Extracting Knowledge Bases from table-structured Web Resources applied to the semantic based Requirements Engineering Methodology SoftWiki

Rubén Navarro Piris

ruben.navarro.piris@gmail.com

A lot of information on the Web is provided as HTML formatted tables and CSV[1] files. Such tables contain semantic information that can be derived from the embedded environment of the table as well from the heading of each column. Often the problem of integrating and linking this information into semantic web applications occurs. One way to solve this is a transformation of these tables into OWL ontologies.

The requirements engineering tool SoftWiki [DAR06] is an example of such a semantic web application. The SoftWiki methodology [LR09], based on a set of tools including the SoftWiki tool, is used to interpret and manipulate data based on the SoftWiki vocabulary. A common task using the SoftWiki methodology is to import data. The developed OntoWiki extension imports the issue tracker of Google Code[2] projects which is provided as a CSV file and transforms it into an ontology.

The paper describes a methodology that is divided into three parts: (1) Analysis of the table structure, (2) Analysis of the vocabulary and (3) Definition of a N3[3] template. In a typical table the structure is always the same: the columns contain the attributes and the rows the entities. The first row is the one that determines the semantic of the other rows, which contain data that follow this model. Because of its simplicity it is very space efficient, but it’s difficult to further analyze or use this information because of the lack of explicit semantic relations.

city	country	Inhabitants	universities
Leipzig	Germany	518,862	Uni Leipzig, HTWK
Barcelona	Spain	1,621,547	UPC, UB, UPF

Table 1: Example of data in a table

We have to take into consideration also that in some case an attribute can be multi-valued, for example the attribute universities of a city[4] (cf. table 1). The table is also a good example of the biggest problem that the semantic data methodology and technologies try to solve: although a person is able to understand the meaning of the

[1] Comma Separated Values

[2] Google Code hosts Open Source Projects and is available at http://code.google.com

[3] Notation 3 (http://www.w3.org/TeamSubmission/n3/), an easy readable RDF/XML syntax alternative.

[4] Source: http://en.wikipedia.org/wiki/Leipzig, http://en.wikipedia.org/wiki/Barcelona

attributes of a city thanks to the context and its own previous knowledge, a computer does not have this information and cannot therefore use it. To solve this problem data can be semantically structured and tagged with the help of semantic ontologies. One of the most common used and simple table formats is CSV and because of that it is the supported format for the data importation. In order to put the new data into the SoftWiki (or OntoWiki) system it has to be adapted to a OWL[5] model using the semantic information of the first row of a table.

As a lot of information that shares the same structure is going to be imported, the definition of a template with an abstract definition of the matching between a row and its corresponding OWL triples is the best way to go. In this template the matching is achieved by using variables with the names of the attributes of the CSV table (or tables) that we want to import. The system is flexible enough to receive tables with attributes that won't be imported or tables with less attributes than expected or with empty attributes but still import all the information possible.

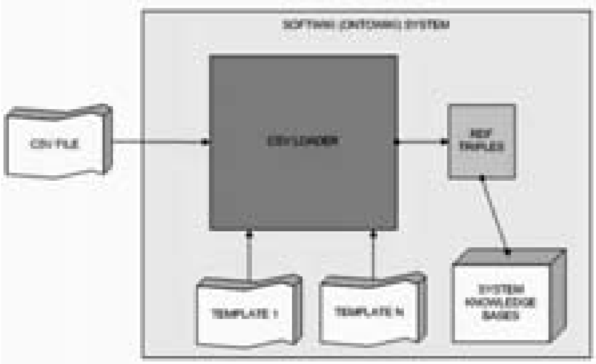

Figure 1: CSV Loader diagram

The importing process (cf. figure 1) would therefore be the following: (1) The CSV Loader application reads the CSV file, (2) According to the information sent by the user, a template is selected and read, (3) RDF triples are created by matching the template with every information row, (4) Finally the triples are stored into the knowledge bases system.

Although the project is focused in the importation of requirements from Google Code Issues to the SoftWiki platform, the application is built in order to be easily configurable. By these means if someone wants to add a template to import other type of data to OntoWiki the following steps are to be followed: (1) Define the template in a .ttlt file and include it in the extension folder, (2) Declare the template in the configuration file, (3) Define the parameters of the template in the configuration file: label, description, url with parameters if used. The application is open source and under development. Its latest version[6] is available online.

