holds data about “Persons” and the process of structuring related information in different aspects was manual. The manual allocation of DBpedia properties over informational aspect in this application is illustrated in figure 1. Based on the similar mapping strategy a Concept Aggregation Framework has been applied over another DBpedia resource type “Organization” [16]. These experiments on “Persons” and “Organization” helped us in studying different issues related to information structuring and presentation and subsequently helped us in automatically mapping properties to related aspects and sub aspects.

B. Automating Property Binding

Data extracted from DBpedia is in semantic form and comprises of RDF tags. The aim is to structure the similar information in groups called aspects. The taxonomy of a class is defined by its ontology. Ontology gives following information of a resource.

a. Property Name
b. Label
c. Domain
d. Range
e. Comments

“Property Name” is the name of property, “Label’ is the descriptive name of the property, “Domain” describes the area or field from which a property belongs, “Range” defines the which type of data the property holds and comments gives some description about the property. Data related to any resource is extracted on the basis of the properties the resource holds. Property Names are used in SPAQL queries for retrieving the desired information. In previous experiments similar type of properties were manually placed in similar aspects for structuring the data.

C. Information Structuring Using Domain & Range

After studying the properties of resource type “Organization” in above mentioned experiment a hypothesis was made that the Domain and Range can be helpful in structuring the data. Similar properties appear on similar level if arranged with respect to Domain and Range. Thus the
domain and range can be used in automating the processing of property binding to aspects.

V. CAFSIAL AUTOMATION

The process of property retrieval can be automated using ontology of resource i.e. Domain and Range as described in previous section. This assumption will be used as a principle for the automation of CAFSIAL application. The steps involved are described ahead.

A. Resource Selection

The resource for the experiment is “Organization”. Since the study of this resource has already been made it will further swift up the task. Moreover, “Organization” spreads to wide range of subclasses like “Airline”, “College”, “Military”, and “Television” etc, it will certainly allow us in testing our new rule on wide range of information.

B. Properties Extraction

Properties names for each sub type of Organization has been extracted and saved in a local database using SPARQL queries. An example query extracting the properties for “College” is

```
SELECT DISTINCT ?p WHERE {
}
```

C. Identifying Domain & Range

Domain and Range of the properties of a resource are identified by using SPARQL query for each property name. For example,

```
SELECT DISTINCT ?domain WHERE {
    property_name http://www.w3.org/2000/01/rdf-schema#domain ?domain
}
```

The above query retrieves the domain of a property named “property_name”. However, for using the above query the property must be of type owl:ObjectProperty. Unfortunately, this is not the case for majority of the DBpedia properties. Other type of properties like dbpedia2:property can’t be used in above query. For the time being, this issued has been resolved by query each property for an instance of resource class, and its domain / range have been decided by analyzing the output of the query. Domain and Range of each property are also saved in a local database.

D. Defining a Threshold

Each DBpedia resource holds a number of properties which varies from resource to resource as described in table below. It can be noted that some of the resource types hold very large number of properties. Moreover, it is important to present only necessary of common aspects / information related to a searched content. This rule is used by “Google”. All of the properties are of no use, only common, important properties holding the key data be displayed to user. In order to limit the number of properties a threshold has been defined. It was noted that the properties less than 10 records are of less importance. Thus filtering the properties with less than 10 records resolves the issue of properties selection which was manually being done in previous experiments. The filtered properties can now help is displaying the most relevant content to the user. A SPARQL query retrieving the number of records for a property named “property_name” against a resource “Company” is given below. The number of records for each property for each resource type has also been saved in the database and is also tabulated in table 1 for simplification.

```
SELECT COUNT(DISTINCT ?s) AS ?counts WHERE {
    ?s property_name ?o .
}
```

TABLE I. PROPERTY COUNT FOR ORGANIZATION

<table>
<thead>
<tr>
<th>Resource Type</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Airline</td>
<td>254</td>
</tr>
<tr>
<td>Band</td>
<td>1007</td>
</tr>
<tr>
<td>Broadcaster</td>
<td>279</td>
</tr>
<tr>
<td>Radio Station</td>
<td>528</td>
</tr>
<tr>
<td>Company</td>
<td>2056</td>
</tr>
<tr>
<td>Record Label</td>
<td>221</td>
</tr>
<tr>
<td>College</td>
<td>133</td>
</tr>
<tr>
<td>School</td>
<td>2723</td>
</tr>
<tr>
<td>Geopolitical</td>
<td>316</td>
</tr>
<tr>
<td>Legislature</td>
<td>101</td>
</tr>
<tr>
<td>Political Party</td>
<td>330</td>
</tr>
</tbody>
</table>

E. Mapping of Related Properties to Aspects

Properties are grouped together to form aspects. Each aspect consists of some properties that are related to it. Maximum number of aspects displayed is 8. Name of the aspects and its details is given below.

- Abstract: Abstract describes the searched term in descriptive form. It gives a brief introduction about the subject.
- Organization Aspect: Properties whose Domain and Organization types are similar are bind aspect named similar to the organization type, for example, if the Domain type is “Airline” and the Organization is also “Airline” the name of aspect will be “Airline”.
- Related Persons: Every organization has some important persons related to it. Properties related to these persons are displayed in this aspect.
- Financial Aspects: Financial aspect is an important aspect. It contains information related to financial issues of the Organizations.
VI. OPERATIONAL MECHANISM

The working of the automated CAFSIAL is explained in this section. CAFSIAL has a simple web based interface, and can be run in any of the web browser. User enters an organization’s name and related names are suggested by the system automatically. User enters the name or selects one form auto-suggested terms and searched button is pressed. Names of the organizations which match the user’s entered term are retrieved along the organization type and are displayed to user. For example, if user searches “Oxford”, he is provided with list of possible college (organization) resources to decide on as shown in figure 2.

![Fig. 2. Screenshot - List of Suggested Organizations](image)

Thus autosuggesting the organization name can help in refining the results. When user clicks at the name of organization in the suggested list e.g. “Christ Church, Oxford” properties related to the searched organization are selected from the database on the basis of organization type and are filtered by the threshold already saved. Each property is queried to DBpedia Server and data is retrieved. It is important to note that no data i.e. DBpedia dump is being stored locally. This is notable improvement since the previous version of CAFSIAL was locally storing the DBpedia dumps. As the data is retrieved, at the same time similar properties are allied together to their relevant aspects using the already stored Domain and Range. The data retrieval and property binding is done in parallel. As soon as similar retrieval of dated of similar group of properties is completed, they are assigned an appropriate aspect name and displayed to the user. The final presentation of the results is illustrated in figure 3.

VII. EVALUATIONS

The initial evaluation of the proposed system with the existing leading similar systems is presented in this section. There are four similar systems such as: Initial CAFSIAL, Facet search, 123people, and Freebase. Facet Precision Search & Find is an online semantic system presented by Open Link Software and uses virtuoso server. 123people is an online search engine dedicated to people search and empowers users to find information about themselves and other people. Whereas Freebase is a community-curated database of well-known people, places and things working under flagship of Google. Comparisons of theses system with CAFSIAL are given below.

<table>
<thead>
<tr>
<th>TABLE II. COMPARISON BETWEEN OLD AND NEW CAFSIAL APPLICATION</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Existing CAFSIAL</strong></td>
</tr>
<tr>
<td>Properties have been manually mapped to related aspects.</td>
</tr>
<tr>
<td>Holds data in a local database</td>
</tr>
<tr>
<td>Works on resource type “Person”</td>
</tr>
<tr>
<td>Output is displayed to user after all the queries are</td>
</tr>
<tr>
<td>processed.</td>
</tr>
</tbody>
</table>

-899-
TABLE III. COMPARISON BETWEEN FACET AND CAFSIAL

<table>
<thead>
<tr>
<th>Facet</th>
<th>CAFSIAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gives option to search a key term, Label and URI semantically</td>
<td>Allows semantic searching by taking input in simple text form.</td>
</tr>
<tr>
<td>Gives information in text format i.e. paragraph about all the searched terms in first step</td>
<td>Shows the names of all organizations that matches the keyword along with its type in first step</td>
</tr>
<tr>
<td>After detailed search shows information in semantic + text form</td>
<td>Displays in simple text which is easily conceivable.</td>
</tr>
<tr>
<td>Uses DBpedia Data.</td>
<td>Takes information from DBpedia server</td>
</tr>
</tbody>
</table>