[5] Web Ontology Language: http://www.w3.org/TR/owl-features/

[6] http://softwiki.de/netzwerk/de/plattform/csv-load-projekt/

Related work

Convert To RDF[7] *"…is a tool for automatically converting delimited text data into RDF via a simple mapping mechanism."* [GGP+02] Interesting features of this tool are:

- Direct mapping of attributes (columns) to object with properties.
- Possibility of choosing the column that will identify the rdf object and the columns that will be mapped as object attributes, together with its rdf syntax.
- Use of a GUI[8] for defining the mapping. There is also a former version8 that uses predefined mapping files (with either a simple own defined syntax or a rdf syntax, based on a predefined ontology).

There are however some features not supported by ConvertToRDF but in CSVLoad:

- Defining multiple objects from a single row. A row does not necessarily contain information of only one object. Our ontology may, for example, include the City and Country objects. In this case, using the information of table 1, we would want to define one City object and one Country object.
- Support for multi-valued attributes. Using again the information of table 1, our ontology may also contain the University object. In this case the system has to be able of splitting the multi-valued attribute and process every attribute separately.
- Support for missing or empty attributes. CSVLoad works with predefined templates, but does not require that the defined attributes appear the exact way in the input CSV table. If an attribute does not exist in the table (or it is empty) it is automatically omitted of the resulting N3 triples. Also if there are attributes in the table, which are not defined in the template, they are omitted too.
- Data conversion. CSVLoad offers some data converting/processing functions. In this case just by adding the specific tag to an attribute the system will execute an specific routine/conversion to the tagged attribute.

The largest part of information on the Web is already stored in structured form, often as data contained in relational databases, but usually published by Web applications only as HTML mixing structure, layout and content [ADL+09]. The Triplify application[9], born with the idea of overcoming the chicken-and-egg dilemma (simultaneously lacking of semantic representations and semantics-conscious Web search facilities) that delays the expansion of the Semantic Web, permits the conversion of web information (extracted from a relational DB) into RDF, JSON and Linked Data. Interesting features of Triplify tool are:

[7] ConvertToRDF is available under: http://www.mindswap.org/ mhgrove/convert/

[8] Graphical User Interface

[9] http://triplify.org/About

- Easy to install and configure (with few SQL knowledge).
- Already pre-configured mappings to several popular Web applications.
- Focused on deploying the information of a Web into the Semantic Web, fact that provides several advantages: (1) Search engines can better evaluate the content and more easily find content, (2) Possibility of create customized data queries, for example, easy searching for a product with certain characteristics.

Triplify and CSVLoad are focusing on the conversion of table-based data to RDF. However, Triplify works directly on SQL while CSVLoad does it with CSV and template-based mapping, in which case the flexibility is higher because of the possibility of defining different templates for different types of tables. As CSVLoad ignores non defined or empty attributes it is easier to define more general purpose input configuration. Again, Triplify does not support data processing or converting.

Perspectives

Considering that every web application organizes and labels its information in different ways and is usually presented in HTML format, obtaining data from CSV files results insufficient. In this context a plug-in for SoftWiki is under development. This plug-in detects if a loaded requirement resource was imported from Google Code Issues and if positive it imports additional information from Google (community comments with author, date and attachments).

References

[ADL+09] Sören Auer, Sebastian Dietzold, Jens Lehmann, Sebastian Hellmann und David Aumueller. Triplify - Lightweight Linked Data Publication from Relational Databases. In Proceedings of the 17th International Conference on World Wide Web, WWW 2009, Madrid, Spain, April 20-24, 2009, Seiten 621–630, 2009.

[DAR06] Sebastian Dietzold, Sören Auer und Thomas Riechert. Kolloborative Wissensarbeit mit OntoWiki. In Proceedings of the INFORMATIK 2006 Workshop: Bildung von Sozialen Netzwerken in Anwendungen der Social Software, 2006.

[GGP+02] Jennifer Golbeck, Michael Grove, Bijan Parsia, Adtiya Kalyanpur und James Hendler. New Tools for the Semantic Web. In Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web, Seiten 23–38. 2002.

[LR09] Kim Lauenroth und Thomas Riechert. Der SoftWiki-Ansatz für verteiltes Requirements Engineering mit großen Stakeholdergruppen. In Sören Auer, Kim Lauenroth, Steffen Lohmann und Thomas Riechert, Hrsg., Agiles Requirements Engineering für Softwareprojekte mit einer großen Anzahl verteilter Stakeholder, Leipziger Informatik-Verbund (LIV), 2009.