TABLE IV. COMPARISON BETWEEN 123 PEOPLE AND CAFSIAL

<table>
<thead>
<tr>
<th>123People</th>
<th>CAFSIAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gives option to search people by their name</td>
<td>Allows searching Organizations (like Airline, TV, Radio, Colleges).</td>
</tr>
<tr>
<td>US database is used.</td>
<td>DBpedia is used.</td>
</tr>
<tr>
<td>Searches data in different categories, i.e. name, organizations, telephone numbers, email addresses</td>
<td>Searches from Organizations Data and returns the searched name to user for the results he/she wants</td>
</tr>
<tr>
<td>Shows names with few details on main page</td>
<td>Displays common aspects and properties on the first page</td>
</tr>
<tr>
<td>Demands fee for viewing complete data of US citizens</td>
<td>CAFSIAL is a Public system for research purposes</td>
</tr>
</tbody>
</table>

The comparison of Freebase and CAFSIAL are given in the Table V. The Freebase is manually created database, however, from the Table V, it is clear that CAFSIAL is comparable in results in some cases; however, in majority of cases, the CAFSIAL has outperformed Freebase.

TABLE V. COMPARISON CAFSIAL AND FREEBASE

<table>
<thead>
<tr>
<th>Organization Type</th>
<th>Airline</th>
<th>College</th>
<th>Hospital</th>
<th>Legislation</th>
<th>Library</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAFSIAL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No of concepts searched</td>
<td>39</td>
<td>27</td>
<td>52</td>
<td>40</td>
<td>24</td>
<td>182</td>
</tr>
<tr>
<td>Min searched</td>
<td>33</td>
<td>30</td>
<td>20</td>
<td>38</td>
<td>25</td>
<td>121</td>
</tr>
<tr>
<td>Max searched</td>
<td>21</td>
<td>26</td>
<td>19</td>
<td>32</td>
<td>7</td>
<td>61</td>
</tr>
<tr>
<td>Avg searched</td>
<td>15.33</td>
<td>21.6</td>
<td>14.34</td>
<td>23.47</td>
<td>14.54</td>
<td>17.87</td>
</tr>
<tr>
<td>FREEBASE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No of concepts searched</td>
<td>45</td>
<td>43</td>
<td>25</td>
<td>10</td>
<td>27</td>
<td>150</td>
</tr>
<tr>
<td>Min searched</td>
<td>4</td>
<td>12</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>24</td>
</tr>
<tr>
<td>Max searched</td>
<td>25</td>
<td>26</td>
<td>17</td>
<td>10</td>
<td>14</td>
<td>92</td>
</tr>
<tr>
<td>Avg searched</td>
<td>15.33</td>
<td>20.1</td>
<td>6.56</td>
<td>5.96</td>
<td>6.28</td>
<td>10.85</td>
</tr>
</tbody>
</table>

VIII. CONCLUSIONS

Linked Open Data provides huge opportunities for knowledge discovery and mash up application development. However to make full use of Linked Data, still there lies many challenges like extraction and presentation of semantic data. In past a Linked Data application named as: CAFSIAL which was based on "Concept Aggregation Framework" was presented. This application was aimed in providing the user with easy to use interface and presentation of the retrieved information in structured way.
using Concept Aggregation Framework. In this application similar type of data was grouped into different aspects and sub-aspects as a profile before presenting it to the users. However the mapping of properties to information aspect was achieved manually. In this research CAFSIAL was extended by introducing the process of automatically binding the properties to their relevant concepts by using information present in the ontology of the resource. This updating of CAFSIAL has been tested and evaluated for the new resource type: “Organization”. The evaluation has proved that the grouping of related properties can be automated with good accuracy. Furthermore, the results are compared to traditional manual systems. It is found that the proposed system is able to achieve good results without manual effort. In future this proof of concept extension will be deployed along with CAFSIAL running system; moreover other approaches for generating new specific aspects and sub aspects will be targeted.

IX. REFERENCES


